School of Social Sciences

Master in Business Administration (MBA)

Postgraduate Dissertation

Customer Churn Analysis in Insurance Aggregators: Predictive Modelling and Retention Strategies between Online and Offline Channels

Loizos Aggelos

Supervisor: Dimitrios Fouskakis

Athens, Greece, June 2026

HELLENIC
OPEN
UNIVERSITY

Customer Churn Analysis in Insurance Aggregators: Predictive Modelling and Retention Strategies between Online and Offline Channels

Loizos Aggelos

Supervising Committee

Supervisor:	Co-Supervisor:
Dimitrios Fouskakis	Athanasios Rentizelas

Athens, Greece, June 2026

"I would like to sincerely thank Mr. Fouskakis for his guidance, valuable input and advice throughout the preparation of my dissertation.

I would like to thank my beloved wife and son whose unwavering love, patience and encouragement became my greatest source of strength and motivation".

Abstract

This dissertation focuses on investigating customer churn for an insurance aggregator that uses online vs offline distribution channels and explores churn drivers, differences between these channels and the performance of various possible predictors. In the aftermath of digital insurance distribution channels development of customer retention is becoming a topic of interest in terms of portfolio stability, customer life-time value and a long term profitability. The purpose of this dissertation is therefore to reveal churn drivers, explore difference between studied channels and test the predictive capabilities of the variables on respective dataset. For an empirical part results are derived from anonymous operating data, collected from the real insurance aggregator. The dataset consists of approximately 50000 customers with 5-year history of policy holding. Churn is defined as the absence of a paid renewal, pay or replace contract issued within 30 days from contract end, in order to correctly distinguish actual churn from any technical/administrative process interruption events. For analytical purposes, we use a combination of descriptive statistics and predictive models, where logistic regression serves as an interpretable baseline model and LightGBM provides a more flexible machine learning approach for model comparison. Results show that churn is related to relationship history, pricing, renewal, contract details and distribution channel. Online customers have higher churn than offline customers, which confirms the view that digital self-service increases price sensitivity and reduces switching costs. Key variables identified as important for predicting retention are tenure, historical renewal, premium and premium change and policy duration. Logistic regression exhibits good interpretability for managers, whilst LightGBM demonstrates clearly superior predictive accuracy this as a result suggests that there may be non-linear effects and interaction effects. In total, this dissertation contributes to the so far, sparse literature surrounding customer churn in insurance aggregators through the utilization of real world data, customer channel segmentation and predictive analytics to inform retention strategy in online insurance markets.

Keywords

customer churn, insurance aggregators, logistic regression, LightGBM, digital insurance, renewal behaviour

Ανάλυση ρυθμού αποχωρήσεων (Churn Rate) πελατών σε ασφαλιστικές πλατφόρμες σύγκρισης: Μοντελοποίηση πρόβλεψης και στρατηγικές διατήρησης μεταξύ ψηφιακών και φυσικών καναλιών

Λοΐζος Άγγελος

Περίληψη

Η παρούσα διπλωματική εργασία επικεντρώνεται στη διερεύνηση της απώλειας πελατών σε μια ασφαλιστική πλατφόρμα σύγκρισης, με έμφαση στη σύγκριση μεταξύ ψηφιακών και φυσικών καναλιών διανομής, καθώς και στην ανάλυση των παραγόντων που επηρεάζουν την αποχώρηση των πελατών και την απόδοση διαφορετικών προβλεπτικών μοντέλων. Στο πλαίσιο της ανάπτυξης των ψηφιακών καναλιών διανομής ασφαλιστικών προϊόντων, η διατήρηση πελατών αποκτά συνεχώς μεγαλύτερη σημασία, καθώς επηρεάζει άμεσα τη σταθερότητα του χαρτοφυλακίου, την αξία κύκλου ζωής του πελάτη και τη μακροπρόθεσμη κερδοφορία. Σκοπός της έρευνας είναι να εντοπίσει τους βασικούς οδηγούς του churn, να εξετάσει τις διαφορές μεταξύ των δύο καναλιών και να αξιολογήσει την ικανότητα διαφορετικών μοντέλων πρόβλεψης στο συγκεκριμένο σύνολο δεδομένων. Το εμπειρικό μέρος της μελέτης βασίζεται σε ανωνυμοποιημένα λειτουργικά δεδομένα που προέρχονται από πραγματική ασφαλιστική πλατφόρμα σύγκρισης. Τα δεδομένα περιλαμβάνουν περίπου 50000 πελάτες και ιστορικό ασφαλιστηρίων πέντε ετών. Η ακυρωσιμότητα ορίζεται ως η απουσία ανανέωσης, πληρωμής ή αντικατάστασης ενός συμβολαίου με νέο πληρωμένο συμβόλαιο εντός 30 ημερών από τη λήξη του, ώστε να διαχωρίζεται η πραγματική αποχώρηση από τεχνικές ή διοικητικές διακοπές της διαδικασίας. Για τους σκοπούς της ανάλυσης, χρησιμοποιείται συνδυασμός περιγραφικής στατιστικής και προβλεπτικών μοντέλων, όπου η λογιστική παλινδρόμηση λειτουργεί ως ένα ερμηνεύσιμο βασικό μοντέλο αναφοράς, ενώ το LightGBM εφαρμόζεται ως μια πιο ευέλικτη προσέγγιση μηχανικής μάθησης για συγκριτική αξιολόγηση. Τα αποτελέσματα δείχνουν ότι το churn δεν εμφανίζεται τυχαία, αλλά συνδέεται συστηματικά με το ιστορικό της σχέσης με τον πελάτη, την τιμολόγηση, τη συμπεριφορά ανανέωσης, τα χαρακτηριστικά του συμβολαίου και το κανάλι διανομής. Οι ψηφιακοί πελάτες εμφανίζουν υψηλότερα ποσοστά ακυρωσιμότητας σε σχέση με τους πελάτες του φυσικού καναλιού, επιβεβαιώνοντας την άποψη ότι τα ψηφιακά περιβάλλοντα αυτοεξυπηρέτησης ενισχύουν την ευαισθησία στην τιμή και μειώνουν τα εμπόδια μετακίνησης. Μεταξύ των σημαντικότερων μεταβλητών για την πρόβλεψη της διατήρησης πελατών αναδεικνύονται η διάρκεια σχέσης με τον πελάτη, το ιστορικό ανανεώσεων, το επίπεδο ασφαλίστρου, η μεταβολή του ασφαλίστρου και η

διάρκεια του συμβολαίου. Η λογιστική παλινδρόμηση προσφέρει χρήσιμη ερμηνευσιμότητα για διοικητική αξιοποίηση, ενώ το LightGBM επιτυγχάνει σημαντικά υψηλότερη προβλεπτική ακρίβεια, γεγονός που υποδηλώνει την ύπαρξη μη γραμμικών σχέσεων και αλληλεπιδράσεων. Συνολικά, η εργασία συμβάλλει στην ακόμη περιορισμένη βιβλιογραφία σχετικά με την ακυρωσιμότητα στις πλατφόρμες σύγκρισης ασφαλειών, αξιοποιώντας πραγματικά λειτουργικά δεδομένα, διαχωρισμό πελατών ανά κανάλι τεχνικές προγνωστικής ανάλυσης, με στόχο την υποστήριξη αποτελεσματικότερων στρατηγικών διατήρησης πελατών στις ψηφιακές ασφαλιστικές αγορές.

Λέξεις - Κλειδιά

απώλεια πελατών, ασφαλιστικη πλατφόρμα σύγκρισης, λογιστική παλινδρόμηση, LightGBM, ψηφιακή ασφάλιση, συμπεριφορά ανανέωσης

Table of Contents

List of Figures

List of Tables

List of Abbreviations & Acronyms

Abbreviation	**Full term**
AUC	Area Under the Curve
CAC	Customer Acquisition Cost
CRM	Customer Relationship Management
GBM	Gradient Boosting Machine
KPI	Key Performance Indicator
LTV	Lifetime Value
OR	Odds Ratio
ROC	Receiver Operating Characteristic
CC	Cubic Centimeters
MLE	Maximum Likelihood Estimation

1. Introduction

1.1 Significance of the Churn Phenomenon in Insurance Aggregators

The insurance market the past 20 years has been transformed due to digital intermediation and the rise of online insurance aggregators. Commonly known as digital platforms that their main role is to intermediate between consumers and insurance companies. Worldwide, insurance aggregators have become the main channel through which customers compare insurance products and purchase insurance policies and mainly in line market such as motor insurance. Customers want convenience transparency of their suppliers and through digital experience through a medium starting 2010 onwards so that they get everything they need as they see them side-by-side in pc and mobiles (Dumm&Hoyt, 2002).

During that period of transformation, the concept of customer churn, which is the non-renewal or the termination of an insurance policy, had gained strategic importance. Churn is an important KPI that affects annual revenues and the long-term value of the portfolio. Insurance Aggregators want to hold relationship with its customer and user. The customers need to be retained since it is directly associated with Customer Acquisition Cost (CAC) and Customer Lifetime Value (LTV) which influences the profits of an organization (Gupta, et al. 2004).

The purpose of this dissertation is to identify the most important factors associated with customer churn for an aggregator of insurances between online and offline customers. An online customer of an aggregator of insurance communicates only with the website of the aggregator. The customer uses only the tools put at his disposal. On the other hand, offline customers are traditional customers who prefer contact with human advisors and brokers, who play the active role in policy management by analyzing customers' needs and proposing insurance policies.

This distinction is important because these two channels have different dynamics such as value, customer experience, behavioral triggers, trust and mostly, cost of service. Therefore, churn is not exclusive to technical or statistical exercise but also is strategically important due to the connection to profitability, maintaining competitive position in market and the sustainability of the aggregator business model.

1.1.1 Role of Aggregators in the Transformation of the Insurance Market

Since the beginning, the insurance sector has relied exclusively on traditional way of acquiring insurance policies through physical intermediaries such as agents and brokers. The insurance aggregators offered a new model which is fundamentally new, where the customers can obtain multiple quotes in seconds through their devices (mobile, laptop, tablet), whenever they decide without time restrictions, compare those products objectively, complete their purchase online and receive insurance policy through their email.

Insurance aggregators new model offers significant advantages to the customers. First of all, transparency, with real-time price comparison of various insurers of the market. Efficiency, by providing tools to faster decision making and policy purchasing (Ellison & Ellison, 2009). A simple onboarding by providing a fully digital customer journey. A wide range of insurers through one platform and of course, lower distribution costs in comparison to offline brokers (Eling & Lehmann, 2018).

That has positives as it means the insurers have more detailed data about their customers, but it also has a negative side as those customers are more sensitive to price as well and therefore likely to churn to a competitor if they see their premium go up.

On the other hand, offline customers tend to have high confidence in the broker, and they are more resilient to the increase in prices. Therefore, lower probability to cancel their insurance policy and longer retention rates. They also tend to show trust to the broker as a professional and develop personal connections.

1.1.2 The Concept of Churn

Churn is the situation where a customer does not renew the insurance contract they hold or they transfer it to another insurance company. In the insurance industry, especially in motor insurance, the products are periodic and renewable and churn is a critical performance KPI that shows, profitability, customer satisfaction, portfolio quality, distribution efficiency and competitive positioning (Reichheld, 2003).

Customers who choose to maintain for longer periods of time with the same policy and company typically have higher LTV, lower loss ratios due to behavioral patterns that are consistent and easily to predict and finally they are more likely to buy another product, cross - sell, or upgrade the existent policy, up - sell (Neslin et al., 2006).

Customers who often change their policies and do not stick with the same policy are generating revenue losses immediately, increasing the CAC due to costs requirements to replace the policy, lower portfolio profits.

The strategic importance of churn is increasing among insurance aggregators where the competition is digital, and the comparison of price is instantly, and the loyalty is lower and therefore switching barriers have almost not existed.

1.1.3 Online versus Offline Customers

Online customers seem to be quite price sensitive in terms of choosing the lowest price that fulfills their needs with change of insurers (Chen and Hitt, 2002). However, there is low level of sensitivity with an exception if the change is applicable to many insurers. Their transactional needs are mostly rational as they do not have any personal attachment, and they are able to change without any obstacle as their customer journey almost never has an interaction with people.

Offline customers, on the other hand, maintain their relationship with the broker. This relationship minimizes uncertainty of the steps required for the issue of the policy. The trust placed in the insurance agent, who guides customers through every step of the insurance process and takes all their needs into account, is perceived as a high-quality service. This, in turn, leads to a significant reduction in churn, even when premium prices increase.

Customers' decisions are affected by reference to last year's premium expecting it to be lower or the increase will be insignificant. Also, there is the fear that they might have chosen the wrong policy or the many options they must choose from creates uncertainty (Iyengar & Lepper 2000). Finally, there is the preference to maintain the current policy as they trust that their first choice covers all their needs and there is no incentive to change.

1.2 Predictive Modelling and Churn Analysis

In insurance industry predictive analytics have a main role in insurance operations, in order to understand the customer behavior, to optimize the pricing methods and forecast renewals probability the industry is using statistical methods and machine learning models.

The most common techniques are logistic regression, decision trees, gradient boosting machines (GBM), random forests and behavioral scoring models. Insurance aggregators have larger datasets compared to traditional brokers and insurers due to the digital interactions of customers and their platforms. They can collect data throughout the customer journey like browsing behavior of customers, quote search patterns, customer-service interactions and other real-time digital engagement metrics which automatically collected during session or any other interaction with them.

In this dissertation we will use logistic regression and GBM method, to calculate the probability of churn and identify the key factors behind it. Logistic regression is one of the most robust and easy to interpret and that makes it an ideal model for strategic decision making in insurance industry.

1.2.1 Research of Churn

Churn is an area where many industries worldwide extensively research such as telecommunications and banking, although studies focusing on insurance aggregators are limited (Neslin et al., 2006). As long as, studies focus on churn behavior between online and offline customers by using real operational data.

This dissertation work aims to explore and understand the major drivers of customer churn. It considers demographic, financial, behavioral and channel variables (online vs offline) and develops a strong logistic regression model for churn prediction. The research seeks to examine and explore the issue of churn based on product type, channel of distribution and insurance company and finally compares customers' behavior in the online and offline channels, with an intent to uncover and distinguish any potential churn patterns existing within these segments.

This dissertation work aims to explore and understand the major drivers of customer churn. It considers demographic, financial, behavioral and channel variables (online vs offline) and develops a strong logistic regression model for churn prediction. The research seeks to examine and explore the issue of churn based on product type, channel of distribution and insurance company and finally compares customers' behavior in the online and offline channels, with an intent to uncover and distinguish any potential churn patterns existing within these segments. This dissertation aims to propose data-driven retention strategies based on customer profiles to minimize churn in insurance aggregators. It is important to consider the differences in how users behave across different channels for developing effective retention strategies (Gensler et al., 2012).

1.2.2 Study Approach and Methodology

The framework of the dissertation combines statistical analysis and strategic interpretation by using anonymized data of approximately 50000 clients of an insurance aggregator and all their policy purchase history for a 5 - year period, descriptive and inferential statistics for the factors, and a logistic regression as a predictive model, the performance metrics such as AUC, recall, precision and feature interpretation based on odds ratio from the model (Hosmer et al., 2013).

The clients of the aggregator are combining of online customers and offline customers. Online customers are those who use the online platform to purchase their policy and offline customers are those who address brokers who use the insurance aggregators platform to propose quotes to them and purchase their policy through them. Both clients have access to the same products, and the main difference is that the first ones choose mostly by themselves using all the available information provided by the insurance aggregator and the second ones are addressing to the broker who choose for them from the platform.

2. Literature Review: Customer Churn in Insurance Aggregators

2.1 Definition and Different Types of Churns

Churn refers to the termination of a customer relationship with a company in most cases by cancelling a contract or not renewing a contract. In the insurance market, the contracts have renewal cycles of one year, six months, three months, and one month and so churn is considered when a customer stops the renewal cycle by not paying the renewal proposal of the insurer. In business, there is a distinction between voluntary churn when customers do not renew their policy intentionally and involuntary churn where the customer stops renewing their policy due to the absence of an insurable risk, such as vehicle theft, vehicle destruction, sale of the vehicle, vehicle immobilization, or the death of the policyholder (Neslin et al., 2006).

In insurance market, churn, mainly the voluntary insurance category examined in this study, is driven by many factors such as price of a policy, trust in the insurer, aggregator or broker, the service quality and claims experience (Huigevoort, 2015).

2.1.1 Churn and Customer Lifetime Value (LTV)

Churn has an inverse association with CLV. Therefore, lower is the Churn rate is more profitable for business in the long-term (Gupta et al., 2004). Retention is better than acquisition in the digital aggregator world with the high cost of CAC (Reinartz & Kumar, 2000). It was shown by Reichheld (2003) that 5% increase of retention rate generates the rise of profit up to 95%.

2.1.2 Predictive Modelling of Churn

This dissertation demonstrates how different supervised learning algorithms - logistic regression, decision trees, random forests, gradient-boosting and neural networks - can be employed to tackle one specific data science task, churn prediction. It highlights the trade-offs between the methods, both in terms of performance on the prediction task and on real-world industry problems. This also illustrates what the differences really mean in terms of actual business outcomes and the strengths and weaknesses of each method.

2.1.3 Churn in the Insurance Industry

Prior research on the insurance industry suggests that both product-specific and relationship-specific factors play a role in churning. Premium increase, vehicle age and frequency of claims have been identified as key drivers in motor insurance but in other industries, insurance for health-related products, trust, service quality and the duration of the relationship seem more relevant. The literature on churn in insurance so points to a mix of pricing, risk, behavior and relationship specific factors though their importance might differ between insurance products.

2.1.4 Online versus Offline Distribution Channels

Online distribution of insurance increases transparency, price sensibility and lower switching costs among its customers. Online aggregator clients also show lower customer experience quality. On the other hand, offline aggregators serve very connected clients who make use of the advisory services, have lower churn rate and less simple products.

2.1.5 Behavioural Economics and Churn

Iyengar and Lepper (2000) in their highly cited study found that when faced with a large choice set, consumers in digital environments tend not to act as perfectly rational agents as theorised in traditional economic theory. This dissertation discusses that the above argument can be applied to contract renewal. The dissertation is structured around three main points to demonstrate that an abundance of choices can lead to two outcomes: a higher willingness to switch the current provider, and/or greater likelihood to avoid deciding by remaining with the current provider.

2.1.6 Research Gap

This dissertation investigates the churn behaviour in insurance aggregator sector, a relatively less studied area. The literature review highlighted that this area has been neglected in the study of churn behaviour and there is a paucity of empirical research, especially in the area of online-offline segmentation of customer and churn prediction. Two published studies by Baesens (2014) and Henao Madrigal (2020) show that most of the churn studies have been done in banking and telecom sector. The dissertation proposal is about online-offline segmentation of customer in the hybrid model and identifying its unique drivers.

3. Methodology and Data Description

In this chapter the methodological approach followed for the analysis and prediction of customer churn in an insurance aggregator is presented. A quantitative and explanatory research design is followed to search for statistical relations between customer, product, financial and channel characteristics and churn probability (Saunders et al., 2019). Specifically, the adopted methodology blends elements from business analytics and from a managerial point of view. Also, the analysis is based on anonymised operational data from a real insurance aggregator and not on surveys that ask for self-reported opinions about renewing, cancelling and switching intentions (Baesens, 2014). The hypotheses and variables of the empirical analysis are taken from previous literature on churn, retention and CLV, with specific focus on variables that are available at renewal (Verbeke et al., 2011). Finally, logistic regression is adopted as baseline model since it is the most adopted technique for binary target variables, like churn and it provides good insights to be used in practice (Hosmer et al., 2013).

3.1 Data Source

The dataset used in this dissertation has been made available by a real online aggregator active in a competitive online insurance market, including fully anonymized transactional data pertaining to the operations of the aggregator and covering more than 50000 customers and the history of their policies over a period of five years. Churn being time dependent, the dataset follows the same customers for multiple renewal events to be able to track changes in behavior over time (Neslin et al., 2006).

The utilization of large-scale transactional data is commonplace in literature relating to churn analysis. Specifically in fields such as banking, insurance and telecommunications, customer's choices are readily observed through renewal, cancellation and switching behavior (Verbeke et al. 2011). They are preferred over surveys due to being more representative of the actual behavior and minimizing respondents' bias.

3.1.1 Data Structure

The dataset which used in the analysis was selected based on the availability of the data at the time of the renewal. The final dataset uses the original data or transformations to better fit and interpretation from the logistic regression. The final variables used to model churn be shown on Table 1.

The dataset combines data that describe the type of insurance policy, namely whether it belongs to the basic category, the fire-theft category, or the comprehensive insurance category (full-casco). The basic category includes the mandatory coverages required by law as well as some optional coverages that the insurance company may include free of charge or at low cost. Fire-theft packages include all the coverages of the basic package, as well as coverages related to damage to the vehicle caused by fire or the partial theft of vehicle components necessary for its operation, for example the steering wheel, and coverage for total theft, where in the event that the risk occurs the policyholder is compensated with the value of the vehicle at that specific moment.

The data also include the start date of the policy, which in the final dataset is transformed into StartSemesterGroup, where the start period is split into two groups: 1st_semester and 2nd_semester. The daily premium amount of the contract in euros is represented by the variable mikta_per_day, which shows how much the policy costs per day for each policy; in this way, policies with different durations become comparable. The value of the vehicle is divided into three distinct categories: Value_up_to_10000, which represents vehicles with values in the interval [0, 10,000] Value_10000.01_to_20000, which represents vehicles with values in the interval (10,000, 20,000] and Value_greater_20000.01, which represents vehicles with values in the interval [20,000, +∞).

Duration of the contract is in days, although in the final dataset is grouped in four categories, quarterly, semi-annual, annually and monthly, represented by variable DurationGroup. Furthermore, there is the percentage change in the value of the insurance policy compared to the previous contract which was also grouped in four categories which are Decrease_gt_20pct indicates that the new policy has 20% or more lower premium than the previous, corresponding to the interval (-∞, -20%], Decrease_20_to_0pct indicates that premium of new policy is decreased between 19,99% up to 0% compared to previous policy, corresponding to the interval (-20%, 0%], Increase_0_to_20pct indicates that premium of new policy is increased between 0,01% up to 19,99% compared to previous policy, corresponding to the interval (0%, 20%), Increase_gt_20pct which indicates that premium has increased 20% or more compared to previous policy, corresponding to the interval [20%, +∞), and represented in the final dataset by PremiumChangeBand variable.

How many active insurance policies the customer has for other vehicles represented in the final dataset by VehiclePoliciesCount variable, and whether the customer has purchased additional coverages from the insurance aggregator as extra contracts concerning optional coverages such as roadside assistance, legal protection, and vehicle glass breakage, represented in the final dataset by ExtraCoversCount variable.

The data also includes the vehicle's engine power in cubic centimetres (CC), which is a continuous variable which transformed in the final dataset in a categorical with 5 categories which are CC_0_499 which are vehicles with engine power up to 499 CC, CC_500_999 which are vehicles with engine power from 500 CC up to 999 CC, CC_1000_1499 which are vehicles with engine power from 1000 CC up to 1499 CC, CC_greater_1500 which are vehicles with engine power from 1500 CC and more.

Use of the vehicle as a car, motorcycle, or truck represented in the final dataset by variable Usage. Year of manufacture of the vehicle shows the date that the vehicle buyed for the 1st time and transformed into the VehicleAge variable which shows the age of vehicle on the start date of the policy.

Customers might have different contact information (email, mobile, landline phone). For the purpose of the modelling, we split the ContactChannelGroup into two categories: - Email_contact: the customer has at least an email address (it does not matter if they also have a mobile or landline phone) - Phone_contact: the customer does not have an email address and the only way to contact them is through a mobile or landline phone This is an important information for an insurance aggregator because email is the main channel for communicating with customers, especially regarding their policies. Policy documents, renewal reminders, payment due notifications and other service-related communication is typically sent via email. This information could play an important role in the communication speed and efficiency during the renewal process and could also be an important factor to consider when analysing the churn.

In the dataset there are 22 insurance companies, due to the large number the insurance companies transformed into variable CompanySizeGroup which shows if the company is large, medium or small depending where large represent companies with frequency in the dataset greater than 10%, corresponding to the interval (10%, +∞), medium represents companies including 5% up to 10% frequency in the dataset, corresponding to the interval [5%, 10%], and small less than 5%, corresponding to the interval [0%, 5%).

The dataset also includes some demographic characteristics of the customer, such as the postal code which shows where the vehicle is insured and transformed into categorical variable RegionGroup, which has 3 categories first is Attiki which shows the vehicles that are insured in Attica region, second is Thessaloniki which shows the vehicles that are insured in Thessaloniki region and third Other_region which represents vehicles that are insured in the rest of Greece region. The birth date of the customer also is included and transformed into variable Age_at_start which is the age of the customer at the start of the policy. The years of tenure with the insurance aggregator is represented by the variable TenureYears which shows how many years is the customer with insurance aggregator at the start of the policy. The number of active vehicles is represented by the variable ActiveVehiclesCount which shows how many different vehicles the customer had at the start of the policy. Lastly, the dataset includes the information where the customer is either offline or online represented by Client_cat variable.

Table 1 : Variables Description

Variable Name	Variable Name (Dataset)	Description
Continuous Variables		
Active Vehicles Count	ActiveVehiclesCount	Number of active vehicles insured by the customer
Customer Age at Start	Age_at_start	Customer age at policy start, calculated from date of birth
Vehicle Age	VehicleAge	Age of the vehicle at policy start date
Premium per Day	mikta_per_day	Daily premium amount of the policy in euros
Customer Tenure	TenureYears	Number of years the customer has been with the aggregator
Number of Vehicle Policies	VehiclePoliciesCount	Total number of vehicle insurance policies held by the customer
Number of Extra Covers	ExtraCoversCount	Number of additional coverages (add-ons) included in the policy
Categorical Variables		
Policy Category	packet_cat	Type of insurance package (basic, fire-theft, full-casco)
Customer Category / Channel	Client_cat	Indicator of customer channel (online vs offline)
Company Size	CompanySizeGroup	Insurance company grouped as Large, Medium, or Small
Contact Channel	ContactChannelGroup	Main contact method grouped as Email_contact, Phone_contact
Vehicle Usage	Usage	Vehicle usage category: Car, Motorcycle, or Truck
Customer Region	RegionGroup	Geographic area based on zip code (Attiki, Thessaloniki, or Other region)
Vehicle Value Band	VALUE	Insured vehicle value band (Value_up_to_10000, Value_10000.01_to_20000, Value_greater_20000.01)
Contract Start Semester	StartSemesterGroup	Contract start period (1st semester or 2nd semester)
Payment / Contract Duration	DurationGroup	Duration grouped (Monthly, Quarterly, Semi-annual, Annually)
Premium Change Band	PremiumChangeBand	Premium change grouped (Decrease_gt_20pct, Decrease_20_to_0pct, Increase_0_to_20pct, Increase_gt_20pct)
Engine Capacity Band	CCGroup	Vehicle cubic centimetres capacity (CC_0_499, CC_500_999, CC_1000_1499, CC_greater_1500)

3.1.2 Definition of the Dependent Variable: Customer Churn

Customer churn is an event when the customer ends his insured relation with the insurance aggregator, being detected by absence of a valid paid renewal and replacement contract within 30 days of the contract expiration. This definition is more sophisticated than the detection of an unpaid renewal proposal since there are cases that at first such activity would be detected for administrative/procedural reasons and later followed by a valid paid contract. That means only renewals, which are not followed by a valid paid contract within 30-day time frame in a production system, are finally labelled as churn. The cases that had payment delays, renewal reissues or prolongations through the start of a new paid contract within 30 operational days are excluded from final analysis.

3.1.3 Independent Variables and Feature Selection

The independent variables considered in this dissertation have been selected from prior research studies on insurance churn/retention. These variables were selected because they were hypothesized to be available at or very close to the time of the renewal decision. Predictors considered for the current study were classified into four categories as Customer profile, Financial, Product and Channel-related variables.

3.1.4 Customer Profile Variable

The customer profile variables describe the relationship between the customer and the aggregator, and include the customer's age upon renewal, the tenure, customer category and customer postcode. Tenure has specific relevance in our sample as customers staying longer in a program are presumed more familiar with and they place greater trust in the provider, both factors influence their likelihood of churning out of the programme (Reinartz & Kumar, 2000). Customer category describes on-line and off-line customers while postal codes may also show geographical differences on risk, price and condition.

3.1.5 Financial Variables

Financial variables describe the financial properties of the policy at the time of renewal and are assumed to be a major driver for churn. They consist of the premium per day, which makes policies with different lengths comparable, as well as the percentage difference between the new and the previous policy premium. Prior studies and market experience lead to the assumption that higher price increases will be more strongly associated with churn behaviour, however, individual customer's tolerance towards premium changes may differ (Peter C. Verhoef, 2003; Reinartz & Kumar, 2000).

3.1.6 Product Variables

Product variables relate to the specific characteristics of the insured risk as well as the offering's value. This includes policy duration, usage type, vehicle age, the number of other vehicles held by the customer, as well as the number of attachments included in the policy, such as roadside assistance or legal protection coverage. The analysis further distinguishes between different package categories: basic coverage, fire-theft coverage and complete (full casco) coverage, as this may be relevant for value perception and retention behavior.

3.1.7 Channel Variable

The channel variable includes online and offline customers. The first type interacts directly with the aggregator's on-line platform to organize their insurance search, comparability and purchase process themselves. The second one deals with an insurance broker or advice who in turn utilizes the system to support the customer through this insurance process. Research has shown that channel choice is associated with price sensitivity, switching, trust and loyalty as well as that it serves as predictor of churn Gensler et al. (2012).

3.2 Data Preprocessing and Preparation

The purpose of this stage was to clean the data to perform a strong churn modelling and to render the empirical analysis reliable. The operation data have been checked for missing, inconsistent, outlying and duplicate or ambiguous values on renewal. This last test has been critical as the renewal’s mechanics, delayed payments and changes of categories might have affected the observed outcomes. Categorical variables were coded where necessary, while continuous variables were kept in their original scale.

3.2.1 Handling Missing Values

All the variables were checked for missing values. Given the minimal number of missing observations in the dataset used in this study, no extensive imputation methods were needed. Less than 100 rows with missing or inconsistent data were removed for data quality purposes.

3.2.2 Outlier Detection

Every continuous variable checked for extreme values. Extreme values identified and assessed and if those values were removed but if there was no data entry error then those values retained to preserve real word variability of the dataset, especially premiums.

3.2.3 Variable Transformation

All variables used in a logistic regression model will be treated as dummy variables, except for one category which serves as a reference group. Continuous variables remain on their original scale.

3.3 Logistic Regression Model Specification

3.3.1 Rationale for Logistic Regression

Logistic regression is a statistical technique to model binary responses, and for this reason it was selected for this analysis. It is the most widely used statistical technique and is extensively applied in research for the analysis of churn (Hosmer et al., 2013). Other statistical methods or machine learning models do not offer clear interpretation of the parameter estimates that offered by logistic regression which can be easily interpreted by managers and decision makers in an organization with no statistical or mathematical background.

3.3.2 Model Formulation

The logistic regression model estimates the probability that customer *i* will churn at renewal:

$$P(Churn_i = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ki})}},$$

where $\boldsymbol{X_i} = (X_{1i}, \dots, X_{ki})$ represents the vector of independent variables for customer i, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$ represents the estimated coefficients of the model.

3.3.4 Interpretation using odds ratios

Odds of an event is defined as the ratio of the probability of the event happening to the probability of the event not happening. While probability is expressed on a scale of 0 to 1, odds express how many times more likely something is to happen than not:

$$odds = \frac{P}{1-P}$$

Odds ratios (OR) are used to interpret the coefficient estimate of a predictor from logistic regression, showing the change in the odds of an event (such as churn) for an one-unit change for continuous variables in the predictor (such as mikta_per_day), while holding all other predictors in the model constant. In other words, it measures how much the dependent variable (churn) is likely to be changed (either a positive or a negative change) given one unit change in an independent variable (such as mikta_per_day), holding all other independent variables constant.

Categorical variables entered the model using dummy variables and the odds ratio represents the change in the odds of churn given that category versus the reference category. If the odds ratio is greater than 1, it increases the odds of churn, if it is less than 1, it decreases the odds of churn. The factor change is equivalent to the odds ratio compared to the omitted reference group.

Also, the interpretation of the effects of the variables on the dependent variable is more direct and intuitive, because besides the effects' direction, there is also an intuitive interpretation for the effect's magnitude. Odds ratios are very used in applied research, especially because they are interpreted by non-technical people such as managers or decision-makers (Hosmer et al., 2013).

3.4 Model Estimation and Evaluation

3.4.1 Model Evaluation and Performance Metrics

Because of the binary nature of the dependent variable, model performance is assessed in terms of several complementary metrics. We measure accuracy to understand the overall rightness of a model, while precision and recall help us to understand how well the model balances false positives vs. false negatives. The performance of the model is then evaluated using a ROC (Receiver Operating Characteristic) curve and Area Under the Curve (AUC), to see how well it separates churners and non-churners at different cut-off points. A confusion matrix provides a more detailed summary of prediction results for each class. It is important to use multiple evaluation criteria, as churn data are typically imbalanced with renewals much more common than cancelations (Verbeke et al., 2011; Baesens, 2014).

Because of the binary nature of the dependent variable, model performance is assessed in terms of several complementary metrics. We measure accuracy to understand the overall rightness of a model, while precision and recall help us to understand how well the model balances false positives vs. false negatives. Model performance evaluation was done based on the ROC curve and the AUC because it aims to differentiate churners and non-churners at all possible cut-off points. A confusion matrix is also given as it presents a more detailed view of our prediction results, divided by class.

The evaluation cannot be based on a single criterion, especially when we deal with imbalanced churn data (i.e., renewals vs. cancelations) (Verbeke et al., 2011; Baesens, 2014).

3.4.2 Methodological Limitations

The proposed approach has few limitations that must be mentioned. First, analysis is done on past data and then we cannot incorporate future changes of customer's behaviour. The second limitation is that there are factors such as customers attitude, emotional part, fairness perception that could not directly measured by analysing only transactional data. For these reasons, the effect of these variables was not explicitly put inside of the model. Last one is that the study was applied on data of one insurance aggregator and results could be difficult generalized to other markets and business cases. The above limitations are taken into consideration when exploring the results. Discussion part highlights the limits as direction for further research.

3.4.3 Summary of the Methodological Framework

In this chapter the research design, data source, variable descriptions and the applied analytical approach to investigate customer churn using an example of an insurance aggregator have been discussed. The pre-processing, the specification of the logit regression model as well as the measures for model assessment have been explained. Altogether, this comprises the setup for the empirical study, which will be conducted in the next chapter.

4. Data Analysis and Empirical Results

4.1 Descriptive Analysis of the Dataset

4.1.1 Overview of the Sample

In this chapter, we will see the presentation of the empirical findings which are based on the dataset described in previous chapter. The conduction of a descriptive analysis of the dataset must proceed with the predictive modelling and the hypothesis analysis. Descriptive analysis will provide an understanding of the consistency of the sample as also the key characteristics which will show the biases might occur or imbalances that need to be identified and resolved at the beginning (Hair et al., 2019).

The dataset is fully anonymized and consists of insurance policy details from an insurance aggregator over a period of five years. Each record corresponds to a policy at a given point in time and includes whether the policy has been renewed or cancelled. This data structure allows us to analyse the binary effect of churn associated with policy renewal decisions and is consistent with the practices used for churn research.

The dataset includes online and offline customers, which shows the dual operating model of the insurance aggregator. Online customers use the aggregator’s online platform and follow the digital process to select the appropriate insurance policy that meets their needs, while offline customers follow the same process but with the support of an agent or broker who will use the platform to provide the appropriate offer based on the customer’s needs. These two groups of customers can be compared for their behaviour across these distribution channels. Table 2 shows that the dataset includes 174116 policy observations covering January 2021 to February 2025. The overall churn rate is 19,10%. Of all the policies included in the dataset, online channels account for 82,25% and offline channels account for 17,75%.

Table 2: Sample overview

Metric	Value
Policy observations	174116
Observation starts	2021-01-01
Observation end	2025-02-28
Overall churn rate (%)	19,10
Online share (%)	82,25
Offline share (%)	17,75

4.1.2 Time Horizon and Observation Window

The data covers the period of customer insurance policies starting from January 2021 to February 2025, which gives us the picture of multiple renewals of a policy within these consecutive years and allows us to see the behaviour of customers and analyse it in these renewal periods. The long period in years allows us to analyse churn in a better way because the specificity of insurance policies is associated with the expiration of the policy and less often with a continuous period.

Using data that covers a long period of time enhances the analysis because changes in customer behaviour can be recorded and thus reduces the risk that the results are due to short-term market fluctuations or other events that can directly affect the purchase of an insurance policy. Studies have shown that in churn modelling, the choice of a long time is required as customer loyalty and sensitivity to fluctuations in insurance policy prices can change as the duration of the relationship increases (Neslin et al., 2006).

4.1.3 Distribution of Customers by Channel (online vs offline)

An important factor in the data is the separation of customers into two independent categories, online and offline. From the initial examination of the data, as shown on Table

2, we observe that online customers constitute most of the portfolio in the sample, over 80% of the registrations. This also shows us the need for the expansion of digital distribution in the insurance market.

On the other hand, while offline customers are less than 20% of the portfolio (Table 3), they have lower churn rates and are less volatile than online customers. Average tenure is similar across the two groups but the reduced variance in offline customers suggests a stronger relationship with the insurer, as shown on Table 3. This supports our idea that human contact contributes to retention Gensler et al. (2012).

Table 3 : Client_cat Tenure and Churn Rate

Client_cat	N	Average of TenureYears	StdDev of TenureYears	Churn
Online	143202	2,47	2,06	19,62%
Offline	30914	2,43	1,87	16,67%

The descriptive data at this stage is not intended to find the reason why this is happening but it is an initial feeling that the source from which the customer obtained their insurance policy is an important indicator and can be linked to the customer's behaviour during the renewal period of their insurance policy and this justifies its importance as an explanatory variable in the models.

4.1.4 Distribution by Insurance Company and Product Type

The dataset consists of contracts from 22 insurance companies. The list was segmented and categorized to ensure a more robust and interpretable analysis purposes. Consequently, the segmenting resulted in three groups: large, medium and small insurance companies, which hold more than 10%, between 5 and 10% and less than 5% of the total policies, respectively. In percentage terms, large companies account for 49,34%, medium companies represent 38,25%, while small companies made up only 12,41% of total policies, as shown on Table 4. The stated segmentation demonstrates that the number of insurance policies is mainly consisted of large companies. Also, it becomes clear that customers are served by insurers unevenly. At a more descriptive level, the segmentation is meaningful from a churn perspective as we observed that there were no important differences in churn rates among small, medium and large companies with an average of 18,77% and 19,31% and 19,01% respectively.

Following, we will proceed with the direct analysis of churn, which will explore the relationship between churn and different variables such as customer channels, product characteristics and pricing-related variables.

Table 4 : Distribution of policies by insurance company

company_code	Count	% of total	Churn Rate
Large_company	85912	49,34%	19,01%
Medium_company	66601	38,25%	19,31%
Small_company	21603	12,41%	18,77%
Grand Total	**174116**	**100,00%**	

Similar behaviour is observed in the type of insurance package. The insurance aggregator data sample shows that most insurance policies consist largely of the basic package at approximately 84%, while the fire and theft package constitutes 12% and the mixed coverage at approximately 4%, as shown on Table 5. Churn also differs between these types of packages with the basic package having the highest churn rate among them.

Table 5 : Distribution of policies by insurance package

packet_cat	Count	% of total	Churn Rate
BASIC	146319	84,04%	19,73%
FULL CASCO	6663	3,83%	16,00%
FIRE-THEFT	21134	12,14%	15,70%
Grand Total	**174116**	**100%**	

Also, the duration of the contract shows a large difference in the churn rate as the data shows that shorter contract durations show a high churn rate compared to longer durations. In general, the above shows us that the content of the contracts greatly affects the churn rate.

Table 6 : Distribution of DurationGroup

DurationGroup	Count	% of total	Churn Rate
Monthly	9202	1,74%	24,36%
Quarterly	44304	16,72%	20,36%
Semi-annual	50210	28,42%	17,61%
Annually	70400	53,13%	18,67%
Grand Total	**174116**	**100,00%**	

Table 6 shows that many of the customers, over 53%, prefer annually duration for their policies, while over than 28% prefer the six months duration and the rest prefer three months and monthly duration policies. On Table 7, we see that the majority of contracts issued on the 1st semester and the churn rate is higher in this semester than the 2nd semester.

Table 7 : Distribution of policy start month

StartMonth	Count	% of total	Churn Rate
1st_semester	102323	58,77%	19,59%
2nd_semester	71793	41,23%	18,39%
Grand Total	**174116**	**100,00%**	

4.1.5 Summary Statistics of Key Variables (tenure, premium, duration, renewals)

The dataset also includes quantitative variables used for modelling which are shown to Table 8. The continuous variables are the policy premium represent by mikta_per_day in euros, the time the customer stays in the insurance aggregator represented by TenureYears, the total contracts the customer has with the insurance aggregator represented by variable VehiclePoliciesCount, the total active vehicles the customer has with the insurance aggregator represented by variable ActiveVehiclesCount, how many extra covers have the customer selected at the start of policy represented by ExtraCoversCount variable as long as the vehicle age and the customers age at start of the policy.

Table 8: Summary statistics

Variable	N	Mean	Std. Dev.	Min	Median	Max
mikta_per_day	174116	0,45	0,28	0,03	0,39	7,60

TenureYears	174116	2,46	2,03	0,00	2,00	11,07
VehiclePoliciesCount	174116	3,63	3,65	0	2	38
ActiveVehiclesCount	174116	1,27	0,85	0	1	20
ExtraCoversCount	174116	0,81	1,14	0	0	4
VehicleAge	174116	15,08	7,76	0	16	71
Age_at_start	174116	45,37	13,58	18,00	45,11	99,00

The descriptive statistics show that there is a large heterogeneity within the dataset. The premium per day show that the average price of the policy is around 0,45 euros, the median is 0,39 euros with a standard deviation of 0,28 euros. These data show us that there is a large dispersion between the policies due to the existence of many different pricing policies within the sample. Based on the descriptive statistics, it is observable that there are important differences in the traits of policies. The variable TenureYears shows that customers tend to have moderate experience with the company (average tenure 2.46 years) but with a large degree of dispersion. This can suggest a wide range of customer segments with varying degrees of tenure. VehiclePoliciesCount and ActiveVehiclesCount reveal that while customers tend to have multiple policies, they tend to have only one vehicle active at any point in time, showing a concentrated portfolio. ExtraCoversCount is relatively low with a median of zero, suggesting that customers tend to be conservative and avoid extra covers. VehicleAge is relatively high with an average of 15 years, which could show a higher likelihood of claims and, as a result, higher premiums. Age_at_start shows a fairly even distribution of customers with a central tendency towards middle-aged people with large dispersion among policyholders.

The general picture from the dataset is that they consist of heterogeneous observations in terms of pricing, duration and customer history. This shows us that there is a need for multivariate analysis as simple averages will not be good indicators to highlight the behavioural differences between parts of the dataset.

In addition to the features of the sample described above, it is important to look at the distribution of key categorical variables to obtain further insights on the structure of the dataset and the composition of the insurance portfolio.

Table 9: Distribution of customers by channel

Client_cat	Count	% of total
Online	143202	82,25%
Offline	30914	17,75%
Grand Total	**174116**	**100,00%**

Table 9 shows that the dataset consists of both online customers and offline customers where online customers are more than 80% of the sample which shows the strong presence of digital distribution in the insurance aggregator.

Table 10 : Distribution of usage

Usage	Count	% of total
Car	123218	70,77%
Motorcycle	44511	25,56%
Truck	6387	3,67%
Grand Total	**174116**	**100,00%**

Table 10 shows that more than 70% of the policies are related to cars and second are the motorcycles with over 25% and the rest referred to trucks

Table 11: Geographic distribution of customers

ZipCode	Count	% of total
Attiki	81741	46,95%
Thessaloniki	23453	13,47%
Other_region	68922	39,58%
Grand Total	**174116**	**100,00%**

Table 11 shows that the majority of policies belong the region of Attiki and Thessaloniki represent more than 60% of the policies while Other_region represents less than 40% of the policies which shows higher penetration into urban areas than rural areas.

Table 12 : Distribution of customer contact types

ContractType	Count	% of total
Email_contact	171836	98,69%
Phone_contact	2280	1,31%
Grand Total	**174116**	**100,00%**

Table 12 shows that over 98% of the customers have an e-mail address and the rest provided a mobile or phone number. Therefore, digital communication channels are the most common way of contacting customers and mobile communication in the form of calls and notifications also plays a key role in this process. The variability in this metric is limited and therefore, its explanatory power in the modelling process expecting to be low. Although it does provide a useful insight into the aggregator's operational communication strategy, it may not differ significantly enough to differentiate churn behaviour of customers.

Table 13 : Distribution of vehicle value

Value	Count	% of total
Value_up_to_10000	132897	76,33%
Value_10000.01_to_20000	33305	19,13%
Value_greater_20000.01	7914	4,55%
Grand Total	**174116**	**100,00%**

Table 13 shows that the majority of the vehicles has a value less than 10000 euros (Value_up_to_10000), corresponding to the interval [0, 10,000], vehicles value from 10001 to 20000 euros consists 19,13% of the dataset (Value_10000.01_to_20000), corresponding to the interval (10,000, 20,000], and over 20000,01 euros are less than 5% of the dataset, corresponding to the interval [20,000, +∞).

Table 14 : Distribution of PremiumChangeBand

PremiumChangeBand	Count	% of total
Decrease_gt_20pct	8314	1,76%
Decrease_20_to_0pct	43619	18,50%
Increase_0_to_20pct	112722	71,71%
Increase_gt_20pct	9461	8,03%
Grand Total	**174116**	**100,00%**

Table 14 shows the frequency of policies according to the change of premium relative to the previous contract: Decrease_gt_20pct: corresponding to the interval (-∞, -20%], Decrease_20_to_0pct: corresponding to the interval (-20%, 0%], Increase_0_to_20pct: corresponding to the interval (0%, 20%) and Increase_gt_20pct: corresponding to the interval [20%, +∞)

Increase_0_to_20pct is the largest group, accounting for 71,71% of the sample, suggesting that many contracts were either not decreased or increased by a maximum of 20% relative to the previous contract. Decrease_20_to_0pct is the second largest group, with 18,50% of all policies and refers to policies with a premium less than the previous contract but a difference of at least 20%. Much fewer policies have experienced a large change in premium at renewal: 1,76% of policies have seen a decrease greater than 20% and 8,03% of policies have increased by more than 20%. Overall, Table 15 suggests that most renewal premia tend to lie close to the original premium, more exactly between -20% and +20%.

Table 15 : Distribution of CCGroup

CCGroup	Count	% of total
CC_0_499	35820	7,66%
CC_500_999	19867	8,50%
CC_1000_1499	81905	52,57%
CC_greater _1500	36524	31,26%
Grand Total	**174116**	**100,00%**

Table 15 shows that the majority of vehicles fell under the 1000 -1499 cc category, making it the core segment of the portfolio. Among the engine sizes, while those were less common in higher categories, there was an important share of them. That means that the portfolio represented a diverse mix of vehicle types.

4.1.6 Initial Insights and Implications for Further Analysis

The descriptive analysis shows us some initial indications that are connected to the central idea of the thesis. Initially, the sample is heterogeneous within the insurance aggregator's portfolio and substantial differentiation is observed between channels, companies and customer characteristics. The overall churn rate is quite high at 19,1% and established within the portfolio which shows that customer churn is a substantial phenomenon that can be further investigated.

Then we observe differentiation between insurers, insurance products, insurance policy duration, insurance policy price level as well as previous contracts held by a customer which shows us that churn cannot be attributed to a single factor but is linked to many dimensions.

The above suggests that we need to move from descriptive statistics to modelling. In the following chapters we will look at logistic regression and the behaviour of all variables against churn before moving on to more robust predictive methods.

4.2 Churn Rate Measurement and Comparison

4.2.1 Definition and Measurement of Churn Rate

An important point in the analysis of customer retention is the precise definition and measurement of the churn rate. In insurance markets, the churn rate is associated with the non-renewal of the insurance policy at the expiration of the previous one or with cancellation for various reasons such as sale or destruction of the vehicle during the term of the contract. However, in practice the definition of churn needs some modification because some contracts may appear unpaid or not renewed for administrative reasons, but the customer continues his relationship with the insurance aggregator or broker through a subsequent new paid contract.

In this dissertation, the churn rate is defined as the failure of a contract to be paid or replaced with a new contract within a period from the expiration of the contract. Specifically, a contract is defined as cancelled if the customer does no have valid paid renewal his contract by the end of the contract or does not replace it with a new contract within 30 days of the end of the contract. If the customer proceeds to purchase a new contract within 30 days, then the proposal to renew the insurance policy that was issued is not counted in the churn because there is a new contract and the renewal case is not calculated. In the opposite case, if there is no new paid-off contract within 30 days, then the renewal case counts in the churn.

The adjusted definition of the churn rate is important for contracts that are related to renewal. In the dataset there are records that initially appear invalid but are followed by paid-off contracts, this means that the relationship between the customer and the insurance aggregator or broker has not changed. If these cases are managed as churn, then this will technically increase and will introduce an error in the analysis. Therefore, only those outstanding renewals without a paid policy within 30 days of expiration are considered churn.

Formally, the churn rate is calculated as:

$$\text{Churn Rate} = \frac{Non\ Paid\ Renewals - Non\ Paid\ Renewals\ followed\ by\ paid\ policy}{\text{Total number of valid renewal } - Non\ Paid\ Renewals\ followed\ by\ paid\ policy},$$

where the numerator defines the policies that have not been renewed or are not followed by a paid policy within 30 days and the denominator all enrolments except those that have an enrolment with a paid policy in 30 days.

This is consistent with the literature on churn rate, which reflects the continuation of the relationship between a customer and a company or its termination despite the existence of an intermediate management status (Neslin et al., 2006). In this way, churn is used in this thesis, which considers the relationship between the customer and the company, excluding technical and procedural interferences.

4.2.2 Overall Churn Rate of the Aggregator

Using the definition of churn described previously, the total churn rate of the aggregator is estimated to be approximately 19,1% of the final sample. Specifically, there are 174116 valid records in the total dataset with start date 01/01/2021 to 28/02/2025, of which 33250 are cancelled contracts and 140866 are paid-off contracts as Table 16 shows. Our result shows that the churn rate is characteristic of an insurance aggregator portfolio. This shows us that one in five contracts do not pay off at maturity or within the specified 30-day period, which is significant enough to justify the creation of a statistical model that can predict.

Table 16 : Overall Churn rate

Class	Count	Share (%)
Non-churn	140866	80,90
Churn	33250	19,10
Total	174116	100,00

4.2.3 Churn Rate Comparison: Online vs Offline Customers

Using the dataset and the Client_cat variable that defines online customers as 0 and offline customers as 1 allows us to make an initial comparison of churn behaviour between them. The descriptive tabulation of the analytical dataset shows us that these two groups, in addition to their size, also differ in their churn rate. Online customers have a churn rate of 19,62% while offline customers have a lower rate of 16,67% as shown on Table 17. The difference of 2,95% gives us the initial hypothesis that the churn rate is not uniform between these 2 groups and may reflect the differences in their behaviour in the policy market, their sensitivity to price changes from insurance companies, and the way they purchase and manage policies.

The above is an indication that supports the central argument of the dissertation regarding customer behaviour depending on the different channels they come from. However, this channel difference should not be considered as a cause and effect may be due to the composition of the insurance aggregator's portfolio, products, pricing or customer characteristics of each channel. The empirical evidence is further examined through multivariate modelling which will be discussed below.

Table 17 : Churn rate by channel

Channel	Observations	Churn_Rate_pct %
Online	143202	19,62
Offline	30914	16,67

4.2.4 Churn Rate by Policy Percentage Change

Table 18 illustrates the connection between premium changes and churn rate. The churn rate remains closer to the portfolio average when the premium change ranges between moderate decreases and moderate increases, namely within the intervals [-20%, 0%) and [0%, 20%]. But a drop of over 20%, which ranges from negative infinity to -20% (-∞, -20%), links to a much smaller number of customers leaving and a rise of over 20%, which ranges from 20% to positive infinity (20%, +∞), links to a much larger number of customers leaving. It appears that the number of customers leaving is most often at its lowest when premiums drop by more than 20% and at its highest when premiums rise by more than 20%. So, it appears that changes in premium level might be influencing whether people decide to renew their policies and this should be investigated further as part of the predictive modelling process.

Table 18 : Churn rate per percentage change in price groups

PremiumChangeBand	Observations	Churn_Rate_pct
Decrease_gt_20pct	8314	11,16%
Decrease_20_to_0pct	43619	21,05%
Increase_0_to_20pct	112722	18,24%
Increase_gt_20pct	9461	27,29%

4.3 Exploratory Analysis of Churn Drivers

4.3.1 Relationship between Churn and Customer Tenure

Customer tenure is one of the central variables of the churn rate analysis model because it shows us the length of time that the customer has been in a relationship with the insurance aggregator or broker. In general, it is observed that customers with a long-term relationship are more familiar with the renewal process and are less likely to change contracts or cancel it, while customers with a shorter relationship are more sensitive and react to temporary dissatisfaction or price change or to more competitive offers.

Our descriptive data shows a significant heterogeneity in tenure between the observations which show that they consist of both relatively new and quite old customers with an established relationship with the insurance aggregator or broker. This variable becomes very important for the analysis of the churn rate as it is directly linked to a low frequency of contract cancellation.

The data suggests that customers with longer periods tend to behave differently in terms of renewal decision compared with new customers. Also, the length of the relation seems to affect customers in deciding at renewal time.

4.3.2 Price Sensitivity and Premium Changes

Price-related variables play an important role in insurance churn rates, often the decision to renew a policy is influenced by whether the price and the value of the coverage package are affordable. In the insurance aggregator environment, customers have the ability to compare offers from many insurance companies very easily, which makes the premium amount and the amount of the renewal change very important factors for changing or cancelling the policy.

The data shows that changes in premium do impact churn, with customers more likely to leave at renewal if premiums are increased and important increases in premium being associated with higher levels of churn. High premium policies may be associated with higher churn.

These findings show us that there is a clear indication of price sensitivity in the portfolio. Customers are very likely to cancel the contract when the price is relatively more expensive or face large increases in renewal compared to the previous one contract. This shows that pricing policy affects not only revenue but also should be a central pillar in sustainability management.

4.3.3 Renewal History and Behavioural Patterns

Renewal history is an important variable in insurance churn rate analysis. The reason is that a customer who has renewed their insurance many times in the past is more likely to continue doing so in the future. Repeated renewals show that the customer is familiar with their insurer and has no particular incentive to switch.

Customers with a strong history of renewing through an aggregator may be more likely to continue to choose them as they continue to find value in their relationship. In other words, renewal history may act as a factor that supports customer retention in the portfolio.

This is also evident from the way the sample has been constructed. Variables such as VehiclePoliciesCount and has_previous_policy have been included precisely because it is important to distinguish customers who are appearing for the first time from those who already have a history with the insurer.

4.3.4 Channel-Specific Behavioural Differences

Our data also shows additional variables related to the behaviour between customers in different purchase or distribution channels of contracts. The variables Client_cat that define whether the customer came from an online process or an offline process as well as the ContractType that shows us the type of package for the contract that the customer chose, are variables that show that the behaviour of the churn rate is not uniform in the acquisition of the customer and in the way they interact.

Comparing online and offline customers, suggests that channel greatly impacted churn with offline customers exhibiting lower churn rates than online customers. This shows us that customers of category 1, i.e. offline customers appear to have a lower probability of churn compared to customers of category 0, i.e. online customers.

4.3.5 Preliminary Insights Before Modelling

The analysis of variables provides important conclusions that will be used or confirmed in the models below. The first and main conclusion is that churn does not occur randomly in the portfolio. On the contrary, it is associated with specific characteristics regarding the customer's relationship with the insurer, the pricing and the segment to which they belong. This shows us that the cancellation of a policy by a customer is not a single factor but a combination of complex combinations of them.

An important part is the long-term relationship of the customer with the insurance aggregator or broker. We observe that customers who have maintained their policies for a long time with them tend to remain in the duration of their policy renewal. Simply put, the older and more active a customer is, the less likely they are to change insurer.

The third factor and logically expected is the amount of the premium. The ability to compare many insurance companies in a short time allows customers to react easily and quickly to changes in insurance companies. At the same time, the channel through which the customers came, i.e. online or offline, shows that the churn rate is affected by this and shows that there is not a uniform distribution of it in the dataset.

All the above shows that it is necessary to proceed to a more complex analysis. In this section, we provided descriptive findings on the motivations that may be related to churn. In the following section we will present the results of a multivariate approach to examine the factors related to churn and their combined effect.

4.4 Logistic Regression Model for Churn Prediction

4.4.1 Model Specification and Dependent Variable

To examine the factors that affect customer churn in a multifactorial environment, we are going to use the logistic regression. In logistic regression, the dependent variable can take two values and in the case of churn, it will take 1 if the contract is categorized as churn or 0 otherwise. This model allows the probability of churn to be expressed as a function of multiple factors.

The final dataset consisted of 174116 observations, of which 33250 were classified as churn. The dataset was split by time for model development. Observations with policy start dates up to 31 December 2024 were kept as the main modelling sample, which resulted in 164524 records. The modelling sample was used for the train-test modelling procedure. The remaining 9592 records belonged to policies with start dates after 31 December 2024. Specifically, 5200 records belonged to January 2025, and 4392 records belonged to February 2025. During the empirical evaluation in the modelling section, the out-of-time test was performed on the January 2025 subset.

Formally, the model estimates the conditional probability that policy i will churn:

$$P(Y_i = 1 \mid X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}},$$

where $\boldsymbol{Y_i} = (Y_{1i}, \dots, Y_{ki})$ denotes the binary churn outcome and $\boldsymbol{X_i} = (X_{1i}, \dots, X_{ki})$ represents the vector of explanatory variables. The estimated coefficients indicate the direction of the relationship between each predictor and churn probability, while them exponentiated form is interpreted through odds ratios.

The logistic regression model is used here as the main interpretable baseline specification. Its purpose is not only predictive classification but also the identification of statistically meaningful associations between customer, pricing, and policy characteristics and the probability of churn.

4.4.2 Independent Variables and Feature Selection

The predictor variables used in the logistic regression model were selected on the basis of their relevance in an insurance setting, data that were readily available at the time of renewal and their expected relationship with customer churn discussed in the preceding sections.

The model considers several variables from 4 main categories: customer profile, financial characteristics, product attributes and channel-related factors. Examples include customer tenure, premium level, premium change, policy duration, vehicle characteristics, renewal history and distribution channel.

Although the model contains several continuous variables, it also includes several categorical predictors. These are encoded as a set of binary indicators, with one category for each variable taken as the base category, to represent contrasts between groups, for example, between different insurance companies, product types, usage types or customer channel.

The final model then combines this set of continuous explanatory variables with a selection of relevant dummy variables and serves to document the heterogeneity across customers, products and insurers. The present dissertation tries to take advantage of this approach in the setting of insurance aggregator data.

A wide range of variables is used to account for the multidimensional nature of churn behaviour because the actions of customers are driven by a mix of monetary, behavioural and product-related factors.

4.4.3 Estimation Results of the Logistic Regression

The logistic regression model was estimated using data as of December 31, 2024. The final dataset consists of 164524 observations and 30 explanatory variables and the model constant. The model ran normally and gave stable results, which means that the estimation process was completed without a problem.

Overall, the regression results, as shown on Figure 1, show that the model has statistical value. The likelihood ratio test is significant (LLR p-value = 0,000), meaning that the variables we used explain churn better than a model containing no factors. The Pseudo R^2 is estimated at 0,0574, which in itself may seem small, but is expected for a phenomenon like churn, which is influenced by many factors, many of which are not easy to measure. The important thing is that the model manages to detect part of the pattern behind customers' decisions.

The results that churn seems to be influenced by a combination of financial, behavioural, product-related and channel-specific factors. Some of the model's results show statistically important relations with the variable of interest.

Figure 1 : Logistic regression results

Logit Regression Results

Dep. Variable	Churn	No. Observations	164524
Model	Logit	Df Residuals	164494
Method	MLE	Df Model	29
Date	Sat, 25 Apr 2026	Pseudo R-squ.	0,0574
Time	13:12:39	Log-Likelihood	-75606,25
converged	TRUE	LL-Null	-80214,48
Covariance Type	nonrobust	LLR p-value	0,0000

Coefficient Table

Variable	Coef (β)	std err	z	p-value	CI 2.5%	CI 97.5%
const	-1,8143	0,0546	-33,2529	0,0000	-1,9212	-1,7073
ActiveVehiclesCount	-0,0729	0,0089	-8,2029	0,0000	-0,0903	-0,0555
Age_at_start	-0,0007	0,0005	-1,3943	0,1632	-0,0017	0,0003
CCGroup_CC_1000_1499	-0,1035	0,0365	-2,8372	0,0046	-0,1751	-0,0320
CCGroup_CC_500_999	0,0602	0,0299	2,0159	0,0438	0,0017	0,1188
CCGroup_CC_greater_1500	-0,0617	0,0389	-1,5860	0,1127	-0,1379	0,0145
Client_cat_1	-0,1770	0,0193	-9,1864	0,0000	-0,2148	-0,1392
CompanySizeGroup_Medium_company	-0,0274	0,0154	-1,7818	0,0748	-0,0575	0,0027
CompanySizeGroup_Small_company	-0,0201	0,0220	-0,9127	0,3614	-0,0632	0,0231
ContactChannelGroup_Phone_contact	0,5639	0,0546	10,3307	0,0000	0,4569	0,6709
DurationGroup_Monthly	0,9006	0,0333	27,0738	0,0000	0,8354	0,9658
DurationGroup_Quarterly	0,5771	0,0193	29,9006	0,0000	0,5393	0,6149
DurationGroup_Semi-annual	0,2386	0,0173	13,8217	0,0000	0,2048	0,2724
ExtraCoversCount	0,2141	0,0060	35,8163	0,0000	0,2023	0,2258
mikta_per_day	0,5504	0,0332	16,5787	0,0000	0,4853	0,6154
packet_cat_FULL CASCO	-0,3384	0,0425	-7,9598	0,0000	-0,4217	-0,2550
packet_cat_FIRE-THEFT	-0,1982	0,0237	-8,3450	0,0000	-0,2447	-0,1516
PremiumChangeBand_Decrease_gt_20pct	-0,9054	0,0389	-23,2821	0,0000	-0,9816	-0,8292
PremiumChangeBand_Increase_0_to_20pct	-0,2991	0,0159	-18,8560	0,0000	-0,3302	-0,2680
PremiumChangeBand_Increase_gt_20pct	0,3096	0,0280	11,0415	0,0000	0,2546	0,3646
RegionGroup_Other_region	0,1279	0,0152	8,4140	0,0000	0,0981	0,1577
RegionGroup_Thessaloniki	0,1226	0,0204	5,9983	0,0000	0,0825	0,1626
StartSemesterGroup_2nd_semester	0,0713	0,0130	5,4845	0,0000	0,0458	0,0968
TenureYears	0,2247	0,0043	52,0605	0,0000	0,2162	0,2331
Usage_Motorcycle	0,3435	0,0338	10,1638	0,0000	0,2773	0,4098
Usage_Truck	-0,2812	0,0406	-6,9259	0,0000	-0,3608	-0,2017
VALUE_Value_greater_20000.01	-0,1441	0,0378	-3,8166	0,0001	-0,2181	-0,0701
VALUE_Value_up_to_10000	0,0599	0,0197	3,0494	0,0023	0,0214	0,0985
VehicleAge	0,0067	0,0009	7,3172	0,0000	0,0049	0,0084
VehiclePoliciesCount	-0,2316	0,0034	-68,0070	0,0000	-0,2383	-0,2249

Table 19 : Odds Ratio per Variable

Variable	coef	**Odds Ratio**
ActiveVehiclesCount	-0,0729	0,9297
Age_at_start	-0,0007	0,9993
CCGroup_CC_1000_1499	-0,1035	0,9016
CCGroup_CC_500_999	0,0602	1,0621
CCGroup_CC_greater_1500	-0,0617	0,9402
Client_cat_1	-0,1770	0,8378
CompanySizeGroup_Medium_company	-0,0274	0,9730
CompanySizeGroup_Small_company	-0,0201	0,9801
ContactChannelGroup_Phone_contact	0,5639	1,7575
DurationGroup_Monthly	0,9006	2,4611
DurationGroup_Quarterly	0,5771	1,7809
DurationGroup_Semi-annual	0,2386	1,2695
ExtraCoversCount	0,2141	1,2387
mikta_per_day	0,5504	1,7339
packet_cat_FULL CASCO	-0,3384	0,7129
packet_cat_FIRE-THEFT	-0,1982	0,8202
PremiumChangeBand_Decrease_gt_20pct	-0,9054	0,4044
PremiumChangeBand_Increase_0_to_20pct	-0,2991	0,7415
PremiumChangeBand_Increase_gt_20pct	0,3096	1,3629
RegionGroup_Other_region	0,1279	1,1365
RegionGroup_Thessaloniki	0,1226	1,1304
StartSemesterGroup_2nd_semester	0,0713	1,0739
TenureYears	0,2247	1,2519
Usage_Motorcycle	0,3435	1,4099
Usage_Truck	-0,2812	0,7548
VALUE_Value_greater_20000.01	-0,1441	0,8658
VALUE_Value_up_to_10000	0,0599	1,0618
VehicleAge	0,0067	1,0067
VehiclePoliciesCount	-0,2316	0,7933

From Figure 1 and Table 19 we see the following results:

1. Variable mikta_per_day has a coefficient of 0,5504 which means that an increase of 1 euro in the daily premium (365 euros in yearly base) is associated with an increase of approximately 73,4% in the odds of churn ($e^{0,4675} \cong 1,734$), holding all other variables constant.
2. Variable PremiumChangePct :
    a. PremiumChangeBand_Decrease_gt_20pct has a coefficient of -0,9054, which means that a decrease in premium greater than 20% is associated with a reduction of approximately 59,6% in the odds of churn compared to customers which premium decreased between -20% up to 0% ($e^{-0,9054} \cong 0,404$), holding all other variables constant. This indicates that large price reductions significantly improve customer retention.
    b. PremiumChangeBand_Increase_0_to_20pct has a coefficient of -0,2991, which means that a moderate increase in premium (0%–20%) is associated with a reduction of approximately 25,9% in the odds of churn compared to customers which premium decreased between -20% up to 0% ($e^{-0,2991} \cong 0,741$), holding all other variables constant. This suggests that small price increases are generally tolerated by customers and do not lead to higher churn.
    c. PremiumChangeBand_Increase_gt_20pct has a coefficient of 0,3096, which means that a large increase in premium greater than 20% is associated with an increase of approximately 36,3% in the odds of churn compared to customers which premium decreased between -20% up to 0% ($e^{0,3096,} \cong 1,363$), holding all other variables constant. This indicates that significant price increases strongly contribute to customer behaviour.
3. Variable VehiclePoliciesCount has a coefficient of -0,2316 which means that each additional policy decrease odds of churn approximately 20,70% ($e^{-0,2316} \cong 0,7933$), holding all other variables constant.
4. Variable TenureYears has a coefficient of 0,2247 which means that each additional year the customers stay in aggregator increase odds of churn approximately 25,2% ($e^{0,2247} \cong 1,252$), holding all other variables constant.

5. Variable DurationGroup :
    a. Variable DurationGroup_Monthly has a coefficient of 0,9006, which means that monthly contracts are associated with an increase of approximately 146% in the odds of churn compared to customers with annual policies ($e^{0,9006} \cong 2{,}461$), holding all other variables constant. This indicates that very short-term contracts are highly prone to customer switching.
    b. Variable DurationGroup_Quarterly has a coefficient of 0,5771, which means that quarterly contracts are associated with an increase of approximately 78,1% in the odds of churn compared to customers with annual policies ($e^{0,5771} \cong 1{,}781$), holding all other variables constant. This suggests that shorter commitment periods lead to higher churn compared to longer-term policies.
    c. Variable DurationGroup_Semi-annual has a coefficient of 0,2386, which means that semi-annual contracts are associated with an increase of approximately 26,9% in the odds of churn compared to customers with annual policies ($e^{0,2386} \cong 1{,}269$), holding all other variables constant. This shows that even medium-term contracts have higher churn compared to annual policies, but to a less than shorter durations.
6. Variable ActiveVehiclesCount has a coefficient of -0,0729 which means that each additional insured vehicle decreases odds of churn approximately 7,03% ($e^{-0,0729} \cong 0{,}927$), holding all other variables constant.
7. Variable ExtraCoversCount has a coefficient of 0,2141 which means that each additional cover increase odds of churn approximately 23,9% ($e^{0,2141} \cong 1{,}239$), holding all other variables constant.
8. Variable Age_at_start has a coefficient of -0,0007 which means that each additional year in customer's age decrease odds of churn approximately 0,1% ($e^{-0,0007} \cong 0{,}993$), holding all other variables constant.
9. Variable VehicleAge has a coefficient of 0,0067 which means that each additional year in vehicle's age increase odds of churn approximately 0,7% ($e^{0,0067} \cong 1{,}0067$), holding all other variables constant.
10. Variable Client_cat_1 has a coefficient of -0,177 which means that offline customers have 16,22% lower in the odds of churn compared to online customers ($e^{-0,177} \cong 0{,}8378$), holding all other variables constant.

11. Variable packet_cat :
    a. packet_cat_FULL CASCO has a coefficient of -0,3384 which means customers with full casco policies have approximately 28,7% lower odds of churn compared to those have basic package ($e^{-0,3384} \cong 0,7129$), holding all other variables constant.
    b. packet_cat_FIRE-THEFT has a coefficient of -0,1982 which means customers with full casco policies have approximately 18% lower odds of churn compared to those have basic package ($e^{-0,1982} \cong 0,8202$), holding all other variables constant.
12. Variable usage:
    a. Usage_Motorcycle has a coefficient of 0,3435 which means customers with motorcycles have approximately 41% higher odds of churn compared to customers with car ($e^{0,3435} \cong 1,4099$), holding all other variables constant.
    b. Usage_Truck has a coefficient of -0,2812which means customers with motorcycles have approximately 24,5% lower odds of churn compared to customers with car ($e^{-0,2812} \cong 0,7548$), holding all other variables constant.
13. Variable VALUE:
    a. VALUE_Value_greater_20000,01 has a coefficient of -0,1441 which means customers with vehicle value greater than 20000 euros have approximately 13,4% less odds of churn compared to customers with vehicle value between 10000 to 20000 euros ($e^{-0,1441} \cong 0,8658$), holding all other variables constant.
    b. VALUE_Value_up_to_10000 has a coefficient of 0,0599 which means customers with vehicle value less than 10000 euros have approximately 6,2% higher odds of churn compared to customers with vehicle value between 10000 to 20000 euros ($e^{0,0599} \cong 1,0618$), holding all other variables constant.
14. Variable StartSemesterGroup:
    a. StartSemesterGroup_2nd_semester has a coefficient of 0,0713 which means customers with start date in the second semester have approximately 7,4% higher odds of churn compared to customers with start date in the first semester ($e^{0,0713} \cong 1,0739$), holding all other variables constant.

15. Variable RegionGroup:
    a. RegionGroup_Other_region has a coefficient of 0,1279 which means customers in other regions have approximately 13,7% higher odds of churn compared to customers in Attiki region ($e^{0,1279} \cong 1,1365$), holding all other variables constant.
    b. RegionGroup_Thessaloniki has a coefficient of 0,1226 which means customers in other regions have approximately 13% higher odds of churn compared to customers in Attiki region ($e^{0,1226} \cong 1,1304$), holding all other variables constant.
16. Variable CCGroup:
    a. CCGroup_CC_1000_1499 has a coefficient of -0,1035 which means customers with vehicle engine between 1000 to 1499 cc have approximately 9,8% less odds of churn compared to customers with vehicle vehicle engine less than 500 cc ($e^{-0,1035} \cong 0,9016$), holding all other variables constant.
    b. CCGroup_CC_500_999 has a coefficient of 0,0602 which means customers with vehicle value less than 10000 euros have approximately 6,2% higher odds of churn compared to customers with vehicle vehicle engine less than 500 cc ($e^{0,0602} \cong 1,0621$), holding all other variables constant.
    c. CCGroup_CC_greater_1500 has a coefficient of -0,0617 which means customers with vehicle engine greater than 1499 cc have approximately 6% less odds of churn compared to customers with vehicle vehicle engine less than 500 cc ($e^{-0,0617} \cong 0,9402$), holding all other variables constant.
17. Variable CompanySizeGroup:
    a. CompanySizeGroup_Medium_company has a coefficient of -0,0274 which means customers in medium companies have approximately 2,7% less odds of churn compared to customers in large companies ($e^{-0,0274} \cong 0,973$), holding all other variables constant.
    b. CompanySizeGroup_Small_company has a coefficient of -0,0201 which means customers in small companies have approximately 2% less odds of churn compared to customers in large companies ($e^{-0,0201} \cong 0,9801$), holding all other variables constant.

18. Variable ContactChannelGroup:
    a. ContactChannelGroup_Phone_contact has a coefficient of 0,5639 which means customers only provide phone or mobile as a contact type have approximately 75,7% higher odds of churn compared to customers provide an email as contact type ($e^{0,5639} \cong 1{,}7575$), holding all other variables constant.

The results show that many variables contribute to the model and highlight the most influential drivers of churn, which include but are not limited to pricing, customer relationship history and product characteristics. These findings show that pricing, customer relationship history and product characteristics jointly shape customer decisions at renewal. Churn is not driven by a single part, and the results demonstrate the complex and multidimensional nature of customer retention in insurance aggregators highlighting how financial, behavioural and product-related elements interact simultaneously.

4.4.4 Statistical Significance of Predictors

The significance of the estimated coefficients shows us that the relationship between the explanatory variables and churn is not the result of random variation within the sample we used from the insurance aggregator. Most of the coefficients of the explanatory variables in the logistic regression are statistically significant, which indicates that the model has reasonable and significant relationships regarding the composition of the policy, the pricing policy, the history between the customer and the company, the type of customer and the probability of churn.

Some variables are statistically significant at the 1% level. This fact confirms that churn is systematically related to many dimensions and does not depend on a single factor. Specifically, variables related to the previous relationship of the customer and the company, the design of the insurance policy, the pricing policy and the characteristics of the customer play a large role in explaining the variation in the churn result.

However, not all variables have the same strength of influence. For example, in category variables like Age_at_start, CompanySizeGroup and VehicleAge are not important. The effect can also vary by level or reference category. So, the effect is not the same for all insurers and segments.

Multicollinearity testing was performed as part of the feature selection criteria and statistical testing (Table 20). Most VIF scores were close to 1. VehiclePoliciesCount had the highest VIF score at around 1,59 while TenureYears had a score around 1,58. All VIF scores were within acceptable thresholds. As such, the multicollinearity results show no important connection between the variables in the final model. The coefficients are stable and each variable independently contributes to the ability of predicting policy churn. This in turn helps to improve the reliability and interpretability of the results.

Table 20 : VIF

	feature	VIF
2	VehiclePoliciesCount	1,597883
1	TenureYears	1,585298
5	VehicleAge	1,057234
0	mikta_per_day	1,054877
6	Age_at_start	1,049438
3	ActiveVehiclesCount	1,028466
4	ExtraCoversCount	1,023881

Overall, the significance pattern supports the empirical usefulness of the selected feature set. At the same time, statistical significance should not be interpreted as equivalent to economic importance. For this reason, the next section evaluates model performance more directly through classification metrics, discrimination measures, and validation tests.

4.5 Model Performance and Validation

4.5.1 Confusion Matrix and Classification Accuracy

To evaluate the predictive performance of the logistic regression model we should create a confusion matrix using the test dataset. This matrix will show as the classification into true positives (TP) and true negatives (TN) results and false positives (FP) and false negatives (FN) results. TP shows that the model predicted the churn correctly and TN shows that the model predicted the no churn correct. On the other hand, FP means that the model predicted churn and the result is no churn and FN means that the model predicted no churn and the result is churn. The main modelling sample includes 164524 observations up to 31 December 2024. From this sample, 49358 observations were kept as the test set to evaluate model performance on unseen historical data.

A threshold of 0,65 rather than 0,5 was selected to increase the overall accuracy and precision of the model, by doing this reducing the number of false positives being classified as churn. Setting the threshold at 0,65 resulted in an accuracy rate of 77,80% and a precision rate of 35,64% for false positives, with the recall numbers slightly lowered at 20,20% as indicated in Table 21. However, the increased level of accuracy outweighs the risk of missing certain instances that may finally fall into the churn category.

Table 21 : Threshold Selection Logistic Regression

	Model	Threshold	Accuracy	Precision	Recall
0	LogReg	0,15	20,70%	19,32%	99,31%
1	LogReg	0,25	24,15%	19,79%	97,37%
2	LogReg	0,35	35,20%	21,78%	92,38%
3	LogReg	0,45	55,68%	26,96%	77,30%
4	LogReg	0,55	71,22%	32,84%	48,55%
5	LogReg	0,65	77,80%	35,64%	20,20%
6	LogReg	0,75	79,66%	32,24%	5,94%
7	LogReg	0,85	80,53%	28,92%	1,37%

Table 22 : Confusion Matrix Logistic Regression

			Predicted: No Churn (0)
Actual: No Churn (0)	36499 (TN)	3436 (FP)	39935
Actual: Churn (1)	7520 (FN)	1903 (TP)	9423
Total	44019	5339	49358

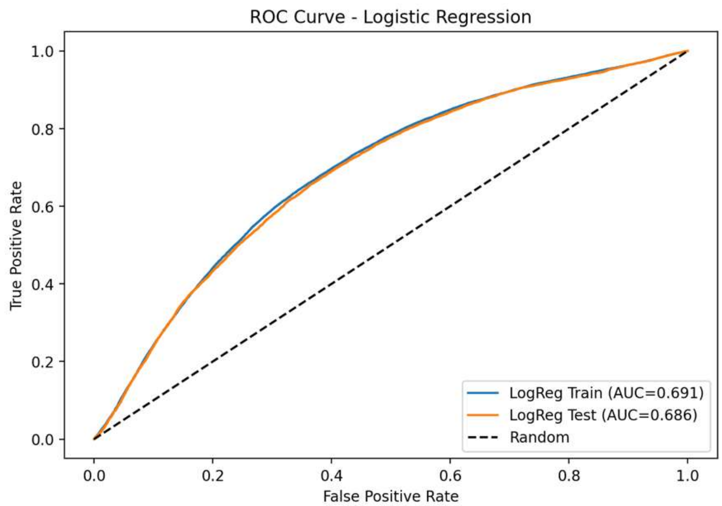
The confusion matrix achieved as shown on Table 22 and we see that the model has achieved 36499 non-churn cases and 1903 churn cases, while it has also miss-classified 3436 non-churn cases and 7520 churn cases. So, the model has an accuracy $\frac{(TN+TP)}{Total} = \frac{36}{49358} = 0{,}778$ which represents the 77,8% of the total of test cases (49358) have been correctly found out by the model. It is clear from the confusion matrix the fact of the logistic regression has performed well classifying the vast majority of non-churn observations. Although recall gives us a value of $\frac{TP}{(FN+TP)} = \frac{1903}{9423} \cong 0{,}202$ which means that the model found only 1903 of 9423 or 20,2% of samples that are real churn cases in our test set.

4.5.2 ROC curve and AUC

Independent of the confusion matrix, discriminative performance of the logistic regression model is assessed in terms of the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). The former presents a graphical plot of true positive rate against false positive rate at each possible classification threshold. In that sense, the ROC offers a threshold-free evaluation measure for the discriminatory ability of the logistic model on the observed outcome variable, such as Churn vs. No-churn.

The logistic regression model trained on the available data has a ROC-AUC of 0,691 on the test sample. For a given pair of randomly chosen observations - one being a churn and another one non-churn - this implies that the fitted model has a 69,1% chance of producing a higher churn score for the true churn case as shown on Figure 2. So, not only does the logistic model show acceptable discriminative performance, it also does not seem to overfit the data heavily despite a relatively limited ROC-AUC value.

Figure 2 : ROC curve / Logistic regression ROC-AUC plot



Out-of-time evaluation on the January 2025 policies, as shown on Table 23, yields a slightly worse result of 0,674 although this difference is not dramatic when compared to in-sample results. So, the ranking performance of the logistic model is also fairly competitive, though predictive performance is weaker when moving slightly further into the future away from the estimation sample. However, as the ROC-AUC measures still convey, the logistic model is somewhat better than a random classifier, although having a relatively small discriminatory power.

Table 23: Out-of-time evaluation

model	n	accuracy	precision	recall	roc_auc
LogRegEval 1/2025	5200	0,756	0,322	0,229	0,674

In summary, the logistic model offers clear advantages as a baseline model, a more flexible machine learning approach will be tested next.

4.6 LightGBM Model and Comparative Performance

4.6.1 Rationale for Using LightGBM

While the logistic regression model offers an interpretable, statistically well-founded baseline for our analysis, its implicit structure places a relatively restrictive form of linearity on the relationship between the predictors and the log-odds of churn. However, insurance churn is very likely to involve more complex patterns including non-linearity, interaction effects in pricing and policy characteristics and threshold-style behavioural responses which could not be appropriately represented by a simple logit specification. Therefore, we extend our analysis using a Light Gradient Boosting Machine (LightGBM) model.

LightGBM is tree-based ensemble learning approach that creates multiple decision trees incrementally, with each subsequent decision tree trying to rectify errors of the previously built tree. It is suitable for our task of churn prediction as it could model non-linear relations as well as variable interactions without the need for pre-defining these. It is also compatible with structured tabular data consisting of customer, policy and price records under consideration in this study (Verbeke et al., 2011; Baesens, 2014).

LightGBM is a suitable choice for datasets with non-linear relationship and interaction, using this to better estimate the relationship between target and features. Also, since it can automatically determine interaction, we can easily get a good model even without tell user which variables to consider. In addition, the LightGBM model also apply some technique which is more efficient for computing complexity. Firstly, it splits data by histogram. If the variable is not continuous, we can easily use LightGBM without any complicated trick to model the variable. Secondly, it grows tree leaf-wise from the bottom. This approach is proven to be more efficient than growing from top. Also, the LightGBM model could model large datasets which is optimal for dealing with huge tabular datasets. Finally, after combining the above 3 points, LightGBM perform well in any kind of analysis and modelling, for example, combining it with logistic regression model to improve modelling ability without sacrificing interpretability.

4.6.2 Model Setup and Training Procedure

The LightGBM model has been trained over the same churn dataset used as the benchmark with logistic regression to keep both modelling approaches comparable. The model-ready sample was filtered by keeping only policy observations with start date on or before 31/12/2024, which produced a total of 164524 samples. This was later split into a training set (115166) and test set (49358), keeping a consistent churn distribution between the two. The churn rate of the training set is 19,1%, which is similar to the test set, facilitating an adequate comparison between both models.

LightGBM was consisted with both numerical and categorical predictors. This richer specification will use not only linear effects but also nonlinear splits among product, customer, price and customer related measures.

LightGBM for binary classification (churn) was trained. It generates an ensemble learning method where models learn from each other. Each node adds on to the previous nodes so that it can keep improving its forecast.

To evaluate the performance of the model we compute predictions using the predicted model for both training and testing set. Also, we perform an out-of-time evaluation on the policy observations whose start date is January 2025. So, the proposed model evaluates not only the standard train-test experiment but also a forward-in-time train-test experiment that is closer to what a practitioner faces when performing churn modelling.

4.6.3 Performance Comparison with Logistic Regression

Comparative results show an important improvement of the LightGBM model on predictive classification over logistic regression baseline. On the test sample (up to 31 December 2024), the logistic regression model achieves accuracy of 0,778, precision of 0,356, recall of 0,202 and ROC-AUC of 0,686. In contrast, the LightGBM model achieves higher performance across all metrics, with accuracy of 0,895, precision of 0,789, recall of 0,613 and ROC-AUC of 0,909 as Table 24 and Figure 3 shows.

Table 24 : Model Comparison

Model / Sample	N	Accuracy	Precision	Recall	ROC-AUC
LightGBM Test	49358	0,895	0,789	0,613	0,909
LogReg Test	49358	0,778	0,356	0,202	0,686

This analysis focuses on improving the model's ability to pick up churn cases. While logistic regression picked up 20,2% of actual churn cases, the LightGBM model picked up 66,1%. This is a important increase in recall. Precision also saw an improvement from 35,6% for logistic regression to 61,3% for the LightGBM model. The boosted tree model not only detects churn cases better, but it also makes more accurate positive predictions.

LightGBM model is more effective in ranking customers for their risk of churn and in translating those ranks into final classification outcomes. It is observed that non-linear modelling approaches perform better in capturing the non-linear relationships involved in customer churn and both discrimination (in terms of ROC-AUC) and prediction also improve.

The prediction threshold implemented for LightGBM is 0.65, translating the continuous probability of churn into a binary classification of churn or non-churn. Metrics like accuracy, precision and recall were 0,895, 0,789 and 0,613 respectively, as shown on Table 25. Other probabilities were tested but 0.65 was chosen because it demonstrated the best combination of recall and precision on the test set, as per the comparison of different classification thresholds in the confusion matrix.

Table 25 : Threshold Selection Logistic Regression

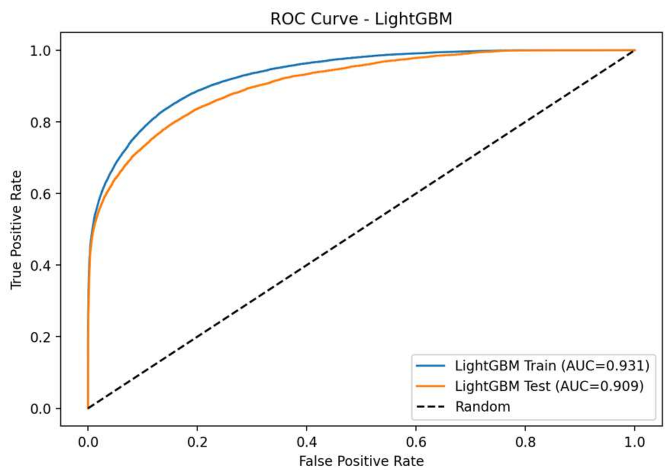
	Model	Threshold	Accuracy	Precision	Recall
0	LightGBM	0,15	44,6%	25,5%	98,9%
1	LightGBM	0,25	61,4%	32,5%	94,9%
2	LightGBM	0,35	74,7%	42,3%	88,9%
3	LightGBM	0,45	82,8%	53,3%	80,7%
4	LightGBM	0,55	87,5%	66,0%	71,0%
5	LightGBM	0,65	89,5%	78,9%	61,3%
6	LightGBM	0,75	89,7%	89,6%	52,4%
7	LightGBM	0,85	89,2%	94,8%	45,7%

Table 26 : Confusion Matrix LightGBM

	Predicted: No Churn (0)	Predicted: Churn (1)	Total
Actual: No Churn (0)	38387(TN)	1548 (FP)	39935
Actual: Churn (1)	3651 (FN)	5772 (TP)	9423
Total	42038	7320	49358

From Table 25 we can see that the model has achieved 38387 non-churn cases and 5772 churn cases, while it has also miss-classified 1548 non-churn cases and 3651 churn cases. So, the model has an accuracy $\frac{(TN+TP)}{Total} = \frac{38387+5772}{49358} = 0{,}895$ which represents the 89,5% of the total of test cases (49358) have been correctly found out by the model. It is clear from the confusion matrix the fact of the LightGBM have performed extremely well classifying the vast majority of non-churn observations. Also, recall gives us a value of $\frac{TP}{(FN+TP)} = \frac{5772}{9423} \cong 0{,}613$ which means that the model found 5772 of 9423 samples that are real churn cases in our test set, which improved by 203% compared to logistic regression at 20,2% recall rate. The ROC performance increases to as shown on Figure 3 compared to logistic regression.

Figure 3 : LightGBM performance comparison plot



A similar pattern is observed in the out-of-time evaluation for January 2025 in Table 27. The logistic regression model achieves accuracy of 0,756, precision of 0,322, recall of 0,229 and ROC-AUC of 0,674. In comparison, the LightGBM model achieves accuracy of 0,877, precision of 0,750, recall of 0,549 and ROC-AUC of 0,875. The consistency of these results confirms that the superior performance of the LightGBM model generalizes well to new data and is not limited to the training period.

Table 27 : Model Comparison out-of-time

Model / Sample	N	Accuracy	Precision	Recall	ROC-AUC
LogReg Eval 01/2025	5200	0,756	0,322	0,229	0,674
LightGBM Eval 01/2025	5200	0,877	0,750	0,549	0,875

LightGBM outperforms at predicting churn than logistic regression. Logistic regression may be an important baseline to check if the data makes sense but for prediction, LightGBM should be used.

4.6.4 Feature Importance and Interpretation

With the feature importance analysis of LightGBM (Table 28), we can provide insights on what variables are most influential on the prediction of churn. This contrasts with a logistic regression approach where we would look at the magnitude and sign of the coefficients. The more value in this approach lies in the fact that feature importance analysis tells us which variables matter the most. Feature importance in LightGBM model refers to the relative importance of each feature in improving the accuracy of prediction by the ensemble of decision trees in the LightGBM model and does not show whether higher values are associated with an increase or decrease in the outcome.

Table 28 : LightGBM feature importance

	feature	importance_gain	importance_pct
2	num__VehiclePoliciesCount	290059,9104	0,3564
1	num__TenureYears	232444,9432	0,2856
3	num__ActiveVehiclesCount	61294,9921	0,0753
24	cat__PremiumChangeBand_Increase_0_to_20pct	42270,6330	0,0519
4	num__ExtraCoversCount	32330,3274	0,0397
22	cat__DurationGroup_Semi-annual	23647,5723	0,0291
21	cat__DurationGroup_Quarterly	22565,0756	0,0277
20	cat__DurationGroup_Monthly	22538,4426	0,0277
0	num__mikta_per_day	19840,7116	0,0244
5	num__VehicleAge	13140,6324	0,0161
6	num__Age_at_start	12445,2509	0,0153
23	cat__PremiumChangeBand_Decrease_gt_20pct	9500,0428	0,0117
13	cat__Usage_Motorcycle	6849,9007	0,0084
10	cat__CompanySizeGroup_Medium_company	5079,9145	0,0062
25	cat__PremiumChangeBand_Increase_gt_20pct	3864,1253	0,0047
11	cat__CompanySizeGroup_Small_company	2243,5085	0,0028
18	cat__VALUE_Value_up_to_10000	1855,9831	0,0023
7	cat__packet_cat_FULL CASCO	1841,2772	0,0023
8	cat__packet_cat_FIRE-THEFT	1612,4785	0,0020

15	cat__RegionGroup_Other_region	1486,9989	0,0018
9	cat__Client_cat_1	1457,3398	0,0018
19	cat__StartSemesterGroup_2nd_semester	1297,8716	0,0016
26	cat__CCGroup_CC_1000_1499	971,8684	0,0012
14	cat__Usage_Truck	740,9698	0,0009
28	cat__CCGroup_CC_greater_1500	666,2081	0,0008
16	cat__RegionGroup_Thessaloniki	572,1398	0,0007
27	cat__CCGroup_CC_500_999	526,9465	0,0006
17	cat__VALUE_Value_greater_20000.01	386,0241	0,0005
12	cat__ContactChannelGroup_Phone_contact	338,9153	0,0004

From Table 28, we see that the most important factor in the model is the number of policies (VehiclePoliciesCount 35,64%), which implies that the churn has a strong link with the policy history, a key variable which included in data since customers go through several policies in a lifetime and interact with the aggregator many times. This shows that customers with a rich policy history are more likely to return and the interaction history of policy renewal might be a key driver of their churn behaviour. This supports the argument that the churn has more to do with the accumulated experience of customers over time and less to do with immediate factors.

Customer tenure, TenureYears (28,56%) was the second most important variable. Customer tenure is the duration between the customer first purchases and a cutoff time. Customer retention or churn behaviour can change over time, and long-term customers may have different motivations for leaving an aggregator due to pricing, service and market alternatives. So, customer tenures may also play a role in determining the churn behaviour of customers.

Customers with more than one active vehicle appear to have significantly different churn behaviour than those with only a single policy. The ActiveVehiclesCount importance is 7,53%, this insight might be leveraged by expanding product offerings to deepen customer relationships or strategically cross-selling to existing customers.

PremiumChangeBand_Increase_0_to_20pct at 5,19%, which captures how customers respond to changes between 0 to 20% increase in premium at renewal. Customers are more sensitive to changes in premium than to absolute levels. Large increases in premium are likely to cause churn, making PremiumChangePct a critical consideration in pricing strategy at renewal and a key lever for retention. This insight highlights the price sensitivity as a basic driver of customer decisions across insurance markets.

4.7 Summary of Empirical Findings

The phenomenon of customer churn has been studied for the insurance aggregator portfolio using descriptive analysis, logistic regression and machine learning techniques. Empirical evidence suggests that churn is important and non-trivial in the investigated portfolio. The analytical sample after refinement exhibits aggregate churn approaching one fifth of valid renewal-eligible observations, which demonstrates that retention is a strategically important question for the aggregator and justifies the thorough empirical investigation presented in this chapter.

Exploratory analysis suggested that churn is not random but is related to multiple dimensions at the same time: relationship, premium, premium changes, contract length, insurance company and customer segment. Variables related to past continuity such as tenure and policy, consistently appear as protective factors. Pricing variables and premium changes are found positively correlated to churn suggesting that attrition can be due to multiple factors like behavioural, economic and structural, instead of one.

The logistic regression model provided us with an interpretable statistical baseline model confirming that several variables are statistically important and directionally informative. However, its predictive performance was only moderately good (Test-Set ROC-AUC = 0,686) and the recall for the churn class was rather poor, 20,2%, so confirming that Logit model can be used for feature extraction and interpretation, but it fails as a standalone classification tool for churn prediction.

A LightGBM model was proposed in this chapter to address the limitations. This gives an important improvement with logistic regression where test-set AUC-ROC scores increase to 0,909 with strong improvements on accuracy +15%, recall + 203,3%, precision +121,2%

compared to logit model and finally has a good performance (Out-Of-Time prediction) in January 2025. Results of feature-importances showed that tenure, previous policies, premium change during renewal, DurationGroup, premium level had strong features in churn prediction.

Overall, this chapter finds that churn in insurance aggregation is understandable and predictable. The logistic regression helps create the general direction and the statistical importance of the underlying churn drivers, whereas the LightGBM yields substantially better performance. These highlight a central claim of this dissertation that there is a systematic difference in churn between customers and policies which can be modelled by a combination of interpretable and high-accuracy models.

5. Interpretation of Empirical Results

5.1 Interpretation of the Main Empirical Findings

5.1.1 Main Determinants of Churn

Results of the empirical studies suggest that churn in an insurance aggregator portfolio happens owing to a blend of behavioural, economic and structural drivers and not a single determinant. Using both logistic regressions and LightGBM we identify churn as influenced by a complex interplay between relationship history, price setting, contractual features and customer-segment characteristics.

Most importantly, customer-policy and prior continuity related variables seem to be among the main determinants. Specifically, the role of customer tenure and prior policy history seem to show reduced churn risk regarding stronger and longer standing customer-policy relationships. This appears reasonable as a well-established customer-policy relation should lead to higher familiarity with renewals, reduced switching tendency and higher attachment to the current insurance situation.

The second main group of determinants pertains to prices. The evidence shows that premium and renewals premiums increase are positively linked to churning. This implies that the customers in the aggregator environment are sensitive to price and that renewal decisions depend not only on the absolute value of the price of coverage but also upon the value of the changes in price in respect of previous policy conditions. So, price would appear to act as one of the economic triggers of churn.

Also, contract structure also plays a role, as do product related characteristics, which suggests that the insurance product design matters for predicting churn. Lastly, customer segment variables or interactions thereof, show that there are differing churn propensities across segments of the portfolio.

Overall, these results show that in addition to taking a broader view of churn, we need to see it as a multi-dimensional output variable. The main factors extracted suggest a broader picture: the dimensionality of retained insurance services aggregates customer retention regarding relational maturity, price pressure, arrangement and customer-type specific behaviour.

5.1.2 Role of Price, Tenure and Renewal Behaviour

Among the drivers of churn, price, tenure and renewal behaviours emerge as particularly important contributors to explaining whether a customer chooses to stay or to leave at renewal period. These three types of variables capture much of the economic and behavioural essence of the insurance relationship and why it is continued by some and ended by others at renewal. Specifically:

1. Price. Logistic regression and the LightGBM model converge on the view that premium and renewal increases are strongly and positively associated with the probability of churn. This reflects the idea that customers within the aggregator setting appear attentive to both levels and changes in policy price, consistent with the relative ease with which they can consider alternatives. In a setting in which customers may easily compare the details of alternative deals, there are likely to be important consequences for renewal from even moderately high degrees of pricing pressure.

2. Tenure. Relationship duration seems to reduce churn probability, suggesting a protective effect for relationship between customer and insurance aggregator as they develop over time. Policy holds time captures several potential mechanisms that may produce this pattern (familiarity and reduced switching costs), although in aggregate it may prove a strong indication of stable customer behaviour. Lengthier tenure will make the customer better accustomed to the renewal process and generally less likely to switch between portfolios.

3. Renewal behaviours and prior policy history. The number of prior policies shows that those customers with a better history of renewal are less likely to churn at renewal. In other words, the strength of their prior policy behaviour appears to be a valid basis upon which to predict their future behaviour. Customers with a strong policy history appear to be more resilient to potential price and product competition than those whose prior record of policy ownership was weak.

Taken together, these three dimensions represent among the strongest patterns uncovered by the analysis.

5.1.3 Importance of Service Channel (online vs offline)

The empirical results show that Customer and Service-channel dimensions contribute non-dominantly to explain Churn. In the current data this part is represented by the variable Client-cat. The more interaction type information is provided in the ContactType. Neither client_cat nor contact_type alone explain a large share of churn, but both contribute to the structure of how the behaviour differs between portfolio segments.

As shown from the logistic regression results, Client-cat is statistically important and negatively related to churn. In the final specification of the logistic regression Client-cat's coefficient is -0,177 with $p < 0,000$, so there is a less than one odds ratio (odds ratio = 0,8378), therefore offline customers, is associated with lower probability of churn relative to the reference group holding all other variables constant.

At the same time the LightGBM results confirm that there is indeed value in the segment and the interaction-type related variables although it has a lower importance rank than the strongest churn drivers. There is still Client-cat present in the feature-importance ranking and at least one of the encoded ContactChannelGroup categories, but they work together with a richer set of much stronger predictors such as Tenure, renewals history, premium changes, duration.

So, these results advise us to consider the customer and service structure dimensions but only in the setting of other factors. It is also important not to interpret the effect only in terms of client and service interaction dimensions, because these can also be reflective of the insurer mixture, customer base, pricing environment and product structures in each segment.

5.2 Comparison with Existing Literature

5.2.1 Consistencies with Prior Churn Studies

Previous studies related to churn prediction particularly in the insurance industry, the findings of this dissertation corroborate previous studies that churn is driven by price. Studies in competitive insurance markets show that customer reaction to premiums change and rate competitiveness do affect cancellations and renewals. Premiums and change in premiums at the time of renewal have been found to affect churn. Therefore, the findings of this study corroborate previous findings (Verschuren, 2022).

Second, the result on tenure and prior relationship history of the present work is in line with previous research in customer churn and customer lifetime value which consistently reports that longer relationships relate to better retention or switching behavior (Van den Poel & Lariviere, 2004, Glady et al., 2009). And in the dissertation, we see that tenure and prior policy continuity are relevant churn related dimensions.

Third, results confirm previous research on churn-prediction which shows that churn is best understood as a multi-dimensional phenomenon. Models with more flexibility in the specification are expected to outperform simple linear specifications if the underlying churn behavior is non-linear (Verbeke et al., 2011, Neslin et al., 2006) and previous benchmarking exercises show how more interpretable baseline models combined with more advanced ML-approaches deliver excellent results in the prediction of churn.

Conclusion the empirical findings in the dissertation do not contradict any of the existing literature but they do point out some clear strengths of the overall ideas. More importantly, it applies these ideas to the specific case of insurance aggregation, considering renewals, price comparison and policy continuity.

5.2.2 Differences from Prior Churn Studies

Empirical results from this dissertation are generally in line with the extant churn literature which considers customer switching as a multidimensional outcome driven by behavioural, economic and relational factors, rather than by a single factor (Neslin et al., 2006, Van den Poel & Lariviere, 2004). Previous literature has identified that a combination of customer-history and/or behavioural and/or value-based factors is superior to any individual set for churn modeling (Neslin et al., 2006, Verbeke et al., 2011).

First, in terms of premium and premium change, results from this study are consistent with previous literature results in that both are positively correlated with churn (Van den Poel & Lariviere, 2004, Glady et al., 2009). This shows that customer attrition is sensitive to the economics of value received and the changes in contractual offers within a competing service market.

Second, as regards tenure and prior relationship with insurer, results of this study are also consistent with those reported in the extant literature in finance (Van den Poel & Lariviere, 2004, Glady et al., 2009). In this setting, customer churn is often found to be closely related to the strengths and tenures of the relationship between customers and their banks. The empirical evidence provided in this dissertation supports the view that the tenure and history of the prior relationship play important protective roles in understanding customer churn.

Third, the overall results from this dissertation support those reported in the extant literature in that a multifactor approach often helps in identifying customer churn (as the current dissertation demonstrates) since a combination of customer-history and/or behavioural and/or value-based factors is superior to any individual set for churn modelling purposes (Neslin et al., 2006, Verbeke et al., 2011). In the current study the standard logistic model provides clear base results, whereas the final LightGBM model shows much stronger predictive ability.

Therefore, the results from this dissertation are not contradictory to the extant literature. The findings from this dissertation support the general conclusions found in the extant literature on churn. However, this dissertation extends the current literature by applying it specifically to the insurance aggregation market where renewals, premium changes and prior policy continuity are the key factors.

5.2.3 Contribution to Insurance and Aggregator Literature

The first contribution of this dissertation is contextual. Customer churn here relates not to exit from a provider offering but to problems of renewal and continuance in a multi-provider intermediary setting where customer choices will be influenced by price competition, facilitated by search and comparison technologies. Differences emerge in the switching incentives across these digital intermediaries: price-comparison sites and online insurance aggregators compared with traditional single-provider marketplaces. They reflect a shift in competitive pressure, the way consumers search and switching costs. (Ronayne, 2021, Hu, T.-I., & Tracogna, A. 2020).

The second contribution of this dissertation is methodological. By integrating an interpretable logistic regression baseline with a better-performing LightGBM approach, we demonstrate that churn in the insurance aggregation marketplace is both explainable but also very non-linear in its interaction between tenure, policy prior continuance, premium change, duration and market-segment. This contribution to the literature on churn-models links back to the existing literatures' recognition of a superiority of more flexible model structures over simple (such as Logistic regression) ones when faced with heterogeneous client behaviours and churn outcomes (Vafeiadis et al., 2015, Verbeke et al., 2011).

Overall, the contribution to the insurance-literature and the aggregator-literature is evidence that retention policy in a digital insurance marketplace needs to be thought of not only in terms of price and policy offerings but in terms of renewal process logic, segment heterogeneity and predictive analytics.

5.3 Strategic Implications for Insurance Aggregators

5.3.1 Data-driven Retention Strategies

These empirical results show that insurance aggregators have the potential to increase their portfolio retention by using not only empirical data and opinions but also by basing decisions on churn data. In this way they can identify customers with high probability of churn by predicting the drivers of churn. Targeted adjustments to reduce churn can be more useful and meaningful. Previous research supports this idea that being proactive is better than using reactive measures and is the key to retaining customers (Jamal and Bucklin, 2006, Gattermann-Itschert et al., 2022).

In this study, the best performing models shown that the risk for churn is no longer random across the portfolio but is instead related to certain (identifiable) factors such as tenure, previous continuity of policies, premium change, policy duration and customer segment. The implication here is therefore that retention efforts to the insurer cannot be applied uniformly to all renewals. We argue instead that the insurer should rely on information about predicted churn probability to augment business criteria, such as premium value, expected future profitability and historical relationship tenure for retention decision making (Jayachandran et al. 2005). Analogous reasoning is also be found in the churn management literature, where prediction-based targeting has been found superior to non-targeted churn management approaches and standard managerial heuristics Jamal and Bucklin (2006), Kim et al., 2012.

A data-driven retention strategy must move beyond a unique decision for every client, instead it should differentiate based on the drivers of churn. Churn related to price sensitivity should focus on pricing adjustments at renewal and churn related to service experience should focus on improve communication and engagement initiatives. Therefore, churn management suggests that predictive models should viewed as information to be proactive and use effective retention actions (Jamal and Bucklin, 2006, Reinartz et al., 2005). Recent studies show that companies should move to that way, involving the decision making based on engaging predictive models rather than descriptive statistics (Hossain et al., 2023, Broby, 2022).

5.3.2 Channel-specific Retention Policies

The dissertation findings imply that retention policy should not be developed for the portfolio but for channels or groups associated with different interaction- and service structure. More generally, the multi-channel literature suggests that retention outcome depends on channel experience, communication mode and customer channel-migration behavior, which implies that retention should be adjusted for particular customer populations instead of being applied uniformly (Neslin et al., 2006, Chang & Tsay 2016).

The current findings related to Client_cat and ContactType suggest that channel / interaction-related customer differences contain some predictive value, even though they may not be the most powerful ones when compared to tenure and pricing / renewal-related ones. This result supports the idea that customers associated with different service structures might need different intervention logics. A solution would be that for online customers should be better served via automated reminders such as emails, SMS, push notifications, with personalization being mandatory and also provide them with a fast and easy online renewal journey while offline customers might benefit from person-to-person interaction communication and advice before the renewal. Those two different segments of customers are both benefit through a different service management approach through convenience and information (Zhao et al., 2022).

It follows that a channel-specific retention policy has its operational benefits. The aggregator can use the predicted values of churn probability along with channel / interaction preferences of cases to create interventions at the segment-level instead of treating all churn risk cases alike. This enhances the effectiveness of retention resources while improving intervention fit and match to customers' interaction behaviour. In the multichannel literature it is often argued that matching retention intervention to customers' channel / interaction behaviour is critical for successful retention and long-term customer development (Neslin et al., 2006, Boehm 2008).

5.3.3 Targeted Interventions for High-Risk Customers

The results show that churn management in aggregating insurance should move away from general retention towards targeted treatment of high-risk people. Targeting retention based upon customers rankings by predicted churn propensity stands directly in line with our existing customer, retention literature, such as discussed by Jamal & Bucklin (2006), Kim et al. (2012), which posits that predictive churn modelling only has managerial value if data used act on intervention activities, not simply just modelling.

Also, models developed here show the churn risk is concentrated in identifiable groups as opposed to equally spread out across customer base. So, variables such as tenure, prior policy continuity, premium changes, contract duration and customer segments significantly distinguishing higher risk from lower risk and therefore targets retention resources selectively. For instance, retention for risky business with premium business potentially generates enough business opportunity where it might be economically workable treating these risky customers with personalization including customised renewal offers, direct contact from service channel to serve customers' needs so that customers do not drop policies and alternative product offerings so as to tailor solutions from exit strategies.

On the other hand low business value class retaining through lower cost automated system may be adopted so the study conforms to existing retention targeting literature suggesting businesses perform better when offering target treatments based on perceived churn/exit risk versus offering undifferentiated retention treatments for all perceived customers exit propensity and aligning with underlying business value that those customers offer to company (Kim et al., 2012).

Importantly these also confirm that not all likely-to-drop business deserves the same treatment, such findings are appropriately inferred by recent development in modelling aspects of uplift models such as how the effect of the treatment with respect to desired outcome affects the response. In that recent literature the management of retention does not concentrate simply on identifying good candidate targets for treatment.

5.3.4 Integration of Churn Models into Renewal Processes

The results presented in this dissertation provide evidence that a churn model cannot be viewed as a separate analytical tool, rather, it should be viewed as part of the renewal decision making process where its full managerial value materializes when a prediction based on churn probability relates to the operations preceding policy renewal. This conclusion supports the related churn management and CRM research stating that data analytics delivers strategic value to firms only when predictive insights relate to an organizational decision process (Wen et al. 2022 and Ledro et al. 2025).

From an operational perspective, embedding a churn model into a renewal decision process requires that each renewal eligible case is considered from both a traditional underwriting and pricing perspective as well as from a retention risk perspective. So, insurance aggregators can group their upcoming renewals by segmenting their renewal pool into operationally relevant renewal cases with minimal, medium and high churn probability. Similarly, the timing and intensity of any renewal related adjustments can be planned based on expected churn behaviour.

Connecting the scorecard to renewal workflow represents a shift from a reactive to a proactive approach to customer retention and contributes to an emerging body of works on predictive and prescriptive analytics in CRM. Another consequence of a process integration of churn scoring into renewal workflow is that it facilitates better inter-functional coordination. As the proposed scorecard is integrated into CRM systems and renewal process dashboards, commercial teams, service teams and other insurer-side decision makers act following one risk signal instead of individual preferences or intuition. In recent AI-in-CRM research, this idea has been similarly recognized by arguing that AI value is determined not by AI capabilities alone but by how these capabilities relate to organizational systems, routines and business processes (Ledro et al., 2025).

Overall, the above results imply that churn modelling needs to be built into an operational fabric of the renewal process where scoring rules and retention interventions are connected. In an insurance aggregator setting, this implies embedding the scorecard into renewal workflow, intervention rules and CRM systems so that retention decisions are made earlier, more consistently and more proactively than today.

5.4 Implications for Insurance Companies

5.4.1 Collaboration Between Insurers and Insurance Aggregators

The results from the dissertation imply that churn management for digital insurance is a challenge that can best be addressed by both insurance aggregators actions and a tighter connection between the insurer and the aggregator. Churn being determined by premium change, renewal time, previous policy continuation and customer segment, we cannot speak only of one party's responsibility in retaining customers. Retention needs to be addressed by the sides that control price and underwriting conditions on one hand which is the insurer and comparison, customers and renewal communication on the other hand which is the insurance aggregator .This stands per existing literature demonstrating that digital insurance intermediation has moved the frontier between different entities involved in the insurance value chain and elevated the strategic relevance of intermediary relations (Robertshaw 2011, Gomber et al. 2022).

First, there will need to be improved information exchange, as insurance aggregators may have valuable information about customer search behaviour, propensity for renewal and segment-level churn risk and insurers will have information about claims history, underwriting outcomes and pricing rules. This can help their joint usage for the purposes of retention targeting and renewal strategy. In related work examining digital transformation in the insurance industry, it has been argued that the future value creation will happen at the level of the ecosystem and no longer be based on individual-firm-level decisions (Gomber et al., 2022, Shahroodi et al., 2024).

The second implication relates to renewal design and portfolio quality. If insurance aggregators can identify customers with increased churn propensity but underwriters do not adjust pricing, offer designs or retention rules, then there would be little benefit from modelling. However, if the insurer and aggregator are better aligned, then churn models could help early pricing reviews, offer changes and differential renewal treatment. Analysis of comparison driven insurance markets supports the contention that price comparison channels influence profit and retention profiles and so requires closer decision alignment between distributor and provider (Robertshaw, 2011).

To summarise, the results suggest that collaboration with an aggregator should be an ability for insurers. Retention in the face of digital intermediation in insurance markets may improve if analytics, pricing and interventions at renewal are coordinated on both sides of the relationship.

5.4.2 Portfolio Quality and Long-term Profitability

Dissertation results suggest that churn should not only be seen as a short run renewal performance measure but also from a long run considering portfolio quality and profitability perspective. In an insurance market setting, retaining customers is important for more than just keeping volume in the long-run economic value is impacted by continuity of business, future cash flows and relationship stability. This aligns with the broader customer-profitability literature suggesting that retention and profitability be examined together rather than separately (Lariviere & Van den Poel, 2005, Jacobs et al., 2001).

In the insurance business, as well as in many other industries, customer retention has been widely considered to be a critical factor for the long-term success of firms. For this reason, numerous companies are allocating large amounts of money to retention efforts. From a conceptual point of view, the importance of churn is however not undisputed. In this piece, we argue that from a managerial point of view, customer retention should not necessarily be viewed as an objective that must be achieved at any price but rather that policy churn should be assessed in the context of the portfolio quality, the expected margin from a renewal decision and the long-term value of customers. Recent research (e.g. Verschuren 2022, Eling et al. 2023) has found that in the long run, retention efforts seem to be more effective if their objectives are aligned with a profitability-oriented renewal decision, as opposed to purely volume-oriented, top-of-the-funnel policies.

This paper shows that churn analytics have two-fold implications for portfolio management: they predict the short-term revenue impact of policy churn but can also provide valuable insights in the quality of the portfolio. Indeed, if policy churn occurs among customers that have historically a high propensity to renew their policy and provide a stable, predictable source of premium revenue, churn analysis can indicate a deteriorating quality of the portfolio. We also show that selecting whom to retain can increase the short-term renewal rates and improve the long-term economic quality of the retained book.

In general, our findings show that the use of churn management techniques for insurance aggregations should measure effectiveness in terms of what customers as well as how many customers remain on the books, especially in so far as it impacts long-term portfolio profitability and portfolio stability. This points to a concept of retention as a portfolio quality-management issue as much as a sales force issue.

5.4.3 Channel Mix Optimisation

It is empirically obvious that the issue of churn management cannot be isolated from the larger problem of channel mix optimisation. In insurance aggregation, customer churn has been observed to be driven by both product-price elements and the structure of the interaction between customers, insurance aggregators and insurers. So, firms may not be justified in assuming that one channel configuration is better than another and should effort towards a channel mix that optimises on the trade-off between acquisition efficiency, retention effectiveness, customer experience and long-term profitability. This is consistent with a stream of research in multichannel and digital distribution literature which highlights how channel configuration influences customer behaviours and firm performance (Neslin et al., 2006).

Some online and comparison-based channels allow for efficient customer acquisition but at the same time also increase the competitive pressure through lower switching frictions, potentially reducing retention. We first document that this "acquisition-retention-trade-off" exists: being acquired through a particular channel does not necessarily correlate with a long-term customer value contribution to the firm that can be above or below the industry average. We then outline reasons why this trade-off can exist, as well as its implications for firms and regulators. Our discussion is informed by recent work by Robertshaw (2011), Ronayne (2021) and others.

A second implication is that channel mix decisions should consider customer heterogeneity. Some customers may be attracted to the convenience, speed and transparency of a digital journey whereas other customers may be more responsive to the advice, reassurance and service interaction provided through face-to-face interactions. The literature on online and offline channels shows that customer performance varies based on how congruent the chosen channel structure is with customer service preferences and service expectations (Zhao et al., 2022) rather than the "Online vs. Offline" channel structure itself.

Overall, the findings point towards insurers and insurance aggregators seeing the channel mix as a design variable. The aim would be to balance the use of online and offline channels to achieve improvements in retention and portfolio quality, not maximize digitization. In such a view, churn analytics may offer a basis for decisions on how best to arrange renewals, servicing and comparisons across channels.

5.5 Managerial and Operational Implications

5.5.1 Use of Predictive Analytics in Decision-making

The dissertation presents evidence that the predictive analytics can serve as a support for managerial decisions of insurance aggregation. The primary contribution of this study is in providing predictive models which enable managers to correctly predict and forecast the decisions on renewal. Thus, with respect to those predictions, the managers can decide what particular action to take in relation to clients with lower probability of renewal. The empirical evidence of the dissertation agrees with the studies of Anshari et al.(2019) and Ledro et al. (2025). This contribution can be seen as a part of the current trend in research of CRM and analytics.

In the context of the present study is on the use of predictive models as a decision-support system to inform renewal interventions to reduce customer churn-risk. It moves beyond reporting descriptive statistics or descriptive analytics and considers decision-support analytics and predictive analytics. Chan et al. show how managers can use churn-risk scores for ranking (Chan et al., 2011, pp. 297-298). A recent survey of predictive analytics literature shows that predictive models are useful only when their recommendations are incorporated into managerial decisions (El Koufi et al., 2024).

Managers can be prone to systematic biases and variability in decision making. We highlight how these issues manifest in a customer renewal context. Our position is that predictive analytics can serve as an important signal that allows managers to focus their attentions, prioritize actions and provide consistency across teams and customer segments. We discuss the dual emphases in the literature, first, on the importance of model explainability and organizational adoption and, second, on the capacity of the manager to translate analytical outputs into actionable interventions. The arguments and findings from Galanti et al. (2023) and Baabdullah et al. (2024) are used to provide supporting evidence.

Generally, this dissertation supports the position that predictive analytics should be viewed as an underpin for business decision making capabilities of insurers and not just a means for building models. Regarding insurance aggregation, the present practitioner interest lies around using predictive modelling to better inform the handling of renewal risks on a day-by-day basis at lower management levels.

5.5.2 CRM Systems and Automation

The results of this dissertation add further evidence to the mounting research which suggests that the real-world success of churn predictive modelling can be boosted by introducing predictive models into CRM software and full automation of the related business process. It has been established by mounting research (Lerdo et al., 2025 and Anshari et al, 2019) that the predictive analytics works best with its outcome supplied to an intelligent system that causes sending communication with customers, setting the precedence, and triggering for manual intervention to take place whereby the CRM fully automated.

Once fully integrated with CRM, the output from the churn risk score can be attached to the renewal calendar or account for the customer. Any account with a renewal date in the next 30 days, for example, that has a risk of churn can automatically be flagged and sent a message to customers perhaps in the form of a discount, contacted by advisers to retain them or offered a price review, for example. This would then be an operationally feasible and repeating routine, not dependent on good, judicious decision-making at the point of action.

CRM automation is the use of technologies for automating customer relationship management (CRM) activities. It allows, among others, for automating retention interventions which, in turn, allows for implementing differentiated retention strategies. This can be done by automating treatments in a function of the estimated churn-risk, customer segment, treatment characteristic, etc. The interest in CRM automation is a part of a broader trend toward more automated and personalized customer analytics solutions, as is illustrated, among others, by Anshari et al., 2019 and El Koufi et al., 2024

In general, the dissertation concludes that CRM systems and automation should be regarded as enablers of Predictive Churn Management. For the case of insurance aggregation, they help the translation of modelling outcomes into operations routines for earlier, more uniform and scalable Renewal Management.

5.5.3 Limitations of Purely Price-based Competition

The insights from the dissertation suggest that a pure price driven retention strategy will not be sufficient in the long term. While the empirical study demonstrates that premium level and premium change have an important impact on the churn behaviour, this should not lead to an overall conclusion that the appropriate counter action is a general decrease in prices across the portfolio. Within competitive services markets, price can be effective for customer acquisition and short-term retention, but a one-dimensional market position based on price can endanger margin, portfolio quality and relationship strength - a claim also supported by previous literature on pricing and retention.

A first implication of "pure" price competition is that it attracts (or retains) comparison-prone customers that may be less desirable for long-term portfolio quality. The reduced search costs and increased price transparency in digital markets lead to more intensive switching behaviour and, so, aggressive price competition may lead to better short-run conversions but at the expense of worse long-run retention. Similar evidence from a price-comparison environment reveals how a comparison mechanism changes competition and profitability, rather than stable retention.

The second limitation is that price captures only one aspect of what drives customer continuation. Empirical evidence collected in this dissertation shows that, among others, tenure, previous policy continuation, contract length and customer segment also matter, for example retention seems to be more than how low (or high) one prices his offer. Work on customer retention and service competition has consistently shown that non-price dimensions play an important role for long-run customer outcomes.

This dissertation questions the current market assumption held within the insurance industry, that pricing is the critical factor in retaining customers. It argues that although price is a useful tool in winning and keeping customers, it is best used as one element in a retention strategy focussed on building customer value through business and service delivery design.

5.6 Broader Implications for Digital Insurance Markets

The results presented in this dissertation have implications beyond predicting churn for the operation of digital insurance markets. They imply that customer behaviour in the digital field for financial services will be influenced by price and product characteristics but also by the level of transparency and comparability and service quality, particularly as delivered by digital intermediation. In such environments, retention and the ability to retain customers based on perceived value, trust and the ability to combine data driven decision making and human interaction will form key success factors for intermediaries and insurers. More broadly, they imply that the long run viability of aggregator business models will depend on their ability to offer support to stronger retention and portfolio quality and more intelligent renewal processes within competitive and challenging digital market conditions (Neslin et al., 2006, Ronayne, 2021, Ledro et al., 2025**).**

6. Conclusions and Recommendations

6.1 Main Conclusions of the Dissertation

The dissertation aimed at exploring customer churn in the setting of an insurance aggregator and to answer whether churn can be explained meaningfully and predicted reliably in a digital, multiprovider setting of insurance. The results allow to confirm that aim.

Firstly, churn emerged as a portfolio level, not marginally relevant phenomenon in the refined sample with a final churn rate of 19,1%, so making it a clearly strategically important issue for the observed business.

Secondly, it has been proven that churn in insurance aggregation is not a random phenomenon but one that relates to a combination of relationship-related, price-related and structure-related variables. Tenure, prior policy continuation, premium, premium change and contract duration emerge as key dimensions of churn-risk, suggesting that customer migration in the investigated business setting is the outcome of a combination of behavioural persistency aspects, economic cost aspects and policy-level aspects, rather than a single factor.

Finally, while showing that there is a statistically strong way to explain churn by means of interpretable statistical methods, the study demonstrates that its prediction calls for more powerful but still comprehensible models. While the fitted logit model allows a convenient, statistically meaningful interpretation of the direction of the main churn drivers but only performs reasonably well at best, the LightGBM model offers an excellent prediction performance, with a model selection based on ROC-AUC 0,923 on test data, even the out-of-time prediction of churn for January 2025 scores a ROC-AUC value of 0,891.

To conclude, the dissertation demonstrates that churn in insurance aggregation, is relevant at the portfolio level, can be related to a combination of relationship-level, pricing-level and structure-level risk-variables and it can be explained in an interpretable sense. In addition, strong, non-linear prediction models can yield an even higher predictive performance. With those results at hand, this dissertation concludes that churn in insurance aggregation can not only be "made sense of " but it can be reliably predicted as well.

6.2 Practical Recommendations

6.2.1 Recommendations for Insurance Aggregators and Insurers

The results of this dissertation have several implications for the practice of insurance aggregators and insurers.

Firstly, renewal management needs to become more data-driven in the sense that churn models are used not only to report on past events but also in the regular assessment of renewal-eligible policies. Churn risk is found to concentrate around certain groups of customers (those characterized by weaker policy continuity, shorter relationship history, more aggressive price sensitivity and shorter contract term), which implies that firms ought to move away from uniform treatment and towards risk-based handling of renewals. For insurance aggregators this means starting to segment clients on the basis of churn risk scores and matching churn score ranges to distinct customer case treatments, high-risk and high-value customer cases might call for bespoke renewal quotes and pre-emptive communication and product offering, while lower-value cases can be handled with less costly digital follow-up journeys.

On the insurer side this means that the pricing and offer logic ought to become more aligned with the retention objective, since (as the analysis has shown) a purely price-oriented approach is unsustainable in the long run because churn is driven by tenure, prior renewal behaviour and portfolio composition rather than just the size of the premium discount.

The second point is the need for greater collaboration between aggregator and insurer, given that the two parties are sitting across different information assets. Insurance aggregators know more about the comparison behaviour, about customer interaction and the customer journey. Insurers know more about their pricing strategy and the underwriting logic and product structure.

The third point is that firms should not view predictive analytics as a modelling activity but a part of the operating architecture of the system. Incorporating the churn score into CRM system, dashboarding, decision making - to make intervention decision earlier, more consistently and at scale. And then determining the optimum (and therefore maximum) retention policy with respect to volume and other profitability and performance metrics.

6.3 Limitations of the Study

The contributions of this dissertation are tempered by some limitations that should be acknowledged, including the following.

First, the analysis is grounded in the setting of one insurance aggregator company and therefore represent one particular market, business process and set of customers, in contrast to other settings in insurance distribution.

Secondly, despite refinement of the churn definition in an effort to reduce label noise (for example, excluding technical renewal-related events and imposition of continuation windows), we will never have perfect representation of contract non-renewal on the actual defection event in insurance distribution.

Third, the analysis is limited to the variables that were available in our data, although the empirical models account for many dimensions expected to matter (such as, tenure, previous policy continuation, premium, premium change, duration, provider and customer segment).

Fourth, the logistic regression model is easy to interpret but has only fair explanatory and predictive power, while higher-performing LightGBM is much less interpretable than a simple linear model in turn, nonetheless, we face the classic trade-off between explanatory power and predictive strength.

Fifth, the analyses here are finally oriented toward predicting churn, rather than causally estimating the effects of a retention intervention, our models will tell us which customers are most prone to churning but it will require further data to tell us which customers are most responsive to a given retention program. For these reasons, we hope this dissertation offers a solid foundation for managerially relevant predictive modelling but not a complete causal toolbox for retention.

6.4 Directions for Future Research

The data here presented gives rise to a series of potential future works in multiple directions. Firstly, it will be interesting to explore a wider range of datasets and/or markets to check if the patterns discovered are representative of wider contexts (insurer/market/regulation/distribution). Secondly, including in the analysis a richer behaviour, service type features, we could potentially benefit from an increase in explanatory and predictive power in terms of claims, complaints and customer service, online behaviour and competitor offers. Thirdly, there exists a transition from mere prediction towards characterising causal effectiveness of retention which could be made with uplift models, field experiments/quasi experimentation in both cases identification and estimation problems for the same "which customer" question but in setting of this treatment's response. Fourthly, the need for both channel and segment specific analyses was highlighted, especially focusing in on distilling the differences and differences in the mechanism of insuring of a digital only customer versus an adviser-based path, particularly important within an insurance aggregation set up affecting both acquisition and retention probabilities. Lastly, the long-term connection between attrition and portfolio profitability, lifetime value, insurer aggregator relationship.

References

Neslin, S. A., Gupta, S., Kamakura, W. A., Lu, J., & Mason, C. H. (2006). *Defection detection: Measuring and understanding the predictive accuracy of customer churn models*. *Journal of Marketing Research, 43*(2), 204-211.

Van den Poel, D., & Larivière, B. (2004). *Customer attrition analysis for financial services using proportional hazard models*. *European Journal of Operational Research, 157*(1), 196-217.

Glady, N., Baesens, B., & Croux, C. (2009). *Modeling churn using customer lifetime value*. *European Journal of Operational Research, 197*(1), 402-411.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). *Building comprehensible customer churn prediction models with advanced rule induction techniques*. *Expert Systems with Applications, 38*(3), 2354-2364.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory, 55*, 1-9.

Broby, D. (2022). *The use of predictive analytics in finance*. *Data Science in Finance and Economics, 2*(1), 145-161.

Gattermann-Itschert, T., Thonemann, U. W., & Föhrenbach, J. (2022). *Proactive customer retention management in a non-contractual business-to-business setting*. *Industrial Marketing Management, 105*, 274-288.

Hossain, M. A., Akter, S., Uddin, M. R., & Hamid, A. B. A. (2023). *Data-driven market effectiveness: The role of a sustained customer analytics capability*. *Technological Forecasting and Social Change, 197*, 122887.

Jamal, Z., & Bucklin, R. E. (2006). *Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach*. *Journal of Interactive Marketing, 20*(3-4), 16-29.

Kim, Y. S., Street, W. N., Russell, G. J., & Menczer, F. (2012). *Measuring the success of retention management models based on lift and profit*. *Expert Systems with Applications, 39*(8), 7323-7331.

Boehm, M. (2008). *Determining the impact of internet channel use on a customer's lifetime*. *Journal of Interactive Marketing, 22*(3), 2-22.

Chang, C. W., & Tsay, Y. C. (2016). *The effects of channel experiences and direct marketing on customer retention in multichannel settings*. *Journal of Interactive Marketing, 35*, 34-48.

Neslin, S. A., Grewal, D., Leghorn, R., Shankar, V., Teerling, M. L., Thomas, J. S., & Verhoef, P. C. (2006). *Challenges and opportunities in multichannel customer management*. *Journal of Service Research, 9*(2), 95-112.

Zhao, Y., Wang, L., Guo, Y., & Jiang, J. (2022). *Exploring the impact of online and offline channel advantages on brand relationship performance: The mediating role of consumer perceived value*. *Frontiers in Psychology, 13*, 1005851.

Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2022). *Understanding the InsurTech dynamics in the insurance value chain*. *Electronic Markets, 32*, 885-907.

Robertshaw, G. (2011). *An examination of the profitability of customers acquired through price comparison sites: Implications for the UK insurance industry*. *Journal of Direct, Data and Digital Marketing Practice, 12*, 216-229.

Shahroodi, K., Rezvani, M., & colleagues. (2024). *Developing strategies to retain organizational insurers through churn prediction and retention strategy design*. *Technological Forecasting and Social Change*.

Eling, M., Kochanski, M., & Mattei, A. (2023). *Including individual customer lifetime value and competing risks in lapse management*. *European Actuarial Journal*.

Jacobs, F. A., Johnston, W., & Kotchetova, N. (2001). *Customer profitability: Prospective vs. retrospective approaches in a business marketing setting*. *Industrial Marketing Management, 30*(4), 353-364.

Larivière, B., & Van den Poel, D. (2005). *Predicting customer retention and profitability by using random forests and regression forests techniques*. *Expert Systems with Applications, 29*(2), 472-484.

Verschuren, R. M. (2022). *Customer price sensitivities in competitive insurance markets*. *Expert Systems with Applications, 202*, 117133.

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). *Customer relationship management and big data enabled: Personalization and customization of services*. *Applied Computing and Informatics, 15*(2), 94-101.

Baabdullah, A. M., Alalwan, A. A., Al-Emran, M., Kizgin, H., & Dwivedi, Y. K. (2024). *The precursors of AI adoption in business: Towards an integrated framework for understanding AI-driven decision-making*. *Industrial Marketing Management, 116*, 1-15.

Chan, S. L., Ip, W. H., & Cho, V. (2011). *A dynamic decision support system to predict the value of customer for new product development*. *Decision Support Systems, 52*(1), 178-188.

El Koufi, N., Bouraoui, M. A., & colleagues. (2024). *Toward a decision-making system based on artificial intelligence for precision marketing in insurance*. *Journal of Innovation & Knowledge / related decision-support outlet*.

Galanti, R., Cominelli, M., De Giovanni, P., & Lucertini, G. (2023). *An explainable decision support system for predictive analytics*. *Engineering Applications of Artificial Intelligence, 122*, 106147.

Ledro, C., Nosella, A., & Vinelli, A. (2025). *Artificial intelligence in customer relationship management*. *Journal of Business Research, 189*, 115293.

Dumm, Randy & Hoyt, Robert. (2002). Insurance Distribution Channels: Markets in Transition. Journal of Insurance Regulation.

Gupta, S., Lehmann, D., & Stuart, J. (2004). *Valuing customers.*

Eling, Martin & Lehmann, Martin. (2018). The Impact of Digitalization on the Insurance Value Chain and the Insurability of Risks. Geneva Papers on Risk and Insurance - Issues and Practice. 43. 359-396. 10.1057/s41288-017-0073-0.

Reichheld, F. F. (2003). The one number you need to grow. Harvard Business Review, 81(12), 46-54.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*(6), 995-1006

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression.*

Gensler, S., Leeflang, P. S. H., & Skiera, B. (2012). Impact of online channel use on customer revenues and costs to serve: Considering product portfolios and self-selection**.** *International Journal of Research in Marketing, 29*(2), 192-201

Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing, 64*(4), 17-35

Ellison, G., & Ellison, S. F. (2009). Search, obfuscation, and price elasticities on the Internet. *Econometrica, 77*(2), 427-452

Chen, P.-Y., & Hitt, L. M. (2002). Measuring switching costs and the determinants of customer retention in Internet-enabled businesses: A study of the online brokerage industry. *Information Systems Research, 13*(3), 255-274.

Jayachandran, S., Sharma, S., Kaufman, P., & Raman, P. (2005). The role of relational information processes and technology use in customer relationship management. *Journal of Marketing, 69*(4), 177-192

Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. Wiley

Huigevoort, C. W. J. M. (2015). *Customer Churn Prediction for an Insurance Company* (Master's thesis, Eindhoven University of Technology)

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students* (8th ed.). Pearson

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th ed.). Cengage

Ronayne, D. (2021). Price comparison websites. *International Economic Review, 62*(3), 1081-1110.

Madrigal, M. H. (2020). *Customer churn prediction in insurance industries* [master's thesis / institutional repository record]. Universidad EAFIT

Wen, X., Li, Y., Tian, L., & Zhou, J. (2022). Three-stage churn management framework based on DCN and bi-level programming. *Expert Systems with Applications, 206*, 117998

Hu, T.-I., & Tracogna, A. (2020). Multichannel customer journeys and their determinants: Evidence from motor insurance. *Journal of Retailing and Consumer Services, 54*, 102022

Appendix A: "Descriptive and Segmentation Tables"

Table 29: Churn rate per year

Year	policy_observations	churn_rate_pct
2021	24933	17,59
2022	36088	18,39
2023	46725	19,30
2024	56778	20,03
2025	9592	19,17

Table 30 : Churn rate per packet category

packet_cat	observations	share_pct	churn_rate_pct
BASIC	146319	84,04	19,73
FIRE-THEFT	21134	12,14	15,70
FULL CASCO	6663	3,83	16,00

Table 31 : Churn rate per Duration

DurationBand	Observations	Churn_Rate_pct
Monthly	9202	24,36%
Semi-annual	50210	17,61%
Quarterly	44304	20,36%
Annually	70400	18,67%

Table 32 : Churn rate per Premium per date

PremiumBand	Observations	Churn_Rate_pct
Q1 (≤0,275)	43530	19,36
Q2 (0,275-0,387)	43528	18,99
Q3 (0,387-0,545)	43529	19,48
Q4 (>0,545)	43529	18,56

Table 33 : Churn rate per Tenure

TenureBand	Observations	Churn_Rate_pct
<1 year	32217	18,19
1-3 years	85381	19,22
3-5 years	34182	20,54
5+ years	22336	17,72

Table 34 : Churn rate per Premium Change

PremiumChangeBand	Observations	Churn_Rate_pct
Decrease_gt_20pct	8314	11,16%
Decrease_20_to_0pct	43619	21,05%
Increase_0_to_20pct	112722	18,24%
Increase_gt_20pct	9461	27,29%

Table 35 : Renewal Behaviour

Indicator	Value	Unit
Mean vehicle policies count	3,63	count
Mean active vehicles count	1,27	count
Mean extra covers count	0,81	count

Table 36 : Channel Profile 1

Channel	Observations	Churn_Rate_pct	Mean_Tenure
Offline	30914	16,67	2,43
Online	143202	19,62	2,47

Table 37 : Channel Profile 2

Channel	Mean_Premium_per_Day	Mean_Premium_Change_pct	Mean_Duration_Days
Offline	0,56	4,65	228,88
Online	0,42	4,85	224,19

Appendix B: "Predictive Modelling Outputs"

Table 38 : Logistic Confusion Matrix Test

Class	Predicted Non-churn	Predicted Churn
Actual Non-churn	36499	3436
Actual Churn	7520	1903

Table 39 : Logistic Sensitivity Specificity

Sample	Sensitivity (Recall)	Specificity
Test (<=2024)	0,202	0,356
Eval Jan 2025	0,229	0,322

Table 40 : LightGBM Confusion Matrix Test

Class	Predicted Non-churn	Predicted Churn
Actual Non-churn	38387	1548
Actual Churn	3651	5772

Table 41 : Model Comparison

Model / Sample	N	Accuracy	Precision	Recall	ROC-AUC
LogReg Test (<=2024)	49358	0,802	0,482	0,474	0,766
LGBM Test (<=2024)	49358	0,895	0,789	0,613	0,909
LogReg Eval Jan 2025	5200	0,756	0,322	0,229	0,674
LGBM Eval Jan 2025	5200	0,877	0,750	0,549	0,875

Author's Statement:

I hereby expressly declare that, according to the article 8 of Law 1559/1986, this dissertation is solely the product of my personal work, does not infringe any intellectual property, personality and personal data rights of third parties, does not contain works/contributions from third parties for which the permission of the authors/beneficiaries is required, is not the product of partial or total plagiarism, and that the sources used are limited to the literature references alone and meet the rules of scientific citations.