ΕΛΛΗΝΙΚΟ
ΑΝΟΙΚΤΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΚΟΙΝΩΝΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ:
Τραπεζική, Χρηματοοικονομική και Χρηματοοικονομική Τεχνολογία (FinTech) (ΤΡΑΧ)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ:

Ανάλυση πιστωτικού κινδύνου στο P2P Lending: Προσδιορισμός παραγόντων που επηρεάζουν την αποπληρωμή δανείων με χρήση τεχνικών επιστήμης δεδομένων

ΠΑΠΑΔΟΠΟΥΛΟΣ ΣΠΥΡΙΔΩΝ
Α.Μ: 166608

Επιβλέπων Α' ΑΝΔΡΙΚΟΠΟΥΛΟΣ ΑΝΔΡΕΑΣ
Επιβλέπων Β' ΠΑΠΑΔΑΜΟΥ ΣΤΕΦΑΝΟΣ

Αθήνα, Μάιος 2026

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Περιεχόμενα

Περίληψη 5

Abstract 7

Κατάλογος διαγραμμάτων 9

Κατάλογος πινάκων 10

Συντομογραφίες & Ακρωνύμια 11

Κεφάλαιο 1: Εισαγωγή 12

1.1 Εισαγωγή στο P2P δανεισμό και το FinTech 12

1.2 Σημασία της ανάλυσης δεδομένων στο P2P Lending 14

1.3 Κεντρικό ερευνητικό ερώτημα και ειδικοί στόχοι της μελέτης 15

1.4 Δομή της εργασίας 16

Κεφάλαιο 2: Θεωρητικό υπόβαθρο και ανασκόπηση βιβλιογραφίας 18

2.1 Το οικοσύστημα του FinTech και ο P2P δανεισμός 18

2.2 Πιστωτικός κίνδυνος και αξιολόγηση δανειοληπτών 20

2.3 Η επιστήμη των δεδομένων στη διαχείριση κινδύνου 21

2.4 Ανασκόπηση βιβλιογραφίας 22

2.5 Κριτική σύνθεση και διαμόρφωση ερευνητικών υποθέσεων 24

Κεφάλαιο 3: Μεθοδολογία 26

3.1 Επιλογή και περιγραφή δεδομένων 26

3.2 Προεπεξεργασία δεδομένων (data preprocessing) στο RStudio 28

3.3 Εργαλεία ανάλυσης 30

Κεφάλαιο 4: Ανάλυση δεδομένων και αποτελέσματα 34

4.1 Περιγραφική ανάλυση δεδομένων 34

4.2 Συγκριτική ανάλυση χαρακτηριστικών δανείων 36

4.3 Ανάλυση συσχετίσεων 38

4.4 Στατιστική μοντελοποίηση 41

Κεφάλαιο 5: Συζήτηση αποτελεσμάτων 44

5.1 Ερμηνεία ευρημάτων 44

5.2 Σύγκριση με τη βιβλιογραφία 45

5.3 Πρακτικές επιπτώσεις 46

Κεφάλαιο 6: Συμπεράσματα και μελλοντική έρευνα 48

6.1 Συμπεράσματα 48

6.2 Περιορισμοί της μελέτης 49

6.3 Προτάσεις για μελλοντική έρευνα 50
Βιβλιογραφία 52
Παράρτημα: Κώδικα προγραμματισμού σε RStudio 55

Περίληψη

Η παρούσα διπλωματική εργασία επικεντρώνεται στη διεξοδική διερεύνηση και ανάλυση του πιστωτικού κινδύνου στο πλαίσιο της αγοράς του Peer-to-Peer (P2P) δανεισμού, ενός εκ των πλέον δυναμικών και καινοτόμων κλάδων της Χρηματοοικονομικής Τεχνολογίας (FinTech). Κεντρικός στόχος της μελέτης είναι ο προσδιορισμός και η αξιολόγηση των κρίσιμων εκείνων παραγόντων που επηρεάζουν καθοριστικά την πιθανότητα επιτυχούς αποπληρωμής ή αθέτησης των δανειακών υποχρεώσεων. Σε ένα περιβάλλον όπου η παραδοσιακή τραπεζική διαμεσολάβηση υποχωρεί, η ανάγκη για έγκυρα μοντέλα πρόβλεψης καθίσταται επιτακτική για τη διασφάλιση της σταθερότητας του εναλλακτικού αυτού χρηματοδοτικού μοντέλου.

Για την επίτευξη των ερευνητικών σκοπών, αξιοποιήθηκε ένα εκτενές σύνολο δευτερογενών, ιστορικών δεδομένων (big data) προερχόμενο από μια κορυφαία παγκοσμίως πλατφόρμα P2P δανεισμού. Η μεθοδολογική προσέγγιση βασίστηκε στην εφαρμογή προηγμένων τεχνικών της επιστήμης των δεδομένων (data science), με επίκεντρο τη στατιστική μοντελοποίηση μέσω της λογιστικής παλινδρόμησης (logistic regression). Η συγκεκριμένη μέθοδος επελέγη για τη δυνατότητά της να παρέχει υψηλή ερμηνευσιμότητα των αποτελεσμάτων, επιτρέποντας την ποσοτικοποίηση της επίδρασης κάθε ανεξάρτητης μεταβλητής στην έκβαση του δανείου.

Τα ερευνητικά ευρήματα αναδεικνύουν ότι η πιθανότητα αθέτησης δεν είναι τυχαία, αλλά συνδέεται άρρηκτα με τη χρηματοοικονομική κατάσταση του δανειολήπτη. Συγκεκριμένα, παράγοντες όπως ο λόγος χρέους προς εισόδημα (DTI) και το ετήσιο εισόδημα παρουσιάζουν σημαντική προβλεπτική ικανότητα. Ωστόσο, ο πιστοληπτικός δείκτης (FICO score) και το επιτόκιο δανεισμού αναδείχθηκαν ως οι ισχυρότεροι προγνωστικοί παράγοντες, επιβεβαιώνοντας τη σημασία της ιστορικής πιστωτικής συμπεριφοράς και της τιμολόγησης βάσει κινδύνου.

Η παρούσα μελέτη συμβάλλει ουσιαστικά στη βελτίωση των στρατηγικών διαχείρισης πιστωτικού κινδύνου για τις ψηφιακές πλατφόρμες, προσφέροντας παράλληλα πρακτικές και εφαρμόσιμες γνώσεις στους επενδυτές για τη βελτιστοποίηση των χαρτοφυλακίων τους. Επιπλέον, αναδεικνύει την προστιθέμενη αξία των μεθοδολογιών της επιστήμης των δεδομένων στον τομέα των χρηματοοικονομικών, γεφυρώνοντας το κενό μεταξύ θεωρητικής χρηματοοικονομικής ανάλυσης και σύγχρονης υπολογιστικής στατιστικής. Τα

συμπεράσματα της εργασίας ενισχύουν τη διαφάνεια και την αξιοπιστία στις P2P συναλλαγές, προωθώντας τη βιώσιμη ανάπτυξη των FinTech οργανισμών στην παγκόσμια αγορά.

Abstract

This Master's thesis conducts a comprehensive investigation and analysis of credit risk within the Peer-to-Peer (P2P) lending market, a primary and transformative sector of the Financial Technology (FinTech) industry. The central objective of this research is to identify, evaluate, and prioritize the critical determinants that significantly influence the probability of successful loan repayment versus default. In a contemporary financial landscape where traditional banking intermediation is being increasingly complemented or bypassed by decentralized digital platforms, the development of robust, data-driven predictive models is imperative for ensuring the long-term stability and viability of this alternative financing paradigm.

To achieve the research objectives, an extensive dataset of secondary, historical "big data" from a major global P2P lending platform was utilized. The methodological framework is grounded in the application of advanced Data Science techniques, specifically focusing on statistical modeling through Logistic Regression. This methodology was selected for its high degree of interpretability, which allows for the precise quantification of each independent variable's impact on the loan's final status. The rigorous data preprocessing and feature selection phases ensure that the model captures the essential dynamics of borrower behavior.

The empirical findings demonstrate that the probability of default is not a stochastic occurrence but is systematically linked to the borrower's financial profile. The results indicate that financial metrics, such as the Debt-to-Income (DTI) ratio and annual income, possess significant predictive power. However, the credit score (FICO) and the assigned interest rate emerged as the most potent predictors of default. This confirms the validity of risk-based pricing mechanisms and underscores the fact that historical credit behavior remains a cornerstone of creditworthiness assessment even in innovative FinTech environments.

This study contributes substantially to the field by enhancing credit risk management strategies for P2P platforms and providing actionable insights for investors seeking to optimize their portfolios. Furthermore, it highlights the practical efficacy of Data Science methodologies in Finance, bridging the gap between classical financial theory and modern computational statistics. The conclusions of this

work promote transparency and trust in digital transactions, fostering the sustainable growth of FinTech organizations within the global financial market.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Κατάλογος διαγραμμάτων

Διάγραμμα 4.3.1: Σχέση σκορ FICO και επιτοκίου

Διάγραμμα 4.3.2: Σπουδαιότητα μεταβλητών

Διάγραμμα 4.3.3: Κατανομή σκορ κινδύνου

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Κατάλογος πινάκων

Πίνακας 3.2.1: Ορισμός και ταξινόμηση μεταβλητών της έρευνας.

Πίνακας 3.3.1: Συνοπτική παρουσίαση στατιστικών εργαλείων και βιβλιοθηκών της R

Πίνακας 3.3.2: Περιγραφικά στατιστικά στοιχεία ποσοτικών μεταβλητών

Πίνακας 4.1.1: Σύγκριση χαρακτηριστικών ανά κατάσταση δανείου

Πίνακας 4.2.1: Πίνακας συντελεστών λογιστικής παλινδρόμησης

Πίνακας 4.2.2: Πίνακας Odds Ratios και διαστημάτων εμπιστοσύνης

Πίνακας 4.3.1: Πίνακας συσχετίσεων ποσοτικών μεταβλητών

Πίνακας 4.3.2: Έλεγχος πολυσυγγραμμικότητας (VIF Values)

Πίνακα 4.4.2: Μέσο σκορ κινδύνου ανά ομάδα

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Συντομογραφίες & Ακρωνύμια

P2P	ιδιώτες δανείζουν χρήματα απευθείας σε άλλους ιδιώτες
DTI	Debt to Income ratio
FICO	πιστοληπτικό σκορ
Big Data	μεγάλα σύνολα δεδομένων
ΗΠΑ	Ηνωμένες Πολιτείες Αμερικής
AI	Artificial Intelligence
FinTech	Χρηματοοικονομική Τεχνολογία
ETF	Διαπραγματεύσιμο Αμοιβαίο Κεφάλαιο
PD	Probability of Default
EAD	Exposure at Default
LGD	Loss Given Default
EAD	Exposure at Default
SVM	Μηχανές Διανυσμάτων Υποστήριξης
OR	Odds Ratios
VIF	Variance Inflation Factor
ESG	Environmental, Social, and Governance
ECSP	European Crowdfunding Service Providers

Κεφάλαιο 1: Εισαγωγή

Η ραγδαία εξέλιξη των χρηματοοικονομικών τεχνολογιών (FinTech) κατά την τελευταία δεκαετία έχει μετασχηματίσει ριζικά τον τρόπο με τον οποίο πραγματοποιούνται οι συναλλαγές, οι επενδύσεις και, κυρίως, ο δανεισμός. Στο επίκεντρο αυτής της αλλαγής βρίσκεται ο Peer-to-Peer (P2P) δανεισμός, ο οποίος επιτρέπει την απευθείας σύνδεση δανειοληπτών και επενδυτών χωρίς την παραδοσιακή τραπεζική διαμεσολάβηση. Η επιτυχία και η βιωσιμότητα αυτού του μοντέλου εξαρτώνται άμεσα από την ικανότητα των πλατφορμών να αξιολογούν με ακρίβεια τον πιστωτικό κίνδυνο, χρησιμοποιώντας προηγμένα στατιστικά εργαλεία και μεγάλα σύνολα δεδομένων (big data).

Στο παρόν κεφάλαιο τίθεται το πλαίσιο της έρευνας, ξεκινώντας από την εννοιολογική προσέγγιση του P2P δανεισμού και τη σημασία της ανάλυσης δεδομένων στην ψηφιακή εποχή. Στη συνέχεια, ορίζονται το κεντρικό ερευνητικό ερώτημα και οι επιμέρους στόχοι της μελέτης, ενώ η εισαγωγή ολοκληρώνεται με την παρουσίαση της δομής της εργασίας, η οποία αποσκοπεί στη διερεύνηση των παραγόντων που καθορίζουν την πιστοληπτική ικανότητα των δανειοληπτών στην πλατφόρμα LendingClub.

1.1 Εισαγωγή στο P2P δανεισμό και το FinTech

Η ραγδαία εξέλιξη της τεχνολογίας κατά τον 21ο αιώνα έχει οδηγήσει σε έναν βαθύ και μη αναστρέψιμο μετασχηματισμό του παγκόσμιου χρηματοπιστωτικού συστήματος. Ο όρος FinTech (financial technology) περιγράφει την τομή μεταξύ της τεχνολογικής καινοτομίας και των χρηματοοικονομικών υπηρεσιών, περιλαμβάνοντας ένα ευρύτατο φάσμα εφαρμογών: από τις ψηφιακές πληρωμές και τα κρυπτονομίσματα, μέχρι την ψηφιακή ασφάλιση (InsurTech) και τη διαχείριση περιουσίας μέσω αυτοματοποιημένων συμβούλων (Robo-advisors). Κεντρική φιλοσοφία του FinTech είναι η αποδιαμεσολάβηση (disintermediation), δηλαδή η στρατηγική παράκαμψη των παραδοσιακών χρηματοπιστωτικών ιδρυμάτων προς όφελος πιο άμεσων, διαφανών και οικονομικότερων ψηφιακών λύσεων που εστιάζουν στην εμπειρία του χρήστη.

Μία από τις πλέον δυναμικές εφαρμογές αυτού του μετασχηματισμού είναι ο Peer-to-Peer (P2P) δανεισμός, γνωστός και ως Marketplace Lending. Ο P2P

δανεισμός ορίζεται ως το μοντέλο όπου διαδικτυακές πλατφόρμες λειτουργούν ως ψηφιακοί μεσάζοντες, συνδέοντας απευθείας ιδιώτες ή θεσμικούς επενδυτές που διαθέτουν πλεονάζουσα ρευστότητα με δανειολήπτες (ιδιώτες ή μικρομεσαίες επιχειρήσεις) που αναζητούν χρηματοδότηση. Το μοντέλο αυτό αναδείχθηκε ως μια ρηξικέλευθη εναλλακτική μετά τη χρηματοπιστωτική κρίση του 2008, όταν η αυστηροποίηση του κανονιστικού πλαισίου και η μείωση της ρευστότητας των τραπεζών δημιούργησαν ένα σημαντικό χρηματοδοτικό κενό στην αγορά.

Η λειτουργία των P2P πλατφορμών βασίζεται στην εκμετάλλευση των μεγάλων δεδομένων (big data) και των προηγμένων αλγορίθμων για την ταχεία αξιολόγηση των αιτήσεων. Σε αντίθεση με τις παραδοσιακές τράπεζες που βασίζονται σε δυσκίνητες γραφειοκρατικές διαδικασίες, οι πλατφόρμες P2P προσφέρουν μια πλήρως ψηφιοποιημένη εμπειρία, μειώνοντας δραστικά το λειτουργικό κόστος. Αυτή η εξοικονόμηση πόρων μεταφράζεται σε «διπλό όφελος»: υψηλότερες αποδόσεις για τους επενδυτές σε σύγκριση με τις κλασικές καταθέσεις και πιο ανταγωνιστικά επιτόκια για τους δανειολήπτες, ειδικά για εκείνους που ενδέχεται να απορρίπτονταν από το παραδοσιακό τραπεζικό σύστημα.

Ωστόσο, η απουσία του τραπεζικού «μαξιλαριού» ασφαλείας μετατοπίζει το βάρος του πιστωτικού κινδύνου (credit risk) απευθείας στους επενδυτές. Στο οικοσύστημα του P2P δανεισμού, ο επενδυτής αναλαμβάνει εξ ολοκλήρου την πιθανότητα αθέτησης πληρωμής (default) από την πλευρά του δανειολήπτη, χωρίς την προστασία κάποιου συστήματος εγγύησης καταθέσεων. Η πρόκληση αυτή γίνεται εντονότερη λόγω της ασύμμετρης πληροφόρησης, όπου ο δανειολήπτης γνωρίζει καλύτερα την πραγματική του οικονομική κατάσταση από ό,τι ο επενδυτής.

Ως εκ τούτου, η βιωσιμότητα του P2P δανεισμού βασίζεται στην ικανότητα της πλατφόρμας να «φιλτράρει» αποτελεσματικά τους αιτούντες. Η ανάπτυξη και εφαρμογή εξελιγμένων υποδειγμάτων πιστοληπτικής αξιολόγησης (credit scoring) δεν αποτελεί απλώς μια τεχνική απαίτηση, αλλά τη θεμελιώδη προϋπόθεση για τη διατήρηση της εμπιστοσύνης των επενδυτών και την εξασφάλιση της μακροπρόθεσμης σταθερότητας του κλάδου. Στο πλαίσιο αυτό, η ανάλυση δεδομένων αναδεικνύεται στον κρισιμότερο παράγοντα για την ακριβή πρόβλεψη της συμπεριφοράς των δανειοληπτών και τον μετριασμό των επισφαλειών.

1.2 Σημασία της ανάλυσης δεδομένων στο P2P Lending

Στο αναδυόμενο και ταχέως εξελισσόμενο οικοσύστημα του P2P δανεισμού, η ανάλυση δεδομένων (data analytics) και οι τεχνικές της επιστήμης των δεδομένων (data science) δεν αποτελούν απλώς ένα υποστηρικτικό εργαλείο, αλλά τον κεντρικό πυλώνα πάνω στον οποίο οικοδομείται η διαχείριση του χρηματοοικονομικού κινδύνου. Ουσιαστικά, οι προηγμένες αναλυτικές μέθοδοι αναλαμβάνουν τον ρόλο που παραδοσιακά κατείχε ο τραπεζικός υπάλληλος και η επιτροπή πιστώσεων, αυτοματοποιώντας και αντικειμενικοποιώντας την αξιολόγηση της πιστοληπτικής ικανότητας. Η μετάβαση από την υποκειμενική ανθρώπινη κρίση στην αλγοριθμική επεξεργασία δεδομένων επιτρέπει στις πλατφόρμες να επεξεργάζονται χιλιάδες αιτήσεις σε πραγματικό χρόνο, εξασφαλίζοντας ταχύτητα και ακρίβεια που θα ήταν αδύνατες με τις συμβατικές μεθόδους, ενώ παράλληλα περιορίζεται η πιθανότητα ανθρώπινου σφάλματος ή μεροληψίας.

Οι P2P πλατφόρμες, όπως η LendingClub, έχουν το προνόμιο της πρόσβασης σε τεράστια σύνολα δεδομένων (big data), τα οποία λειτουργούν ως η απαραίτητη πρώτη ύλη για την προγνωστική μοντελοποίηση. Η πρόσβαση αυτή επιτυγχάνεται μέσω της ψηφιακής διασύνδεσης (APIs) με γραφεία πιστωτικής διαβάθμισης (Credit Bureaus), της αξιοποίησης του Open Banking για την άμεση επαλήθευση τραπεζικών κινήσεων, αλλά και της εθελοντικής παροχής στοιχείων από τους χρήστες κατά τη διαδικασία της αίτησης (Jagtiani & Lemieux, 2019). Αυτά τα δεδομένα περιλαμβάνουν μια πολυδιάστατη απεικόνιση του δανειολήπτη, καλύπτοντας από τα κλασικά οικονομικά χαρακτηριστικά, όπως το ετήσιο εισόδημα, ο λόγος χρέους προς εισόδημα (DTI) και το πιστοληπτικό σκορ (FICO), μέχρι τα ειδικά χαρακτηριστικά του δανείου, όπως το ύψος του κεφαλαίου, ο σκοπός της χρηματοδότησης και η διάρκεια αποπληρωμής. Η σημασία της ανάλυσης αυτών των δεδομένων έγκειται στην ικανότητά της να γεφυρώνει το χάσμα της ασύμμετρης πληροφόρησης, προσφέροντας στους επενδυτές μια σαφή και ποσοτικοποιημένη εικόνα του κινδύνου που καλούνται να αναλάβουν, ενισχύοντας έτσι την εμπιστοσύνη σε μια αγορά που στερείται φυσικής επαφής.

Κατά συνέπεια, η συστηματική αξιοποίηση αυτού του όγκου πληροφοριών καθιστά δυνατή την εφαρμογή εξελιγμένων στατιστικών υποδειγμάτων και αλγορίθμων μηχανικής μάθησης (machine learning). Η διαδικασία αυτή ενισχύει την προγνωστική ικανότητα μέσω της κατασκευής μοντέλων ταξινόμησης, όπως η

λογιστική παλινδρόμηση (logistic regression), η οποία επιτρέπει τον ακριβή υπολογισμό της πιθανότητας αθέτησης (probability of default) για κάθε δάνειο. Ο εντοπισμός των κρίσιμων μεταβλητών που λειτουργούν ως προειδοποιητικά σήματα κινδύνου επιτρέπει τον αποτελεσματικό διαχωρισμό των αξιόπιστων δανειοληπτών από εκείνους με υψηλή πιθανότητα αποτυχίας, γεγονός που θωρακίζει το σύστημα έναντι επισφαλειών (Miller, 2015· Jagtiani & Lemieux, 2019).

Παράλληλα, οι επενδυτές αποκτούν ένα σημαντικό πλεονέκτημα στη βελτιστοποίηση του χαρτοφυλακίου τους, καθώς μπορούν να μεταβούν από την εμπειρική επιλογή δανείων στη στρατηγική διαχείριση κεφαλαίων. Χρησιμοποιώντας τα αποτελέσματα των μοντέλων, είναι σε θέση να διαμορφώσουν διαφοροποιημένα χαρτοφυλάκια που εξισορροπούν τον κίνδυνο με την αναμενόμενη απόδοση, μεγιστοποιώντας την αποδοτικότητα της επένδυσής τους. Υπό αυτό το πρίσμα, η παρούσα εργασία δεν περιορίζεται σε μια θεωρητική προσέγγιση, αλλά μέσω της εφαρμογής μεθοδολογίας επιστήμης των δεδομένων σε πραγματικά δεδομένα της αγοράς, αναδεικνύει την πρακτική χρησιμότητα αυτών των τεχνικών στο σύγχρονο FinTech περιβάλλον, συμβάλλοντας στην ενίσχυση της διαφάνειας και της ασφάλειας του οικοσυστήματος.

1.3 Κεντρικό ερευνητικό ερώτημα και ειδικοί στόχοι της μελέτης

Η κατανόηση των κρίσιμων παραγόντων που καθορίζουν την αποπληρωμή ή την αθέτηση ενός δανείου P2P αποτελεί την κινητήρια δύναμη της παρούσας μελέτης, καλύπτοντας ένα σημαντικό κενό γνώσης στον τομέα της χρηματοοικονομικής τεχνολογίας.

Η έρευνα επικεντρώνεται στον πιστωτικό κίνδυνο και έχει ως κύριο στόχο να απαντήσει με συστηματικό και ποσοτικό τρόπο στο ακόλουθο κεντρικό ερευνητικό ερώτημα: Ποιοι είναι οι κυριότεροι παράγοντες, αναφορικά με τα χαρακτηριστικά του δανειολήπτη και του δανείου, που επηρεάζουν σημαντικά την πιθανότητα αποπληρωμής των δανείων στην αγορά του P2P δανεισμού;

Για την πλήρη απάντηση του κεντρικού αυτού ερωτήματος, τίθενται πολλαπλοί ειδικοί στόχοι που καθοδηγούν κάθε στάδιο της μεθοδολογίας. Αρχικά, είναι απαραίτητη η θεωρητική θεμελίωση του θέματος, η οποία περιλαμβάνει μια εις βάθος βιβλιογραφική ανασκόπηση του οικοσυστήματος P2P δανεισμού και των κλασικών χρηματοοικονομικών μοντέλων αξιολόγησης πιστωτικού κινδύνου, σε

συνδυασμό με τις βασικές έννοιες της επιστήμης δεδομένων. Στη συνέχεια, ο πρώτος πρακτικός στόχος αφορά την περιγραφική ανάλυση των ιστορικών δεδομένων δανείων. Μέσα από την ανάλυση και την οπτικοποίηση, επιδιώκεται ο εντοπισμός βασικών τάσεων, κατανομών και συσχετίσεων που ενδέχεται να υπάρχουν μεταξύ των χαρακτηριστικών των δανειοληπτών (π.χ. εισόδημα, DTI, σκορ FICO) και της τελικής έκβασης του δανείου.

Ο κύριος αναλυτικός στόχος της εργασίας έγκειται στην μοντελοποίηση κινδύνου. Συγκεκριμένα, η μελέτη στοχεύει στην εφαρμογή κατάλληλων τεχνικών στατιστικής μοντελοποίησης, όπως η λογιστική παλινδρόμηση ή άλλα μοντέλα ταξινόμησης, για την ποσοτικοποίηση της επίδρασης των ανεξάρτητων μεταβλητών στην πιθανότητα αποπληρωμής ή αθέτησης. Η διαδικασία αυτή θα οδηγήσει στην αξιολόγηση της προβλεπτικής ικανότητας των μοντέλων που θα κατασκευαστούν, προσδιορίζοντας με ακρίβεια ποιες μεταβλητές είναι στατιστικά σημαντικές και ποιες έχουν την ισχυρότερη επίδραση στον πιστωτικό κίνδυνο.

Τέλος, ο απώτερος στόχος είναι η εξαγωγή πρακτικών συμπερασμάτων, μέσα από την ερμηνεία των ευρημάτων. Τα συμπεράσματα αυτά πρόκειται να διατυπωθούν σε πρακτικές προτάσεις που θα μπορούν να αξιοποιηθούν από τους επενδυτές P2P Lending για τη βελτιστοποίηση των χαρτοφυλακίων τους και από τις ίδιες τις πλατφόρμες για τη βελτίωση των αλγορίθμων αξιολόγησης και διαχείρισης κινδύνου.

1.4 Δομή της εργασίας

Η παρούσα διπλωματική εργασία είναι δομημένη σε έξι κύρια κεφάλαια, ακολουθώντας την καθιερωμένη ακαδημαϊκή μεθοδολογία των εμπειρικών μελετών που συνδυάζουν τη χρηματοοικονομική ανάλυση με την επιστήμη των δεδομένων. Η διάρθρωση της ύλης έχει σχεδιαστεί με τέτοιο τρόπο ώστε να καθοδηγεί τον αναγνώστη από το γενικό θεωρητικό πλαίσιο στην εξειδικευμένη στατιστική επεξεργασία και, τελικά, στην εξαγωγή πρακτικά εφαρμόσιμων συμπερασμάτων.

Το πρώτο κεφάλαιο θέτει τις βάσεις της μελέτης, ορίζοντας το ραγδαία αναπτυσσόμενο οικοσύστημα του P2P Lending και τη θέση του στον ευρύτερο κλάδο του FinTech. Αναλύεται η θεμελιώδης σημασία της ανάλυσης δεδομένων για τον μετριασμό του πιστωτικού κινδύνου, ενώ οριοθετούνται με σαφήνεια το κεντρικό ερευνητικό ερώτημα και οι επιμέρους στόχοι που διέπουν την έρευνα.

Το δεύτερο κεφάλαιο είναι αφιερωμένο στη θεωρητική θεμελίωση και την κριτική ανάλυση της υφιστάμενης γνώσης. Εμβαθύνει στις έννοιες του πιστωτικού κινδύνου και στις σύγχρονες μεθοδολογίες της επιστήμης δεδομένων, ενώ ολοκληρώνεται με μια ανασκόπηση προγενέστερων εμπειρικών μελετών, αναδεικνύοντας τους παράγοντες που έχουν ταυτοποιηθεί διεθνώς ως καθοριστικοί για την αθέτηση πληρωμών σε ψηφιακές πλατφόρμες.

Το τρίτο κεφάλαιο περιγράφει αναλυτικά τη διαδικασία της έρευνας, ξεκινώντας από την προέλευση και την τεχνική φύση των δευτερογενών δεδομένων (LendingClub). Ιδιαίτερη έμφαση δίνεται στις τεχνικές προεπεξεργασίας (data preprocessing), όπως ο καθαρισμός και ο μετασχηματισμός των μεταβλητών, καθώς και στην αιτιολόγηση της επιλογής της λογιστικής παλινδρόμησης ως του καταλληλότερου εργαλείου για την επίτευξη των ερευνητικών σκοπών.

Το τέταρτο κεφάλαιο αποτελεί τον πυρήνα του πρακτικού μέρους της εργασίας. Παρουσιάζονται τα ευρήματα της περιγραφικής στατιστικής και οι οπτικοποιήσεις που αναδεικνύουν τις εσωτερικές συσχετίσεις των δεδομένων. Ακολουθεί η παρουσίαση των αποτελεσμάτων της στατιστικής μοντελοποίησης, όπου προσδιορίζονται οι συντελεστές επιρροής κάθε μεταβλητής και αξιολογείται η προβλεπτική ικανότητα του παραχθέντος υποδείγματος.

Το πέμπτο κεφάλαιο εστιάζει στην ερμηνεία των ευρημάτων υπό το πρίσμα της θεωρίας που αναπτύχθηκε στα προηγούμενα στάδια. Πραγματοποιείται σύγκριση των αποτελεσμάτων με τη διεθνή βιβλιογραφία και αναλύονται οι πρακτικές επιπτώσεις (practical implications) για τους συμμετέχοντες στην αγορά του P2P δανεισμού, αναδεικνύοντας την προστιθέμενη αξία της μελέτης.

Το έκτο κεφάλαιο ολοκληρώνει την εργασία με την ανακεφαλαίωση των βασικών πορισμάτων και την απάντηση στο κεντρικό ερευνητικό ερώτημα. Ταυτόχρονα, αναγνωρίζονται με ειλικρίνεια οι περιορισμοί που συνόδευσαν την έρευνα και διατυπώνονται συγκεκριμένες προτάσεις για μελλοντική διερεύνηση, λαμβάνοντας υπόψη τις τεχνολογικές τάσεις που διαμορφώνουν το μέλλον του FinTech.

Κεφάλαιο 2: Θεωρητικό υπόβαθρο και ανασκόπηση βιβλιογραφίας

Η θεμελίωση μιας εμπεριστατωμένης εμπειρικής έρευνας προϋποθέτει την εις βάθος κατανόηση του θεωρητικού πλαισίου και των τεχνολογικών εξελίξεων που διέπουν το αντικείμενο μελέτης. Το παρόν κεφάλαιο επιδιώκει να αναλύσει τις βασικές συνιστώσες που συνθέτουν το περιβάλλον του Peer-to-Peer (P2P) δανεισμού, ξεκινώντας από την ανάδυση του οικοσυστήματος FinTech και τον ρόλο του στη σύγχρονη χρηματοοικονομική διαμεσολάβηση. Η ανάλυση επεκτείνεται στις θεμελιώδεις έννοιες του πιστωτικού κινδύνου και της διαχείρισης αυτού, ενώ παράλληλα εξετάζονται οι αρχές της επιστήμης των δεδομένων που καθιστούν εφικτή τη σύγχρονη προγνωστική μοντελοποίηση.

Μέσω της κριτικής ανασκόπησης της διεθνούς βιβλιογραφίας και των προγενέστερων εμπειρικών μελετών, επιχειρείται η χαρτογράφηση των παραγόντων που επηρεάζουν την πιστοληπτική ικανότητα των δανειοληπτών. Η σύνθεση αυτής της γνώσης αποτελεί τη βάση για τη διαμόρφωση των ερευνητικών υποθέσεων της εργασίας, επιτρέποντας τη σύνδεση της υφιστάμενης θεωρίας με τα πραγματικά δεδομένα που θα αναλυθούν στα επόμενα κεφάλαια.

2.1 Το οικοσύστημα του FinTech και ο P2P δανεισμός

Η εμφάνιση και η ραγδαία ανάπτυξη της χρηματοοικονομικής τεχνολογίας, ευρύτερα γνωστής ως FinTech, αντιπροσωπεύει μία από τις σημαντικότερες δομικές μεταβολές στο σύγχρονο παγκόσμιο οικονομικό σύστημα. Ο όρος FinTech δεν περιορίζεται απλώς στην ψηφιοποίηση των υπαρχουσών τραπεζικών λειτουργιών, αλλά περιγράφει μια ριζική ενσωμάτωση προηγμένων τεχνολογικών λύσεων —όπως η τεχνητή νοημοσύνη, τα μεγάλα δεδομένα (big data) και οι αλγόριθμοι μηχανικής μάθησης— στις χρηματοπιστωτικές υπηρεσίες (Arner et al., 2015). Στρατηγικός στόχος αυτής της σύγκλισης είναι η βελτίωση της επιχειρησιακής αποτελεσματικότητας, η δραστική μείωση του λειτουργικού κόστους και η ενίσχυση της χρηματοοικονομικής συμπερίληψης (financial inclusion) για χρήστες που παραδοσιακά υποεξυπηρετούνταν από το συμβατικό τραπεζικό σύστημα (Gomber et al., 2017).

Στο επίκεντρο αυτού του νέου οικοσυστήματος βρίσκεται ο P2P δανεισμός (Peer-to-Peer Lending), ο οποίος λειτουργεί ως ένας μηχανισμός απευθείας σύζευξης

μεταξύ δανειοληπτών και επενδυτών μέσω εξειδικευμένων ψηφιακών πλατφορμών.

Η θεμελιώδης ειδοποιός διαφορά του P2P δανεισμού από το παραδοσιακό τραπεζικό μοντέλο έγκειται στην έννοια της χρηματοοικονομικής αποδιαμεσολάβησης (financial disintermediation). Ενώ οι εμπορικές τράπεζες λειτουργούν ως κλασικοί χρηματοοικονομικοί διαμεσολαβητές που μετασχηματίζουν τις καταθέσεις σε δάνεια αναλαμβάνοντας τον κίνδυνο στον ισολογισμό τους, οι πλατφόρμες P2P λειτουργούν ως τεχνολογικοί διευκολυντές (enablers). Σύμφωνα με τους Milne & Parboteeah (2016), οι πλατφόρμες αυτές παρέχουν την απαραίτητη υποδομή για τη σύναψη της συμφωνίας, την αυτοματοποιημένη αξιολόγηση της πιστοληπτικής ικανότητας και τη διαχείριση της ροής των πληρωμών, χωρίς όμως να διατηρούν τα δάνεια στο ενεργητικό τους. Με αυτόν τον τρόπο, ο πιστωτικός κίνδυνος μεταφέρεται απευθείας στους επενδυτές, οι οποίοι σε αντάλλαγμα προσδοκούν υψηλότερες αποδόσεις από αυτές των παραδοσιακών αποταμιευτικών προϊόντων.

Η ιστορική αφετηρία της ανόδου του P2P δανεισμού εντοπίζεται στην περίοδο μετά την παγκόσμια χρηματοπιστωτική κρίση του 2008. Η αυστηροποίηση του κανονιστικού πλαισίου (όπως οι συμφωνίες της Βασιλείας ΙΙΙ) και η ανάγκη των τραπεζών για απομόχλευση οδήγησαν σε σημαντική μείωση των χορηγήσεων προς ιδιώτες και μικρομεσαίες επιχειρήσεις (Thakor, 2020). Αυτό το «κενό χρηματοδότησης» λειτούργησε ως καταλύτης για τις FinTech εταιρείες, οι οποίες εκμεταλλεύτηκαν την τεχνολογική τους ευελιξία για να προσφέρουν ταχύτερες και διαφανέστερες διαδικασίες δανειοδότησης. Παράλληλα, η άνοδος του κλάδου υποβοηθήθηκε από το περιβάλλον ιστορικά χαμηλών επιτοκίων, το οποίο ώθησε τους επενδυτές στην αναζήτηση εναλλακτικών περιουσιακών στοιχείων (alternative assets).

Με την πάροδο των ετών, ο κλάδος μετεξελίχθηκε από μια «κοινωνική» μορφή δανεισμού μεταξύ ιδιωτών σε μια ώριμη αγορά, συχνά αποκαλούμενη και ως Marketplace Lending. Σε αυτή τη φάση, η συμμετοχή θεσμικών επενδυτών (hedge funds, ασφαλιστικές εταιρείες) έχει αυξηθεί κατακόρυφα, προσφέροντας μεγαλύτερη ρευστότητα. Η ωρίμανση αυτή συνοδεύτηκε από τη χρήση εξελιγμένων εργαλείων ανάλυσης δεδομένων, τα οποία στοχεύουν στον μετριασμό της ασύμμετρης πληροφόρησης (adverse selection). Καθώς ο δανειολήπτης διαθέτει εγγενώς περισσότερες πληροφορίες για την οικονομική του κατάσταση από ό,τι ο επενδυτής, οι πλατφόρμες επιστρατεύουν προηγμένα πιστωτικά σκορ (credit scoring) που

υπερβαίνουν τα παραδοσιακά κριτήρια, ενσωματώνοντας εναλλακτικά δεδομένα για την ακριβέστερη τιμολόγηση του κινδύνου και τη διασφάλιση της μακροπρόθεσμης ευστάθειας του συστήματος.

2.2 Πιστωτικός κίνδυνος και αξιολόγηση δανειοληπτών

Ο πιστωτικός κίνδυνος (credit risk) συνιστά τη θεμελιώδη πρόκληση στον κλάδο του P2P δανεισμού και ορίζεται ως η ενδεχόμενη αδυναμία ενός δανειολήπτη να ανταποκριθεί στις συμβατικές του υποχρεώσεις για την καταβολή του κεφαλαίου και των δεδουλευμένων τόκων. Στο πλαίσιο της χρηματοοικονομικής θεωρίας, ο κίνδυνος αυτός αναλύεται συνήθως μέσω τριών παραμέτρων: της πιθανότητας αθέτησης (probability of default - PD), του μεγέθους της απώλειας σε περίπτωση αθέτησης (Loss Given Default - LGD) και του εκτεθειμένου ποσού κατά τη στιγμή της αθέτησης (Exposure at Default - EAD). Ενώ στο παραδοσιακό τραπεζικό σύστημα οι απώλειες αυτές απορροφώνται από τα ίδια κεφάλαια και τις προβλέψεις του ιδρύματος, στο μοντέλο του P2P δανεισμού ο κίνδυνος μετακυλίεται απευθείας στον επενδυτή, καθιστώντας τα συστήματα πιστοληπτικής αξιολόγησης (credit scoring) τον μοναδικό μηχανισμό προστασίας του κεφαλαίου.

Η διαδικασία αξιολόγησης της πιστοληπτικής ικανότητας εστιάζει στην ανάλυση της οικονομικής φερεγγυότητας του αιτούντος, χρησιμοποιώντας έναν συνδυασμό ιστορικών και τρεχόντων δεδομένων. Κεντρικό ρόλο παίζει το πιστωτικό σκορ (π.χ. FICO score), το οποίο αποτελεί μια στατιστική σύνοψη της προγενέστερης πιστωτικής συμπεριφοράς. Ωστόσο, η σύγχρονη ανάλυση στον τομέα του FinTech υπερβαίνει το FICO, ενσωματώνοντας δείκτες που αντικατοπτρίζουν τη δυναμική της τρέχουσας οικονομικής κατάστασης, όπως ο λόγος χρέους προς εισόδημα (Debt-to-Income ratio - DTI), η σταθερότητα της απασχόλησης και ο ειδικός σκοπός του δανείου. Η συνδυαστική μελέτη αυτών των παραμέτρων αποσκοπεί στον εντοπισμό μοτίβων που υποδηλώνουν αυξημένη πιθανότητα επισφάλειας, ακόμη και όταν το πιστωτικό σκορ φαίνεται ικανοποιητικό.

Κυρίαρχη πρόκληση σε αυτό το στάδιο αποτελεί το φαινόμενο της ασύμμετρης πληροφόρησης, το οποίο εκδηλώνεται με δύο μορφές: τη δυσμενή επιλογή (adverse selection) πριν από τη σύναψη του δανείου και τον ηθικό κίνδυνο (moral hazard) κατά τη διάρκεια της αποπληρωμής. Για την άμβλυνση αυτών των φαινομένων, οι πλατφόρμες εφαρμόζουν εξελιγμένους αλγορίθμους ταξινόμησης, οι

οποίοι κατατάσσουν τους δανειολήπτες σε διαβαθμισμένες κατηγορίες κινδύνου (risk grades). Η διαδικασία αυτή, γνωστή ως τιμολόγηση βάσει κινδύνου (risk-based pricing), διασφαλίζει ότι το επιτόκιο δανεισμού είναι ανάλογο του αναλαμβανόμενου κινδύνου.

Η ακρίβεια της κατηγοριοποίησης αυτής είναι ζωτικής σημασίας για τη μακροπρόθεσμη ευστάθεια του συστήματος. Μια υποεκτίμηση του κινδύνου οδηγεί σε χαμηλά επιτόκια που δεν καλύπτουν τις επερχόμενες απώλειες, απογοητεύοντας τους επενδυτές, ενώ μια υπερεκτίμηση οδηγεί σε αδικαιολόγητα υψηλό κόστος δανεισμού, ωθώντας τους αξιόπιστους δανειολήπτες προς τον ανταγωνισμό. Έτσι, η συνεχής βελτιστοποίηση των μοντέλων αξιολόγησης μέσω της επιστήμης των δεδομένων αποτελεί τη δικλείδα ασφαλείας για την ισορροπημένη λειτουργία της αγοράς P2P.

2.3 Η επιστήμη των δεδομένων στη διαχείριση κινδύνου

Η ανάδυση της επιστήμης των δεδομένων και της μηχανικής μάθησης (machine learning) έχει αναδιαμορφώσει ριζικά το παράδειγμα της διαχείρισης του πιστωτικού κινδύνου στον P2P δανεισμό. Σε αντίθεση με τις παραδοσιακές στατιστικές μεθόδους, οι οποίες συχνά περιορίζονταν από γραμμικές παραδοχές και μικρά δείγματα, οι τεχνικές της επιστήμης των δεδομένων επιτρέπουν την επεξεργασία τεράστιων όγκων αδόμητων και δομημένων πληροφοριών (big data). Η ικανότητα αυτών των συστημάτων να εντοπίζουν σύνθετα πρότυπα συμπεριφοράς και κρυφές συσχετίσεις μέσα στα δεδομένα είναι καθοριστική για την αναβάθμιση της προβλεπτικής ικανότητας, επιτρέποντας μια πιο δυναμική και εξατομικευμένη προσέγγιση του δανειολήπτη.

Η μεθοδολογική προσέγγιση βασίζεται στην ανάπτυξη μοντέλων ταξινόμησης (classification models), τα οποία λειτουργούν ως συναρτήσεις απόφασης με στόχο την πρόβλεψη της δυαδικής έκβασης ενός δανείου: της πλήρους αποπληρωμής ή της αθέτησης. Μία από τις πλέον θεμελιώδεις και δοκιμασμένες μεθόδους σε αυτό το πεδίο, την οποία υιοθετεί και η παρούσα εργασία, είναι η λογιστική παλινδρόμηση (logistic regression). Η συγκεκριμένη τεχνική χρησιμοποιεί τη λογιστική συνάρτηση (sigmoid function) για να μετασχηματίσει έναν γραμμικό συνδυασμό μεταβλητών σε μια τιμή πιθανότητας στο διάστημα [0, 1]. Παρά την εμφάνιση πιο πολύπλοκων αλγορίθμων, η λογιστική παλινδρόμηση παραμένει το «χρυσό πρότυπο» στον

τραπεζικό κλάδο λόγω της υψηλής ερμηνευσιμότητάς της. Επιτρέπει στον ερευνητή να υπολογίσει τους λόγους πιθανοτήτων (odds ratios), κατανοώντας με ακρίβεια πώς μια μοναδιαία μεταβολή σε χαρακτηριστικά όπως το εισόδημα ή ο δείκτης DTI επηρεάζει τις πιθανότητες αθέτησης (Hosmer et al., 2013).

Ωστόσο, η σύγχρονη επιστήμη των δεδομένων εισάγει και την έννοια της προεπεξεργασίας και της επιλογής χαρακτηριστικών (feature engineering), η οποία είναι εξίσου κρίσιμη με τον ίδιο τον αλγόριθμο. Μέσα από διαδικασίες όπως ο καθαρισμός ακραίων τιμών (outliers) και η αντιμετώπιση της πολυσυγγραμμικότητας, διασφαλίζεται ότι το μοντέλο είναι στατιστικά εύρωστο. Παρόλο που η βιβλιογραφία αναδεικνύει ότι μοντέλα όπως τα τυχαία δάση (random forests) ή οι μηχανές διανυσμάτων υποστήριξης (SVM) μπορούν να επιτύχουν οριακά υψηλότερη ακρίβεια εκμεταλλευόμενα μη γραμμικές σχέσεις, η λογιστική παλινδρόμηση προσφέρει την απαραίτητη διαφάνεια που απαιτείται από τις κανονιστικές αρχές (explainable AI), αποφεύγοντας τη λογική του «μαύρου κουτιού». Οι Lessmann et al. (2015) σε μια εκτενή σύγκριση αλγορίθμων για πιστωτική βαθμολόγηση, κατέδειξαν ότι αν και οι προηγμένες τεχνικές μηχανικής μάθησης προσφέρουν οριακά καλύτερες επιδόσεις, οι παραδοσιακές στατιστικές μέθοδοι παραμένουν εξαιρετικά εύρωστες και προτιμητέες σε περιβάλλοντα όπου η ερμηνεία της απόφασης είναι κρίσιμη.

Στον P2P δανεισμό, η εφαρμογή αυτών των μεθοδολογιών επεκτείνεται πέρα από την αρχική έγκριση, συμβάλλοντας στη δυναμική τιμολόγηση και στη βελτιστοποίηση των εισπράξεων. Η αυτοματοποίηση της αξιολόγησης χιλιάδων αιτήσεων σε πραγματικό χρόνο μειώνει δραστικά το λειτουργικό κόστος και τον κίνδυνο ανθρώπινης προκατάληψης (bias). Με αυτόν τον τρόπο, η επιστήμη των δεδομένων λειτουργεί ως η τεχνολογική γέφυρα που επιτρέπει στις P2P πλατφόρμες να ανταγωνίζονται επί ίσοις όροις τα παραδοσιακά ιδρύματα, προσφέροντας ένα περιβάλλον όπου ο κίνδυνος δεν είναι απλώς μια αβεβαιότητα, αλλά μια μετρήσιμη και διαχειρίσιμη παράμετρος.

2.4 Ανασκόπηση βιβλιογραφίας

Η διεθνής βιβλιογραφία γύρω από τον P2P δανεισμό έχει σημειώσει εκθετική ανάπτυξη την τελευταία δεκαετία, ακολουθώντας την ανοδική πορεία του κλάδου του FinTech. Η ερευνητική κοινότητα επικεντρώνεται συστηματικά στον προσδιορισμό των μεταβλητών εκείνων που διαθέτουν την υψηλότερη προγνωστική ικανότητα

αναφορικά με την αθέτηση πληρωμών. Η ελεύθερη διαθεσιμότητα εκτενών συνόλων δεδομένων από ηγετικές πλατφόρμες, όπως η LendingClub και η Prosper, επέτρεψε στους ερευνητές να πειραματιστούν με σύνθετα στατιστικά υποδείγματα, αναδεικνύοντας τη σημασία της πολυπαραγοντικής ανάλυσης στην αξιολόγηση του πιστωτικού κινδύνου.

Στις απαρχές της έρευνας, μελέτες όπως αυτή των Emekter et al. (2015) αποτέλεσαν ορόσημο, καθώς ανέλυσαν τη σχέση μεταξύ της εσωτερικής πιστοληπτικής διαβάθμισης (grades) που αποδίδουν οι πλατφόρμες και της πραγματικής απόδοσης των δανείων. Τα ευρήματά τους κατέδειξαν ότι, ενώ το πιστωτικό σκορ FICO αποτελεί ισχυρό θεμέλιο, δεν επαρκεί για την πλήρη αποτύπωση του κινδύνου. Αντιθέτως, μεταβλητές που σχετίζονται με τη μόχλευση του δανειολήπτη, όπως ο λόγος χρέους προς εισόδημα (DTI) και η διάρκεια του δανείου, αποδείχθηκαν κρισιμότερες για την πρόβλεψη της επισφάλειας. Στην ίδια κατεύθυνση, οι Serrano-Cinca et al. (2015) εμβάθυναν στη δυναμική του επιτοκίου, συμπεραίνοντας ότι αν και λειτουργεί ως άμεσος δείκτης κινδύνου, η αποτελεσματικότητά του ενισχύεται δραστικά όταν συνδυάζεται με ποιοτικά χαρακτηριστικά, όπως ο σκοπός του δανείου. Διαπίστωσαν, για παράδειγμα, ότι δάνεια για αναχρηματοδότηση χρέους ή κάλυψη πιστωτικών καρτών παρουσιάζουν διαφορετικά προφίλ κινδύνου σε σχέση με δάνεια για επιχειρηματικούς σκοπούς.

Η βιβλιογραφία αναδεικνύει επίσης μια σταδιακή μεθοδολογική μετατόπιση προς πιο σύνθετες υπολογιστικές τεχνικές. Οι Malekipirbazari και Aksakalli (2015) πραγματοποίησαν μια εκτενή σύγκριση μεταξύ της κλασικής λογιστικής παλινδρόμησης και αλγορίθμων μηχανικής μάθησης, όπως τα τυχαία δάση (random forests). Τα αποτελέσματά τους υπέδειξαν ότι οι μη γραμμικοί αλγόριθμοι μπορούν να εντοπίσουν δάνεια υψηλού κινδύνου που οι παραδοσιακές μέθοδοι συχνά παραλείπουν, αν και η λογιστική παλινδρόμηση παραμένει ανώτερη ως προς την ερμηνευσιμότητα των αποτελεσμάτων. Επιπλέον, έρευνες (π.χ. Lin et al., 2013) έχουν καταδείξει ότι η εργασιακή εμπειρία και το ετήσιο εισόδημα, ενώ είναι απαραίτητα στοιχεία, συχνά φέρουν λιγότερο προγνωστικό βάρος από ό,τι η συνολική οικονομική έκθεση του δανειολήπτη τη στιγμή της αίτησης.

Τέλος, ένα σημαντικό κομμάτι της βιβλιογραφίας εστιάζει στην κοινωνική διάσταση και την ασύμμετρη πληροφόρηση. Ερευνητές έχουν μελετήσει πώς εναλλακτικά δεδομένα, όπως οι κοινωνικοί δεσμοί (Lin κ.α., 2013) ή ακόμη και το

κείμενο της περιγραφής και οι αφηγήσεις του δανειολήπτη (Herzenstein κ.α., 2011), μπορούν να λειτουργήσουν ως συμπληρωματικοί δείκτες φερεγγυότητας. Η χρήση τέτοιων "ήπιων" πληροφοριών (soft information) επιτρέπει στους επενδυτές να εξάγουν συμπεράσματα για την ποιότητα του δανειολήπτη που συχνά διαφεύγουν από τα τυπικά πιστωτικά σκορ (Iyer κ.α., 2015). Η σύνθεση των παραπάνω ευρημάτων υπογραμμίζει ότι η πρόβλεψη της αθέτησης στον P2P δανεισμό απαιτεί μια ολιστική προσέγγιση, η οποία θα συνδυάζει την αυστηρότητα των οικονομικών δεικτών με την ευελιξία των σύγχρονων αναλυτικών εργαλείων, θέτοντας έτσι το πλαίσιο για την εμπειρική ανάλυση που ακολουθεί.

2.5 Κριτική σύνθεση και διαμόρφωση ερευνητικών υποθέσεων

Η συστηματική ανασκόπηση της διεθνούς βιβλιογραφίας και των θεωρητικών υποδειγμάτων αναδεικνύει ότι η αθέτηση πληρωμών στο οικοσύστημα του P2P δανεισμού δεν αποτελεί ένα στοχαστικό ή τυχαίο γεγονός. Αντιθέτως, συνιστά το αποτέλεσμα μιας πολυδιάστατης αλληλεπίδρασης μεταξύ της οικονομικής ισχύος του δανειολήπτη, της πρότερης πιστωτικής του συμπεριφοράς και των δομικών χαρακτηριστικών της δανειακής σύμβασης. Παρά τις μεθοδολογικές διαφοροποιήσεις που εντοπίζονται μεταξύ των ερευνητών, παρατηρείται μια ισχυρή σύγκλιση ως προς το ποιες μεταβλητές διαθέτουν τη μεγαλύτερη ερμηνευτική ισχύ. Με βάση αυτή την κριτική σύνθεση, η παρούσα εργασία διαμορφώνει τέσσερις κεντρικές ερευνητικές υποθέσεις, οι οποίες θα τεθούν υπό έλεγχο στο εμπειρικό μέρος της μελέτης.

Η πρώτη ερευνητική υπόθεση (Υ1) εστιάζει στην ιστορική πιστοληπτική αξιοπιστία. Αναμένεται ότι το πιστωτικό σκορ FICO και η εσωτερική διαβάθμιση (Grade) της πλατφόρμας θα λειτουργούν ως αντίστροφοι δείκτες του κινδύνου αθέτησης. Η θεωρητική βάση της υπόθεσης αυτής εδράζεται στην πεποίθηση ότι η πρότερη συνέπεια ενός δανειολήπτη αποτελεί τον ισχυρότερο προγνωστικό παράγοντα για τη μελλοντική του συμπεριφορά. Συνεπώς, υποτίθεται ότι υψηλότερα σκορ FICO και ανώτερες διαβαθμίσεις (π.χ. Grade A ή B) θα συνδέονται στατιστικά με σημαντικά χαμηλότερη πιθανότητα αθέτησης (Probability of Default).

Η δεύτερη υπόθεση (Υ2) αφορά τη χρηματοοικονομική μόχλευση του αιτούντος, όπως αυτή εκφράζεται μέσα από τον λόγο χρέους προς εισόδημα (DTI). Η υπόθεση αυτή υποστηρίζει ότι υπάρχει θετική συσχέτιση μεταξύ του δείκτη DTI και

της πιθανότητας αθέτησης. Στο πλαίσιο της διαχείρισης κινδύνου, ένας υψηλός δείκτης DTI υποδηλώνει περιορισμένα περιθώρια ρευστότητας, καθιστώντας τον δανειολήπτη ευάλωτο σε τυχόν αρνητικούς οικονομικούς κλυδωνισμούς που μπορεί να δυσχεράνουν την εξυπηρέτηση του πρόσθετου δανεισμού.

Η τρίτη υπόθεση (Υ3) εξετάζει τον ρόλο του ετήσιου εισοδήματος ως παράγοντα ανάσχεσης του κινδύνου. Εκτιμάται ότι το ύψος του εισοδήματος λειτουργεί ως προστατευτικό «μαξιλάρι», παρέχοντας τη δυνατότητα απορρόφησης απρόβλεπτων εξόδων. Ως εκ τούτου, αναμένεται ότι οι δανειολήπτες που ανήκουν σε υψηλότερες εισοδηματικές κατηγορίες θα επιδεικνύουν μεγαλύτερη σταθερότητα και χαμηλότερα ποσοστά επισφάλειας, επιβεβαιώνοντας τη σημασία της εισοδηματικής επάρκειας στη διασφάλιση της αποπληρωμής.

Τέλος, η τέταρτη υπόθεση (Υ4) επικεντρώνεται στις παραμέτρους της δανειακής σύμβασης, και συγκεκριμένα στο επιτόκιο και τη διάρκεια. Σύμφωνα με τη θεωρία της τιμολόγησης βάσει κινδύνου (risk-based pricing), τα υψηλότερα επιτόκια αντανακλούν την προσπάθεια της πλατφόρμας να αποζημιώσει τους επενδυτές για την ανάληψη αυξημένου κινδύνου (Einav et al., 2012).

Παράλληλα, η μεγαλύτερη διάρκεια (π.χ. 60 μήνες έναντι 36) αυξάνει την έκθεση στην αβεβαιότητα του χρόνου. Αναμένεται, λοιπόν, ότι τα δάνεια με υψηλότερα επιτόκια και μεγαλύτερο χρονικό ορίζοντα θα εμφανίζουν αυξημένη τάση προς την αθέτηση. Η εμπειρική διερεύνηση αυτών των υποθέσεων μέσω του μοντέλου της λογιστικής παλινδρόμησης στο τέταρτο κεφάλαιο, θα επιτρέψει όχι μόνο την επιβεβαίωση ή απόρριψή τους, αλλά και την ποσοτικοποίηση της σχετικής βαρύτητας κάθε παράγοντα, συμβάλλοντας στην εξαγωγή ασφαλών συμπερασμάτων για τη δυναμική του κινδύνου στην ψηφιακή αγορά δανεισμού.

Κεφάλαιο 3: Μεθοδολογία

Η επίτευξη των ερευνητικών στόχων και ο έλεγχος των υποθέσεων που διατυπώθηκαν στο προηγούμενο κεφάλαιο απαιτούν μια αυστηρά δομημένη μεθοδολογική προσέγγιση. Το παρόν κεφάλαιο αναλύει τα στάδια της εμπειρικής διαδικασίας, ξεκινώντας από την επιλογή και την αναλυτική περιγραφή του συνόλου δεδομένων που χρησιμοποιήθηκε. Η μεθοδολογία εστιάζει στην αξιοποίηση πραγματικών δεδομένων από την πλατφόρμα LendingClub, η οποία αποτελεί σημείο αναφοράς για την αγορά του P2P δανεισμού παγκοσμίως.

Κεντρικό κομμάτι της διαδικασίας αποτελεί η προεπεξεργασία των δεδομένων (data preprocessing) στο περιβάλλον RStudio, μια φάση κρίσιμη για τη διασφάλιση της ποιότητας και της στατιστικής εγκυρότητας των αποτελεσμάτων. Στη συνέχεια, παρουσιάζονται τα εργαλεία ανάλυσης και η στατιστική μεθοδολογία της λογιστικής παλινδρόμησης, η οποία επιλέχθηκε ως η καταλληλότερη για την ταξινόμηση του πιστωτικού κινδύνου. Μέσω αυτής της διαδρομής, η μελέτη μετασχηματίζει τον μεγάλο όγκο πρωτογενών πληροφοριών σε ένα δομημένο αναλυτικό πλαίσιο, ικανό να προσφέρει απαντήσεις στα ερευνητικά ερωτήματα που τέθηκαν.

3.1 Επιλογή και περιγραφή δεδομένων

Η παρούσα μελέτη υιοθετεί μια αυστηρή ποσοτική ερευνητική προσέγγιση με αντικείμενο τη διερεύνηση των προσδιοριστικών παραγόντων του πιστωτικού κινδύνου στο περιβάλλον του Peer-to-Peer (P2P) δανεισμού. Η μεθοδολογική δομή της εργασίας εδράζεται στην ανάπτυξη και εφαρμογή ενός υποδείγματος δυαδικής ταξινόμησης, το οποίο αποσκοπεί στην εκτίμηση της πιθανότητας αθέτησης πληρωμής μέσω της στατιστικής ανάλυσης ιστορικών δεδομένων. Το πρωτογενές υλικό της έρευνας αντλήθηκε από το διεθνώς αναγνωρισμένο αποθετήριο ανοικτών δεδομένων Kaggle, το οποίο μπορείτε να βρεί στο σύνδεσμο https://www.kaggle.com/datasets/wordsforthewise/lending-club, και αφορά τα επίσημα στοιχεία της πλατφόρμας LendingClub για την περίοδο 2007-2018. Η επιλογή της χρονικής υστέρησης 2007-2018 κρίθηκε ως η πλέον ενδεδειγμένη, καθώς επιτρέπει την εξέταση δανείων που έχουν ολοκληρώσει τον συμβατικό τους κύκλο, παρέχοντας ασφαλή στοιχεία για την τελική τους έκβαση (Fully Paid vs Charged Off).

Επιπλέον, η συγκεκριμένη περίοδος καλύπτει έναν πλήρη οικονομικό κύκλο —από την παγκόσμια χρηματοπιστωτική κρίση έως την περίοδο σταθεροποίησης— προσφέροντας ένα "stress-tested" δείγμα που αντικατοπτρίζει τη συμπεριφορά των μεταβλητών σε διαφορετικές οικονομικές συνθήκες ενώ παράλληλα, εξασφαλίζει τη σταθερότητα των αποτελεσμάτων μακρυά από τις στρεβλώσεις που επέφερε η πανδημία του COVID-19 στον δανεισμό μετά το 2019.

Τέλος, το εύρος του δείγματος και η πληθώρα των παρεχόμενων μεταβλητών διασφαλίζουν τη διενέργεια μιας στατιστικά ισχυρής ανάλυσης, ελαχιστοποιώντας το σφάλμα γενίκευσης και επιτρέποντας τη μελέτη πραγματικών συνθηκών αγοράς για τη διαμόρφωση ενός αξιόπιστου προφίλ κινδύνου που παραμένει θεωρητικά επίκαιρο και ακαδημαϊκά έγκυρο.

Κεντρικό σημείο της μοντελοποίησης αποτελεί η μεταβλητή της κατάστασης του δανείου, η οποία μετασχηματίζεται σε δυαδική μορφή προκειμένου να εξυπηρετήσει τις ανάγκες της οικονομετρικής επεξεργασίας. Ειδικότερα, το δείγμα περιορίζεται αποκλειστικά σε δάνεια που έχουν ολοκληρώσει τον κύκλο ζωής τους, κατηγοριοποιούμενα σε περιπτώσεις αθέτησης πληρωμής και περιπτώσεις πλήρους αποπληρωμής. Ο περιορισμός αυτός κρίνεται μεθοδολογικά αναγκαίος για την αποφυγή μεροληψίας που θα εισήγαγαν τα τρέχοντα δάνεια, η τελική έκβαση των οποίων παραμένει αβέβαιη κατά τη στιγμή της ανάλυσης. Η ερευνητική πορεία διαρθρώνεται σε τρεις διακριτές και αλληλένδετες φάσεις, ξεκινώντας από την εκκαθάριση και τον μετασχηματισμό των δεδομένων, όπου πραγματοποιείται η επιλογή των χαρακτηριστικών βάσει της θεωρητικής τεκμηρίωσης, η διαχείριση των ελλιπών τιμών και η αριθμητική κωδικοποίηση των ποιοτικών μεταβλητών. Ακολουθεί η φάση της περιγραφικής και επαγωγικής στατιστικής για τη χαρτογράφηση των συσχετίσεων εντός του δείγματος, ενώ η διαδικασία ολοκληρώνεται με την εφαρμογή του υποδείγματος της λογιστικής παλινδρόμησης.

Η επιλογή της συγκεκριμένης μεθόδου εδράζεται στην υψηλή ερμηνευτική της ικανότητα, καθώς επιτρέπει τον ακριβή προσδιορισμό της επίδρασης κάθε μεταβλητής στην πιθανότητα αθέτησης, προσφέροντας έτσι τη δυνατότητα επιστημονικού ελέγχου των ερευνητικών υποθέσεων που διαμορφώθηκαν στο πλαίσιο της βιβλιογραφικής ανασκόπησης.

3.2 Προεπεξεργασία δεδομένων (data preprocessing) στο RStudio

Η διαδικασία επιλογής των ανεξάρτητων μεταβλητών (feature selection) αποτέλεσε ένα κρίσιμο στάδιο της ερευνητικής πορείας, καθώς η ποιότητα των δεδομένων εισόδου καθορίζει άμεσα την ερμηνευτική ισχύ και την προγνωστική ακρίβεια του τελικού υποδείγματος. Από το πρωτογενές σύνολο δεδομένων της LendingClub, το οποίο περιλαμβάνει ένα εξαιρετικά ευρύ φάσμα πληροφοριών, απομονώθηκαν δεκαοκτώ προσδιοριστικοί παράγοντες. Η επιλογή αυτή βασίστηκε σε έναν συνδυασμό βιβλιογραφικής τεκμηρίωσης και στατιστικής καταλληλότητας, με κύριο στόχο την αποφυγή του φαινομένου της πολυδιάστατης πολυσυγγραμμικότητας (multicollinearity) και της υπερπροσαρμογής του μοντέλου (overfitting). Οι επιλεγμένες μεταβλητές κατηγοριοποιούνται σε τρεις κύριους άξονες: στα χαρακτηριστικά της δανειακής σύμβασης, στα οικονομικά στοιχεία του δανειολήπτη και στους δείκτες πιστωτικής συμπεριφοράς και ιστορικού.

Προκειμένου το δείγμα να καταστεί συμβατό με τις απαιτήσεις της λογιστικής παλινδρόμησης στο περιβάλλον RStudio, πραγματοποιήθηκαν εκτεταμένοι μεθοδολογικοί μετασχηματισμοί. Αρχικά, οι μεταβλητές που έφεραν λεκτικές πληροφορίες αλλά εμπεριείχαν εγγενή ιεράρχηση, όπως η διάρκεια του δανείου (term) και η εργασιακή εμπειρία (emp_length), μετασχηματίστηκαν σε αριθμητικές κλίμακες. Συγκεκριμένα, η μεταβλητή της προϋπηρεσίας κωδικοποιήθηκε σε μια διαβαθμισμένη κλίμακα από το 0 έως το 10, επιτρέποντας στον αλγόριθμο να αντιληφθεί την επίδραση της επαγγελματικής σταθερότητας στην πιθανότητα αθέτησης. Ιδιαίτερη βαρύτητα δόθηκε στη διαχείριση του πιστωτικού σκορ FICO. Καθώς το σύνολο δεδομένων παρείχε τα ανώτατα και κατώτατα όρια της αξιολόγησης, κρίθηκε επιστημονικά ορθότερη η δημιουργία μιας νέας μεταβλητής (fico_avg), η οποία προέκυψε από τον υπολογισμό του αριθμητικού μέσου όρου των δύο τιμών, προσφέροντας έτσι μια πιο σταθερή και αντιπροσωπευτική μέτρηση της πιστοληπτικής ικανότητας του ατόμου.

Παράλληλα, η κατηγορική μεταβλητή (categorical variable) Κατάσταση του δανείου, υπέστησει προεπεξεργασία ώστε να είναι δυνατή η μετέπειτα κωδικοποίησή της σε ψευδομεταβλητή (dummy variable). Η νέα μεταβλητή “default_ind” μετράπηκε σε μια δυαδική, όπου 1 έχει γίνει αθέτηση του δανείου και όπου 0 έχει εξοφληθεί. Η διαδικασία αυτή είναι θεμελιώδης, καθώς επιτρέπει την ποσοτικοποίηση ποιοτικών χαρακτηριστικών που παραδοσιακά επηρεάζουν τον

πιστωτικό κίνδυνο. Τέλος, το αρχικό δείγμα των δεδομένων περιλάμβανε 2.260.701 παρατηρήσεις, όπου κάθε παρατήρηση αντιστοιχεί σε μια μεμονωμένη περίπτωση δανειολήπτη και την αντίστοιχη καταγραφή της δανειακής του σύμβασης στην πλατφόρμα. Μετά την εφαρμογή της μεθόδου πλήρους διαγραφής (listwise deletion) για τον χειρισμό των ελλιπών στοιχείων, το τελικό δείγμα προς ανάλυση διαμορφώθηκε σε 1.265.976 παρατηρήσεις. Η απώλεια 994.725 παρατηρήσεων (ποσοστό 45%) κρίθηκε αποδεκτή, καθώς το εναπομείναν δείγμα παραμένει στατιστικά επαρκές για την εξαγωγή ασφαλών συμπερασμάτων, αποφεύγοντας τις μεροληψίες που συχνά εισάγουν οι τεχνικές τεχνητής συμπλήρωσης τιμών. Η αναλυτική χαρτογράφηση των μεταβλητών, οι ορισμοί τους και η κωδικοποίηση που χρησιμοποιήθηκε στην R, αποτυπώνονται συγκεντρωτικά στον Πίνακα 3.2.1. Ο Πίνακας αυτός αποτυπώνει το σύνολο των μεταβλητών που επιλέχθηκαν για την εμπειρική ανάλυση, οι οποίες έχουν ομαδοποιηθεί στρατηγικά ώστε να καλύπτουν όλες τις πτυχές του προφίλ του δανειολήπτη στην αγορά των ΗΠΑ. Οι μεταβλητές του πιστωτικού ιστορικού παρέχουν μια διαχρονική εικόνα της οικονομικής συμπεριφοράς του δανειολήπτη, με κεντρικό άξονα το FICO Score, που αποτελεί την τυποποιημένη σύνοψη της αξιοπιστίας του στις ΗΠΑ. Η τρέχουσα έκθεση σε χρέος αξιολογείται μέσω των ανοιχτών πιστωτικών γραμμών (open_acc) και του ποσοστού χρησιμοποίησης ορίων (revol_util), όπου υψηλές τιμές υποδηλώνουν οικονομική πίεση, ενώ οι συνολικές γραμμές (total_acc) αντικατοπτρίζουν τη μακροχρόνια εμπειρία του στο πιστωτικό σύστημα. Τέλος, το πιστωτικό υπόλοιπο (revol_bal) ποσοτικοποιεί το υφιστάμενο ανεξόφλητο χρέος, επιτρέποντας στο μοντέλο να εκτιμήσει τη δυνατότητα του αιτούντος να διαχειριστεί πρόσθετες δανειακές υποχρεώσεις χωρίς να οδηγηθεί σε αθέτηση.

Κατηγορία	Μεταβλητή	Περιγραφή	Τύπος
Εξαρτημένη Μεταβλητή			
Εξαρτημένη	**default_ind**	Κατάσταση δανείου (1 = Αθέτηση/Charged Off, 0 = Εξόφληση)	Binary
Χαρακτηριστικά Δανείου			
Χαρακτηριστικά Δανείου	**loan_amnt**	Συνολικό ποσό δανείου που ζητήθηκε	Numeric
Χαρακτηριστικά Δανείου	**term**	Διάρκεια δανείου (36 ή 60 μήνες)	Integer
Χαρακτηριστικά Δανείου	**int_rate**	Επιτόκιο δανείου	Numeric
Χαρακτηριστικά Δανείου	**installment**	Μηνιαία δόση πληρωμής	Numeric
Χαρακτηριστικά Δανείου	**grade**	Κατηγορία κινδύνου (A-G) βάσει Lending Club	Factor
Χαρακτηριστικά Δανείου	**purpose**	Σκοπός δανείου (π.χ. χρέος, αγορά)	Factor
Δημογραφικά & Οικονομικά Στοιχεία			
Δημογραφικά/Οικονομικά	**emp_length**	Έτη εργασιακής εμπειρίας (0-10)	Integer
Δημογραφικά/Οικονομικά	**home_ownership**	Καθεστώς στέγασης (Rent, Mortgage, Own)	Factor
Δημογραφικά/Οικονομικά	**annual_inc**	Ετήσιο δηλωθέν εισόδημα	Numeric
Δημογραφικά/Οικονομικά	**verification_status**	Πιστοποίηση εισοδήματος από την πλατφόρμα	Factor
Δημογραφικά/Οικονομικά	**addr_state**	Πολιτεία κατοικίας του δανειολήπτη	Factor
Πιστωτικό Ιστορικό			
Πιστωτικό Ιστορικό	**fico_avg**	Μέσο σκορ FICO (Πιστοληπτική ικανότητα)	Numeric
Πιστωτικό Ιστορικό	**dti**	Λόγος χρέους προς εισόδημα (Debt-to-Income)	Numeric
Πιστωτικό Ιστορικό	**delinq_2yrs**	Αριθμός καθυστερήσεων >30 ημερών τελευταία 2 έτη	Integer
Πιστωτικό Ιστορικό	**open_acc**	Αριθμός ανοιχτών πιστωτικών γραμμών	Integer
Πιστωτικό Ιστορικό	**pub_rec**	Αριθμός υποτιμητικών δημόσιων εγγραφών	Integer
Πιστωτικό Ιστορικό	**revol_util**	Ποσοστό χρησιμοποίησης πιστωτικού ορίου	Numeric
Πιστωτικό Ιστορικό	**total_acc**	Συνολικός αριθμός πιστωτικών γραμμών	Integer
Πιστωτικό Ιστορικό	**log_annual_inc**	Λογάριθμος ετήσιου εισοδήματος (για κανονικοποίηση)	Numeric

Πίνακας 3.2.1: Ορισμός και ταξινόμηση μεταβλητών της έρευνας.

3.3 Εργαλεία ανάλυσης

Η υλοποίηση της εμπειρικής ανάλυσης βασίστηκε στη χρήση του στατιστικού προγράμματος R, μέσω του ολοκληρωμένου περιβάλλοντος ανάπτυξης RStudio. Η επιλογή του συγκεκριμένου εργαλείου κρίθηκε ιδανική λόγω της εξειδίκευσής του στη διαχείριση μεγάλων συνόλων δεδομένων (big data) και της διαθεσιμότητας προηγμένων βιβλιοθηκών, όπως οι tidyverse και caret, οι οποίες διευκολύνουν τον μετασχηματισμό και τη μοντελοποίηση των δεδομένων. Τα κυριότερα πακέτα λογισμικού που χρησιμοποιήθηκαν για τον μετασχηματισμό, την περιγραφική στατιστική και τη μοντελοποίηση των δεδομένων παρουσιάζονται συγκεντρωτικά στον Πίνακα 3.3.1.

Βιβλιοθήκη (Library)	Κύρια Λειτουργία	Χρήση στην Έρευνα
readr	Εισαγωγή & Ανάγνωση δεδομένων	Γρήγορη ανάγνωση αρχείων .csv, .tsv και .txt.
dplyr	Επεξεργασία & Μετασχηματισμός	Φιλτράρισμα, ομαδοποίηση (group_by) και σύνοψη (summarize).
stringr	Διαχείριση κειμένου (Strings)	Καθαρισμός κειμένων, εύρεση προτύπων και αντικατάσταση χαρακτήρων.
knitr	Δυναμική δημιουργία αναφορών	Μετατροπή του κώδικα σε επαγγελματικά έγγραφα (PDF/HTML).
kableExtra	Προχωρημένη μορφοποίηση πινάκων	Προσθήκη στυλ, χρωμάτων και γραμμών σε πίνακες του RMarkdown.
tidyr	Τακτοποίηση δεδομένων (Data Tidying)	Μετατροπή πινάκων από 'wide' σε 'long' μορφή και αντίστροφα.
car	Στατιστικοί έλεγχοι παλινδρόμησης	Έλεγχος πολυσυγγραμμικότητας (VIF) και ANOVA (Type II/III).
ggplot2	Οπτικοποίηση δεδομένων (Γραφήματα)	Δημιουργία σύνθετων γραφημάτων με το σύστημα layers.

Πίνακας 3.3.1: Συνοπτική Παρουσίαση Στατιστικών Εργαλείων και Βιβλιοθηκών της R

Το κεντρικό εργαλείο της ανάλυσης είναι το υπόδειγμα της Λογιστικής Παλινδρόμησης (logistic regression), το οποίο αποτελεί την ενδεδειγμένη μέθοδο για τη μελέτη δυαδικών εξαρτημένων μεταβλητών. Η μέθοδος αυτή επιτρέπει την εκτίμηση της πιθανότητας αθέτησης πληρωμής συναρτήσει των επεξηγηματικών μεταβλητών, χρησιμοποιώντας τη συνάρτηση logit για τη μετατροπή των γραμμικών συνδυασμών σε πιθανότητες εντός του διαστήματος [0, 1].

Πριν την εφαρμογή του επαγωγικού μοντέλου, η έρευνα εστιάζει στην ποσοτική χαρτογράφηση του δείγματος. Μέσα από τη χρήση της βιβλιοθήκης psych, εξήχθησαν οι κεντρικές τάσεις και οι δείκτες διασποράς για το σύνολο των ποσοτικών μεταβλητών. Τα αποτελέσματα αυτά αποτυπώνονται στον Πίνακα 3.3.2, ο οποίος προσφέρει μια ολοκληρωμένη εικόνα της οικονομικής κατάστασης και του πιστωτικού προφίλ των δανειοληπτών.

Μεταβλητή	Μέσος Όρος	Διάμεσος	Τυπ. Απόκλιση	Ελάχιστο	Μέγιστο
Ποσό Δανείου	14602.69	12075.00	8745.42	500.00	40000.00
Διάρκεια (Μήνες)	41.92	36.00	10.34	36.00	60.00
Επιτόκιο (%)	13.23	12.74	4.77	5.31	30.99
Μηνιαία Δόση	442.97	379.45	262.41	4.93	1719.83
Έτη Προϋπηρεσίας	5.97	6.00	3.69	0.00	10.00
Ετήσιο Εισόδημα	77887.02	65000.00	71018.53	33.00	10999200.00
Δείκτης DTI	18.13	17.52	9.57	-1.00	999.00
delinq_2yrs	0.32	0.00	0.88	0.00	39.00
Ανοιχτοί Λογαριασμοί	11.68	11.00	5.49	1.00	90.00
Δημόσιες Εγγραφές	0.21	0.00	0.60	0.00	86.00
Χρήση Πιστωτικού Ορίου (%)	52.05	52.50	24.48	0.00	892.30
Συνολικοί Λογαριασμοί	25.07	23.00	12.01	2.00	176.00
Σκορ FICO (MO)	698.12	692.00	31.66	627.00	847.50
Δείκτης Αθέτησης (0/1)	0.20	0.00	0.40	0.00	1.00

Πίνακας 3.3.2: Περιγραφικά Στατιστικά Στοιχεία Ποσοτικών Μεταβλητών

Από την επεξεργασία των δεδομένων στον Πίνακα 3.3.2, προκύπτει ότι ο μέσος δανειολήπτης διαθέτει πιστωτικό σκορ FICO 698,12, τιμή που υποδηλώνει ικανοποιητική πιστοληπτική ικανότητα. Το μέσο ετήσιο εισόδημα ανέρχεται σε 77.887,02 δολάρια, ωστόσο η σημαντική απόκλιση από τη διάμεσο τιμή (65.000 δολάρια) υποδεικνύει την παρουσία θετικής ασυμμετρίας στην κατανομή, με την ύπαρξη υψηλών εισοδημάτων που επηρεάζουν τον αριθμητικό μέσο. Παράλληλα, ο μέσος δείκτης χρέους προς εισόδημα (DTI) διαμορφώνεται στο 18,13, επίπεδο που θεωρείται εντός των ορίων χρηματοοικονομικής υγείας. Τέλος, ο δείκτης αθέτησης (default rate) του δείγματος ανέρχεται σε 20%, προσφέροντας την απαραίτητη διακύμανση για την ανάλυση που ακολουθεί.

Η ανάλυση των παραπάνω στατιστικών μεγεθών, σε συνδυασμό με την εξέταση των συχνοτήτων της εξαρτημένης μεταβλητής, επιβεβαιώνει την καταλληλότητα του δείγματος για την περαιτέρω οικονομετρική επεξεργασία. Η

σαφής εικόνα των κατανομών και των αποκλίσεων που παρουσιάζονται στους πίνακες αυτούς, αποτελεί την απαραίτητη προϋπόθεση για την ορθή ερμηνεία των συντελεστών παλινδρόμησης που ακολουθούν στο επόμενο κεφάλαιο.

Κεφάλαιο 4: Ανάλυση δεδομένων και αποτελέσματα

Στο παρόν κεφάλαιο παρουσιάζονται τα αποτελέσματα της εμπειρικής έρευνας, η οποία βασίστηκε στην ανάλυση του δείγματος της LendingClub. Στόχος της ανάλυσης είναι ο προσδιορισμός των παραγόντων που επηρεάζουν την πιθανότητα αθέτησης των δανειακών υποχρεώσεων και η αξιολόγηση της προγνωστικής ικανότητας του προτεινόμενου υποδείγματος. Η παρουσίαση δομείται σε τρεις άξονες: αρχικά, πραγματοποιείται μια εκτενής περιγραφική ανάλυση των δεδομένων για την κατανόηση των χαρακτηριστικών του δείγματος. Στη συνέχεια, παρατίθενται τα αποτελέσματα της λογιστικής παλινδρόμησης, όπου εξετάζεται η στατιστική σημαντικότητα των ανεξάρτητων μεταβλητών. Τέλος, το κεφάλαιο ολοκληρώνεται με τον έλεγχο της αξιοπιστίας και της ευστάθειας του μοντέλου μέσω των κατάλληλων διαγνωστικών δοκιμασιών.

4.1 Περιγραφική ανάλυση δεδομένων

Η ανάλυση των δεδομένων στο παρόν στάδιο επικεντρώνεται στη συγκριτική αξιολόγηση των χαρακτηριστικών μεταξύ των συνεπών δανειοληπτών (Fully Paid) και εκείνων που παρουσίασαν αθέτηση πληρωμής (Charged Off). Η διαφοροποίηση των μέσων τιμών στις δύο αυτές κατηγορίες, όπως αποτυπώνεται στον Πίνακα 4.1.1, δεν αποτελεί απλώς μια αριθμητική παράθεση, αλλά συνιστά την πρώτη ουσιαστική ένδειξη για την ερμηνευτική ισχύ των προσδιοριστικών παραγόντων που θα εξεταστούν στη συνέχεια μέσω του μοντέλου λογιστικής παλινδρόμησης.

Εξετάζοντας το πιστωτικό σκορ FICO, παρατηρείται ότι οι συνεπείς δανειολήπτες εμφανίζουν υψηλότερο μέσο όρο, ο οποίος ανέρχεται σε 700,17 μονάδες, έναντι 689,65 μονάδων της ομάδας αθέτησης. Η διαφορά αυτή, αν και αριθμητικά φαίνεται περιορισμένη, είναι στατιστικά κρίσιμη στο πλαίσιο της πιστωτικής διαστρωμάτωσης. Υποδηλώνει ότι οι δανειολήπτες που τοποθετούνται σε υψηλότερα επίπεδα πιστοληπτικής διαβάθμισης διατηρούν μια εγγενή τάση προς τη συνέπεια, επιβεβαιώνοντας ότι το FICO παραμένει ένας αξιόπιστος, αν και όχι ο μοναδικός, δείκτης για την αρχική πρόβλεψη της συμπεριφοράς του οφειλέτη.

Κατάσταση Δανείου	Μέσο FICO	Μέσο Επιτόκιο (%)	Μέσο Εισόδημα	Πλήθος Δανείων
Εξόφληση (Fully Paid)	700.17	12.62	79209.63	1018716
Αθέτηση (Charged Off)	689.65	15.75	72437.81	247260

Πίνακας 4.1.1: Σύγκριση Χαρακτηριστικών ανά Κατάσταση Δανείου

Εντονότερη και εξαιρετικά ενδιαφέρουσα διαφοροποίηση καταγράφεται στο επιτόκιο δανεισμού (interest rate). Η ομάδα των δανειοληπτών που αθέτησαν επιβαρύνεται με μέσο επιτόκιο ύψους 15,75%, σε σαφή αντίθεση με το 12,62% που αντιστοιχεί στους συνεπείς δανειολήπτες. Η απόκλιση αυτή αναδεικνύει το επιτόκιο ως έναν από τους πλέον καθοριστικούς παράγοντες κινδύνου. Από χρηματοοικονομική σκοπιά, το εύρημα αυτό αντανακλά την αποτελεσματικότητα της τιμολόγησης κινδύνου (risk-based pricing) που πραγματοποιεί η πλατφόρμα: οι δανειολήπτες που το σύστημα ορθώς αναγνώρισε ως υψηλότερου κινδύνου, επιβαρύνθηκαν με υψηλότερα επιτόκια, τα οποία τελικά λειτούργησαν και ως πρόσθετο βάρος στην ικανότητα αποπληρωμής τους.

Παράλληλα, η οικονομική επιφάνεια των δανειοληπτών, όπως εκφράζεται μέσω του ετήσιου εισοδήματος, επιβεβαιώνει τη θεωρία που συνδέει τους διαθέσιμους πόρους με την ανθεκτικότητα έναντι της επισφάλειας. Συγκεκριμένα, οι δανειολήπτες που εξόφλησαν το δάνειό τους διαθέτουν μέσο εισόδημα 79.209,63 δολάρια, ποσό αισθητά υψηλότερο από τα 72.437,81 δολάρια της ομάδας αθέτησης. Η διαφορά αυτή υπογραμμίζει ότι το εισόδημα λειτουργεί ως «μαξιλάρι» ασφαλείας, επιτρέποντας στον δανειολήπτη να ανταπεξέλθει σε απρόβλεπτες δαπάνες χωρίς να διακυβεύεται η εξυπηρέτηση του χρέους του.

Συνοψίζοντας, η εικόνα των περιγραφικών μεγεθών καταδεικνύει ότι το προφίλ του δανειολήπτη υψηλού κινδύνου συντίθεται από έναν συνδυασμό χαμηλότερου πιστωτικού σκορ, περιορισμένης εισοδηματικής ροής και αυξημένης δανειακής επιβάρυνσης. Τα ευρήματα αυτά παρέχουν μια ισχυρή εμπειρική βάση και προετοιμάζουν το έδαφος για την εφαρμογή του επαγωγικού μοντέλου παλινδρόμησης, το οποίο θα επιτρέψει την απομόνωση της επίδρασης κάθε μεταβλητής και τον προσδιορισμό της στατιστικής της σημαντικότητας σε ένα πολυπαραγοντικό πλαίσιο.

4.2 Συγκριτική ανάλυση χαρακτηριστικών δανείων

Μετά την ολοκλήρωση της περιγραφικής επισκόπησης, η ανάλυση προχωρά στην εκτίμηση του υποδείγματος λογιστικής παλινδρόμησης (logistic regression), προκειμένου να προσδιοριστεί η στατιστική σημαντικότητα και η ένταση της επίδρασης των ανεξάρτητων μεταβλητών στην πιθανότητα αθέτησης πληρωμής. Η λογιστικής παλινδρόμησης που χρησιμοποιήθηκε είναι η

$$z = 0{,}9881 - 0{,}0061 * (fico\_avg) + 0{,}1093 * (int\_ate) + 0{,}00002 * (loan\_mnt) \\ - 0{,}000003(annual\_nc) + 0{,}0139 * (dti) - 0{,}0117 * (emp\_ength)$$

η οποία και κρίνεται ως η βέλτιστη επιλογή, καθώς επιτρέπει την απομόνωση της επίδρασης κάθε παράγοντα, διατηρώντας τις υπόλοιπες μεταβλητές σταθερές (ceteris paribus). Τα αποτελέσματα των συντελεστών του μοντέλου, όπως αυτά προέκυψαν από τη στατιστική επεξεργασία στο RStudio, παρουσιάζονται αναλυτικά στον Πίνακα 4.2.1. Το πρώτο και πλέον αξιοσημείωτο εύρημα είναι ότι το σύνολο των επιλεγμένων μεταβλητών παρουσιάζει p-value εξαιρετικά κοντά στο μηδέν ($p < 0{,}001$). Το γεγονός αυτό υποδηλώνει ότι όλοι οι εξεταζόμενοι παράγοντες διαθέτουν υψηλή στατιστική σημαντικότητα σε επίπεδο εμπιστοσύνης 99%, επιβεβαιώνοντας ότι η επιλογή των μεταβλητών εδράζεται σε ισχυρή θεωρητική και εμπειρική βάση.

	Συντελεστής (Beta)	Τυπικό Σφάλμα	z-value	P-Value	
(Intercept)	0.9881002	0.0681693	14.49480	< 0.001	***
fico_avg	-0.0060881	0.0000948	-64.24176	< 0.001	***
int_rate	0.1092793	0.0005269	207.38966	< 0.001	***
loan_amnt	0.0000191	0.0000003	61.96150	< 0.001	***
annual_inc	-0.0000026	0.0000001	-40.93511	< 0.001	***
dti	0.0138815	0.0002711	51.20109	< 0.001	***
emp_length	-0.0116899	0.0006368	-18.35775	< 0.001	***

Πίνακας 4.2.1: Πίνακας Συντελεστών Λογιστικής Παλινδρόμησης

Εστιάζοντας στην κατεύθυνση της επίδρασης, οι αρνητικοί συντελεστές στο πιστωτικό σκορ FICO (-0,0061), στο ετήσιο εισόδημα (<-0,001) και στα έτη

προϋπηρεσίας (-0,0117) υποδηλώνουν μια αντίστροφη σχέση με την εξαρτημένη μεταβλητή. Συγκεκριμένα, η βελτίωση της πιστοληπτικής ικανότητας ή η ενίσχυση των εισοδηματικών πόρων και της εργασιακής σταθερότητας του δανειολήπτη δρουν ανασταλτικά ως προς τον πιστωτικό κίνδυνο, μειώνοντας την πιθανότητα εμφάνισης επισφάλειας. Αντιθέτως, το επιτόκιο δανεισμού (0,1093) και ο δείκτης DTI (0,0139) εμφανίζουν θετικούς συντελεστές. Η διαπίστωση αυτή επιβεβαιώνει ότι η αύξηση του κόστους δανεισμού και η υψηλή χρηματοοικονομική μόχλευση του νοικοκυριού λειτουργούν ως καταλύτες που ενισχύουν σημαντικά την πιθανότητα μη αποπληρωμής.

Προκειμένου να καταστεί δυνατή η άμεση ποσοτική ερμηνεία των αποτελεσμάτων, πραγματοποιήθηκε ο μετασχηματισμός των συντελεστών σε λόγους πιθανοτήτων (odds ratios), οι οποίοι παρατίθενται στον Πίνακα 4.2.2. Η ανάλυση των odds ratios αποκαλύπτει ότι το επιτόκιο (interest rate) αποτελεί τον πλέον καθοριστικό προσδιοριστικό παράγοντα του μοντέλου. Με odds ratio 1,115, προκύπτει ότι για κάθε ποσοστιαία μονάδα αύξησης του επιτοκίου, οι πιθανότητες αθέτησης αυξάνονται κατά 11,5%, γεγονός που αναδεικνύει την ευαισθησία των δανειοληπτών P2P στο κόστος χρήματος.

Παρομοίως, μια μονάδα αύξησης στον δείκτη DTI αυξάνει τις πιθανότητες αθέτησης κατά 1,4% (OR: 1,014), ενώ στον αντίποδα, κάθε επιπλέον έτος προϋπηρεσίας μειώνει την πιθανότητα αθέτησης κατά περίπου 1,2% (OR: 0,988). Ιδιαίτερο ενδιαφέρον παρουσιάζει το εύρος των διαστημάτων εμπιστοσύνης, τα οποία εμφανίζονται εξαιρετικά στενά, με το κατώτερο και το ανώτερο όριο συχνά να ταυτίζονται στο τρίτο δεκαδικό ψηφίο. Η στατιστική αυτή σύμπτωση αποτελεί άμεση συνέπεια του μεγάλου μεγέθους του δείγματος (N>1.000.000), το οποίο ελαχιστοποιεί το τυπικό σφάλμα των εκτιμήσεων, προσδίδοντας στο υπόδειγμα εξαιρετική στατιστική ακρίβεια. Η στατιστική αυτή συνέπεια, σε συνδυασμό με τα στενά διαστήματα εμπιστοσύνης που καταγράφονται, υπογραμμίζει την αξιοπιστία και την ευρωστία του υποδείγματος Τα ευρήματα αυτά παρέχουν μια σαφή εικόνα για την ιεράρχηση των κινδύνων, προσφέροντας στις P2P πλατφόρμες και στους επενδυτές ένα ισχυρό εργαλείο για τη λήψη τεκμηριωμένων αποφάσεων στο πλαίσιο της πιστωτικής τους πολιτικής και της διαχείρισης των χαρτοφυλακίων τους.

	Μεταβλητή	Odds Ratio	Κατώτερο Δ.Ε. (95%)	Ανώτερο Δ.Ε. (95%)
fico_avg	Σκορ FICO (ΜΟ)	0.994	0.994	0.994
int_rate	Επιτόκιο (%)	1.115	1.114	1.117
loan_amnt	Ποσό Δανείου	1.000	1.000	1.000
annual_inc	Ετήσιο Εισόδημα	1.000	1.000	1.000
dti	Δείκτης DTI	1.014	1.013	1.015
emp_length	Έτη Προϋπηρεσίας	0.988	0.987	0.990

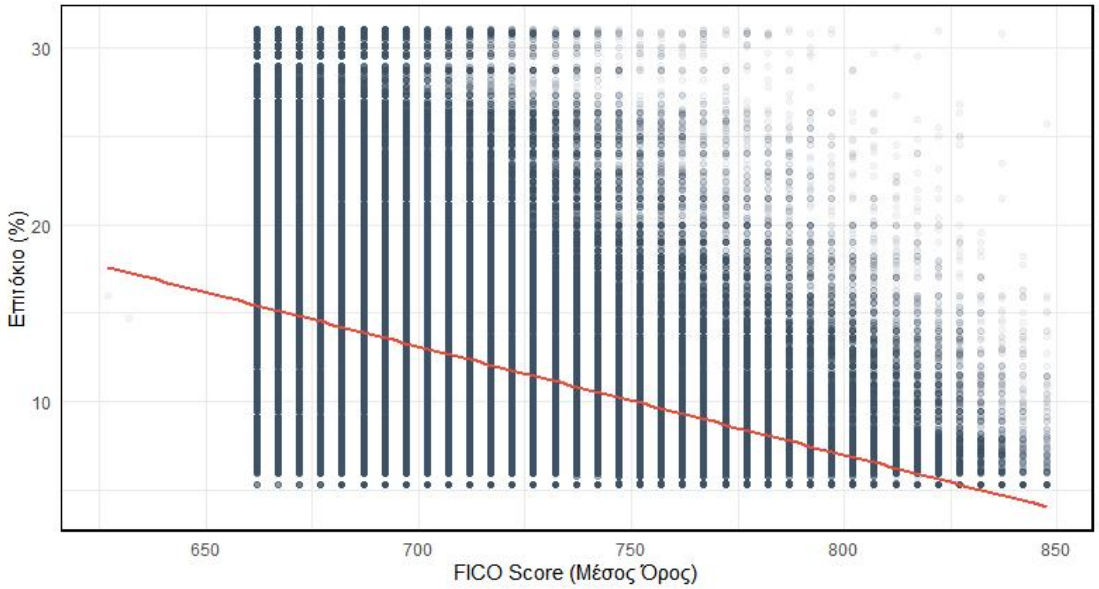
Πίνακας 4.2.2: Πίνακας Odds Ratios και Διαστημάτων Εμπιστοσύνης

4.3 Ανάλυση συσχετίσεων

Η ανάλυση των συσχετίσεων αποτελεί ένα αναπόσπαστο στάδιο της μεθοδολογικής προσέγγισης, καθώς διασφαλίζει τη στατιστική ευστάθεια και την αξιοπιστία του υποδείγματος. Μέσω του πίνακα συσχετίσεων (correlation matrix) που παρουσιάζεται στον Πίνακα 4.3.1, εξετάζεται η ύπαρξη γραμμικής εξάρτησης μεταξύ των επεξηγηματικών μεταβλητών, παράγοντας που θα μπορούσε να αλλοιώσει την ακρίβεια των εκτιμητών. Η ισχυρότερη αρνητική συσχέτιση εντοπίζεται μεταξύ του πιστωτικού σκορ FICO και του επιτοκίου δανεισμού (r = -0,407), εύρημα που επιβεβαιώνεται οπτικά από το Διάγραμμα 4.3.1. Η αρνητική κλίση της γραμμής παλινδρόμησης στο συγκεκριμένο γράφημα διασποράς (scatter plot) αποτελεί εμπειρική απόδειξη ότι η πλατφόρμα ακολουθεί μια αυστηρή πολιτική risk-based pricing. Με απλά λόγια, το κόστος κεφαλαίου μειώνεται γραμμικά καθώς βελτιώνεται η πιστοληπτική ικανότητα του δανειολήπτη, αντανακλώντας τη χαμηλότερη απαίτηση ασφαλίστρου κινδύνου από την πλευρά των επενδυτών.

	Σκορ FICO (ΜΟ)	Επιτόκιο (%)	Ποσό Δανείου	Ετήσιο Εισόδημα	Δείκτης DTI
Σκορ FICO (ΜΟ)	1.000	-0.407	0.102	0.073	-0.071
Επιτόκιο (%)	-0.407	1.000	0.147	-0.071	0.166
Ποσό Δανείου	0.102	0.147	1.000	0.305	0.036
Ετήσιο Εισόδημα	0.073	-0.071	0.305	1.000	-0.156
Δείκτης DTI	-0.071	0.166	0.036	-0.156	1.000

Πίνακας 4.3.1: Πίνακας Συσχετίσεων Ποσοτικών Μεταβλητών



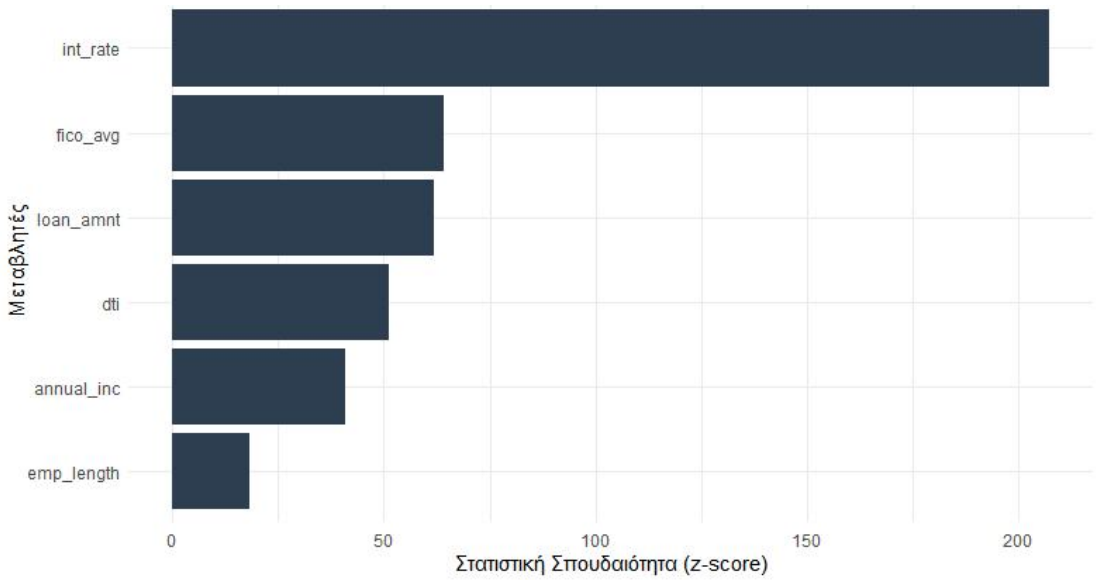
Γράφημα 4.3.1: Σχέση Σκορ FICO και Επιτοκίου

Πέρα από την απλή διμερή συσχέτιση, η έρευνα προχωρά σε έναν πιο αυστηρό έλεγχο της πολυσυγγραμμικότητας μέσω του δείκτη VIF (Variance Inflation Factor). Ο δείκτης αυτός μετρά πόσο αυξάνεται η διακύμανση ενός εκτιμώμενου συντελεστή παλινδρόμησης λόγω της συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών. Τα αποτελέσματα στον Πίνακα 4.3.2 είναι ιδιαίτερα ικανοποιητικά: όλες οι τιμές VIF κυμαίνονται σε χαμηλά επίπεδα, με μέγιστη τιμή το 1,408 (loan amount). Καθώς οι τιμές αυτές υπολείπονται σημαντικά του κρίσιμου ορίου του 5 (ή του 10), επιβεβαιώνεται ότι δεν υφίσταται πρόβλημα αλληλεπικάλυψης της πληροφορίας. Συνεπώς, κάθε μεταβλητή προσφέρει μοναδική ερμηνευτική αξία στο μοντέλο, επιτρέποντας την ασφαλή εξαγωγή συμπερασμάτων για τους μεμονωμένους συντελεστές.

Μεταβλητή	Δείκτης VIF
Σκορ FICO (ΜΟ)	1.155
Επιτόκιο (%)	1.201
Ποσό Δανείου	1.408
Ετήσιο Εισόδημα	1.389
Δείκτης DTI	1.082
Έτη Προϋπηρεσίας	1.016

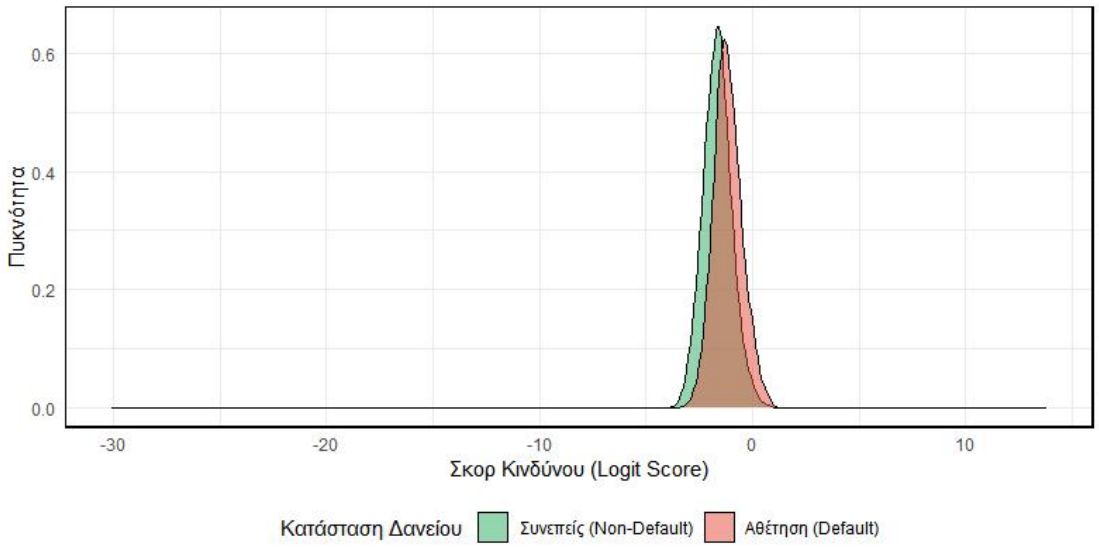
Πίνακας 4.3.2: Έλεγχος Πολυσυγγραμμικότητας (VIF Values)

Η ανάλυση ολοκληρώνεται με την ιεράρχηση της σπουδαιότητας των μεταβλητών (variable importance) στο Γράφημα 4.3.2. Το επιτόκιο δανεισμού (int_rate) αναδεικνύεται ως η κυρίαρχη μεταβλητή με το υψηλότερο στατιστικό z-score, ακολουθούμενο από το σκορ FICO και το ποσό του δανείου. Η ιεράρχηση αυτή αναδεικνύει ότι, στον P2P δανεισμό, οι όροι του δανείου είναι εξίσου σημαντικοί -αν όχι σημαντικότεροι- από τα δημογραφικά στοιχεία του δανειολήπτη.



Γράφημα 4.3.2: Σπουδαιότητα Μεταβλητών

Τέλος, η αξιολόγηση της διαχωριστικής ικανότητας του μοντέλου αποτυπώνεται στο Γράφημα 4.3.3, το οποίο συγκρίνει τις καμπύλες πυκνότητας (density plots) των προβλεπόμενων σκορ κινδύνου. Η σαφής μετατόπιση της καμπύλης της ομάδας αθέτησης (default) προς τα δεξιά, με μέση τιμή logit score -1,167 έναντι -1,655 των συνεπών πελατών, αποτελεί ισχυρή ένδειξη ότι το μοντέλο «αντιλαμβάνεται» τον κίνδυνο πριν αυτός εκδηλωθεί. Η διαφορά αυτή στις μέσες τιμές των προβλεπόμενων πιθανοτήτων επιβεβαιώνει ότι το υπόδειγμα λογιστικής παλινδρόμησης διαθέτει την απαραίτητη ευαισθησία για να διακρίνει τα ποιοτικά χαρακτηριστικά που οδηγούν στην αθέτηση, καθιστώντας το ένα αξιόπιστο εργαλείο για την προληπτική διαχείριση του πιστωτικού κινδύνου.



Γράφημα 4.3.3: Κατανομή Σκορ Κινδύνου

4.4 Στατιστική μοντελοποίηση

Η ολοκλήρωση της στατιστικής ανάλυσης απαιτεί τον αυστηρό έλεγχο της προγνωστικής ευστοχίας του υποδείγματος, προκειμένου να διαπιστωθεί ο βαθμός στον οποίο οι επιλεγμένες ανεξάρτητες μεταβλητές διαθέτουν την απαραίτητη διακριτική ικανότητα για τον διαχωρισμό των συνεπών από τους μη συνεπείς δανειολήπτες. Η αξιολόγηση αυτή πραγματοποιείται μέσω του πίνακα σύγχυσης (confusion matrix), ο οποίος αποτελεί το εργαλείο αναφοράς για την αποτύπωση της ικανότητας ταξινόμησης του μοντέλου σε ένα σύνολο δεδομένων που δεν χρησιμοποιήθηκε κατά την εκπαίδευση.

Όπως προκύπτει από τα αποτελέσματα που παρατίθενται στον Πίνακα 4.4.1, το μοντέλο επιτυγχάνει μια ιδιαίτερα ικανοποιητική συνολική ακρίβεια (accuracy) της τάξης του 80,46%. Συγκεκριμένα, το υπόδειγμα κατάφερε να ταυτοποιήσει ορθά 1.005.015 περιπτώσεις συνεπών δανειοληπτών (true negatives). Ωστόσο, η πρόβλεψη της πραγματικής αθέτησης (true positives), η οποία ανήλθε σε 13.640 περιπτώσεις, αναδεικνύει μια εγγενή πρόκληση των χρηματοοικονομικών δεδομένων: την ασυμμετρία των κλάσεων. Το γεγονός ότι το μοντέλο εμφανίζεται "συντηρητικό" υποδηλώνει μια τάση προς την αποφυγή λανθασμένων συναγερμών (false positives), προτιμώντας να ταξινομεί δάνεια ως συνεπή παρά να αποκλείει δανειολήπτες χωρίς ισχυρές ενδείξεις κινδύνου.

	Προβλέψεις Μοντέλου	
	Πρόβλεψη: Εξόφληση (0)	**Πρόβλεψη: Αθέτηση (1)**
Πραγματικό: Εξόφληση (0)	1005015	13701
Πραγματικό: Αθέτηση (1)	233620	13640

Πίνακας 4.4.1: Σύγχυσης (confusion Matrix) - Συνολικής Ακρίβειας 80,46%

Η διαχωριστική ικανότητα του μοντέλου ενισχύεται περαιτέρω από την ανάλυση της κατανομής του προβλεπόμενου σκορ κινδύνου (logit score) ανά ομάδα. Σύμφωνα με τα ευρήματα του Πίνακα 4.4.2, παρατηρείται μια σαφής και στατιστικά σημαντική διαφοροποίηση των μέσων τιμών. Οι συνεπείς δανειολήπτες παρουσιάζουν μέσο σκορ -1,655, ενώ η ομάδα αθέτησης εμφανίζει αισθητά υψηλότερο μέσο σκορ, ίσο με -1,167.

Η μετατόπιση αυτή των μέσων τιμών, σε συνδυασμό με τη συμπεριφορά των καμπυλών πυκνότητας, επιβεβαιώνει ότι το μοντέλο αναγνωρίζει και αποδίδει υψηλότερη πιθανότητα κινδύνου στα δάνεια που τελικά οδηγήθηκαν σε επισφάλεια. Από τη σκοπιά της επιστήμης των δεδομένων, η απόσταση μεταξύ αυτών των δύο μέσων τιμών αποτελεί ένδειξη της ισχύος του σήματος (signal strength) που εκπέμπουν οι μεταβλητές όπως το επιτόκιο και το σκορ FICO.

Default (0=Όχι, 1=Ναι)	**Μέσο Σκορ Z**
0	-1.655
1	-1.167

Πίνακα 4.4.2: Μέσο Σκορ Κινδύνου ανά Ομάδα

Συμπερασματικά, η στατιστική μοντελοποίηση αναδεικνύει ότι η συνδυαστική μελέτη των οικονομικών και πιστωτικών χαρακτηριστικών προσφέρει μια στέρεη και επιστημονικά τεκμηριωμένη βάση για τη διαχείριση του κινδύνου. Παρά την πολυπλοκότητα των σύγχρονων χρηματοοικονομικών συνθηκών και την εγγενή αβεβαιότητα της ανθρώπινης συμπεριφοράς, η επίδοση του 80,46%, η οποία προκύπτει από τον λόγο των ορθών ταξινομήσεων προς το σύνολο των παρατηρήσεων και ως εξίσωση γράφεται:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Όπου:
TP (True Positives): Ορθές προβλέψεις αθέτησης.
TN (True Negatives): Ορθές προβλέψεις αποπληρωμής.
FP (False Positives): Λανθασμένες προβλέψεις αθέτησης (Σφάλμα Τύπου Ι).
FN (False Negatives): Λανθασμένες προβλέψεις αποπληρωμής (Σφάλμα Τύπου ΙΙ).

καθιστά το προτεινόμενο υπόδειγμα ένα αξιόπιστο εργαλείο υποστήριξης αποφάσεων (decision support system). Το μοντέλο παρέχει τη δυνατότητα στις πλατφόρμες P2P να βελτιστοποιήσουν τις διαδικασίες χορήγησης, εξισορροπώντας την ανάγκη για ανάπτυξη με την επιτακτική ανάγκη για διασφάλιση των επενδυτικών κεφαλαίων.

Κεφάλαιο 5: Συζήτηση αποτελεσμάτων

Σε αυτό κεφάλαιο παρουσιάζονται τα βασικά συμπεράσματα που προέκυψαν από την εμπειρική ανάλυση της πιστοληπτικής αξιολόγησης στην πλατφόρμα LendingClub.

5.1 Ερμηνεία ευρημάτων

Η παρούσα έρευνα κατέδειξε ότι η αθέτηση των δανειακών υποχρεώσεων στην πλατφόρμα LendingClub δεν συνιστά ένα στοχαστικό ή τυχαίο γεγονός, αλλά αποτελεί το αποτέλεσμα συστηματικών επιδράσεων ενός συγκεκριμένου συνόλου χρηματοοικονομικών και δημογραφικών παραμέτρων. Η ερμηνεία των αποτελεσμάτων του υποδείγματος λογιστικής παλινδρόμησης αναδεικνύει το επιτόκιο δανεισμού (interest rate) ως τον ισχυρότερο προγνωστικό παράγοντα κινδύνου. Η θετική συσχέτιση του επιτοκίου με την πιθανότητα αθέτησης -όπως υποδεικνύεται από τον υψηλό λόγο πιθανοτήτων (OR: 1,115)- επιβεβαιώνει τη θεωρητική βάση της «δυσμενούς επιλογής» (adverse selection). Το εύρημα αυτό ευθυγραμμίζεται πλήρως με τη μελέτη των Serrano-Cinca et al. (2015), υποδηλώνοντας ότι οι δανειολήπτες που αποδέχονται υψηλά επιτόκια είναι συχνά εκείνοι που στερούνται πρόσβασης στο παραδοσιακό τραπεζικό σύστημα, φέροντας εγγενώς υψηλότερο πιστωτικό κίνδυνο.

Παράλληλα, η σημασία του πιστωτικού σκορ FICO αναδείχθηκε ως θεμελιώδης, επικυρώνοντας την υπόθεση ότι η ιστορική πιστωτική συμπεριφορά παραμένει ένας από τους πλέον αξιόπιστους δείκτες για τη μελλοντική συνέπεια των πληρωμών. Ωστόσο, η έρευνα αποκάλυψε ότι το FICO score, παρά την ισχύ του, δεν αποτελεί πανάκεια για την αξιολόγηση. Η στατιστικά σημαντική επίδραση του ετήσιου εισοδήματος και, κυρίως, του δείκτη χρέους προς εισόδημα (DTI), αναδεικνύει την ανάγκη εξέτασης της τρέχουσας ρευστότητας. Το εύρημα αυτό έρχεται να ενισχύσει τα συμπεράσματα των Emekter et al. (2015), καταδεικνύοντας ότι η υπερχρέωση λειτουργεί ως καταλύτης για την επισφάλεια, ακόμα και σε περιπτώσεις δανειοληπτών με ικανοποιητικό πιστωτικό παρελθόν. Η υπερβολική μόχλευση μειώνει την ικανότητα απορρόφησης οικονομικών κλυδωνισμών, μετατρέποντας μια παροδική στενότητα σε οριστική αθέτηση.

Ιδιαίτερο ενδιαφέρον παρουσιάζει το εύρημα σχετικά με την επαγγελματική προϋπηρεσία. Η αρνητική συσχέτιση μεταξύ των ετών εργασίας και της πιθανότητας

αθέτησης (OR: 0,988) υποδηλώνει ότι η επαγγελματική σταθερότητα λειτουργεί ως προστατευτικός παράγοντας. Από χρηματοοικονομική σκοπιά, η μακροχρόνια απασχόληση συνδέεται με μεγαλύτερη προβλεψιμότητα των μελλοντικών ταμειακών ροών (cash flows), μειώνοντας την αβεβαιότητα που αντιμετωπίζει ο επενδυτής. Το γεγονός αυτό υπογραμμίζει την αξία των «μαλακών» δεδομένων (soft data) στην ενίσχυση της ακρίβειας των μοντέλων ταξινόμησης.

Συνολικά, η υψηλή προγνωστική ακρίβεια του μοντέλου (80,46%) υποστηρίζει το συμπέρασμα ότι ο συνδυασμός της τιμολόγησης βάσει κινδύνου (risk-based pricing) και της πολυπαραγοντικής αξιολόγησης αποτελεί την πλέον αποτελεσματική στρατηγική για τον μετριασμό των απωλειών. Η παρούσα μελέτη αποδεικνύει ότι στο περιβάλλον του P2P δανεισμού, όπου η ασύμμετρη πληροφόρηση είναι έντονη, η επιστήμη των δεδομένων προσφέρει τα απαραίτητα εργαλεία για τη μετατροπή της αβεβαιότητας σε μετρήσιμο και διαχειρίσιμο κίνδυνο, θωρακίζοντας την αξιοπιστία της πλατφόρμας έναντι των επενδυτών.

5.2 Σύγκριση με τη βιβλιογραφία

Τα ευρήματα της παρούσας μελέτης παρουσιάζουν αξιοσημείωτη σύγκλιση με την υπάρχουσα διεθνή βιβλιογραφία που διερευνά τον πιστωτικό κίνδυνο εντός του οικοσυστήματος του P2P δανεισμού. Η κυρίαρχη σημασία του επιτοκίου και του πιστωτικού σκορ FICO, όπως αναδείχθηκε από το μοντέλο μας, επικυρώνει τα συμπεράσματα των Serrano-Cinca et al. (2015). Οι ερευνητές αυτοί υποστήριξαν ότι τα παραδοσιακά χρηματοοικονομικά κριτήρια όχι μόνο διατηρούν την ερμηνευτική τους ισχύ στις ψηφιακές πλατφόρμες, αλλά αποτελούν και τον πυρήνα της διαδικασίας αποδιαμεσολάβησης. Η διαπίστωση ότι το επιτόκιο συνιστά τον ισχυρότερο προγνωστικό παράγοντα αθέτησης επαληθεύει τη υπόθεση ότι η LendingClub επιτυγχάνει μια αποτελεσματική τιμολόγηση κινδύνου, ενσωματώνοντας την πληροφορία της επισφάλειας απευθείας στο κόστος δανεισμού.

Επιπρόσθετα, η στατιστική σημαντικότητα του δείκτη DTI (Debt to Income) και του ετήσιου εισοδήματος που προέκυψε από την ανάλυσή μας, προσφέρει επιπλέον εμπειρικά δεδομένα που ενισχύουν τις θέσεις των Lin et al. (2013). Οι συγγραφείς αυτοί τόνισαν ότι η τρέχουσα χρηματοοικονομική ικανότητα (financial capacity) και η ρευστότητα του δανειολήπτη αποτελούν πιο άμεσους δείκτες κινδύνου από ό,τι το μακροχρόνιο πιστωτικό ιστορικό, ειδικά σε περιβάλλοντα όπου

οι οικονομικές συνθήκες μεταβάλλονται ραγδαία. Η συνολική ακρίβεια του υποδείγματος (80,46%) κινείται στα ανώτερα επίπεδα αντίστοιχων ερευνών που εφάρμοσαν λογιστική παλινδρόμηση σε εκτενή σύνολα δεδομένων (big data), αποδεικνύοντας ότι η μέθοδος αυτή παραμένει στατιστικά εύρωστη και συγκρίσιμη σε σχέση με πιο υπολογιστικά απαιτητικούς αλγορίθμους μηχανικής μάθησης.

Μια ενδιαφέρουσα απόκλιση ή μάλλον μια διαφοροποίηση έμφασης εντοπίζεται στη χρήση των ποιοτικών χαρακτηριστικών. Ενώ μέρος της σύγχρονης βιβλιογραφίας εστιάζει στην άνοδο των "εναλλακτικών δεδομένων" (soft data), όπως ο σκοπός του δανείου ή η γεωγραφική θέση, η δική μας ανάλυση κατέδειξε τη συντριπτική υπεροχή των ποσοτικών μεταβλητών (hard data). Αυτό υποδηλώνει ότι, παρά την τεχνολογική εξέλιξη, οι "σκληροί" χρηματοοικονομικοί δείκτες παραμένουν η «ραχοκοκαλιά» της πιστωτικής αξιολόγησης. Η διαπίστωση αυτή προσδίδει μια ρεαλιστική διάσταση στην έρευνα, υποδεικνύοντας ότι η αποτελεσματική διαχείριση κινδύνου στο FinTech πρέπει να παραμένει αγκυρωμένη στις θεμελιώδεις αρχές της τραπεζικής ανάλυσης.

Η σύγκλιση αυτή των αποτελεσμάτων μας με την καθιερωμένη διεθνή βιβλιογραφία δεν προσδίδει μόνο εγκυρότητα στα συμπεράσματα της παρούσας εργασίας, αλλά επιβεβαιώνει και τη δυνατότητα γενίκευσης (generalizability) του προτεινόμενου μοντέλου. Η μελέτη καταφέρνει να γεφυρώσει το κενό μεταξύ της παραδοσιακής οικονομικής θεωρίας και της σύγχρονης ανάλυσης δεδομένων, προσφέροντας ένα πλαίσιο που είναι ταυτόχρονα επιστημονικά ακριβές και πρακτικά εφαρμόσιμο από τους συμμετέχοντες στην αγορά του P2P δανεισμού.

5.3 Πρακτικές επιπτώσεις

Τα αποτελέσματα της παρούσας μελέτης υπερβαίνουν το πλαίσιο της στατιστικής ανάλυσης, καθώς προσφέρουν άμεσες πρακτικές εφαρμογές για τη λειτουργία των ψηφιακών πλατφορμών και τη στρατηγική διαχείριση του πιστωτικού κινδύνου. Η διαπίστωση ότι το επιτόκιο δανεισμού και το σκορ FICO αποτελούν τους κυρίαρχους προσδιοριστικούς παράγοντες της αθέτησης, επικυρώνει τη λειτουργική επάρκεια των μηχανισμών τιμολόγησης βάσει κινδύνου (risk-based pricing). Για τους διαχειριστές των πλατφορμών, η διαπίστωση αυτή υποδηλώνει ότι η περαιτέρω βελτιστοποίηση των αλγορίθμων για τα δάνεια υψηλού επιτοκίου -όπου η πιθανότητα αθέτησης αυξάνεται κατά 11,5% ανά μονάδα επιτοκίου- μπορεί να οδηγήσει σε

δραστική μείωση των επισφαλειών, θωρακίζοντας την απόδοση των επενδυτικών κεφαλαίων.

Επιπλέον, η ανάδειξη της στατιστικής βαρύτητας του δείκτη DTI και της επαγγελματικής προϋπηρεσίας υπογραμμίζει την ανάγκη για μια ολιστική προσέγγιση στην αξιολόγηση της πιστοληπτικής ικανότητας. Οι πλατφόρμες P2P οφείλουν να ενσωματώσουν δυναμικά δεδομένα που αντικατοπτρίζουν την τρέχουσα οικονομική ανθεκτικότητα του δανειολήπτη, πέρα από το στατικό πιστωτικό ιστορικό. Οι επενδυτές μπορούν να αξιοποιήσουν αυτά τα ευρήματα για τη διαμόρφωση στρατηγικών επιλεκτικής τοποθέτησης (selective lending). Δίνοντας προτεραιότητα σε δανειολήπτες με εργασιακή σταθερότητα και χαμηλή μόχλευση, οι επενδυτές μπορούν να επιτύχουν βελτιωμένες προσαρμοσμένες ως προς τον κίνδυνο απόδοσης (risk-adjusted returns), αποφεύγοντας δάνεια που, παρά το ελκυστικό επιτόκιο, φέρουν δυσανάλογα υψηλή πιθανότητα επισφάλειας.

Τέλος, η υψηλή προγνωστική ακρίβεια του υποδείγματος (80,46%) αναδεικνύει την επιστήμη των δεδομένων ως το κεντρικό εργαλείο για την αυτοματοποίηση της πιστωτικής αξιολόγησης. Η δυνατότητα πρόβλεψης της αθέτησης με υψηλά ποσοστά ευστοχίας επιτρέπει τη μείωση του λειτουργικού κόστους μέσω της αυτοματοποιημένης διαλογής (screening) των αιτήσεων. Η ενσωμάτωση τέτοιων μοντέλων στη διαδικασία λήψης αποφάσεων δεν ενισχύει μόνο τη χρηματοοικονομική ευστάθεια των FinTech οργανισμών, αλλά αποτελεί και κρίσιμο παράγοντα για την οικοδόμηση εμπιστοσύνης (trust building) στην αγορά. Σε ένα περιβάλλον όπου η ταχύτητα και η ακρίβεια αποτελούν τα κύρια ανταγωνιστικά πλεονεκτήματα, η παρούσα μεθοδολογία προσφέρει έναν οδικό χάρτη για τη μετάβαση σε ένα πιο διαφανές και αποτελεσματικό σύστημα εναλλακτικού δανεισμού.

Κεφάλαιο 6: Συμπεράσματα και μελλοντική έρευνα

Στο έκτο και καταληκτικό κεφάλαιο της παρούσας εργασίας επιχειρείται μια συνολική ανακεφαλαίωση των κυριότερων συμπερασμάτων που προέκυψαν από τη μελέτη της πιστοληπτικής αξιολόγησης στον τομέα του Peer-to-Peer δανεισμού. Μέσα από τη σύνθεση των θεωρητικών αναζητήσεων και των εμπειρικών αποτελεσμάτων, διατυπώνονται οι τελικές διαπιστώσεις σχετικά με τη λειτουργία των μοντέλων πρόβλεψης κινδύνου. Παράλληλα, αναγνωρίζονται οι περιορισμοί που εντοπίστηκαν κατά τη διάρκεια της ερευνητικής διαδικασίας και προτείνονται συγκεκριμένες κατευθύνσεις για τη μελλοντική διερεύνηση του θέματος, λαμβάνοντας υπόψη τις συνεχείς τεχνολογικές και θεσμικές μεταβολές στο διεθνές χρηματοοικονομικό περιβάλλον.

6.1 Συμπεράσματα

Η ολοκλήρωση της παρούσας διπλωματικής εργασίας επιτρέπει τη διατύπωση ορισμένων θεμελιωδών συμπερασμάτων σχετικά με τους μηχανισμούς λειτουργίας και τις παραμέτρους του πιστωτικού κινδύνου στον ψηφιακό δανεισμό. Το πρωταρχικό συμπέρασμα που αναδεικνύεται είναι ότι η λογιστική παλινδρόμηση, παρά την ανάδυση πιο πολύπλοκων αρχιτεκτονικών μηχανικής μάθησης, παραμένει ένα εξαιρετικά ισχυρό, ερμηνεύσιμο και αξιόπιστο εργαλείο για την πρόβλεψη της αθέτησης υποχρεώσεων. Η επίτευξη συνολικής ακρίβειας (accuracy) της τάξης του 80,46% καταδεικνύει ότι τα δεδομένα που συλλέγονται από τις πλατφόρμες P2P διαθέτουν την απαραίτητη πληροφοριακή πυκνότητα για τη δημιουργία εύρωστων μοντέλων αξιολόγησης, ικανών να υποκαταστήσουν αποτελεσματικά τις παραδοσιακές διαδικασίες τραπεζικής διαμεσολάβησης.

Αναφορικά με τους προσδιοριστικούς παράγοντες της επισφάλειας, η έρευνα επικύρωσε ότι το επιτόκιο δανεισμού και το πιστωτικό σκορ FICO συνιστούν τους ακρογωνιαίους λίθους της πιστοληπτικής αξιολόγησης. Η υψηλή στατιστική σημαντικότητα αυτών των μεταβλητών υπογραμμίζει ότι η αγορά του P2P δανεισμού διέπεται από ορθολογικά χρηματοοικονομικά κριτήρια, ενσωματώνοντας τον κίνδυνο απευθείας στην τιμολογιακή πολιτική (risk-based pricing). Ταυτόχρονα, η επιρροή παραμέτρων όπως ο δείκτης DTI και η εργασιακή προϋπηρεσία αναδεικνύει μια κρίσιμη διάσταση: η συνέπεια των πληρωμών δεν είναι μόνο συνάρτηση του

παρελθόντος, αλλά κυρίως της τρέχουσας χρηματοοικονομικής ανθεκτικότητας και της επαγγελματικής σταθερότητας του δανειολήπτη.

Συμπερασματικά, η μελέτη αποδεικνύει ότι ο εκδημοκρατισμός της πίστης και η διεύρυνση της πρόσβασης σε κεφάλαια μέσω των FinTech πλατφορμών μπορεί να συνυπάρξει αρμονικά με τη χρηματοοικονομική ασφάλεια. Η επιτυχία αυτού του μοντέλου βασίζεται στην εφαρμογή αυστηρών στατιστικών κριτηρίων και στη συνεχή παρακολούθηση των δεδομένων. Το προτεινόμενο υπόδειγμα προσφέρει μια ισορροπημένη και επιστημονικά τεκμηριωμένη προσέγγιση, η οποία μπορεί να αποτελέσει τη βάση για τη λήψη αποφάσεων τόσο από τους διαχειριστές των πλατφορμών όσο και από τους επενδυτές. Εν κατακλείδι, η αξιοποίηση της επιστήμης των δεδομένων στον εναλλακτικό δανεισμό δεν αποτελεί απλώς μια τεχνολογική επιλογή, αλλά την απαραίτητη δικλείδα ασφαλείας για τη μακροπρόθεσμη βιωσιμότητα και την αξιοπιστία του σύγχρονου χρηματοπιστωτικού συστήματος.

6.2 Περιορισμοί της μελέτης

Παρά την υψηλή προγνωστική ικανότητα του υποδείγματος και τη στατιστική ευρωστία των αποτελεσμάτων, η παρούσα μελέτη υπόκειται σε ορισμένους περιορισμούς, η κατανόηση των οποίων είναι απαραίτητη για την ορθή πλαισίωση των ευρημάτων.

Ο πρωταρχικός περιορισμός εστιάζεται στην εξωτερική εγκυρότητα (external validity) των δεδομένων. Η ανάλυση βασίστηκε αποκλειστικά σε δευτερογενή στοιχεία της πλατφόρμας LendingClub, γεγονός που συνεπάγεται ότι τα συμπεράσματα αντικατοπτρίζουν τις ιδιαιτερότητες της αγοράς των ΗΠΑ. Λόγω των διαφορών στα κανονιστικά πλαίσια, στις καταναλωτικές συνήθειες και στα συστήματα πιστωτικής αναφοράς (credit reporting), η γενίκευση των αποτελεσμάτων σε αναδυόμενες αγορές ή στην ευρωπαϊκή αγορά P2P δανεισμού πρέπει να γίνεται με προσοχή.

Επιπλέον, η μελέτη υιοθέτησε μια προσέγγιση βασισμένη σε «σκληρά» δεδομένα (hard data), εστιάζοντας σε ποσοτικές και χρηματοοικονομικές μεταβλητές. Αυτή η επιλογή, αν και προσφέρει αντικειμενικότητα, αφήνει εκτός ανάλυσης ποιοτικούς παράγοντες και συμπεριφορικές μεταβλητές (behavioral variables), όπως η ψυχολογία του δανειολήπτη, ο βαθμός χρηματοοικονομικού εγγραμματισμού ή η επίδραση κοινωνικών δικτύων. Παράλληλα, το μοντέλο δεν ενσωμάτωσε

μακροοικονομικές μεταβλητές (π.χ. μεταβολές στο ΑΕΠ ή τα επιτόκια της κεντρικής τράπεζας), οι οποίες αποδεδειγμένα επηρεάζουν τη συστημική πιθανότητα αθέτησης.

Από μεθοδολογικής πλευράς, η επιλογή της λογιστικής παλινδρόμησης προσφέρει μέγιστη ερμηνευσιμότητα, ωστόσο ως παραμετρικό μοντέλο βασίζεται στην παραδοχή γραμμικών σχέσεων. Είναι πιθανό το υπόδειγμα να μην αποτυπώνει πλήρως τις πολύπλοκες, μη γραμμικές αλληλεπιδράσεις μεταξύ των μεταβλητών, τις οποίες θα μπορούσαν ενδεχομένως να ανιχνεύσουν αλγόριθμοι μηχανικής μάθησης, όπως τα Gradient Boosting Trees ή τα Νευρωνικά Δίκτυα.

Τέλος, η παρούσα ανάλυση αποτελεί μια στατική απεικόνιση (cross-sectional analysis) του πιστωτικού κινδύνου. Καθώς τα δεδομένα αφορούν μια συγκεκριμένη χρονική υστέρηση, δεν καθίσταται δυνατή η δυναμική παρακολούθηση της πιστοληπτικής ικανότητας σε πραγματικό χρόνο (real-time credit monitoring). Η έλλειψη δεδομένων για τη συμπεριφορά των δανειοληπτών υπό συνθήκες ακραίων οικονομικών κλυδωνισμών περιορίζει τη δυνατότητα διεξαγωγής ασκήσεων προσομοίωσης ακραίων καταστάσεων (stress testing), οι οποίες θα παρείχαν μια πιο σφαιρική εικόνα της ανθεκτικότητας του P2P οικοσυστήματος.

6.3 Προτάσεις για μελλοντική έρευνα

Η παρούσα εργασία θέτει τις βάσεις για μια συστηματική διερεύνηση του πιστωτικού κινδύνου στον δυναμικά αναπτυσσόμενο τομέα του Peer-to-Peer δανεισμού. Ωστόσο, η ραγδαία εξέλιξη των χρηματοοικονομικών τεχνολογιών και η διαθεσιμότητα νέων μορφών πληροφόρησης ανοίγουν νέους ορίζοντες για την ερευνητική κοινότητα.

Μια κεντρική κατεύθυνση για μελλοντική έρευνα αποτελεί η εφαρμογή και συγκριτική αξιολόγηση μη παραμετρικών αλγορίθμων μηχανικής μάθησης (non-parametric machine learning). Η χρήση μοντέλων όπως τα random forests, το XGBoost ή τα βαθιά νευρωνικά δίκτυα (deep learning) θα μπορούσε να αναδείξει πολύπλοκες, μη γραμμικές αλληλεπιδράσεις μεταξύ των μεταβλητών που η λογιστική παλινδρόμηση αδυνατεί να αποτυπώσει. Ιδιαίτερη έμφαση θα πρέπει να δοθεί στη βελτίωση της ευαισθησίας (sensitivity) των μοντέλων, ώστε να εντοπίζονται με μεγαλύτερη ακρίβεια οι περιπτώσεις αθέτησης (true positives), μειώνοντας το κόστος του πιστωτικού κινδύνου για τους επενδυτές.

Επιπλέον, παρουσιάζει εξαιρετικό ενδιαφέρον η ενσωμάτωση εναλλακτικών πηγών δεδομένων (alternative data). Μελλοντικές μελέτες θα μπορούσαν να εξετάσουν την επίδραση μεταβλητών που σχετίζονται με το ψηφιακό αποτύπωμα των χρηστών, τη συμπεριφορά τους στο ηλεκτρονικό εμπόριο ή ακόμα και τη χρήση ψυχομετρικών τεστ για την αξιολόγηση της συνέπειας. Η προσθήκη δεδομένων που αφορούν τα κριτήρια ESG (Environmental, Social, and Governance) θα μπορούσε επίσης να αναδείξει αν οι κοινωνικά υπεύθυνοι δανειολήπτες επιδεικνύουν υψηλότερα ποσοστά αποπληρωμής, συνδέοντας τη βιωσιμότητα με την πιστοληπτική ικανότητα.

Παράλληλα, η μετάβαση από τη στατική ανάλυση στη δυναμική μοντελοποίηση (dynamic credit scoring) κρίνεται επιβεβαιωμένη. Η ενσωμάτωση μακροοικονομικών δεικτών, όπως ο πληθωρισμός, το ποσοστό ανεργίας και οι μεταβολές των διατραπεζικών επιτοκίων, θα επέτρεπε τη μελέτη της ανθεκτικότητας των P2P χαρτοφυλακίων σε διαφορετικές φάσεις του οικονομικού κύκλου. Μια τέτοια προσέγγιση θα διευκόλυνε τη διενέργεια δυναμικών stress tests, προσφέροντας στους διαχειριστές πλατφορμών τη δυνατότητα να αναπροσαρμόζουν την τιμολογιακή τους πολιτική σε πραγματικό χρόνο.

Τέλος, η διενέργεια συγκριτικών μελετών μεταξύ διαφορετικών γεωγραφικών και κανονιστικών περιβαλλόντων (π.χ. σύγκριση της ώριμης αγοράς των ΗΠΑ με την αναδυόμενη ευρωπαϊκή αγορά υπό το πρίσμα του κανονισμού ECSP) θα παρείχε πολύτιμα συμπεράσματα για την επίδραση των θεσμικών παραγόντων στον πιστωτικό κίνδυνο. Η συνεχής επικαιροποίηση των μοντέλων αξιολόγησης είναι επιτακτική, διασφαλίζοντας ότι η καινοτομία στον δανεισμό θα παραμείνει ευθυγραμμισμένη με την ανάγκη για χρηματοοικονομική σταθερότητα και την προστασία των συμμετεχόντων στην ψηφιακή οικονομία.

Βιβλιογραφία

Arner, D. W., Barberis, J., & Buckley, R. P. (2015). The evolution of Fintech: A new post-crisis paradigm. Georgetown Journal of International Law, 47, 1271.

Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for financing of small and medium enterprises. Journal of Banking & Finance, 30(11), 2945-2966.

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. The Review of Financial Studies, 25(8), 2455-2484.

Einav, L., Jenkins, M., & Levin, J. (2012). The impact of credit scoring on consumer lending. The RAND Journal of Economics, 43(2), 249-274.

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending platform. International Review of Financial Analysis, 43, 54-70.

Gomber, P., Koch, J. A., & Siering, M. (2017). Digital Finance and FinTech: current research and future research directions. Journal of Business Economics, 87(5), 537-580.

Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell me a good story and I will lend you money: The role of narratives in peer-to-peer lending decisions. Journal of Marketing Research, 48(SPL), S138-S149.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. 3rd Edition. John Wiley & Sons.

Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2015). Screening peers softly: Inferring the quality of small borrowers. Management Science, 62(6), 1554-1577.

Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub platform. Financial Management, 48(4), 1009-1029.

LendingClub (2024). LendingClub Statistics and Data Downloads. [Online] Available at: https://www.lendingclub.com/info/download-data.action [Accessed 20 May 2024].

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Social networks and adverse selection in online peer-to-peer lending. Management Science, 59(1), 17-35.

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in peer-to-peer lending via predictive modeling. Decision Support Systems, 78, 1-14.

Milne, A., & Parboteeah, P. (2016). The Business Models and Economics of Peer-to-Peer Lending. SSRN Electronic Journal.

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit-scoring models in peer-to-peer lending. Decision Support Systems, 84, 31-44.

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. PloS One, 10(10), e0139427.

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. The American Economic Review, 71(3), 393-410.

Thakor, A. V. (2020). Fintech and banking: What do we know? Journal of Financial Intermediation, 41, 100833.

West, D. (2000). Neural network credit scoring models. Computers & Operations Research, 27(11-12), 1131-1152.

Zhang, J. J., & Liu, P. (2012). Rational herding in microloan markets: Evidence from an online lender. Management Science, 58(5), 892-912.

ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Παράρτημα: Κώδικας προγραμματισμού σε RStudio

```
#Οι βιβλιοθήκες που χρειάστηκα
library(readr)
library(dplyr)
library(stringr)
library(knitr)
library(kableExtra)
library(tidyr)
library(car)
library(ggplot2)
```

```
#εισαγωγή δεδομένων
accepted_2007_to_2018Q4<-
read_csv("C:/Users/spiro/Desktop/archive(2)/accepted_2007_to_2018q4.csv/accepted
_2007_to_2018Q4.csv")
```

```
# Ορίζουμε τις μεταβλητές που θέλουμε να κρατήσουμε
vars_to_keep <- c("loan_status", "loan_amnt", "term", "int_rate", "installment",
                  "grade", "emp_length", "home_ownership", "annual_inc",
                  "verification_status", "purpose", "dti", "delinq_2yrs",
                  "fico_range_low", "fico_range_high", "open_acc", "pub_rec",
                  "revol_util", "total_acc", "addr_state")

# Δημιουργούμε το νέο, μικρότερο dataframe (df)
df <- accepted_2007_to_2018Q4[, vars_to_keep]
```

```
# Μια γρήγορη ματιά για να βεβαιωθούμε ότι όλα είναι εντάξει
dim(df)       # Θα πρέπει να σου δείξει τον αριθμό των γραμμών και 20 στήλες
summary(df)   # Στατιστική εικόνα των 20 μεταβλητών

# Φιλτράρισμα του loan_status (Κρατάμε μόνο τις ολοκληρωμένες περιπτώσεις)
df <- df %>% filter(loan_status %in% c("Fully Paid", "Charged Off"))

# Μετατροπή του term σε αριθμό (αφαιρούμε το " months")
df$term <- as.numeric(str_extract(df$term, "\\d+"))

# Καθαρισμός του emp_length (Μετατροπή σε αριθμητική κλίμακα 0-10)
df <- df %>% mutate(emp_length = case_when(
  emp_length == "< 1 year" ~ 0,
  emp_length == "1 year" ~ 1,
  emp_length == "2 years" ~ 2,
  emp_length == "3 years" ~ 3,
  emp_length == "4 years" ~ 4,
  emp_length == "5 years" ~ 5,
  emp_length == "6 years" ~ 6,
  emp_length == "7 years" ~ 7,
  emp_length == "8 years" ~ 8,
  emp_length == "9 years" ~ 9,
  emp_length == "10+ years" ~ 10,
  TRUE ~ NA_real_ # Για τις κενές τιμές
))

# Δημιουργία ενός μέσου όρου για το FICO Score
# (Αντί για low και high, κρατάμε έναν αριθμό)
df$fico_avg <- (df$fico_range_low + df$fico_range_high) / 2

# Δημιουργία της δυαδικής μεταβλητής (Target Variable)
# Ορίζουμε 1 όσους "φέσωσαν" (Charged Off) και 0 όσους πλήρωσαν
df <- df %>% mutate(default_ind = ifelse(loan_status == "Charged Off", 1, 0))
```

```
# Αφαίρεση των παλιών στηλών
# Πετάμε τα low/high FICO και το αρχικό κείμενο του loan_status
df <- df %>% select(-fico_range_low, -fico_range_high, -loan_status)

# Αφαίρεση των γραμμών που έχουν ακόμα κενά (NAs) για να έχουμε καθαρό δείγμα
df <- na.omit(df)


################ Τέλος επεξεργασίας δεδομένων################
```

```
#############Πίνανας για την παρουσίαση μεταβλητών ####################

# Δημιουργία των δεδομένων για όλες τις 20 μεταβλητές
full_vars_table <- data.frame(
  "Κατηγορία" = c(rep("Εξαρτημένη", 1), rep("Χαρακτηριστικά Δανείου", 6), rep("Δημογραφικά/Οικονομικά", 5), rep("Πιστωτικό Ιστορικό", 8)),
  "Μεταβλητή" = c(
    "default_ind",
    "loan_amnt", "term", "int_rate", "installment", "grade", "purpose",
    "emp_length", "home_ownership", "annual_inc", "verification_status", "addr_state",
    "fico_avg", "dti", "delinq_2yrs", "open_acc", "pub_rec", "revol_util", "total_acc", "log_annual_inc"
  ),
  "Περιγραφή" = c(
    "Κατάσταση δανείου (1 = Αθέτηση/Charged Off, 0 = Εξόφληση)",
    "Συνολικό ποσό δανείου που ζητήθηκε",
    "Διάρκεια δανείου (36 ή 60 μήνες)",
    "Επιτόκιο δανείου",
    "Μηνιαία δόση πληρωμής",
```

```
    "Κατηγορία κινδύνου (A-G) βάσει Lending Club",
    "Σκοπός δανείου (π.χ. χρέος, αγορά)",
    "Έτη εργασιακής εμπειρίας (0-10)",
    "Καθεστώς στέγασης (Rent, Mortgage, Own)",
    "Ετήσιο δηλωθέν εισόδημα",
    "Πιστοποίηση εισοδήματος από την πλατφόρμα",
    "Πολιτεία κατοικίας του δανειολήπτη",
    "Μέσο σκορ FICO (Πιστοληπτική ικανότητα)",
    "Λόγος χρέους προς εισόδημα (Debt-to-Income)",
    "Αριθμός καθυστερήσεων >30 ημερών τελευταία 2 έτη",
    "Αριθμός ανοιχτών πιστωτικών γραμμών",
    "Αριθμός υποτιμητικών δημόσιων εγγραφών",
    "Ποσοστό χρησιμοποίησης πιστωτικού ορίου",
    "Συνολικός αριθμός πιστωτικών γραμμών",
    "Λογάριθμος ετήσιου εισοδήματος (για κανονικοποίηση)"
  ),
  "Τύπος" = c(
    "Binary", "Numeric", "Integer", "Numeric", "Numeric", "Factor", "Factor",
    "Integer", "Factor", "Numeric", "Factor", "Factor",
    "Numeric", "Numeric", "Integer", "Integer", "Integer", "Numeric", "Integer",
"Numeric"
  )
)

# Δημιουργία του πίνακα με μορφοποίηση για την διπλωματική εργασία
full_vars_table %>%
  kbl(caption = "Ορισμός και Ταξινόμηση Μεταβλητών της Έρευνας") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), full_width =
F) %>%
  pack_rows("Εξαρτημένη Μεταβλητή", 1, 1, label_row_css = "background-color:
#f2f2f2; color: #d9534f;") %>%
  pack_rows("Χαρακτηριστικά Δανείου", 2, 7, label_row_css = "background-color:
#f2f2f2; color: #337ab7;") %>%
```

```
  pack_rows("Δημογραφικά & Οικονομικά Στοιχεία", 8, 12, label_row_css = "background-color: #f2f2f2; color: #337ab7;") %>%
  pack_rows("Πιστωτικό Ιστορικό", 13, 20, label_row_css = "background-color: #f2f2f2; color: #337ab7;") %>%
  column_spec(2, bold = TRUE)
```

```
############## Πίνανας για την παρουσίαση βιβλιοθήκης ####################

# Δημιουργία των δεδομένων για τις επιπλέον βιβλιοθήκες
tools_table_extended <- data.frame(
  "Βιβλιοθήκη (Library)" = c("readr", "dplyr", "stringr", "knitr", "kableExtra", "tidyr", "car", "ggplot2"),
  "Κύρια Λειτουργία" = c(
    "Εισαγωγή & Ανάγνωση δεδομένων",
    "Επεξεργασία & Μετασχηματισμός",
    "Διαχείριση κειμένου (Strings)",
    "Δυναμική δημιουργία αναφορών",
    "Προχωρημένη μορφοποίηση πινάκων",
    "Τακτοποίηση δεδομένων (Data Tidying)",
    "Στατιστικοί έλεγχοι παλινδρόμησης",
    "Οπτικοποίηση δεδομένων (Γραφήματα)"
  ),
  "Χρήση στην Έρευνα" = c(
    "Γρήγορη ανάγνωση αρχείων .csv, .tsv και .txt.",
    "Φιλτράρισμα, ομαδοποίηση (group_by) και σύνοψη (summarize).",
    "Καθαρισμός κειμένων, εύρεση προτύπων και αντικατάσταση χαρακτήρων.",
    "Μετατροπή του κώδικα σε επαγγελματικά έγγραφα (PDF/HTML).",
    "Προσθήκη στυλ, χρωμάτων και γραμμών σε πίνακες του RMarkdown.",
    "Μετατροπή πινάκων από 'wide' σε 'long' μορφή και αντίστροφα.",
    "Έλεγχος πολυσυγγραμμικότητας (VIF) και ANOVA (Type II/III).",
```

```
    "Δημιουργία σύνθετων γραφημάτων με το σύστημα layers."
  ),
  check.names = FALSE
)

# Δημιουργία πίνακα με μορφοποίηση για την διπλωματική εργασία
tools_table_extended %>%
  kbl(caption = "Συνοπτική Παρουσίαση Στατιστικών Εργαλείων και Βιβλιοθηκών της R", align = "l") %>%
  kable_styling(
    bootstrap_options = c("striped", "hover", "bordered"), # "bordered" για να έχεις γραμμές παντού
    full_width = F,
    position = "left"
  ) %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50") %>% # Σκούρο μπλε/γκρι επικεφαλίδα
  column_spec(1, bold = TRUE, width = "4cm") %>%
  column_spec(2, width = "7cm") %>%
  column_spec(3, width = "8cm")
```

```
################## Πίνανας για τα περιγραφικά δεδομένα ####################

# Επιλέγουμε τις αριθμητικές στήλες
df_num <- df %>% select(where(is.numeric))

# Υπολογίζουμε τα στατιστικά συμπεριλαμβάνοντας τον Διάμεσο
summary_numeric <- data.frame(
  Variable = names(df_num),
  Mean   = sapply(df_num, mean, na.rm = TRUE),
```

```
  Median = sapply(df_num, median, na.rm = TRUE), # Προσθήκη Διαμέσου
  SD     = sapply(df_num, sd, na.rm = TRUE),
  Min    = sapply(df_num, min, na.rm = TRUE),
  Max    = sapply(df_num, max, na.rm = TRUE)
)

# Αντιστοίχιση των ονομάτων στα Ελληνικά
labels_map <- c(
  "loan_amnt" = "Ποσό Δανείου",
  "int_rate" = "Επιτόκιο (%)",
  "installment" = "Μηνιαία Δόση",
  "annual_inc" = "Ετήσιο Εισόδημα",
  "dti" = "Δείκτης DTI",
  "fico_avg" = "Σκορ FICO (MO)",
  "open_acc" = "Ανοιχτοί Λογαριασμοί",
  "pub_rec" = "Δημόσιες Εγγραφές",
  "revol_util" = "Χρήση Πιστωτικού Ορίου (%)",
  "total_acc" = "Συνολικοί Λογαριασμοί",
  "default_ind" = "Δείκτης Αθέτησης (0/1)",
  "emp_length" = "Έτη Προϋπηρεσίας",
  "term" = "Διάρκεια (Μήνες)",
  "log_annual_inc" = "Λογάριθμος Εισοδήματος"
)

# Εφαρμογή των ονομάτων
summary_numeric$Variable <- ifelse(summary_numeric$Variable %in%
names(labels_map),
                                   labels_map[summary_numeric$Variable],
                                   summary_numeric$Variable)

# Δημιουργία του τελικού πίνακα με την επιπλέον στήλη
summary_numeric %>%
  kbl(caption = "Συγκεντρωτικός Πίνακας Περιγραφικής Στατιστικής",
```

```
    digits = 2, align = "lccccc", row.names = FALSE,
    col.names = c("Μεταβλητή", "Μέσος Όρος", "Διάμεσος", "Τυπ. Απόκλιση",
"Ελάχιστο", "Μέγιστο")) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "bordered"),
                full_width = F) %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
```

```
# Εξαγωγή των αποτελεσμάτων από το μοντέλο
coef_summary <- as.data.frame(summary(logit_model)$coefficients)
coef_summary$Variable <- rownames(coef_summary)
rownames(coef_summary) <- NULL

# Μορφοποίηση του p-value (η μαγεία γίνεται εδώ)
coef_summary <- coef_summary %>%
  mutate(`Pr(>|z|)` = ifelse(`Pr(>|z|)` < 0.001, "< 0,001", sprintf("%.4f", `Pr(>|z|)`)))

# Ταξινόμηση στηλών και Ελληνικά ονόματα
coef_summary <- coef_summary[, c("Variable", "Estimate", "Std. Error", "z value",
"Pr(>|z|)")]

labels_map <- c(
  "(Intercept)" = "Σταθερά (Intercept)",
  "fico_avg" = "Σκορ FICO (MO)",
  "int_rate" = "Επιτόκιο (%)",
  "loan_amnt" = "Ποσό Δανείου",
  "annual_inc" = "Ετήσιο Εισόδημα",
  "dti" = "Δείκτης DTI",
  "emp_length" = "Έτη Προϋπηρεσίας"
)

coef_summary$Variable <- ifelse(coef_summary$Variable %in% names(labels_map),
                                labels_map[coef_summary$Variable],
```

```
                    coef_summary$Variable)

# Δημιουργία του πίνακα
coef_summary %>%
  kbl(caption = "Πίνακας Συντελεστών Λογιστικής Παλινδρόμησης (Διορθωμένος)",
      digits = 4, align = "lcccc",
      col.names = c("Μεταβλητή", "Συντελεστής (Estimate)", "Τυπ. Σφάλμα", "z-value", "p-value")) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "bordered"),
                full_width = F) %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
```

```
# Εκτέλεση Λογιστικής Παλινδρόμησης (Logit Model)
# Χρησιμοποιούμε το default_ind ως εξαρτημένη μεταβλητή
logit_model <- glm(default_ind ~ fico_avg + int_rate + loan_amnt + annual_inc + dti + emp_length,
                   data = df,
                   family = "binomial")

# Εξαγωγή των αποτελεσμάτων σε καθαρό dataframe
coef_summary <- as.data.frame(summary(logit_model)$coefficients)

# Καθαρισμός ονομάτων γραμμών και στηλών
coef_summary$Variable <- rownames(coef_summary)
rownames(coef_summary) <- NULL

# Μεταφορά της στήλης Variable στην αρχή
coef_summary <- coef_summary[, c(5, 1, 2, 3, 4)]

# Αντιστοίχιση ονομάτων στα Ελληνικά
labels_map <- c(
  "(Intercept)" = "Σταθερά (Intercept)",
```

```
  "fico_avg" = "Σκορ FICO (MO)",
  "int_rate" = "Επιτόκιο (%)",
  "loan_amnt" = "Ποσό Δανείου",
  "annual_inc" = "Ετήσιο Εισόδημα",
  "dti" = "Δείκτης DTI",
  "emp_length" = "Έτη Προϋπηρεσίας"
)

coef_summary$Variable <- ifelse(coef_summary$Variable %in% names(labels_map),
                                labels_map[coef_summary$Variable],
                                coef_summary$Variable)

# Δημιουργία πίνακα με μορφοποίηση για την διπλωματική εργασία
coef_summary %>%
  kbl(caption = "Πίνακας Συντελεστών Λογιστικής Παλινδρόμησης",
      digits = 4, align = "lcccc",
      col.names = c("Μεταβλητή", "Συντελεστής (Estimate)", "Τυπ. Σφάλμα", "z-value", "p-value")) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "bordered"),
                full_width = F) %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50") %>%
  # Επισήμανση των στατιστικά σημαντικών (p < 0.05) με έντονα γράμματα
  column_spec(5, bold = ifelse(coef_summary[,5] < 0.05, TRUE, FALSE))
```

```
# Υπολογισμός Odds Ratios και Διαστημάτων Εμπιστοσύνης (95%)
model_odds <- exp(coef(logit_model))
model_ci <- exp(confint.default(logit_model))

# Προετοιμασία του Dataframe
or_summary <- data.frame(
  Variable = names(model_odds),
  Odds_Ratio = model_odds,
```

```
  Lower_CI = model_ci[,1],
  Upper_CI = model_ci[,2]
)

# Αφαίρεση της Σταθεράς (Intercept) αν δεν τη χρειάζεσαι για την ερμηνεία
or_summary <- or_summary %>% filter(Variable != "(Intercept)")

# Αντιστοίχιση ονομάτων στα Ελληνικά
labels_map <- c(
  "fico_avg" = "Σκορ FICO (MO)",
  "int_rate" = "Επιτόκιο (%)",
  "loan_amnt" = "Ποσό Δανείου",
  "annual_inc" = "Ετήσιο Εισόδημα",
  "dti" = "Δείκτης DTI",
  "emp_length" = "Έτη Προϋπηρεσίας"
)

or_summary$Variable <- ifelse(or_summary$Variable %in% names(labels_map),
                              labels_map[or_summary$Variable],
                              or_summary$Variable)

# Δημιουργία πίνακα με μορφοποίηση για την διπλωματική εργασία
or_summary %>%
  kbl(caption = "Πίνακας Odds Ratios και Διαστημάτων Εμπιστοσύνης",
      digits = 3, align = "lccc",
      col.names = c("Μεταβλητή", "Odds Ratio", "Κατώτερο Δ.Ε. (95%)", "Ανώτερο
Δ.Ε. (95%)")) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "bordered"),
                full_width = F) %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
```

```
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
```

```
  # Επαναδημιουργία Μοντέλου
  logit_model <- glm(default_ind ~ fico_avg + int_rate + loan_amnt + annual_inc + dti + emp_length,
                     data = df,
                     family = "binomial")

  # ΠΙΝΑΚΑΣ VIF (Πολυσυγγραμμικότητα)
  vif_values <- car::vif(logit_model)
  vif_df <- data.frame(
    Variable = names(vif_values),
    VIF = as.numeric(vif_values)
  )

  # Ελληνικά Ονόματα Μεταβλητών
  labels_map <- c(
    "fico_avg" = "Σκορ FICO (ΜΟ)", "int_rate" = "Επιτόκιο (%)",
    "loan_amnt" = "Ποσό Δανείου", "annual_inc" = "Ετήσιο Εισόδημα",
    "dti" = "Δείκτης DTI", "emp_length" = "Έτη Προϋπηρεσίας"
  )

  vif_df$Variable <- ifelse(vif_df$Variable %in% names(labels_map), labels_map[vif_df$Variable], vif_df$Variable)

  vif_table <- vif_df %>%
    kbl(caption = "Έλεγχος Πολυσυγγραμμικότητας (VIF Values)", digits = 3, align = "lc", col.names = c("Μεταβλητή", "Δείκτης VIF")) %>%
    kable_styling(bootstrap_options = c("striped", "hover", "bordered"), full_width = F) %>%
    row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
```

```
# Πίνακας Συσχετήσεων (Correlation Matrix)
cor_vars <- df %>% select(fico_avg, int_rate, loan_amnt, annual_inc, dti)
cor_matrix <- cor(cor_vars, use = "complete.obs")
cor_df <- as.data.frame(cor_matrix)

# Ονοματοδοσία γραμμών/στηλών στα Ελληνικά
colnames(cor_df) <- labels_map[colnames(cor_df)]
rownames(cor_df) <- labels_map[rownames(cor_df)]

cor_table <- cor_df %>%
  kbl(caption = "Πίνακας Συσχετίσεων Ποσοτικών Μεταβλητών", digits = 3, align = "c") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "bordered"), full_width = F) %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")

# Γράφημα συσχέτισης (Scatter Plot)
scatter_plot <- ggplot(df, aes(x = fico_avg, y = int_rate)) +
  geom_point(alpha = 0.05, color = "#34495e") +
  geom_smooth(method = "lm", color = "#e74c3c", se = TRUE) +
  labs(title = "Σχέση Σκορ FICO και Επιτοκίου Δανεισμού",
       x = "FICO Score (Μέσος Όρος)",
       y = "Επιτόκιο (%)") +
  theme_minimal() +
  theme(panel.border = element_rect(colour = "black", fill=NA, size=1))

# Εκτύπωση όλων
vif_table
cor_table
print(scatter_plot)
```

```
# Υπολογισμός σπουδαιότητας (βασισμένος στο z-statistic)
importance <- as.data.frame(abs(summary(logit_model)$coefficients[,3]))
colnames(importance) <- "Importance"
importance$Variable <- rownames(importance)
importance <- importance[importance$Variable != "(Intercept)", ]

# Οπτικοποίηση
ggplot(importance, aes(x = reorder(Variable, Importance), y = Importance)) +
  geom_bar(stat = "identity", fill = "#2c3e50") +
  coord_flip() +
  labs(title = "Σπουδαιότητα Μεταβλητών στο Μοντέλο",
       x = "Μεταβλητές", y = "Στατιστική Σπουδαιότητα (z-score)") +
  theme_minimal()
```

```
# Εμφάνιση των συντελεστών για την εξίσωση
coefficients <- coef(logit_model)
print(coefficients)
 # 1. Υπολογισμός του γραμμικού σκορ (Z)
df$score_z <- predict(logit_model, type = "link")

# Δημιουργία του γραφήματος με σωστά χρώματα και στυλ
library(ggplot2)

ggplot(df, aes(x = score_z, fill = as.factor(default_ind))) +
  # Χρησιμοποιούμε geom_density για ομαλές καμπύλες
  geom_density(alpha = 0.5, color = "white", size = 0.3) +

  # Χειροκίνητος ορισμός χρωμάτων (0 = Πράσινο, 1 = Κόκκινο)
  scale_fill_manual(values = c("0" = "#27ae60", "1" = "#e74c3c"),
```

```
                labels = c("Συνεπείς (Paid)", "Αθέτηση (Default)")) +

  # Προσθήκη τίτλων και ετικετών
  labs(title = "Κατανομή Σκορ Κινδύνου (Logit Scores)",
       subtitle = "Διαχωρισμός συνεπών και μη συνεπών δανειοληπτών",
       x = "Σκορ Z (Πιο δεξιά = Υψηλότερος Κίνδυνος)",
       y = "Πυκνότητα Πληθυσμού",
       fill = "Κατηγορία:") +

  # Καθαρό θέμα για διπλωματική
  theme_minimal() +
  theme(legend.position = "bottom",
        plot.title = element_text(face = "bold", size = 14),
        axis.title = element_text(size = 10),
        panel.grid.minor = element_blank())
```

```
df %>%
  group_by(default_ind) %>%
  summarise(Average_Score = mean(score_z)) %>%
  kbl(caption = "Μέσο Σκορ Κινδύνου ανά Ομάδα", digits = 3, col.names = c("Default (0=Όχι, 1=Ναι)", "Μέσο Σκορ Z")) %>%
  kable_styling(bootstrap_options = c("striped", "bordered"), full_width = F) %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50")
```

```
# Δημιουργία του data frame από το μοντέλο
z_score_df <- as.data.frame(summary(logit_model)$coefficients)
colnames(z_score_df) <- c("Estimate", "Std_Error", "z_value", "P_Value")

# Μορφοποίηση του P-Value για να λέει < 0.001
```

```
z_score_df$P_Value_Clean <- ifelse(z_score_df$P_Value < 0.001, "< 0.001",
                                   as.character(round(z_score_df$P_Value, 4)))

# Προσθήκη Σημαντικότητας
z_score_df$Significance <- ifelse(z_score_df$P_Value < 0.001, "***",
                                  ifelse(z_score_df$P_Value < 0.05, "**",
                                         ifelse(z_score_df$P_Value < 0.1, "*", " ")))

# Επιλογή και μετονομασία στηλών για τον τελικό πίνακα
final_table <- z_score_df[, c("Estimate", "Std_Error", "z_value", "P_Value_Clean",
"Significance")]
colnames(final_table) <- c("Συντελεστής (Beta)", "Τυπικό Σφάλμα", "z-value", "P-
Value", " ")

# Δημιουργία του κανονικού πίνακα
final_table %>%
  kbl(caption = "Αποτελέσματα Λογιστικής Παλινδρόμησης (Coefficients)",
      align = "lcccc") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "bordered"),
                full_width = F,
                position = "center") %>%
  row_spec(0, bold = TRUE, color = "white", background = "#2c3e50") %>%
  column_spec(1, bold = TRUE) %>%
  column_spec(5, color = "red", bold = TRUE) # Τα αστέρια με κόκκινο
```

```
# Υπολογισμός προβλέψεων (πιθανότητες)
probabilities <- predict(logit_model, type = "response")

# Μετατροπή σε 0 ή 1 με βάση το 0.5
predictions <- ifelse(probabilities > 0.5, 1, 0)
```

```
# Δημιουργία του πίνακα σύγχυσης
conf_data <- table(Actual = df$default_ind, Predicted = predictions)

# Μετατροπή σε data frame για το kable
conf_df <- as.data.frame.matrix(conf_data)
rownames(conf_df) <- c("Πραγματικό: Εξόφληση (0)", "Πραγματικό: Αθέτηση (1)")
colnames(conf_df) <- c("Πρόβλεψη: Εξόφληση (0)", "Πρόβλεψη: Αθέτηση (1)")

# Υπολογισμός Ακρίβειας (Accuracy) για τον τίτλο
accuracy <- sum(diag(conf_data)) / sum(conf_data) * 100

# Εμφάνιση πίνακα με μορφοποίηση για την διπλωματική εργασία
conf_df %>%
  kbl(caption = paste("Πίνακας Σύγχυσης (Confusion Matrix) - Συνολική Ακρίβεια:",
round(accuracy, 2), "%"),
      align = "c") %>%
  kable_styling(bootstrap_options = c("striped", "bordered", "hover"),
                full_width = F,
                position = "center") %>%
  add_header_above(c(" " = 1, "Προβλέψεις Μοντέλου" = 2)) %>%
  column_spec(1, bold = TRUE, background = "#f8f9fa")


  #Accuracy
    probabilities <- predict(logit_model, type = "response")
    predictions <- ifelse(probabilities > 0.5, 1, 0)
    conf_data <- table(Actual = df$default_ind, Predicted = predictions)
    accuracy <- sum(diag(conf_data)) / sum(conf_data)
```