Customer Churn Prediction using Quotation Data

Master Thesis

Submitted on: November 29, 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 **1**

2 Data and Methodology **2**

2.1 Understanding the Problem 2

2.2 Modelling Approach 3

2.3 Data Preparation 4

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

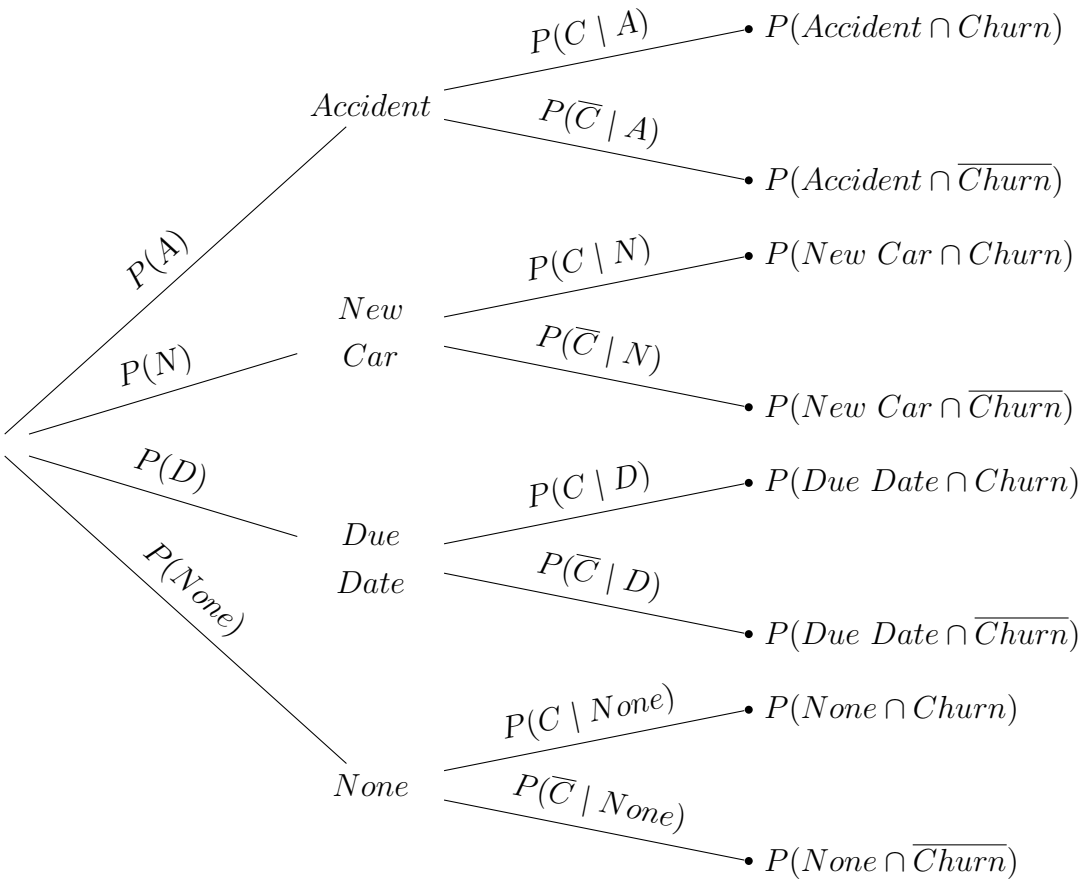
For this work we use customer data from a big insurance company in Germany. Due to data pretection the data is anonymized which does not affect the model accuracy and interpretability in any form. We focus on the product of automobile liability insurance, which is by law a mandatory service every car 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. In rare cases both sides agree on a contract with a due date during the year. If the contract does not get terminated it is automatically extended for another year. Besides the option to terminate the contract at the due date there is also an option to terminate it earlier in a few special cases. These cases mainly involve car accidents and vehicle changes of the contractor. To sum up, here are the three cases in which a churn can (but not must) occur during a year:

Event A:	Contractor is involved in an Accident.
Event N:	Contractor buys a new Car.
Event D:	Due date during the year.

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 (why noch ausführen). In this work we will focus on predicting churns occuring during the year. The purpose is to build a model which can be used at any time t of the year besides on January the 1st to predict the probability of a costumer leaving the company in the next period of time $(t, t + s]$.
It can be argued that in order to provide a model with maximized utility for production one would want to keep s small. For example a company would highly benifit from a model, which can predict the churn-probability of tomorrow or the next week. However we will see that having a small s will decrease the accuracy of our models (drastically?), creating a trade-off situation between model accuracy and the benifits of a small s. With a smaller period the classes of the data become more imballanced, creating a higher challange of preprocessing the training data and feeding the algorithm enough information on potential churners. Furthermore, a small s decreases the scope of action for a company to retain potential customers leaving.

2.2 Modelling Approach

Figure 1 (richtiger Verweis) illustrates how the probability of a churn during the year on can be decomposed using the Events A (Accident), N (New Car), D (Due date during the year) 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, an accident or a due date is very small and can be omitted. Therefore we leave these cases out of our data (?). Also, the probability $P(D)$ can only take values 0 and 1, as either the due date of a customer lies in the next period of time $(t, t+s]$ or not. What we are interested in predicting is the overall probability of a churn, which can be rewritten as:

$$\begin{aligned} P(C) &= P(A \cap C) + P(N \cap C) + P(D \cap C) \\ &= P(A)P(C \mid A) + P(N)P(C \mid N) + P(D)P(C \mid D) \end{aligned}$$

One idea would be to model the three branches of A, N and D seperately. The idea behind this is that different models and sets of explanatory variables have the best fit for the probabilities of the three branches. Not only the the three branches but also the unconditional and conditional probabilities of a single branch may vary in their best modelling approach. For prediciting an accident a model of type (XY) is more suitable, whereas (YZ) would go better with modelling the probability of churn given an accident occured. In the course of this work we will begin with a general modelling approach trying to predict the probability $P(C)$. At a later stage we will compare the accuracy outcomes

of seperately predicting the branch probabilities with the baseline approach.

2.3 Data Preparation

To build the models we use historical data of the insurance company. More specifically we pick (one or many?) timestamp(s) t in the past and collect all the active contracts to that(these) timestamp(s). One row corresponds to one active contract. To each row we merge the status of that contract in $t + s$. Furthermore, we join the number of requests corresponding to that contract in the period $[t - m, t]$. The period length m will be another parameter to tune.
Part of this work will also be to evaluate if the churn probability of customers is time invariant. Therefore we statistically test the equality of monthly and yearly (and weekly?) mean churn rates using the ANOVA (or other?). We will see that it is time variant (or not?). Due to that result it is necessary to include (or not) different timestamps containing all years/months(/weeks) in the data for the model to learn the time differences (or not). (These will simply be appended to the as additional rows creating a panel dataset.)

-Statt "STORNO" im data-load sql query zu definieren, definieren wir es einfach in dem Feature-Engineering Teil? -Test if customer churn probability is time invariant or time variant!!! Is it enough to argue with the Aggregate Churn rate? Show statistically (Voll gute Idee)

Bibliography