Predicting the time of Insurance Policy Cancellation

Stefanela Stevanović[1]

Abstract

In this project, an attempt was made to predict the cancellation time of the insurance policy with two approaches: by predicting the final duration of the policy, and by simplifying the problem to the one of classification nature, where it's predicted if the policy will be cancelled prematurely, at its expiration or be extended. For those approaches different regression and classification models were tested, for which the performance was evaluated and compared. While no production-level results were achieved, the best performing models outperformed the naive estimators by 19.5% in the first approach and by 25.3% in the second approach.

Keywords

Predictive Modeling, Policy Cancellation, Regression, Classification, Decision Tree, Random Forest

[1] *ss51676@student.uni-lj.si, 63220492*

Introduction

This project is mainly centred around predicting the time of the policy cancellation on the data provided by Zurich Insurance Company. Predicting the cancellation time of the policy is of extreme importance for the company. Knowing which policies have a high probability of being cancelled soon allows the company to devote more attention to customers of those policies and potentially take some measures to prevent the cancellation.

This problem was approached in two ways, by formulating two subproblems: predicting the policy's duration in months, and predicting whether the policy will be cancelled prematurely, at the moment of its expiration, or be extended. The first one is of a regression nature, while the second is of a classification nature, and therefore these two problems required different approach in the terms of using different predictive models and evaluation metrics.

All the coding was done in Python programming language and the code for reproducing the results from this report is available in the provided Jupyter notebook.

Methods

In this part of the report, it is explained how the data was prepared, which models were used, and how they were trained, tested and evaluated. Since prediction of the duration of the policy is a regression problem, and prediction of whether the policy will be closed prematurely, at its expiration or will be extended,is a classification problem, these two problems require a different approach, in terms of using different predictive models and evaluation metrics.

Data Preprocessing

The main data set used for predictive modeling of both problems is the policy data data set, which contains the following columns: underwriting date, first end date, cancellation or end date, policy id, sales channel, customer id, premium, status, line, product name and product group.

For the active policies (status = 1), it is not possible to calculate the final duration in months, which is the target value we want to predict, and therefore it is not possible to use these policies for training and testing of the models. For this reason, all active policies were removed from the data set. However, it should be mentioned that active policies make up 28.8% of the total number of policies available in the data set and that their removal may lead to bias.

Since the target variables of both problems are not available in the original data set, they had to be calculated based on the available features. For this reason, rows without underwriting date, first end date or cancellation end date had to be removed, which led to the removal of an additional 5.7% of the data. The target variable for the first problem, which is the final duration of policy in months, was calculated by subtracting the underwriting date from the cancellation or end date for each policy and was labelled with 'duration'. The target variable of the second (classification) problem was labelled with 'duration status' and was calculated by comparing the first end date with the cancellation or end date. Duration status takes the following values:

- 0 if the policy was closed prematurely
- 1 if the policy was closed at its expiration
- 2 if the policy was extended

Some of the columns, which were assumed to be irrelevant

for prediction, were removed from the data set. Also, customer type information (business, natural person male or natural person female) from the customer data data set was added to the prepared data set in the form of dummy variables. The final list of attributes used to predict the target variables for both problems (duration and duration status):

- customer ID: unique for each customer
- underwriting year: just a year from underwriting date
- first end duration: the difference in months between the underwriting date and first end date
- premium: price paid by customer
- line: 0,1 or 2
- product group: classes with labels 0 to 158
- business: 1 if true, 0 if false
- natural person male: 1 if true, 0 if false
- natural person female: 1 if true, 0 if false

Cross-Validation

To ensure that the model is accurately evaluated, it's important to test it on an unseen data (data that it was not trained on). This means that part of the available data should be held out during the fitting of a prediction model, which is achieved by splitting the data into training and testing sets. To avoid the separation bias and ensure that all data is equally considered, 5-fold cross-validation approach is implemented. The data was divided into five folds, which means that five iterations are required to fully complete training and testing of the model. During each iteration, one fold is considered for testing and the rest is for training, changing the testing fold at each iteration. In the end, each data point was used once in a test set and four times in the training. The reported performance measure of the predictive model is the mean performance measure of these five models. This process is visualized in Figure 1.

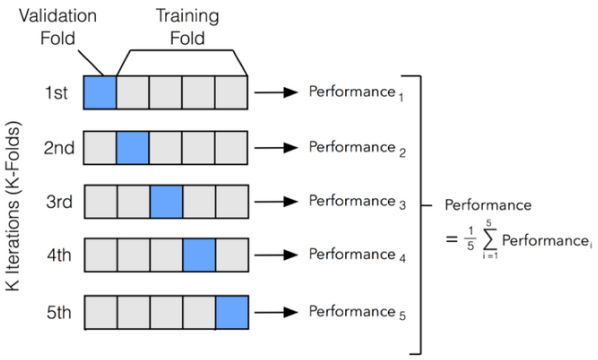

Figure 1. 5-fold cross-validation In each iteration 20% of the data, on which the model is tested, is withheld from the training. Reported performance measure of the predictive model is the mean performance measure of these five models.

For the classification problem the stratified 5-fold cross-validation was implemented, which is just a modification of the classic 5-fold approach. Stratified 5-fold cross validation ensures that the training and test data in each fold reflects the imbalanced distribution of the target variable value in the original data set. This means that the folds are made by preserving the percentage of samples of each class.

Models

For the prediction of the policy duration in months, as a regression problem, Gradient Boosting Regression, Decision Tree Regression and Random Forest Regression were used. For the comparison purpose, Mean and Median Dummy Regressors were used as a baseline models.

Since predicting the duration status is a classification problem, Dummy Classifier, which always predicts the most frequent class in the training data, was used as a baseline model. Among actual models, Multinomial Logistic Regression, Decision Tree Classifier, Random Forest Classifier and Multi-layer Perceptron neural network were used.All of the mentioned models are available as a part of sklearn library.

Many of the mentioned regression and classification models require feature scaling. In order to prevent data leakage (some information from the training data being revealed to the testing data), the scaler was fitted on the training data and then used to transform the test data. For the purpose of this project MinMaxScaler() from sklearn library was used to normalized the data. However, experiments have shown that Decision Tree and Random forest models, which don't necessarily require feature scaling, have slightly better performance on an unscaled data. For this reason, the data was not normalized when implementing these models.

Evaluation Metrics

Mean absolute error (MAE) was used for evaluating the regression models for predicting the policy duration. Mean absolute error is calculated as:

$$MAE = \frac{\sum_{i=1}^{n} |y_{pred} - y_{test}|}{n}, \quad (1)$$

where y_pred and y_test denote predicted and observed target variable values, respectively. The smaller the mean absolute error, the better the performance of the model.

To evaluate the quality of the classifier outputs, standard classification accuracy score was used:

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}, \quad (2)$$

A higher accuracy score indicates a greater number of correct predictions, which means better model performance.

Results

This project is centred around predicting the time of the policy cancellation, for which two problems, that require different approach, were formulated: predicting the policy's duration in months, and predicting whether the policy will be cancelled prematurely, at the moment of its expiration, or be extended. In this part of the paper, the results obtained by the described methodology for both problems are discussed and presented in the form of tables and visualizations.

Problem 1: Predicting the policy's duration in months

When formulating this problem, the following fact was considered: if it is possible to predict the final duration of the policy, it is also possible to predict the moment when it will be cancelled. The predicted time of the policy's cancellation would be obtained by adding the predicted duration of the policy to it's underwriting date.

As the defined problem is of the regression nature, Decision Tree Regressor with 5-fold cross-validation was firstly implemented. Experiments have shown that the optimal value of the maximum tree depth parameter is 20, as shown in Table 1. Maximum tree depth is a limit to stop further splitting of nodes when the specified tree depth has been reached during the building of the initial decision tree. If the maximum tree depth is set at 30, we see that mean absolute error on the testing data starts increasing, meaning that the decision tree is starting to overfit the training data without capturing useful patterns.

Table 1. Mean absolute error across splits for different maximum depths of the Decision Tree Regressor.

Decision Tree Regression	
Maximum Depth	Mean Absolute Error
5	12.89
10	11.95
20	11.84
30	12.67

Mean absolute error across five splits for Decision Tree Regressor (max depth = 20), Random Forest Regressor (max depth = 20) and Gradient Boosting Regressor are shown in Figure 2. We see that for all of these regression models, the mean absolute error is around 12 months, which indicates that with used features it is not possible to accurately predict the duration of the policy. Also, we can see that the value of mean absolute error doesn't vary much between five testing sets. An error bars, a black lines in the Figure 2, represent the standard deviation of the MAE in the five testing sets. However, although the performance of regression models is not satisfactory, all the models outperform Mean and Median Dummy, that have mean absolute error of ≈ 17.81 and ≈ 13.91, respectively. The model with the lowest mean absolute error ≈ 11.64 is the Random Forest Regressor. In order to see which features had the greatest impact on predicting the duration of the policy, the feature importance for this model is shown in Figure 3.

Problem 2: Predicting if the policy will be cancelled prematurely/at its expiration/be extended

Given that predicting the duration/time of policy cancellation by regression did not give good results, an effort was made to simplify this problem to the one of classification nature. In this part of the project we try to find the model that is able to accurately predict one of the three possible labels for each

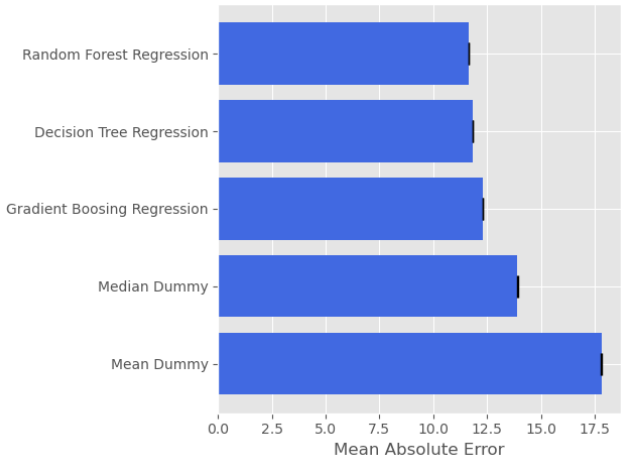

Figure 2. Performance of the Regression Models. This figure shows that the Random Forest Regressor performs best, with a mean absolute error of ≈ 11.64, outperforming the Median Dummy by 19.5%.

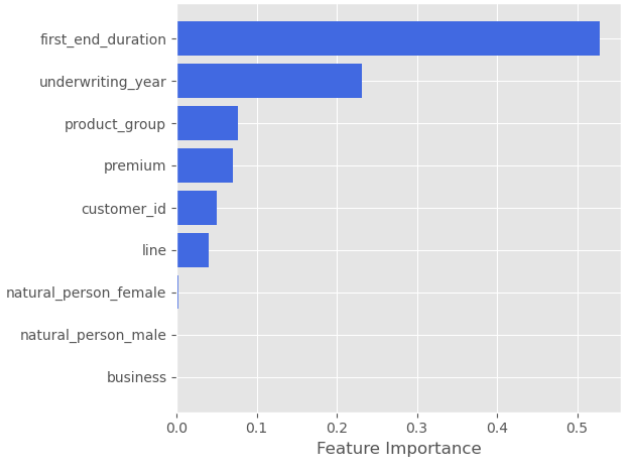

Figure 3. Feature importance of the Random Forest Regressor. This figure shows that the final end duration and the year in which the policy was underwritten had the greatest influence on the predicted output.

active policy: 0 - policy being cancelled before expiration, 1 - policy being cancelled at its expiration and 2 - policy being extended. In order to obtain such a model, training and testing had to be done on inactive policies that have information about cancellation of the policy.

Experiments were performed with different classifier models using stratified 5-fold cross validation. Experiments on the Decision Tree Classifier again showed that 20 is the optimal maximum tree depth, as shown in Table 2. The average accuracies for all used classification models are shown in Figure 4. The Decision Tree and the Random Forest Classifier, for which the maximum tree depth was set to 20, had the highest average accuracy, 62.4 % and 62.9%, respectively.Although these accuracies are far from desirable, we can notice that all

trained classification models outperform the Dummy Classifier, which always predicts label 1 as the most frequent class, with an accuracy of 50.2%.

Table 2. Average accuracy across splits for different maximum depths of the Decision Tree Classifier.

Decision Tree Classification	
Maximum Depth	Average Accuracy
5	0.56
10	0.61
20	0.62
30	0.60

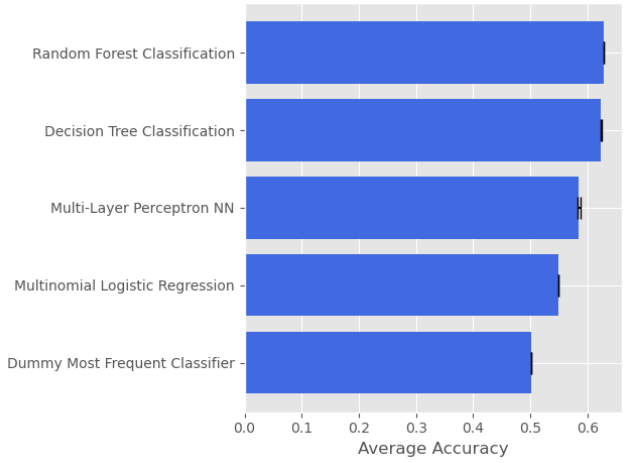

Figure 4. Performance of the Classification Models. Comparison of the average accuracy of classification models shows that the Random Forest Classifier performs best, with an average accuracy of ≈ 0.629, outperforming the Dummy Classifier by 25.3%.

For the Random Forest Classifier, as the classifier with best performance, the importance of the features on the output of the model was shown in Figure 5. Comparing these results with Figure 3, we see that the distribution of feature importances differs for the Random Forest Regressor and Classifier. However, in both cases first end duration, underwriting year and product group happen to be the most important features.

Furthermore, we wanted to show with what probability label 0, 1 or 2 is predicted as the most probable label for each policy with the Random Forest Classifier. For this purpose, the *predict_proba* method of the sklearn library was used, which returns the class probabilities for each data point (three probabilities in this case, for each label). The maximum class probability was extracted for each policy and the frequency distribution of these probabilities was shown in Figure 6.

Discussion

In this project we implemented and tested different regression and classification models for the purpose of predicting policy

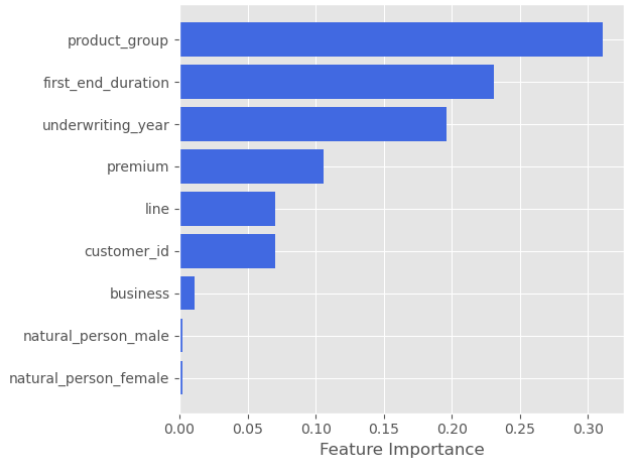

Figure 5. Feature importance of the Random Forest Classifier. This figure shows that the product group, final end duration and the year in which the policy was underwritten had the greatest influence on the predicted output.

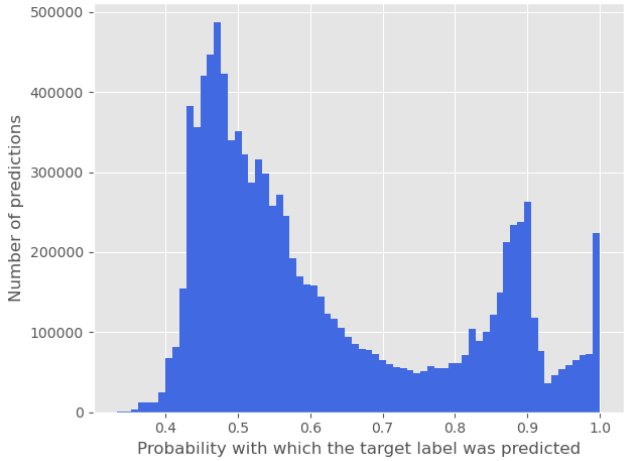

Figure 6. Frequency distribution of the probabilities with which the labels were predicted using Random Forest Classifier. This figure shows that the highest number of labels were predicted with a probability ≈ 0.47, which indicates that the model in most cases predicts a label with low confidence.

cancellation. Although these models outperform naive estimators, their performance is far from satisfactory for direct use. Possible improvements with the available data could come from experimenting with additional features in the training set. Some of the possible features that could motivate the customer to cancel the policy and that should be taken into consideration are the value of the last claim before the policy cancellation and the period for which the customer waited for the claim payment. Another possible idea is to include the number of policies the customer bought, as we can assume that loyal customers are satisfied with the company and thus less likely to cancel the policy before the expiration date.