UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Churn Prediction Model and Segmentation in Insurance Industry

Author:
Viktor TYTARENKO

Supervisor:
Farnoush RESHADI

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Viktor TYTARENKO, declare that this thesis titled, "Churn Prediction Model and Segmentation in Insurance Industry" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Nothing is more powerful than an idea whose time has come."

Victor Hugo

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Churn Prediction Model and Segmentation in Insurance Industry

by Viktor TYTARENKO

Abstract

Increasing competition in US Insurance Industry pushes companies to raise marketing costs and decrease prices to get new customers. As a result, customer acquisition costs skyrocketed, and the only way for companies to be profitable is to keep stable retention and reduce churn. Without data-driven decisions, companies struggle to reduce churn and eventually lose their momentum in a hardly competitive industry. The main goal of this paper is to analyze customer data, describe churn behavior, and develop actionable recommendations to decrease customer churn. We developed a churn prediction model and segmented churned customers. Segmentation combined with model results were used to develop segment-specific recommendations for the business. The business implication of this research is a churn reduction strategy designed specifically for each customer segment.

Acknowledgements

I am very grateful to my supervisor, Farnoush Reshadi from Worcester Polytechnic Institute, for her help and guidance throughout the research, for listening to my ideas, and for improving them. Thanks to Yulia Kleban, Head of the IT and Business Analytics program, for her help and support for these four years. Also, I want to thank Ukrainian Catholic University, Lviv Business School, and the Faculty of Applied Sciences for always being helpful, giving honest feedback, and helping me to think big. Finally, I want to thank Ukrainian Armed Forces for making this diploma possible.

Contents

List of Figures

List of Tables

To my brother Vladyslav

Chapter 1

Introduction

1.1 US Insurance Market Overview

According to Insurance Information Institute, there are three main insurance sectors: property/casualty (P/C), mainly auto, home, and commercial insurance; life/annuity, primarily life insurance and annuity products; and private health insurance, written by insurers whose primary business is health insurance. In 2020, PC Insurance was estimated at $652.8 billion in the US, according to SP Global Market Intelligence. Homeownership in the US is estimated to be 66% of the population. Zillow indicates that $231700 is the median home value in the US. So, it's not surprising that with a relatively low interest rate of an average of 3.99% on a 30-year mortgage and increasing property prices, most property buyers prefer bank loans. According to the US Census, the total percentage of mortgage-free homeowners in the US is 35.2%. So, today about 141 million people in the US have home mortgages. Technically there is no law obligation to have home insurance to get a mortgage, but in reality, lenders require home insurance before agreeing to grant the loan. So, the home insurance industry is the backbone of the housing market.

1.2 Home Insurance Sales Process

The typical initial sales process in the insurance industry is manual. The initial purchase is the hardest in the industry because it requires the biggest man-time per customer. There are two types of professionals involved in initial sales: Insurance Agent and ISR (Inside Sales Representative). The difference between the two is that Agent is required by the law to get an appropriate certification, and only the Agent is authorized to sell an insurance policy. ISR is the person who does not have accreditation and cant complete the sale. Due to the lack of Insurance Agents, ISRs are often the first person with whom customer talks. With ISR, customers review insurance propositions, discuss prices, get policy quotes, and clarify critical aspects. Then after ISR, the customer is transferred to the Agent to complete the sale. Also, Agent can perform the role of ISR and complete the deal by itself. The average industry conversion rate from all customers who want to buy a policy to actual policyholders is about 20%, with the average time to complete a purchase around 45 minutes. To get one policyholder, ISRs and Agents need to talk to five people and spend 3 hours 45 minutes per successful sale. With such high customer acquisition costs and effort required, it's no doubt that insurance company needs to have high retention. The second touchpoint between policyholder and company is during renewal, the process of renewing the policy. It typically occurs around the time of policy expiration; most policies have 1-year tenure. As in the initial sale, the customer talks with ISR

and then signs a deal with the Agent. During this phase, customers get new quotes and may have changes in the policy premium, both bigger and lower.

1.3 Problem

Except for the costly sales process, the insurance industry is very competitive. Due to the nature of the industry, insurance demand is limited to 141 million customers. Therefore, every year the cost of getting customer attention is rising. Also, during the initial purchase phase, customer acquisition costs can be five times higher than during policy renewal. With low conversion at initial purchase [10-20%] and much higher at renewal [60-70%], keeping stable retention and reducing churn is crucial. In this industry, it's hard to be engaged with a customer outside initial sale and policy renewal, so price becomes the main competition feature. So, with the expensive sales process and intense price competition, companies put the main effort into retaining existing customers. So, it's important to predict potential churn and act to prevent it. This research aims to predict and segment customer churn in the insurance industry.

Chapter 2

Literature review

Churn Rate is one of the most important metrics across multiple industries that rely on continuous payments. From Netflix subscriptions and TV/Internet/Mobile providers to Banking and Insurance. As customer acquisition costs increase, more resources are invested in keeping existing customers. Multiple papers and researchers across these industries were covered. In this chapter, the main reserach approaches are covered.

2.1 Churn Prediction

Churn is divided into two sections: involuntary churn and voluntary churn. When a customer churns for an unavoidable reason, such as transaction failure, or code bugs, it's considered involuntary churn. On the other hand, if the customer intends to churn, it's his decision, then it's a voluntary churn. The focus of this paper is on voluntary churn. In [1], Burez and Van den Poel (2007) indicate two types of targeted approaches to managing customer churn: reactive and proactive. When a company adopts a reactive approach, it waits until customers ask the company to cancel its service relationship. In this situation, the company will offer the customer an incentive to stay. On the other hand, when a company adopts a proactive approach, it tries to identify customers who are likely to churn before they do so. The company then provides special programs or incentives for these customers to keep them from churning. Targeted proactive programs have the potential advantages of having lower incentive costs. It is crucial to building a precise customer churn model because if this model is inaccurate, companies will waste incentive money on customers who will not churn.

In [6] authors compared the performance of Random Forest, Naive Bayes, Decision Tree, SVM Radial, Logistic Regression, and SVM Linear for Customer Churn Prediction for a Motor Insurance Company. The random forest algorithm turns out to be a very effective model for forecasting customer churn, reaching an accuracy rate of 91.18%. On the other hand, survival analysis was used to model time until churn. It was concluded that approximately 90% of the policyholders survived for the first five years, while most of the policyholders survived until the end of the policy period.

2.2 Customer Segmentation

Customer segmentation is a categorization of customers into groups based on shared qualities so that businesses may market to each group effectively. It usually includes demographic characteristics such as location, marital status, age, or gender. It allows to allocate better resources such as marketing budgets, RD investments into

the product, etc. Jandaghi, G., Moradpour, Z. (2015) [3] developed segmentation of life insurance customers based on their profile using fuzzy clustering. The authors used fuzzy clustering on 1071 life insurance customers from March to October 2014. FCM is one of the most popular clustering techniques. FCM clustering involves two main steps: 1. Compute cluster centers and allocate points to the center using the Euclidean distance. The process is repeated continuously to stabilize the cluster centers. The algorithm has a membership value to items for clusters in the range of 0 to 1 and a fuzzy parameter in the range [1, n] which determines the degree of fuzziness of the cluster. Results show that the optimal number of clusters was 2, named "investment" and "life safety." Some suggestions are presented to improve the performance of the insurance company.

2.3 Integrated Churn Prediction and Segmentation

Few studies have looked at churn prediction as well as client segmentation. In [2], exploratory data analysis was used to investigate which variables in the dataset have led to customer churn. To create prediction models, two strategies were used. The first author looks at the churn rate across multiple loyalty segments. Researchers used the K-means algorithm to segment customers into five segments, and then a decision tree model was built in each cluster to predict customer churn. The other method used the complete dataset and Back Propagation Neural Network (BPN) followed by a Decision Tree to forecast customer churn. In this paper, Hung et al developed only a churn prediction model without a meaningful explanation for root causes.

In [7], the author developed a methodology to predict customer churn and conduct customer profiling by combining churn prediction and client segmentation. They used Random Forest to estimate client churn in the beginning. Then, to better understand the significant factors that lead to customer churn, Attribute Selected Classifier was used to conduct factor identification. They then pulled all churn data that Random Forest correctly predicted and used customer profiling to evaluate how similar these churn customers were. Finally, specific retention techniques and recommendations were offered based on the results of customer profiling.

Chapter 3

Data Review

In this section, there is a brief overview of data transformation, sources, and feature engineering.

3.1 Dataset description

The dataset used in this work originated from "Université de Technologie de Troyes (UTT), France." it's a real US insurance company dataset. This Insurance Company dataset includes policies between 2007 and 2012. Each policy consists of some significant policy characteristics, the building's features, the zone, the privileges, the faults, some risk indicators, etc., and contains approximately 250000 home insurance policies. The dataset size is (256136, 66); among 66 columns, 42 have categorical type and 24 - float, respectively.

3.2 Data Preparation

There are four steps in the preparation phase: cleaning, removing missing values, correlation check, encoding, and splitting. Before proceeding to the first step, it is important to mention that the target variable for churn prediction is POL_STATUS - policy status, which has the following values in this dataset: Live - 132160, Lapsed - 52534, Cancelled - 4311, Unknown - 16, NULL - 67115. From this Churned feature was constructed with: 0 - (pol_status=Live), and 1 - (pol_status=Lapsed or Canceled). In this dataset, 26% of records did not contain sufficient data in the target variable, policy status, so they were removed, leaving a dataset with 189,005 rows and 66 columns. Determining whether or not a policy is churned is a classification problem. The class labels for each policy are Churned or Not Churned. In this dataset, 132160 records are labeled "Not Churned," and the remaining 56861 records are labeled "Churned."

After the correlation check, four pairs of features were found with a high correlation (>0.6) in Figure 1. Correlations were checked only between numerical variables. The following variables have a high (0.6>) correlation: RISK_RATED_AREA_B and RISK_RATED_AREA_C, SUM_INSURED_BUILDINGS and NCD_GRANTED_YEARS B, SUM_INSURED_CONTENTS and NCD_GRANTED_YEARS_C, SPEC_SUM INSURED and SPEC_ITEM_PREM.

The next step is to clean the dataset from missing variables [Figure 3.2], remove unnecessary features, create new features and transform variables when necessary. During the feature engineering process, we developed a new age feature, which is customer age; it's computed as the difference in years between a customer's date of birth and current date. A similar computation was applied to property build date and insurance cover start date to create property_age and cover_length features.

The next step was to remove correlated features [Figure 3.1] because they violated assumptions described in the methodology sections. Finally, features that have more than 60% of missing values [Figure 3.2] were removed. The rest of the missing features were imputed with mean values of that feature or, in some cases, with 0.

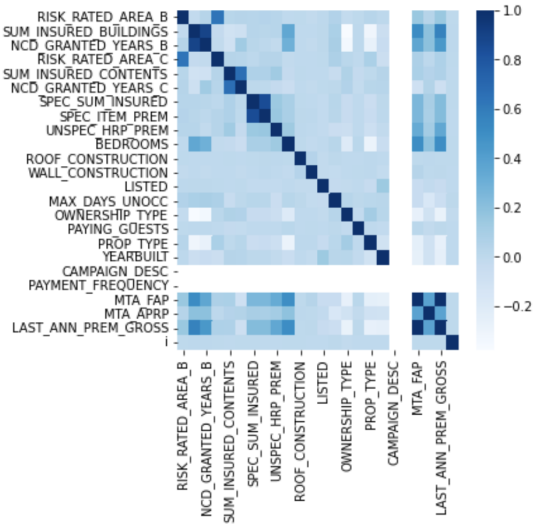


FIGURE 3.1: Feature Correlation

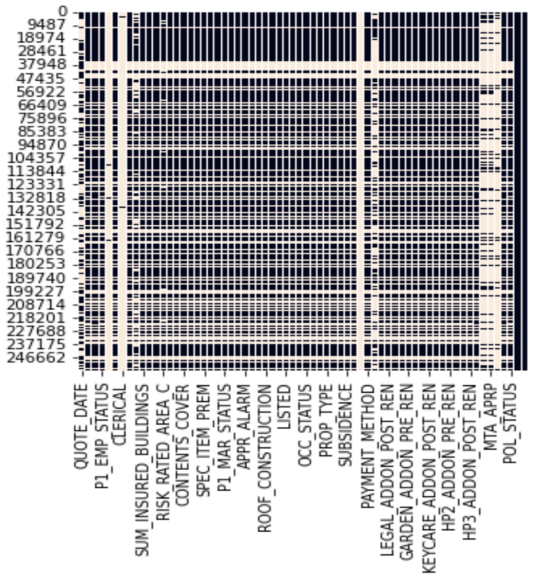


FIGURE 3.2: Missing values in dataset

3.3 Complete List of Features

Table 1 shows the complete list of numerical features in the data set (8 features), and Table 2 shows the complete list of features used in the model (40 features). We eliminated 26 features because of a large number of missing values or the removal of highly correlated features.

Chapter 4

Methodology

In this paper, we will run churn prediction with the variables shown in table X and the predictor x_is_churned, where 1 - churned user, and 0 - not churned user. Predictor x_is_churned is calculated based on policy status; if the status is "lapsed," the user is considered churned. We will segment churned users in the second part to understand the churn segments. Finally, we will combine the model and segmentation results and estimate model performance on different churn segments. In this section, we will discuss methods used to build the model and segment churned users.

4.1 Classification Algorithm Selection

Determining customer churn in any domain is generally a classification problem with Churned and Not Churned classes. There is no specific recommended model in available literature for the insurance industry. Also, "No Free Lunch Theorems for Optimization" (1997) implies that there is not a single machine learning model that performs the best for predictive modeling problems. It states that when optimization algorithms' performance is averaged across all possible problems, they perform equally well. So, we used several methods to select the best one. The classification methods used were Gaussian Naive Bayes; Catboost, Random Forest, Adaboost, Gradient Boosting Classifier, XGBoost, Lightgbm, and Voting Classifier. We took preprocessed data for this experiment, split it into train and test samples in a 70:30 ratio, and applied the abovementioned methods on the same train and test set. Then we collected performance data for each model and compared model performance. But, before going to the final results below, we will briefly discuss the methods used.

4.1.1 Gaussian Naive Bayes

The Bayes Theorem is used to create a statistical categorization approach known as Naive Bayes. Gaussian is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. The Naive Bayes classifier assumes that a feature's effect on a class is independent of other features. This assumption (class conditional independence) is deemed naive since it simplifies computation.

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \tag{4.1}$$

where P(h | D) - posterior probability, P(D | h) - likelihood, P(h) - class prior probability, P(D) - predictor prior probability. The probability of an event is calculated in the following steps: calculate the prior probability for given class labels, and find the likelihood probability with each attribute for each class. Then use these values

in Bayes Formula and calculate posterior probability. Compare which class has a higher likelihood, given the input belongs to the higher probability class.

4.1.2 Random Forest

Random Forest consists of many individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction, and the class with the most votes becomes the model's prediction. Random forests create decision trees on randomly selected data samples, get a prediction from each tree, and select the best solution utilizing voting.

4.1.3 Adaboost

Yoav Freund and Robert Schapire proposed Adaboost, or Adaptive Boosting, in 1996. It combines many classifiers to improve classifier accuracy using an iterative ensemble approach. Adaboost's core principle is to select the weights of classifiers and train the data sample in each iteration so that reliable predictions of uncommon observations may be made. The AdaBoost classifier creates a robust classifier by combining several low-performing classifiers, resulting in the high accuracy of a robust classifier. The process works like this: Adaboost picks a training subset at random and then trains the model iteratively by selecting the training set based on the last training's accurate prediction. The classifier assigns weight based on its accuracy; moreover, the more precise it is, the higher the weight. Expect weighting classifiers, it also weights observations, but vice versa, incorrectly categorized observations have a higher weight. The process is finished when training data fits perfectly or until the maximum number of estimators is reached.

4.1.4 Gradient Boosting

Like AdaBoost, Gradient Boosting trains many models in a gradual, additive, and sequential manner. While the AdaBoost model uses high weight data points to identify shortcomings, gradient boosting uses gradients in the loss function (y=ax+b+e, where e is the error term). The loss function is a metric that indicates how well the coefficients of a model fit the underlying data. In classification tasks, the loss function is a measure of how good the predictive model is at classifying the target class label. One of the most significant benefits of using Gradient Boosting is that it allows for optimization of a user-specified cost function rather than a loss function, which typically provides less control and does not correspond to real-world applications.

4.1.5 XGBoost

XGBoost is an optimized version of the Gradient Boosting Machine. The main difference in the algorithm is the parallelization and speed of this method because of its implementation in C++. Prediction is based on trees based on a series of binary questions, and the final decision output is on the leaf. These trees are combined into the ensemble; trees are built iteratively until a stopping criterion is met. XGBoost uses CART(Classification and Regression Trees) Decision trees. CART is the trees that contain a real-valued score in each leaf, regardless of whether they are used for classification or regression. Real-valued scores can then be converted to categories for classification, if necessary. XGBoost delivers high performance as compared to Gradient Boosting. Its training is high-speed and can be parallelized across clusters.

4.1.6 LightGBM

LightGBM [Light Gradient Boosting Machine] is a distributed gradient boosting system for machine learning created by Microsoft. It is used for ranking, classification, and other machine learning applications and is based on decision tree algorithms. If we compare this method XGBoost, we'll notice their similarity. They are both tree-based, but LightGBM has another tree construction technique. Instead of constructing it row by row, it uses leaf-wise tree growing. The other significant difference in the learning algorithm, LightGBM, uses an optimized histogram-based decision tree learning algorithm instead of the traditional sorted-based approach. It provides powerful performance and memory savings.

4.1.7 Catboost

Catboost is an open-source gradient boosting on decision trees library with categorical features support out of the box. Catboost's ability to integrate a wide range of data types is one of its key features. However, unlike the bulk of other machine learning algorithms, Catboost has a unique approach to dealing with categorical data, requiring only a tiny amount of categorical feature transformation. The transition from a non-numeric state to numeric values can be time-consuming and difficult in feature engineering. However, Catboost eliminates this step. In particular, in [5], Prokhorenkova used a new way of dealing with high cardinality categorical variables. For low cardinality categorical variables, Catboost uses one-hot encoding, the same as in other boosting methods. Catboost is based on decision tree and gradient boosting theory. The primary principle behind boosting is to successively integrate multiple weak models (models that perform marginally better than chance) to generate a robust competitive prediction model using greedy search. Because gradient boosting sequentially fits decision trees, the fitted trees will learn from previous trees' failures and reduce errors. Adding additional functions to old ones is repeated until the loss function chosen cannot be minimized. Catboost does not use similar gradient boosting models in the decision tree growth process. Instead, Catboost creates oblivious trees, which have high CPU efficiency.

4.1.8 Voting Classifier

The Voting Classifier is a method that trains on an ensemble of many models and predicts an output based on the highest probability of the outcome being the chosen class. The result of each model is aggregated, and then these results are passed into Voting Classifier. The output is predicted based on the majority of voting models. The idea is to create separate dedicated models and find the accuracy for each of them. We created a unified model which trains by these models. Its prediction is based on their combined majority of voting for each output class. In this paper, we built the Voting Classifier based on the following models: Catboost, Random Forest, Adaboost, Gradient Boosting, Lightgbm, and XGBoost.

4.1.9 Evaluation Metrics

The following evaluation metrics were used to compare the different classification models: Accuracy, Precision, Recall, F1 score, true positive, true negative, false positive, and false negative are the measurements TP, TN, FP, and FN, respectively. Precision, recall, and specificity measurements provide greater insight into a classifier's

real performance inside classes. These metrics will be used to compare model performances.

- TP - [true positive] - outcome where the model correctly predicts the positive class.
- TN - [true negative] - outcome where the model correctly predicts the negative class.
- FP - [false positive] - outcome where the model incorrectly predicts the positive class.
- FN - [false negative] - outcome where the model incorrectly predicts the negative class.
- Accuracy - a fraction of predictions the model got right,

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

- Precision - the proportion of positive labels was correct,

$$precision = \frac{TP}{TP + FP} \tag{4.3}$$

- Recall - the proportion of actual positives was identified correctly,

$$recall = \frac{TP}{TP + FN} \tag{4.4}$$

- F1-score - is a measure of model accuracy, defined as the harmonic mean of the precision and recall,

$$f1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4.5}$$

- classification threshold (also called the decision threshold) - in binary classification, a value above that threshold indicates predicted class 1; a value below indicates class 0.
- ROC curve - [receiver operating characteristic curve] - a graph that shows the performance of a classification model at different classification thresholds. This plot has two axes: True Positive Rate and False Positive Rate. True Positive Rate (TPR) is the synonym for recall, and False Positive Rate (FPR) is defined as

$$FPR = \frac{FP}{FP + TN} \tag{4.6}$$

- TPR vs. FPR at various categorization criteria is plotted on a ROC curve. As the classification threshold is lowered, more items are classified as positive, increasing both False Positives and True Positives.
- AUC - [Area under the ROC Curve] - AUC measures the area underneath the entire ROC curve. AUC shows the aggregate measure of performance of a classification model at different classification thresholds. AUC can be interpreted

as the probability that the model ranks a random positive example more highly than a random negative example.

4.1.10 Feature Importance - SHAP Values

The SHAP (SHapley Additive exPlanations) is a method of model interpretability developed by Lundberg and Lee (2016). The goal is to explain the output of any machine learning model. Interpretability is not equal to causality. Rather, it helps to uncover how the model works and allows businesses to adopt the model and bring it "out of the black box." In general, SHAP values can show how much each predictor contributes, either positively or negatively, to the target variable. It's also beneficial that SHAP values show explain each individual case, not like traditional importance methods that explain the entire sample. Another important consideration is that Catboost is a tree-based algorithm, and SHAP values support tree-based models.

4.1.11 Result Discussion

After training classifiers based on the methods described above, in table 4, the results of the test sample were obtained.

Model Name	Accuracy	Precision	Recall	F1-score
Catboost	0.75	0.73	0.75	0.72
Random Forest	0.73	0.71	0.73	0.69
Gaussian NB	0.32	0.69	0.32	0.18
Adaboost	0.72	0.70	0.72	0.67
GradientBoosting	0.73	0.71	0.73	0.68
XGBoost.	0.73	0.71	0.73	0.68
Lightgbm.	0.74	0.73	0.74	0.71
Voting Classifier	0.73	0.72	0.73	0.68

TABLE 4.1: Performance comparison on test sample

There is almost no significant difference between methods used, except Gaussian NB, which has the lowest performance. The reason for low performance is many categorical variables in the dataset. To fit the model, categorical variables were converted to dummy variables. Since Gaussian NB has the assumption that all variables are independent, dummy variables violate it, so the performance is poor. On the other hand, all tree-based models give approximately the same results. It's expected that Catboost will have the best results since, among all methods, it has the most sophisticated processing of categorical variables. Surprisingly, Voting Classifier does not improve results. An ensemble of models excludes Gaussian NB, so the impact of this method is irrelevant. In general, Voting Classifier gives average results among all models. Catboost was selected to build the final churn prediction model.

4.2 Churn Segmentation

The dataset used for churn segmentation contains mixed data, both numerical and categorical, so traditional K-Means cannot be used. This paper uses another method - a combination of Factor Analysis of Mixed Data and K-means clustering to segment churned users.

4.2.1 Factor Analysis of Mixed Data (FAMD)

According to [4], Factor analysis of mixed data (FAMD) is a principal component method dedicated to analyzing a dataset containing both quantitative and qualitative variables. It makes it possible to analyze the similarity between individuals by considering mixed types of variables. FAMD algorithm is a mix between principal component analysis (PCA) and multiple correspondence analysis (MCA). In other words, it works as PCA with quantitative variables and as MCA for qualitative variables. Quantitative and qualitative variables are normalized during the analysis to balance the influence of each set of variables.

4.2.2 K-Means Clustering

The most often used unsupervised machine learning approach for clustering a dataset is K-means clustering, where k stands for the number of clusters. The method allows to work with unlabeled datasets and make your inferences from them. K-Means is a type of centroid-based clustering approach, where clusters are represented by a center vector or a centroid in centroid-based clustering. Centroid means data point at cluster center and may not be part of the dataset. Centroid-based clustering is an iterative technique that determines similarity by how close a data point is to the cluster's centroid.

4.2.3 Elbow method

The Elbow method is used in modification to use the within-cluster difference to determine the optimal number of clusters. From the results of plotting within-cluster differences for various values, the principle of the Elbow method takes the value of k at the point when the value does not decrease significantly with the addition of the value of k.

$$WCD = \sum_{j=1}^{k} \sum_{i=1}^{m} d_1(x_i, y_c) \tag{4.7}$$

where WCD is the within-cluster difference, k is the number of clusters, m is the number of observations in each cluster, c is the centroid, and d1 is the simple dissimilarity measure.

Chapter 5

Churn Prediction and Segmentation Results

5.1 Churn Prediction Model

From the previous Chapter, "Classification algorithm selection," we decided to use Catboost for churn prediction. Since features constructed are numerical and categorical, Catboost is the best boosting method because it supports categorical feature handling out of the box. Another important consideration is excellent Catboost performance with default hyperparameters.
Target class X_IS_CHURNED:

- 1: churned user
- 0: not churned user

Preprocessed dataset was split into train and test in ratio 7 : 3. Then Catboost Classification Model was built on train data. In Figure 5.1, we can see the results of the model after cross-validation in this classification report:

```
catboost: Recall w/all features on test data 0.3315:
              precision    recall  f1-score   support

           0       0.76      0.93      0.84     39698
           1       0.67      0.33      0.44     17004

    accuracy                           0.75     56702
   macro avg       0.72      0.63      0.64     56702
weighted avg       0.74      0.75      0.72     56702

[[36971  2727]
 [11368  5636]]
------------------------------------------------------
```

FIGURE 5.1: Classification Report after Cross-Validation

The classification accuracy of the model is 75%. However, it may be due to predicting the most frequent class label. Since classes are unbalanced, "not churned" - 39698, "churned" - 17004, let's focus on the confusion matrix, which breaks down the classification results, showing a summary of the performance of the predictions against the actuals and metrics breakdown by class. Class "churned" has 33% recall and 67% precision, meaning that the model can correctly classify 33% of all churned users with 67% accuracy. In our experiment model correctly classified 5636 churned

users (out of 17004), but the cost was 2727 "not churned" user that was classified as 'churned).

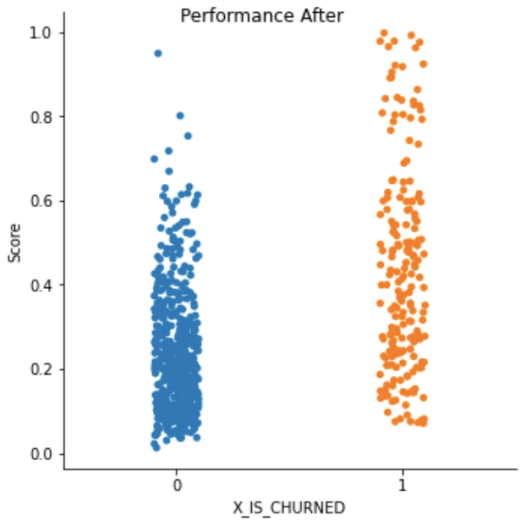


FIGURE 5.2: Sample of model scores per class

In Figure 5.2, there is a sample of the test dataset. This figure shows how the model predicted the score per class label.

Classification report provided with default threshold of 0.5. Threshold impact precision and recall of our model. With threshold, we can control the precision-recall ratio. It allows us to lower precision and increase recall or vice versa, depending on business problem requirements. In Figure 5.3, we built the ROC-AUC curve to examine model TPR and FPR combinations possible. With this threshold, we can control the recall and precision of the model. Depending on the business requirements model can be adjusted. If the business cost of outcome where the model incorrectly predicts the positive class is lower than True Positive, we can increase the threshold and vice versa.

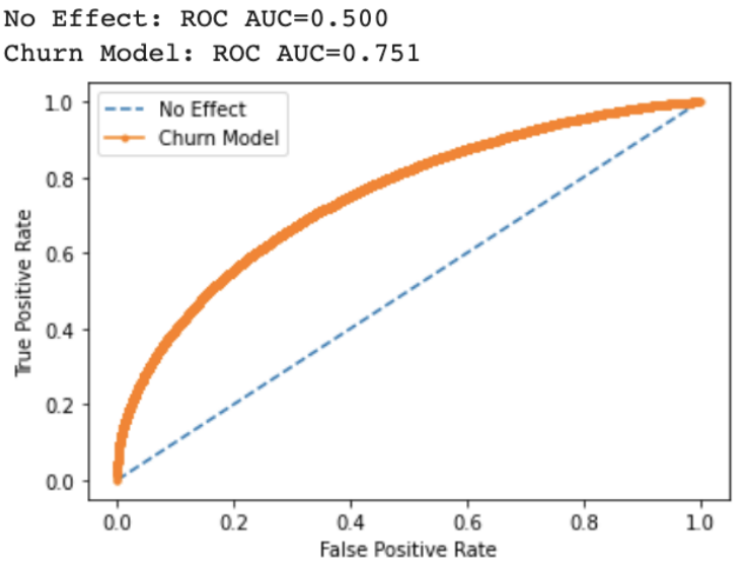


FIGURE 5.3: ROC-AUC Curve

5.2 Feature Importance

Figure 5.4 shows the feature importance - approach that assigns a score to each feature. A higher score indicates that the feature will significantly impact the model.

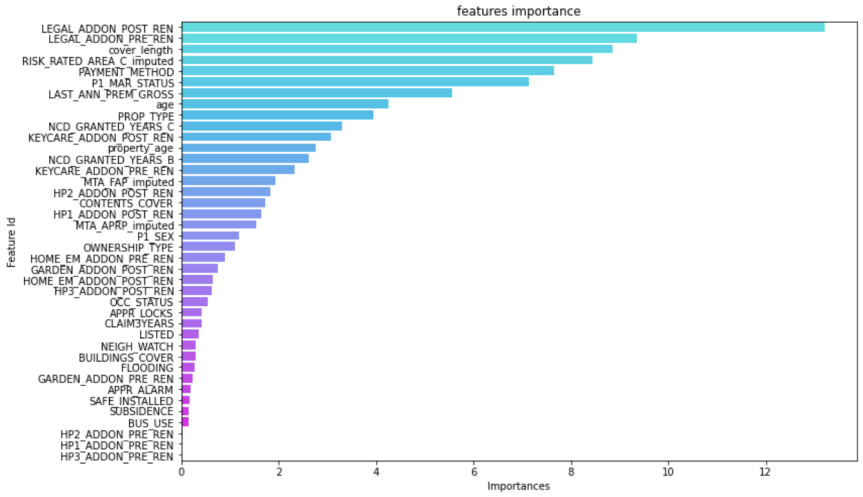


FIGURE 5.4: Feature Importance

5.2.1 SHAP feature importance

To further investigate important features, SHAP values were used. Figure 5.5 illustrates the SHAP feature importance plot that shows how a feature and its value impact model output. For example, feature='property_age', from the plot, we can see that the higher the property age, the more probable this policy will be churned. Please note that this plot is suitable only for numerical features (colored on the plot); categorical cannot be interpreted (colored with grey).

Figure 5.5 plot interpretation:

- Original value: for that observation, the color indicates whether the variable is high (in red) or low (in blue).
- Feature importance: Variables are listed in descending order.
- Impact: horizontal placement indicates whether that value's effect is linked to a greater or lower prediction.
- Correlation: A high property_age has a high and positive impact on the churn. The red color represents "high," while the X-axis represents "positive" influence. Similarly, age is negatively correlated with the target variable.

To check the overall feature impact SHAP, there is a waterfall plot 3 in appendix.

The partial dependence plot was also used to demonstrate the marginal impact of one or two features on a machine learning model's predicted outcome. It indicates if the target-feature connection is linear, monotonic, or more complex.

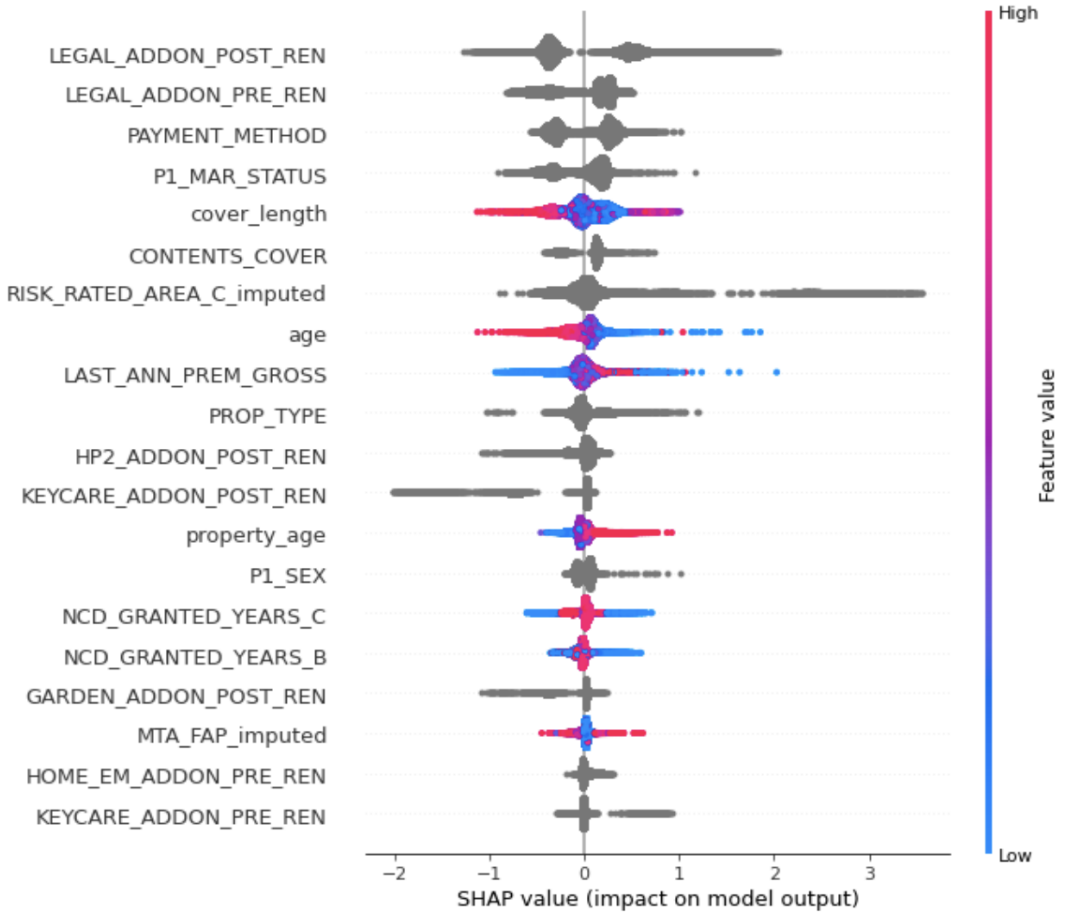


FIGURE 5.5: SHAP Feature Importance

Figure 5.6 shows a negative trend between "NCD_GEANTED_YEARS_B" and the target variable. "NCD_GEANTED_YEARS_B" is the duration of the premium bonus for building cover in years. The more extended bonus, the less likely user will be classified as "churned." Combination with "NCD_GEANTED_YEARS_C," duration of premium bonus for building content cover in years, at the value -1.5 shows that clients with no building and content bonuses most likely will be "churned."

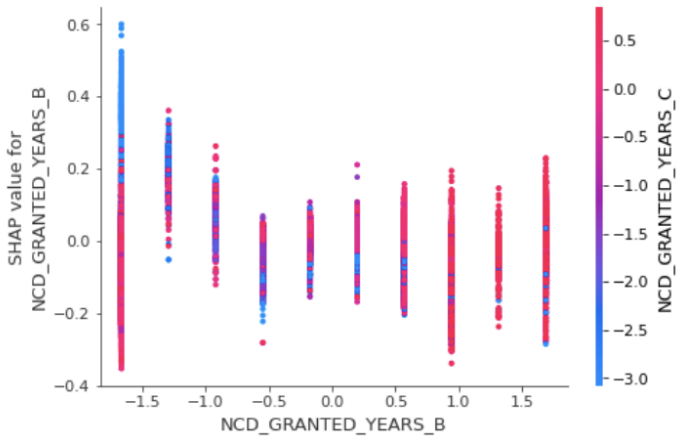


FIGURE 5.6: Partial dependence plot 1

Figure 5.7 shows a positive trend between "LAST_ANN_PREM_GROSS" and the target variable. The higher the last annual premium, the higher the probability client will be classified as churn. Variable "LEGAL_ADDON_POST_REN" means the legal coverage option was included after 1st renewal. There is a relationship that clients with this option (red) are less likely to be classified as churn than those who don't include it. Price sensitivity can explain this; users who can purchase additional optional legal coverage are perhaps less price sensitive since they can spend more.

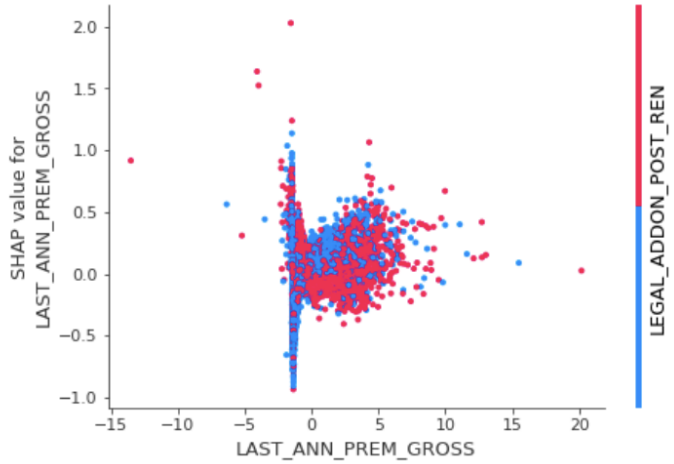


FIGURE 5.7: Partial dependence plot 2

Figure 5.8 shows a strong negative trend between "age" and the target variable. Younger policyholders are more likely to be classified as churn. Younger clients with the NonDD payment method are more likely to be churned than those with other methods, and vice versa, older clients with PureDD, are more likely to churn than older with NonDD or DD-Other.

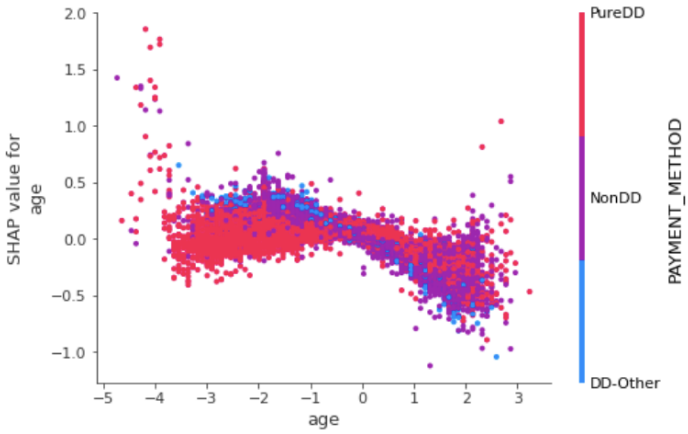


FIGURE 5.8: Partial dependence plot 3

Figure 5.9 shows a trend between "cover_length" and the target variable. The distribution is bell-shaped, meaning that policyholders at the beginning of their lifecycle are less likely to churn than those in the middle. After reaching a peak, there is a positive, strong negative trend: the longer client has with the company, the less likely he'll be classified as churn. Variable "LEGAL_ADDON_POST_REN" means the legal coverage option was included after 1st renewal. If true, then the client will

be able to spend money on optional add-ons to the policy. We see that those who purchase an additional legal option for police are less likely to churn than those who don't.

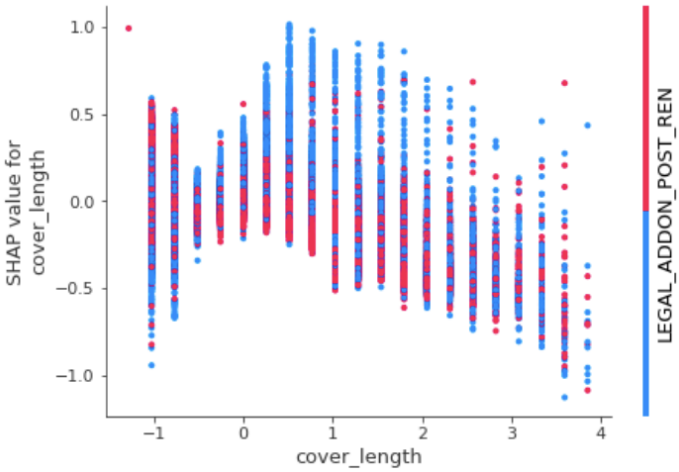


FIGURE 5.9: Partial dependence plot 4

To further illustrate how the model classifies, by using SHAP values, the typical decision path for both classes Figure 2 was built. We can see that each new feature value may increase or decrease the model score.

5.2.2 Tree visualization

To better understand how the machine learning model works, the decision tree that explains this process was built. Figure 5.10 presents part of the tree (full tree in appendix 5.6). The tree's inner vertices represent splits and specify factor names. This is a greedy method. Features are selected in order and their splits for substitution in each leaf. Leaf vertices contain raw values predicted by the tree (RawFormulaVal). RawFormulaVal is a number resulting from applying the model. It's not probabilities, RawFormulaVal, can be described as [Class1 = RawFormulaVal > 0]. Zero value in a leaf usually means that no train objects passed to this leaf. So, any positive RawFormulaVal means - "Churned," negative - "Not Churned."

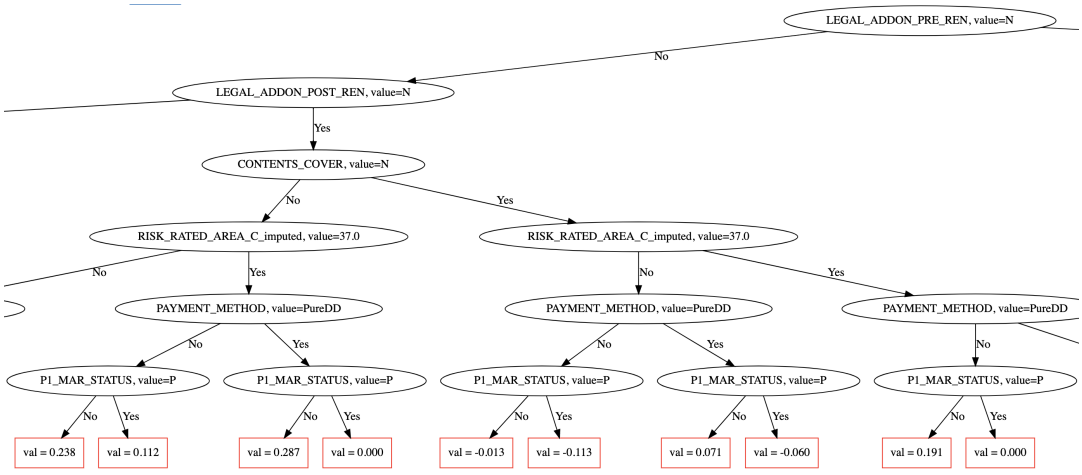


FIGURE 5.10: Catboost Tree Sample

In Figure 1, an entire tree is converted to the table form, with corresponded node values and final score; green color in RawFormulaVal row, when positive, red - negative. In this table, each column is a separate path from the root to the leaf; each row means a decision node (feature). For example, the first column in Figure 1 means that if P1_MAR_STATUS is not 'O', LEGAL_ADDON_PRE_REN=Y, LEGAL_ADDON_POST_REN=Y,LAST_ANN_PREM_GROSS<-1.014, PAYMENT_METHOD is not PureDD, and P1_MAR_STATUS is not 'P' then model classifies as 'Not Churned.'

5.3 Churn Segmentation

5.3.1 Segmentation

Dataset of churned policies consists of categorical and numerical columns. The combination of Factor Analysis of Mixed Data and K-Means clustering was used to segment churn policies. Firstly, the standardized data was fitted using FAMD. Then, a cumulative variance plot was used to decide how many FAMD components to keep.

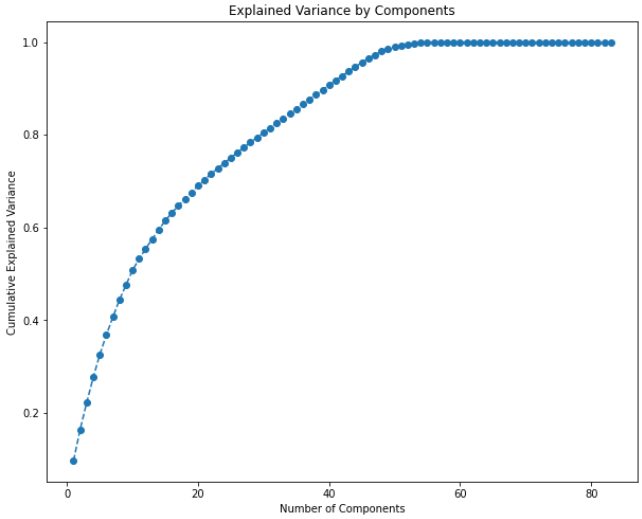


FIGURE 5.11: Cumulative Variance Plot

Figure 5.11 shows the amount of variance captured (on the y-axis) depending on the number of components we include (the x-axis). The general rule is to preserve around 80% of the variance. So, according to the graph, 30 components were selected, which explains 80.24% of the variance.

Then, scores obtained by FAMD were used to fit K-Means for segmentation. K-Means was used with a different number of clusters to determine the number of clusters. The Elbow method determined an optimal number of clusters by determining the Within Cluster Sum of Squares or WCSS for each solution. Based on the values of the WCSS plotted against the number of components [Figure 5.12], the decision on a number of clusters was made. The approach consists of looking for a part of the graph before the elbow would decline steeply, while the part after it – is much smoother. In this instance, the 5 clusters were selected.

The next step is to visualize the segments with respect to the first two components, where the x-axis represents the 2nd component, and y represents the first component. From Figure 5.13, we can observe the separate segments.

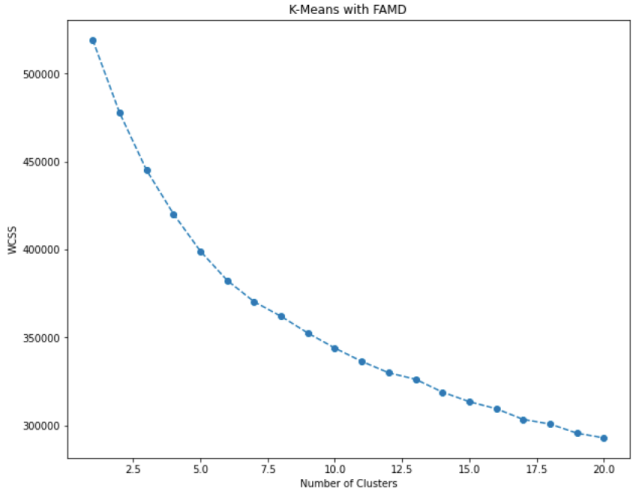


FIGURE 5.12: Elbow Plot

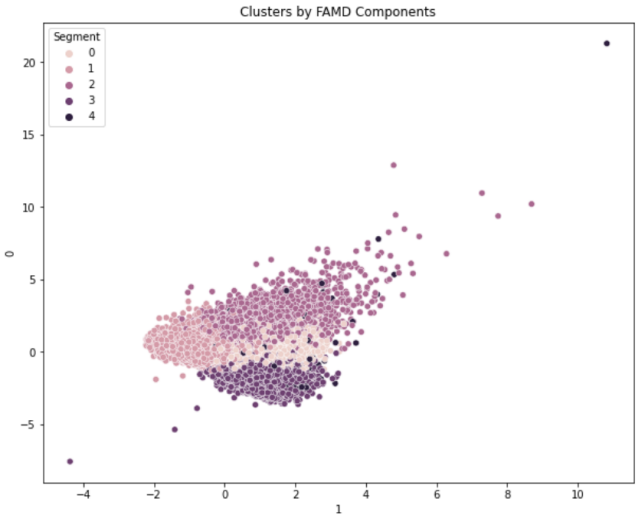


FIGURE 5.13: Churn Segments

Next, we'll examine the difference between each cluster. Table 5.1 presents the results for each cluster, where categorical values - are the most popular value, and numerical values are the mean.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Total Policies	6318	35496	6651	8073	307
Appropriate Lock	Y	Y	Y	Y	N
Listed (in Heritage List)	3	3	3	3	2
Contents Cover	Y	Y	Y	N	Y
Policyholder Sex	M	M	M	F	M
Legal Fees included before 1st renewal	Y	Y	Y	N	Y
Legal Fees included after 1st renewal	N	Y	Y	N	Y
NCD Granted Years B	4.2	6.0	5.8	0.0	4.3
NCD Granted Years C	1.1	6.1	5.8	5.9	5.2
Age	70.2	72.0	70.4	72.0	70.7
Last Annual Premium	220.2	203.1	279.0	81.9	269.3
Property Age	71.2	68.4	72.7	63.6	143.4
Cover Length	5.5	4.6	6.4	6.0	3.9
MTA FAP	59.9	33.3	277.7	21.8	94.5
MTA APRP	8.7	1.5	196.2	9.6	32.3

TABLE 5.1: Cluster Description

5.3.2 Cluster Interpretation

Based on data in Table 5, we analyzed the means of the variables in each cluster and made the following cluster interpretation:

Cluster 1 - Lost Discount
This cluster represents 11.1% of the total churned customers. These customers are price sensitive and lost their home policy discount. NCD Granted Years B - malus bonus for building, a premium discount for customers based on claim history. This segment has the highest difference between NCD Granted Years B (4.2) and Cover Length (5.5), 1 year 3 months. We can interpret it as a customer who had a discount for a long period and lost it, so the following policy cycle will include a price without a discount, which may be the reason for churn. Another important factor is that these customers added Legal Fees only once at the initial policy purchase, and after the first renewal, they never added it again. Legal Fees Addon is typically a small (10-20 $) optional payment covering legal expenses when the property is damaged. Not adding legal fees highlight that this segment may be price sensitive.

Cluster 3 - Betrayed Customer
This cluster represents 11.7% of the total churned customers. This segment has the biggest mean annual policy premium ($279) among all segments. Also, they have the highest MTA FAP and MTA APRP, which corresponds to the bonus before mid-term adjustment and mid-term adjustment premium. Mid-term adjustment is the change to policy price in the middle of the policy cycle. After discounting MTA FAP and MTA APRP by cover length, we can interpret the segment with this story. The customer gets an initial proposition to purchase a policy for $322. Still, during the sales process, it turned out that a $43 bonus was available for him, and the customer purchase policy for $279. Then, in 6 months, he received a message that his policy premium was adjusted, and he needed to pay an additional $30.8 (MTA APRP). This

adjustment makes him angry, and the next time during renewal, the customer leaves this insurance company because he feels betrayed. However, the nature of this price and bonus volatility is unknown. Whether changing risk circumstances or intended sales strategy, the customer feel this initial bonus was a trick to make him buy a policy.

Cluster 4 - Extreme Price-Sensitive
This cluster represents 14.2% of the total churned customers. These customers are extremely price-sensitive and purchase the most basic policy without content cover and any addons. This cluster is the only one that does not cover the content of the building. Moreover, they never add the legal addon to insurance during the initial purchase or after. They purchase insurance in the most basic combination; they just cover the building. Also, these policyholders do not receive a building discount (NCD Granted Years B = 0), but they are offered a discount for content insurance to motivate them to purchase content cover. We can see that they have the smallest policy premium ($81.9) among all clusters.

Cluster 5 - National Heritage Buildings
This cluster is the smallest and represents only 0.6% of the total churned customers. These customers own the oldest property, which includes in National Heritage List, and have the second-highest policy premium among all clusters. Also, we can notice that they have the lowest cover length, meaning that they are not likely to stay long with the current insurer.

Cluster 2 - Regular Churn
This cluster is the largest and represents 62.4% of churned customers. This cluster purchase policy includes content cover and legal fees with all possible add-ons. They have a long bonus for building and content and an average policy premium. So, they don't have volatile premium as 'Betrayed Customer,' 'Lost Disctount' as Segment 1, and any bonus problems. However, this large chunk of customers leaves early compared to others. For this segment, the clustering does not show any evidence of any problem. So, they are classified as regular churn.

5.4 Integrated Churn Prediction and Segmentation

Classification model results on the test sample were combined with clustering to understand insurance churn further. The test sample was fitted to the catboost model and the KMeans model. In Table 5.2, there are results of model performance split by Segment.

The results show that the model doesn't have equal performance among all segments. We can see that the smallest (0.6%) "National Heritage Building" was not even detected by the model. There is a significant difference between recall of the model, which varies between 12.7% to 40.6% among segments. On the other hand, precision does not have such a big spread, with a range of 65.0% to 71.8%. These results help make proper churn management decisions. The traditional approach would be to use a classification model without any further segmentation, meaning that all "churn" is considered the same. This difference in performance by segment allows us to use only part of the model so that the lowest-performing clusters can be excluded from consideration during business decisions.

	Extreme Price Sensitive	Lost Discount	Regular Churn	Betrayed Customer
TP	239	1363	3821	488
FP	1646	2519	5597	1440
TN	8310	8849	15017	4447
FN	94	591	2054	227
Recall	12.7%	35.1%	40.6%	25.3%
Precision	71.8%	69.8%	65.0%	68.3%
F1-Score	21.6%	46.7%	50.0%	36.9%

TABLE 5.2: Model performance per churn segment

5.5 Business Implication

The only way to grow the insurance business is to improve retention and reduce churn. Retention is the ratio of customers who return to purchase from a company. On the other hand, churn rate is the rate of lost customers divided by the number of customers. By default, the company with a high retention rate has a low churn rate and vice versa. For the company increasing retention and decreasing churn are the same goal but in different terms. There are two to learn reasons behind churn:

1. Churn Surveys

   Sending brief surveys to individuals who cancel their policies is one approach is learn from customer churn, similar to how someone unsubscribing from an email list could have a few alternatives to select as to why they're unsubscribing. This type of qualitative survey is not covered in this paper. However, a quantitative approach should not be used with the qualitative survey. Talking to customers is a must-have step to validate quantitative research successfully.

2. Analyze Churn Behaviour

   With quantitive research covered in this paper, we can uncover churn trends and similar groups that tend to churn. Price sensitivity, customer loyalty, and personal characteristics highly impact churn. Although the model does not fully describe churn phenomena, it still gives us an understanding of important factors that affect customer churn. Price sensitivity - is a measure of customer willingness to pay or save on insurance. Several factors describe a willingness to pay. If the customer wants to add additional optional services to insurance like legal insurance, garden, and cover the house's content, he is willing to spend more. On the other hand, customers who purchase the most basic insurance configuration and are sensitive to bonuses are considered those who have a low budget or willingness to pay. The length of the insurance cover describes customer loyalty. We can see the trend that the longer customer stays with the same insurers, the less churn. The last factor is personal characteristics, like age, marital status, building type, etc. All of these features are about risk. Similarly, as in banking, we can see the trends that older customers or married customers are both considered less risky and therefore have a lower churn.

Recommended Approach for each Churn Segment

The recommendation is based on churn segmentation and model performance per segment. Based on the classification, five churn segments were identified: Extreme Price-Sensitive, Lost Discount, Betrayed Customer, Regular Churn, and National Heritage Buildings.

- Extreme Price-Sensitive and Lost Discount Segments

  These two segments have the lowest willingness to pay and churn once the price is not suitable. The problem for the "Extreme Price-Sensitive" segment is that they don't have a bonus for building discount from the policy premium for building coverage, and they represent 14.2% of all churn. For the "Lost Discount" segment, which is 11.1%, the problem is that they lost their discount for insurance that they had for a relatively long period. The recommended strategy for both segments is to reduce the profit margin. Insurance policy price consists of technical price (break-even price) and profit margin. The % discount from profit should be determined by constant ab testing. The conversion rate of churned population will impact the decision on this margin. For segment "Lost Discount," it's recommended to decrease the price in the form of a bonus, as they had before, and for "Extreme Price-sensitive" in the form of a policy discount. The price range for the "Extreme Price-Sensitive" segment for the testing should be $76.9 to $81.9 and for "Lost Discount" from $206.8 to $220.2. This is the range between break-even price and the current segment price. The other approach is to use better marketing communication. The first approach is price match (guarantee), to match the price if competitors propose a better price. Except for price match, another strategy is to provide a comparison table of prices with other competitors.Price Testing Example:Conduct Pricing Test by reducing profit margin by 50%. Considering the typical insurance profit margin is 6.5%, the profit margin is set to 3.25%, with an expected conversion rate of "churned" users - 50

Extreme Price-Sensitive:

Model performance for this segment: recall - 12.7% precision - 71.8% at a 0.5 threshold. This means 239 policies are correct, and 94 policies are incorrect. The average policy price (premium) is 81.9$, the profit from 1 policy is 5$ (6.5%), break-even price=76.9$. So, considering the test setup, this segment will have a policy price=79.4$ and a profit 2.5$ (3.25%). From 239 churned, half will purchase insurance, so we'll have 80 policyholders with revenue 80x2.5$=200$. For the "not churned," we'll have 313x2.5$=782.5$. So, total profit = 1777.5$, without test, profit will be 313x5$=1565$. So, the test resulted in a 13.6% ($212.5) increase in profit for the "Extreme Price-Sensitive" Segment.

Lost Discount:

Model performance for this segment: recall - 35.1% precision - 69.8% at a 0.5 threshold. This means 4543 policies are correct, and 1970 policies are incorrect. The average policy price (premium) is 220.2$, the profit from 1 policy is 13.4$ (6.5%), break-even price=206.8$. So, considering the test setup, this segment will have a policy price=213.5$

and a profit of 6.7$ (3.25%). From 4543 churned, half will purchase insurance, so we'll have 2272 policyholders with revenue 2272x6.7$=15222.4$. For the "not churned," we'll have 1970x6.7$=13199$. So, total profit = 28421.4$, without test, profit will be 1970x13.4$=26398$. So, the test resulted in a 7.7% ($2023.4) increase in profit for the "Lost Discount" Segment.

Price solution covers two segments and 25.3% of all churned users.

- Betrayed Customer Segment

  This cluster represents 11.7% of the total churned customers. These customers get an initial discount to purchase a policy, and in the middle of the policy period, they are asked by an insurance company to pay additional money because their policy price was adjusted. So, this segment gets a false feeling of being "betrayed" by the company. The problem is with the promotion of insurance. Inaccurate advertising and promotion strategies created a wrong expectation for this segment. The solution is to improve messaging to this customer, so they'll understand that their bonus is not absolute and may be changed in the future. The company needs to establish a new process for insurance agents to solve this miscommunication. Another strategy is to communicate better why we increase or change the price in the middle of the policy term. Better communication can help reduce this churn rate in this segment.

- Regular Churn Segment

  This cluster is the largest and represents 62.4% of churned customers. This segment does not have specific insights that reveal churn reasons from the available data. Compared to "Betrayed Customer," "Lost Discount," and "Extreme Price-Sensitive" segments, there are no marketing or pricing issues that may explain this cluster. The only significant difference is that this customer churned earlier than the rest. On average, they spend 24% less time with the company. We don't have enough data to identify the problem for this segment. The company needs to set up a survey collection process, or interview churned customers to collect more data. We will be able to find the broken process for this segment. For example, we have a problem with the renewal process. A historically significant part of churn is happening at the moment of renewal. So, if the renewal procedure is overly complicated, in that case, customers may believe that they would be better off switching insurers. The insurance business depends on agents and personal skills in serving customers. This personal service impacts churn, and this segment may have a problem with the renewal process or other agency processes. The solution may be the automation of the renewal process or the training of agents to improve customer service. But it's also important to keep in touch with customers and send them recommendations and reminders so they will have loyalty to the brand. The solution for this segment is to improve customer service, but firstly to narrow the problem by conducting user interviews or another qualitative survey to discover what process is broken.

- National Heritage Buildings Segment

  This segment is only 0.6% of all churned users. These customers are the smallest segment since they have an old property listed on National Heritage List. Among all segments, they churn the earliest. It's typical for

the insurance industry that insurance products cannot fulfill all customer segments. There is no product-market fit for this segment, as the company does not satisfy the demand. The company needs to conduct market research and understand whether this segment has high investment potential. The solution will be a dedicated product for historic buildings if there is enough demand. If there is no goal to benefit and market for developing a new insurance product, the recommendation will be not to target such users as they are not profitable. On the marketing side, we can improve pricing argumentation. Marketing can explain better why the price is high, because of a hazardous and old property. For this segment, we can propose a price guarantee. These policyholders will be less likely to churn since they need to go through a complicated and costly property inspection process to switch to competitors.

Chapter 6

Conclusions and future work

In this work, we addressed the problem of churn in the insurance industry. We focused on the prediction and segmentation of customer churn. To achieve this, we firstly analyzed and processed US Insurance Company data, created new features, imputed missing values, and removed correlated features.

In the methodology part, we researched different classification methods and evaluation metrics, discussed the performance of other models, and selected Catboost as a model to use for the final solution. Also, we researched the methodology of segmenting data with mixed data types, and we decided to use Factor Analysis of Mixed Data in combination with K-Means for clustering. In the practical part, we built a churn prediction model, segmented churned customers combined the results of the model and segmentation, and provided sufficient segment descriptions and recommendations to businesses. We used Catboost to build a prediction model, then discussed model results after cross-validation. Also, we discussed the rules of how prediction is made with decision tree visualizations and estimated feature importance using SHAP values. In the segmentation part, we combined Factor Analysis of Mixed Data with K-Means for clustering. Firstly, we determined an optimal number of components for FAMD, fitted the data, and then using the Elbow method, determined the optimal number of clusters for the K-Means algorithm. Then we visualized churn segments and made cluster interpretations by means of the variables in each group. Finally, we studied model performance among all churn segments. We found out that churn segments have different prediction power by our model. The final section of the practical part is dedicated to business implications. In this part, we proposed solutions for each churn segment based on insights uncovered in segment description and integrated prediction model and segmentation parts.

One of the main contributions of this work is the estimation of model results on different churn segments. We considered customer churn a complex customer behavior feature and not a boolean feature. With this, we understand the weakness of our model. For "Extreme Price-Sensitive" and "Lost Discount" segments, a business can utilize this model directly to determine appropriate customers to whom decrease price. For the "Betrayed Customer Segment," this model should be used as a targeting tool, and for the rest segments, the model doesn't have utility in decreasing churn.

In the next step of this research, we will focus on the following improvements:

1. Improve feature engineering and extend data. More data, such as call logs, marketing data, and loan data, are needed to improve the model.
2. Experiment with the model by underfitting and overfitting using SMOTE (Synthetic Minority Overlapping Technique)
3. Narrow target variable by excluding churn segments for which model is inappropriate.
4. Research multivariate classification to classify churn per segment by default.

Appendix A

Feature Name	Mean	Std	Min	25%	50%	75%	max
ncd granted years B	4.5	2.7	0.0	3.0	6.0	6.0	9.0
ncd granted years C	5.5	1.8	0.0	6.0	6.0	6.0	9.0
last annual premium	186.7	99.5	-1152	-123.4	177.3	235.0	4632
age	72.7	10.9	21.0	65.0	73.0	81.0	123.0
property age	68.0	28.9	13.0	53.0	67.0	93.0	264.0
cover length	6.0	3.9	1.0	3.0	5.0	8.0	21.0
MTA FAP	58.8	108.5	-1152	0.0	0.0	97.1	4632
MTA APRP	26.5	77.8	-423.1	0.0	0.0	0.0	1449

TABLE 1: Numerical features description

Feature Name	Type	Description	Transformation
ncd granted years b	numerical	Bonus Malus - Building	scaled
ncd granted years c	numerical	Bonus Malus - Personal Items	scaled
last ann prem gross	numerical	Premium - Total for the previous year	scaled
age	numerical	Client Age	scaled, constructed from date of birth and current date difference
property age	numerical	Property Age	scaled, constructed from property build date and current date difference
cover length	numerical	Insurance policy duration	scaled, constructed from policy start date and current date difference
mta fap imputed	numerical	Bonus up to date of Adjustment	scaled, missing values filled with 0
mta aprp imputed	numerical	Adjustment of the premium for Mid-Term Adjustment	scaled, missing values filled with 0
claim3years	categorical	3 last years loss	
bus use	categorical	Commercial use indicator	
contents cover	categorical	Personal Objects is covered by insurance	
buildings cover	categorical	Building is covered by insurance	
p1 mar status	categorical	Marital status of the client	

p1 sex	categorical	customer sex	
appr alarm	categorical	client has an appropriate alarm	
appr lock	categorical	client has an appropriate lock	
flooding	categorical	House susceptible to floods	
listed	categorical	building is listed in National Heritage List	
neigh watch	categorical	Vigils of proximity present	
occ status	categorical	Occupancy status	
ownership type	categorical	Type of ownership	
prop type	categorical	Type of property	
safe installed	categorical	Safe installs	
subsidence	categorical	Subsidence indicator	
payment method	categorical	Method of payment	
legal addon pre ren	categorical	Option "Legal Fees"included before 1st renewal	
legal addon post ren	categorical	Option "Legal Fees" included after 1st renewal	
home em pre ren	categorical	"Emergencies" option included before 1st renewal	
home em post ren	categorical	"Emergencies" option included after 1st renewal	
garden addon pre ren	categorical	Option "Gardens" included before 1st renewal	
garden addon post ren	categorical	Option "Gardens" included after 1st renewal	
keycare addon pre ren	categorical	Option "Replacement of keys" included before 1st renewal	
keycare addon post ren	categorical	Option "Replacement of keys" included after 1st renewal	
HP1 addon pre ren	categorical	Option "HP1" included before 1st renewal	
HP1 addon post ren	categorical	Option "HP1" included after 1st renewal	
HP2 addon pre ren	categorical	Option "HP2" included before 1st renewal	
HP2 addon post ren	categorical	Option "HP2" included after 1st renewal	
HP3 addon pre ren	categorical	Option "HP3" included before 1st renewal	
HP3 addon post ren	categorical	Option "HP3" included after 1st renewal	
risk rated area C imputed	categorical	Geographical Classification of Risk Personal Objects	missing values filled with 0 category

TABLE 2: Complete List of Features

P1_MAR_STATUS=O	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
LEGAL_ADDON_PRE_REN=N	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
LEGAL_ADDON_POST_REN=N	n	n	n	n	n	n	n	n	y	y	y	y	y	y	y	y
LAST_ANN_PREM_GROSS>-1.014	n	n	n	n	y	y	y	y	n	n	n	n	y	y	y	y
PAYMENT_METHOD=PureDD	n	n	y	y	n	n	y	y	n	n	y	y	n	n	y	y
P1_MAR_STATUS=P	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y
RawFormulaVal	-0.09	-0.12	-0.16	-0.21	-0.03	-0.1	-0.09	-0.16	-0.05	-0.24	0.04	-0.04	0.08	-0.12	0.15	-0.03
X_IS_CHURNED	n	n	n	n	n	n	n	n	n	n	y	n	y	n	y	n
P1_MAR_STATUS=O	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
LEGAL_ADDON_PRE_REN=N	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
LEGAL_ADDON_POST_REN=N	n	n	n	n	n	n	n	n	y	y	y	y	y	y	y	y
LAST_ANN_PREM_GROSS>-1.014	n	n	n	n	y	y	y	y	n	n	n	n	y	y	y	y
PAYMENT_METHOD=PureDD	n	n	y	y	n	n	y	y	n	n	y	y	n	n	y	y
P1_MAR_STATUS=P	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y
RawFormulaVal	-0.14	-0.24	-0.10	-0.14	-0.17	-0.3	-0.11	-0.08	-0.07	-0.13	-0.15	-0.23	-0.02	-0.1	-0.07	-0.18
X_IS_CHURNED	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
P1_MAR_STATUS=O	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
LEGAL_ADDON_PRE_REN=N	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
LEGAL_ADDON_POST_REN=N	n	n	n	n	n	n	n	n	y	y	y	y	y	y	y	y
LAST_ANN_PREM_GROSS>-1.014	n	n	n	n	y	y	y	y	n	n	n	n	y	y	y	y
PAYMENT_METHOD=PureDD	n	n	y	y	n	n	y	y	n	n	y	y	n	n	y	y
P1_MAR_STATUS=P	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y
RawFormulaVal	-0.13	0.0	-0.18	0.0	-0.099	0.0	-0.12	0.0	-0.01	0.0	0.08	0.0	0.11	0.0	0.2	0.0
X_IS_CHURNED	n	n	n	n	n	n	n	n	n	n	y	n	y	n	y	n
P1_MAR_STATUS=O	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
LEGAL_ADDON_PRE_REN=N	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
LEGAL_ADDON_POST_REN=N	n	n	n	n	n	n	n	n	y	y	y	y	y	y	y	y
LAST_ANN_PREM_GROSS>-1.014	n	n	n	n	y	y	y	y	n	n	n	n	y	y	y	y
PAYMENT_METHOD=PureDD	n	n	y	y	n	n	y	y	n	n	y	y	n	n	y	y
P1_MAR_STATUS=P	n	y	n	y	n	y	n	y	n	y	n	y	n	y	n	y
RawFormulaVal	-0.17	0.0	-0.24	0.0	-0.26	0.0	-0.22	0.0	-0.12	0.0	-0.14	0.0	-0.05	0.0	-0.08	0.0
X_IS_CHURNED	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n

FIGURE 1: Catboost Tree in Table format

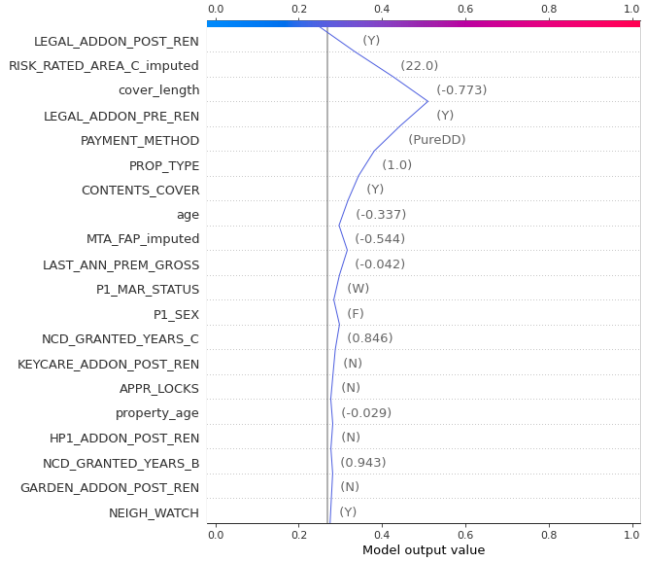


FIGURE 2: SHAP: Decision Example

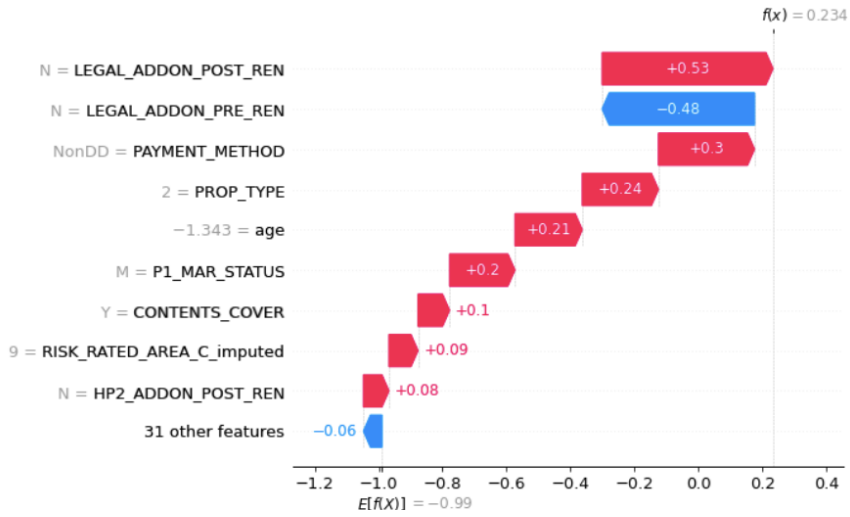


FIGURE 3: SHAP Cumultive Feature Importance Plot

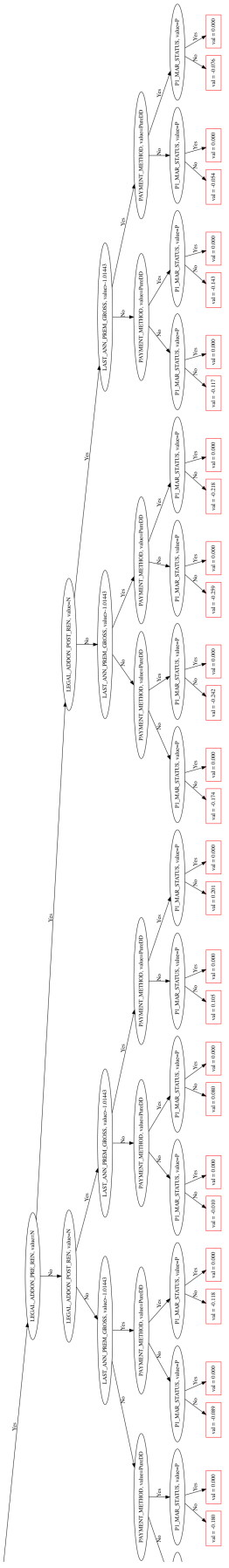
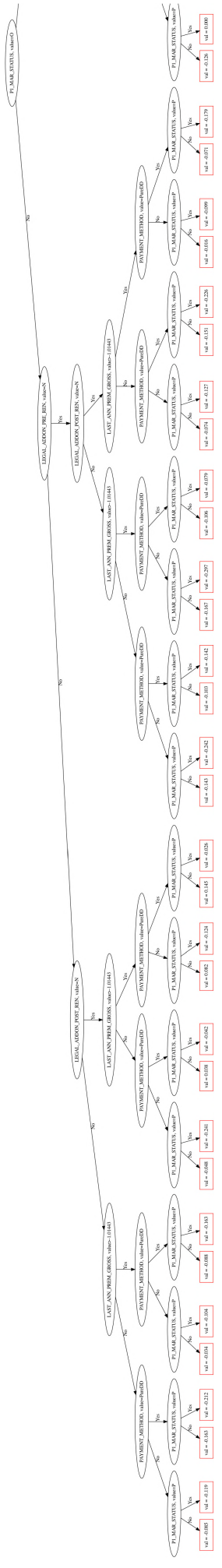


FIGURE 4: Catboost Decision Tree


PJ_MAR_STATUS. value=0
No
LEGAL_ADDON_PRE_REN. value=N
Yes
LEGAL_ADDON_POST_REN. value=N
Yes
LAST_ANN_PREM_GROSS. value<=110.443
Yes
PAYMENT_METHOD. value=Prem30
Yes
PJ_MAR_STATUS. value=F
Yes
val = 0.126
No
val = 0.066
No
PJ_MAR_STATUS. value=P
Yes
val = 0.071
No
val = 0.179
No
PAYMENT_METHOD. value=Prem60
Yes
PJ_MAR_STATUS. value=F
Yes
val = 0.099
No
val = 0.016
No
PJ_MAR_STATUS. value=P
Yes
val = 0.131
No
val = 0.226
No
LAST_ANN_PREM_GROSS. value>110.443
Yes
PAYMENT_METHOD. value=Prem30
Yes
PJ_MAR_STATUS. value=F
Yes
val = 0.074
No
val = 0.127
No
PJ_MAR_STATUS. value=P
Yes
val = 0.106
No
val = 0.079
No
PAYMENT_METHOD. value=Prem60
Yes
PJ_MAR_STATUS. value=F
Yes
val = 0.187
No
val = 0.297
No
PJ_MAR_STATUS. value=P
Yes
val = 0.103
No
val = 0.142
No
No
LEGAL_ADDON_POST_REN. value=N
Yes
LAST_ANN_PREM_GROSS. value<=110.443
Yes
PAYMENT_METHOD. value=Prem60
Yes
PJ_MAR_STATUS. value=F
Yes
val = 0.145
No
val = 0.252
No
PJ_MAR_STATUS. value=P
Yes
val = 0.082
No
val = 0.124
No
PAYMENT_METHOD. value=Prem30
Yes
PJ_MAR_STATUS. value=F
Yes
val = 0.145
No
val = 0.026
No
LAST_ANN_PREM_GROSS. value>110.443
Yes
PAYMENT_METHOD. value=Prem30
Yes
PJ_MAR_STATUS. value=F
Yes
val = 0.038
No
val = 0.042
No
PJ_MAR_STATUS. value=P
Yes
val = 0.038
No
val = 0.241
No
PAYMENT_METHOD. value=Prem60
Yes
PJ_MAR_STATUS. value=F
Yes
val = 0.088
No
val = 0.163
No
PJ_MAR_STATUS. value=P
Yes
val = 0.034
No
val = 0.304
No
PAYMENT_METHOD. value=Prem30
Yes
PJ_MAR_STATUS. value=F
Yes
val = 0.183
No
val = 0.212
No
PJ_MAR_STATUS. value=P
Yes
val = 0.085
No
val = 0.119
No

Bibliography

[1] Jonathan Burez and Dirk Van den Poel. "Handling class imbalance in customer churn prediction". In: *Expert Systems with Applications* 36.3 (2009), pp. 4626–4636.

[2] Shin-Yuan Hung, David C Yen, and Hsiu-Yu Wang. "Applying data mining to telecom churn management". In: *Expert Systems with Applications* 31.3 (2006), pp. 515–524.

[3] Gholamreza Jandaghi and Zahra Moradpour. "Segmentation of life insurance customers based on their profile using fuzzy clustering". In: *International Letters of Social and Humanistic Sciences* 61 (2015), pp. 17–24.

[4] Jérôme Pagès. *Multiple factor analysis by example using R*. CRC Press, 2014.

[5] Liudmila Prokhorenkova et al. "CatBoost: unbiased boosting with categorical features". In: *Advances in neural information processing systems* 31 (2018).

[6] Maria Spiteri and George Azzopardi. "Customer churn prediction for a motor insurance company". In: *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*. IEEE. 2018, pp. 173–178.

[7] Irfan Ullah et al. "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector". In: *IEEE access* 7 (2019), pp. 60134–60149.