# Προγραμματιστική Εργασία Πρόβλεψη κόστους ασφάλισης οχημάτων

Χαρά Τσίρκα, Πρόδρομος Αβραμίδης, and Γεώργιος Γεροντίδης

{ctsirka, pavramidis, ggerontidis}@e-ce.uth.gr
8 εξάμηνο


Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Πανεπιστήμιο Θεσσαλίας, Βόλος

**Εξόρυξη Δεδομένων 2023-24**
Διδάσκον: Μ.Βασιλακόπουλος

Μάιος 2024

# 1 Εισαγωγή

Η εργασία μας επικεντρώνεται στην πρόβλεψη του κόστους ασφάλισης μηχανοκίνητων οχημάτων. Η ανάλυση αυτή αποτελεί ένα κρίσιμο ζήτημα στον τομέα της ασφάλισης, καθώς επιτρέπει στους ασφαλιστές να προσδιορίζουν με μεγαλύτερη ακρίβεια τα ασφαλιστικά ασφάλιστρα, λαμβάνοντας υπόψη διάφορους παράγοντες που επηρεάζουν το κόστος. Η διαδικασία της εργασίας ξεκινά με την προ-επεξεργασία των δεδομένων, κατά την οποία πραγματοποιήθηκε εξερευνητική ανάλυση (exploratory analysis) για τον προσδιορισμό των κριτηρίων διαχωρισμού των δεδομένων. Κατά τη διάρκεια αυτής της ανάλυσης, μετρήθηκε ο βαθμός επίδρασης κάθε χαρακτηριστικού (feature) του συνόλου δεδομένων στα αποτελέσματα. Με τη βοήθεια διαγραμμάτων, καταφέραμε να επιλέξουμε τον κατάλληλο διαχωρισμό των δεδομένων για περαιτέρω ανάλυση. Στη συνέχεια, η εργασία θα προχωρήσει στη δημιουργία και αξιολόγηση των μοντέλων πρόβλεψης, λαμβάνοντας υπόψη την είσοδο των χρηστών, ενώ θα ακολουθήσει η οπτικοποίηση και η αξιολόγηση των αποτελεσμάτων.

# 2 Περιγραφή dataset

Το dataset το οποίο επιλέξαμε αποτελείται από 30 μεταβλητές (columns) και 105555 εγγραφές. Στους παρακάτω πίνακες δίνεται μία σύντομη περιγραφή της κάθε μεταβλητής:

| Μεταβλητή | Περιγραφή |
|---|---|
| **ID** | Εσωτερικός αριθμός αναγνώρισης που εκχωρείται σε κάθε ετήσια σύμβαση που επισημοποιείται από έναν ασφαλισμένο. Κάθε ασφαλισμένος μπορεί να έχει πολλές σειρές στο σύνολο δεδομένων, που αντιπροσωπεύουν διαφορετικές προσόδους του προϊόντος. |
| **Date_start_contract** | Ημερομηνία έναρξης του συμβολαίου (HH/MM/EEEE). |
| **Date_last_renewal** | Ημερομηνία τελευταίας ανανέωσης του συμβολαίου (HH/MM/EEEE). |
| **Date_next_renewal** | Ημερομηνία επόμενης ανανέωσης του συμβολαίου (HH/MM/EEEE). |
| **Distribution_channel** | Κανάλι μέσω του οποίου έγινε το ασφαλιστήριο, 0: για Πράκτορα, 1: για Ασφαλιστικοί μεσίτες. |
| **Date_birth** | Ημερομηνία γέννησης του ασφαλισμένου που δηλώνεται στο ασφαλιστήριο (HH/MM/EEEE). |
| **Date_driving_licence** | Ημερομηνία έκδοσης της άδειας οδήγησης του ασφαλισμένου (HH/MM/EEEE). |

| Μεταβλητή | Περιγραφή |
|---|---|
| **Seniority** | Συνολικός αριθμός ετών που ο ασφαλισμένος έχει συνδεθεί με την ασφαλιστική οντότητα, υποδεικνύοντας το επίπεδο αρχαιότητάς του. |
| **Policies_in_force** | Συνολικός αριθμός συμβολαίων που κατείχε ο ασφαλισμένος στην ασφαλιστική οντότητα κατά την περίοδο αναφοράς. |
| **Max_policies** | Μέγιστος αριθμός συμβολαίων που είχε ποτέ σε ισχύ ο ασφαλισμένος με τον ασφαλιστικό φορέα. |
| **Max_products** | Μέγιστος αριθμός προϊόντων που κατέχει ο ασφαλισμένος ταυτόχρονα σε οποιαδήποτε δεδομένη χρονική στιγμή. |
| **Lapse** | Αριθμός πολιτικών που ο πελάτης έχει ακυρώσει ή έχει ακυρωθεί λόγω μη πληρωμής κατά το τρέχον έτος λήξης, εξαιρουμένων αυτών που έχουν αντικατασταθεί από άλλο συμβόλαιο. |
| **Date_Lapse** | Ημερομηνία ακύρωσης της σύμβασης (HH/MM/EEEE). |
| **Payment** | Τελευταία μέθοδος πληρωμής της πολιτικής<br>1: εξαμηνιαία πληρωμή, 0: ετήσια πληρωμή |
| **Premium** | Καθαρό ποσό ασφαλίστρου που σχετίζεται με το ασφαλιστήριο συμβόλαιο κατά τη διάρκεια του τρέχοντος έτους. |
| **Cost_claims_year** | Συνολικό κόστος ζημιών που πραγματοποιήθηκαν για το ασφαλιστήριο συμβόλαιο κατά τη διάρκεια του τρέχοντος έτους. |
| **N_claims_year** | Συνολικός αριθμός ζημιών που πραγματοποιήθηκαν για το ασφαλιστήριο συμβόλαιο κατά τη διάρκεια του τρέχοντος έτους. |
| **N_claims_history** | Συνολικός αριθμός απαιτήσεων που υποβλήθηκαν καθ' όλη τη διάρκεια του ασφαλιστηρίου συμβολαίου. |
| **R_Claims_history** | Παρέχει μια ένδειξη του ιστορικού συχνότητας αξιώσεων του ασφαλιστηρίου. |
| **Type_risk** | Τύπος κινδύνου για κάθε όχημα<br>1: μοτοσικλέτα, 2:μικρά φορτηγά,<br>3: επιβατικά οχήματα, 4: αγροτικά οχήματα |
| **Area** | 0: αγροτική περιοχή, 1: αστική περιοχή (>30.000 κάτοικοι όσον αφορά τις κυκλοφοριακές συνθήκες) |
| **Second_driver** | 1: περισσότεροι από ένας δηλωμένοι οδηγοί<br>0: μόνο ένας δηλωμένος οδηγός |
| **Year_matriculation** | Έτος καταχώρησης οχήματος (EEEE) |
| **Power** | Ίπποι δύναμης οχήματος |
| **Cylinder_capacity** | Χωρητικότητα κυλίνδρων του οχήματος |
| **Value_vehicle** | Αξία αγοράς οχήματος στις 31/12/2019 |
| **N_doors** | Αριθμός θυρών οχήματος |
| **Type_fuel** | Τύπος καυσίμου, P: πετρέλαιο, D: ντίζελ |
| **Length** | Μήκος του οχήματος σε m |
| **Weight** | Βάρος του οχήματος σε kg |

# 3 Data preproccesing

Το πρώτο βήμα για την προεπεξεργασία των δεδομένων ήταν να κρατήσουμε μία γραμμή για κάθε 'ID'. Σε ένα 'ID' μπορεί να αντιστοιχούν περισσότερες από μία γραμμές που αντιπροσωπεύουν το ίδιο συμβόλαιο του ίδιου πελάτη για διαφορετική χρονική περίοδο. Έτσι, για κάθε 'ID' κρατάμε την τελευταία ανανέωση του συμβολαίου, δηλαδή την γραμμή με το μεγαλύτερο χρονολογικά 'last_renewal_date'. Έπειτα, στη θέση του 'premium' υπολογίζουμε και τοποθετούμε τον μέσο όρο των 'premium' όλων των γραμμών με κοινό 'ID'.

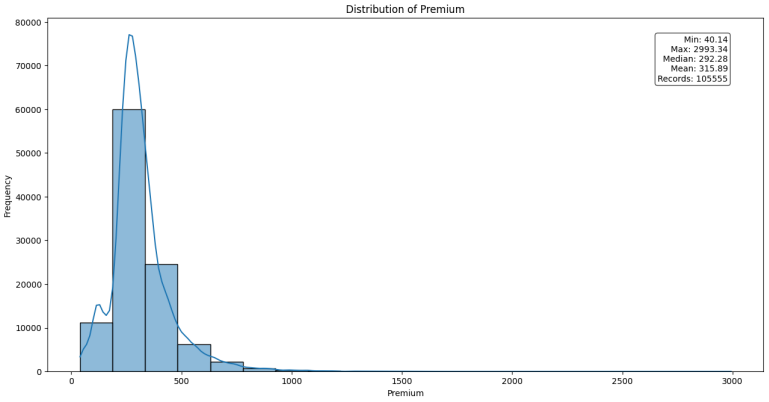
Το δεύτερο βήμα ήταν η επεξεργασία όλων των ημερομηνιών. Ειδικότερα, οι στήλες 'Date_birth', 'Date_driving_license', 'Date_start_contract', 'Date_last_renewal', 'Date_next_renewal', 'Date_lapse' δίνονται στην μορφή HH/MM/EEEE. Αρχικά, για κάθε μία από αυτές τις μεταβλητές κρατήσαμε το έτος (EEEE) και στην συνέχεια πραγματοποιώντας τις κατάλληλες αφαιρέσεις δημιουργήσαμε νέες στήλες στο dataset που πήραν την θέση αυτών που αναφέρθηκαν νωρίτερα. Έτσι, δημιουργήσαμε τις στήλες: 'Age' που προσδιορίζει την ηλικία του πελάτη, 'Years_driving' που προσδιορίζει πόσα χρόνια οδηγεί ο πελάτης, 'Year_on_road' που προσδιορίζει πόσα χρόνια κυκλοφορεί το κάθε όχημα υπό την κατοχή συγκεκριμένου πελάτη, 'Policy Duration' που υποδεικνύει την διάρκεια του εκάστοτε συμβολαίου σε χρόνια και 'Years_on_policy' που προσδιορίζει πόσα χρόνια ο πελάτης βρίσκεται στον ίδιο τύπο συμβολαίου. Πρέπει να σημειωθεί πως το dataset περιέχει δεδομένα μέχρι και το 2019. Για να έχουμε μια σωστή εικόνα των χρονολογιών σε όλες αυτές τις μεταβλητές που δημιουργήσαμε, χρησιμοποιήσαμε ως σημείο αναφοράς την χρονολογία τελευταίας ανανέωσης του συμβολαίου. Για παράδειγμα η μεταβλητή 'Age' προκύπτει από την αφαίρεση: 'Age' = 'Date_last_renewal' - 'Date_birth'.

Επιπλέον, δημιουργήσαμε μία ακόμη νέα στήλη με όνομα 'accidents' για να υπάρχει μια συσχέτιση μεταξύ των αριθμών των ατυχημάτων με τα χρόνια που ένας πελάτης είναι ασφαλισμένος στην εταιρεία.

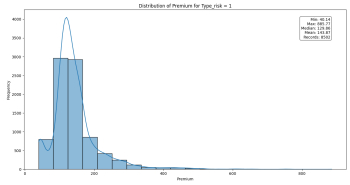
Διαχειριστήκαμε την απουσία τιμών με δύο τρόπους. Στην στήλη 'Length' αντικαταστήσαμε τα κενά πεδία με τον μέσο όρο των τιμών της στήλης. Στην στήλη 'Type_fuel' αντικαταστήσαμε τα κενά πεδία με την τιμή 'Unknown'.

Μετά από δοκιμές διαπιστώσαμε πως κάποιες μεταβλητές του dataset δεν συνεισέφεραν καθόλου στην βελτίωση της απόδοσης και παραλείφθηκαν. Οι στήλες που χρησιμοποιήθηκαν τελικά είναι οι: 'Seniority', 'Premium', 'Type_risk', 'Area', 'Power', 'Second_driver', 'Years_on_road', 'R_claims_history', 'Years_on_policy', 'accidents', 'Value_vehicle, 'Age', 'Years_driving', 'Distribution_channel', 'N_claims_history', 'Cylinder_capacity', 'Weight', 'Length', 'Type_fuel', 'Payment', 'Contract_year', 'Policies_in_force', 'Lapse'.

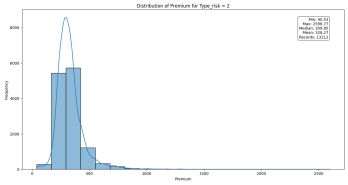
# 4 Διαγράμματα



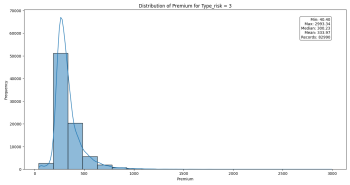

Τα διαγραμματα



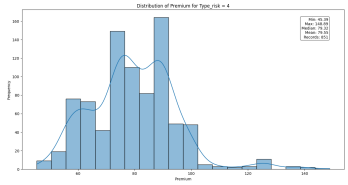
(α΄) Motorbikes



(β΄) Vans



(γ΄) Passenger cars



(δ΄) Agricultural Vehicles

Εικ. 1: Comparison of Different Vehicle Types

**Παρατήρηση 1** *Στο πρώτο διάγραμμα παρουσιάζεται η κατανομή των ασφαλίστρων για όλες τις καταχωρίσεις. Τα επόμενα διαγράμματα δείχνουν την κατανομή των ασφαλίστρων για κάθε κατηγορία οχημάτων ξεχωριστά.*

*Εύκολα διαπιστώνεται από το πρώτο ολικό διάγραμμα ότι οι ασφαλιστικές τιμές πάνω από 500 είναι ελάχιστες και δεν επηρεάζουν σημαντικά το τελικό αποτέλεσμα. Ωστόσο, όταν κατηγοριοποιήσαμε τα δεδομένα, παρατηρήσαμε ότι η διασπορά των τιμών στα αγροτικά οχήματα ήταν μεγαλύτερη, με αποτέλεσμα η διαφοροποίηση του μοντέλου ανά κατηγορία οχήματος να είναι απαραίτητη.*

# 5 User Interface

Το user interface της εφαρμογής αναπτύχθηκε με χρήση της Python και ειδικότερα του framework Kivy, καθώς και της συλλογής από γραφικά στοιχεία KivyMD