Customer Churn Prediction using Quotation Data

Master Thesis

Submitted on: November 22, 2021

at the University of Cologne

Name:	Abdurahman Maarouf
Adress:	Schulstrasse 31
Postcode, Area:	53332, Bornheim
Country:	Germany
Matriculation number:	736481
Supervisor:	Prof. Dr. Dominik Wied

Contents

1 Introduction

Predicting customer churn in order to retain customers has become one of the most important issues for companies. The goal is to estimate probabilities for a costumer churning in the next period of time, in order to be able to detect potential churners before they leave the company. To tackle this issue, more and more advanced Machine-Learning-Algorithms are used guaranteeing high accuracy in their out-of-sample predictions.
Fortunately for most of the companies, churn rates from one period to another are very small. However in classification models predicting a rare event can become challenging. In this so called "Imballanced Classes" issue certain arrangements to the underlying training data need be made. Without these arrangements and with highly imballanced classes, a poor algorithm will simply never predict the outcome of the minority class. In a dataset with 1000 customers containing 5 churners for example, this loss-minimizing algorithm would have an in-sample accuracy of 99.5%.
In order to avoid the high amount of "False-Negative" classifications there are many methods ranging from upsampling the minority class or downsampling the majority class to more advanced techniques. In this work we will present and compare the different methods while applying them to the underlying problem.
We also want to emphasize (or not) the importance of using quotation data for predicting customer churn. A company can track (potential) customer behavior on their distribution channels. Nowadays, in most cases the products or services are offered online on websites, which makes it easy to track website visitor data. In the context of dealing with customer churn this data can be matched to the customers already having a product or contract of this company. We believe (?) that the number of visits of a current customer in the last period (?) plays a big role in predicting the probability of that customer leaving in the next period. (Coming from high correlation between Nvisits and churn)
In order to evaluate the importance of not only the number of website visits but also the other explanatory variables there is typically a trade-off during model selection. The trade-off is between the model complexity or corresponding accuracy and the model interpretability. Deep neural networks or boosted trees belong to the complex models which are famous for their high accuracy in the fields of computer vision and natural language processing. Understanding and interpreting the model is of no big interest in these areas. However in the topic of this work and in many other areas understanding which variables lead to the resulting outcome of the model becomes desirable. The most transparent models in terms of interpretability are linear or logistic models. There the magnitude and sign of the corresponding coefficients (after being testet for significance) illustrate the changes of the outcome for a change in the specific explanatory variable. These models however lack in terms of accuracy when being compared to the comlex ones. In this work we will present the accuracy and interpretability of "Explainable Boosting Machines" developped by (?) for predicting customer churn. It aims to combine the high accuracy of complex models on the one hand and the interpretability of linear models on the other hand.

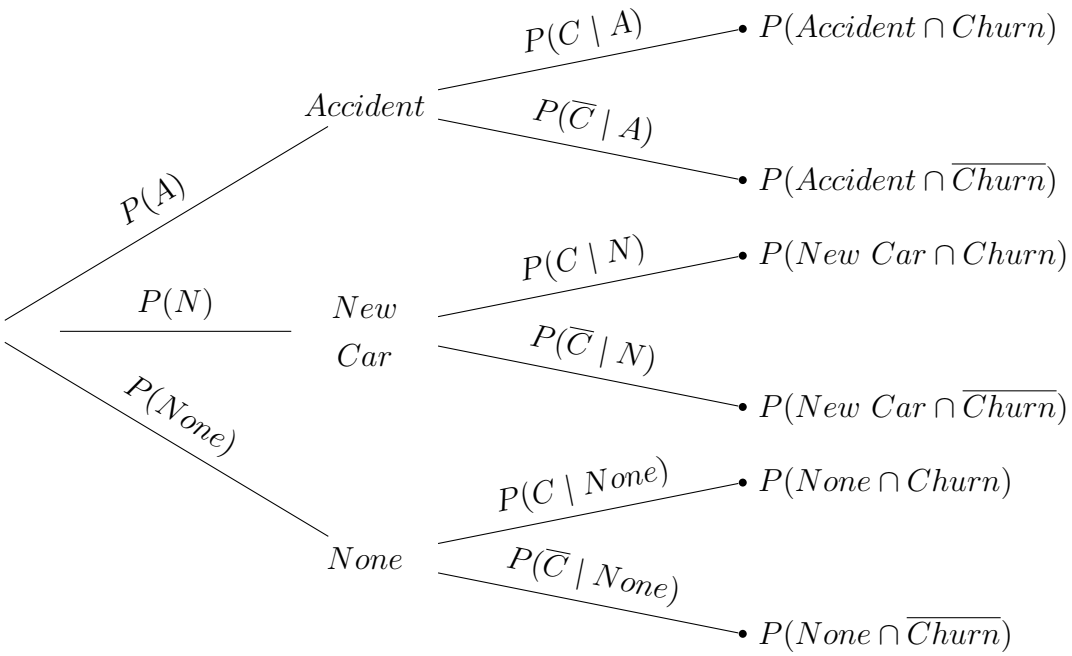
2 Data and Methodology

2.1 Understanding the Problem

For this work we use customer data from a big insurance company in Germany. Due to data pretection the data is anonymized which does not affect the model accuracy and interpretability in any form. We focus on the product of automobile liability insurance, which is by law a mandatory service every car owner must hold in Germany.
Typically car owners close a deal with an insurance company which can be terminated by the end of each year. If the contract does not get terminated it is automatically extended for another year. In rare cases both sides agree on a contract with a due date during the year (We omit this case in this work). Besides the option to terminate the contract at the end of each year there is also an option to terminate it during the year under two circumstances:

Event A: Contractor causes an Accident.
Event N: Contractor buys a new Car.

Therefore the problem of modelling customer churn needs to be seperated into the probability of a costumer leaving during the year and at the end of a year. In this work we will focus on predicting churns occuring during the year. Figure 1 (richtiger Verweis) illustrates how this probability on costumer level can be decomposed using the Events A (Accident), N (New Car) and C (Churn).



By assumption we set $P(C \mid None) = 0$ as the amount of terminated contracts during the year which are not being caused by a new car or an accident is very small and can be

omitted. Therefore we leave these cases out of our data (?). What we are interested in predicting is the overall probability of a churn, which can be rewritten as:

$$P(C) = P(A \cap C) + P(N \cap C) = P(A)P(C \mid A) + P(N)P(C \mid N)$$

Idea: Predict $P(A)$, $P(N)$, $P(C \mid A)$, $P(C \mid N)$ seperately?

Bibliography