ETH zürich

A Churn Model for Swiss Mandatory Health Insurance

An approach from a Pricing Persepective

Master's Thesis

Lena Schütte

Submission Date:

24.01.2022

Supervisor

Prof. Dr. Patrick Cheridito

Department of Mathematics

ETH Zurich

Abstract

In this thesis, we investigate the use of churn models in an actuarial pricing context. We fit logistic regression, classification tree and gradient boosting machine models to a large data set of a swiss health insurer. Here, the actuarial premium is implicitly based on an assumed portfolio structure, which is predicted by the churn model. We therefore develop a pricing loss function that measures the impact of the churn prediction error on the predicted profits and can be seen as a proxy for the error of the actuarial premium resulting from the error in the churn model. Each model's performance is then compared with respect to the Pricing Loss, the Binomial Deviance and the AUC.

As pricing is linked to setting a market premium, we aim to incorporate the impact of the insurer's premium in a competitive market in the churn model. To do this, we include the insurer's premium, premium changes and the premium of the main competitors as explanatory variables. For logistic regression and gradient boosting machine, we then deduce an approximation of the premium sensitivity of the insured.

Acknowledgements

I would like to express my gratitude to Prof. Dr. Patrick Cheridito for accepting to supervise this thesis and to Dr. Sebastian Becker for giving his valuable feedback.
Many thanks also to all involved parties at the insurer for providing the data as well as helpful feedback. I am particularly grateful to Even Meier from Azenes for providing the flexibility and the financial resources to pursue this undertaking. Of course, I also owe thanks to the whole Azenes team for their support and their tolerance with my unavailabilities: Dr. René Dahms, Micha Villing, Peter Matijasic, Adrian von Escher, Pascal Schmid, Stefan Purtschert, Meghna Lakshminarayanan, Aline Schillig and Caroline Ronner.
Finally, my deepest gratitude goes to my family, Richard and Arthur, who sacrificed many weekends, nights and holidays to make my Master studies and this thesis possible.

Contents

Chapter 1

Introduction

Motivation

Customer churn refers to customers deliberately leaving a company, for example due to a better offer from the competition or because they are unsatisfied with the current company. A churn model aims to model this behaviour of the clients by assigning a category (churn/no churn) or a probability of churn to each individual present in a portfolio.
For insurance companies, modeling and predicting churns is relevant for two key business aspects: profit and risk.
Gaining new customers is much more expensive than keeping existing ones [1]. For that reason, it is of interest to identify customers that are likely to leave and to take successful measures to prevent this. This is why churn models have been increasingly popular in marketing and are considered an important part of the Customer Relationship Management (CRM). While this topic receives much attention in marketing, this does not seem to be the case for actuarial models. This neglect seems somewhat unjustified, because churns are not only directly linked to profit, but also to risk. For instance, often, actuarial premium are not purely risk-adjusted in the sense that they reflect exactly the expected claims for each individual in the portfolio. Instead, they include some wanted solidarities between the risks, such as between genders or age groups. Calculation of expected claims of the portfolio (or a "solidarity group") hence implicitly contains assumptions on the structure of this group. Deviations of that structure will then lead to deviations of the expected result, and also has implications on the reserves and the solvency situation due to deviations in the required

risk-bearing capital. Therefore, the implicit assumptions on portfolio structure and volume deserve attention from actuaries. In fact, a full actuarial claims model should include two components: a model to estimate the expected claims per risk (i.e. a claims model) and a model to estimate the expected number of insureds with that respective risk (i.e. a portfolio model). A portfolio model can be understood as the composition of a model for new business and a model for churns. The latter will be adressed in this thesis.
Another aspect of churning behaviour is also important: Often, the premium itself is the decisive factor for customers to leave or to stay in the portfolio. In that sense, not only do churns determine the premium, but also the premium impacts the churns. This means, that premium are also a tool to steer the risk structure and volume of the portfolio (and with that, again, the actuarial result). Modelling the effects of premium changes on the portfolio should therefore be an integral part of the pricing process.

The previous considerations show that churn models are at the intersection of actuarial and marketing purposes, which raises the question whether one can develop a churn model that fulfills the requirements of both. More precisely, it raises the question how to decide what a "good model" is und how its performance can be assessed with regard to its purposes. At first sight, criterions used for models employed in CRM (such as the number of correct guesses of individual churners versus the number of total effective churners) might not be relevant for actuaries, who would rather be interested in a reliable estimator on the whole portfolio-level than on individual policy level.
In this thesis, we aim to develop a model for churns for a Swiss mandatory health insurer. This setting is special for the following reasons:
First, the treatments covered are defined by law, hence the products cannot compete over coverage. The most obvious tool of competition seems to be the price, so we are particularly interested in understanding the consequences of premium setting and expect a strong dependence between churn rates and premiums.
Second, Swiss mandatory health insurance is by law not allowed to make a profit, while at the same time, risk-differentiation of premium is only allowed to a very limited extend. Simply put, we have unitary premiums for adults (per canton) and products should be priced in such a way that they cover exactly the expected claims, with no implicit risk margin. While there are random fluctuations around the mean that can't be controlled nor predicted, changes in the portfolio will also lead to deviatons from the mean, but can to some extent be controlled and (as we try in this thesis) be predicted.

Having said this, it should be added that there is a risk compensation scheme in place ("Risikoausgleich") to balance changes in the risk structure, including adverse selection. However, this scheme does not provide a perfect hedge against structure changes and it itself needs to be estimated, requiring a projection of future portfolio structure. And of course, also from a more general management view, it is of interest to the insurance company to understand and prepare for changes in the portfolio structure and volume, especially when setting the premium.
Third, due to the compulsory nature of the Swiss health insurance scheme, a process between insurers is implemented, which tracks where the insured switch to. When changing the insurer, the insured has to inform the current insurer about his new insurer. This is to ensure that no insured falls out of the mandatory health insurance plan, while, of course, exceptions such as moving abroad are permitted. This procedure results in the unique situation that insurance companies have information on the previous and (in case of churn) the following insurers of their customers, for which they can deduce the corresponding premium. Theoretically, this data allows to set up a churn model that considers not only churn probabilities with respect to changes of the own insurer's premium, but also with respect to the premium of the competition. We should therefore be able to draw helpful conlusions on price elasticity and price sensitivity in a managed competitive environment.

In this thesis, we have the data set of a Swiss health insurer available, containing information on the churn decisions of the insureds over 6 years of observation as well as various attributes of the insured.
We aim to fit different churn models and investigate their suitability for both CRM and pricing purposes.
Our research focuses on the following questions:
1.) What are adequate metrics to assess the perfomance of such a model with resepect to its purposes? Related to this is the question of the choice of the appropriate loss function for optimization.
2.) What are the effects of premium changes on the portfolio? We aim to gain insights on price sensitivity. For instance, we are interested in quantifying the expected effect of premium changes on the expected churn rates.

A practical issue of classification problems like ours is the unbalanced data, meaning that a very large part of the observations is assigned to one class and a very small part to the

other. This affects the informative value of some performance measures and the optimal fitting with respect to some loss functions. One way to tackle this issue would be to adapt the underlying data set (e.g. by over- or undersampling), but we will instead adapt the loss function used for optimizing, where necessary.

Outline

We will adress these two research questions in this thesis by proceeding the following way:

- We start by providing a comprehensive *literature overview*, which finds that a churn model in Swiss mandatory health insurance has not been published yet and that evaluation of model perfomance with respect to pricing risk is not subject of research in the context of insurance churn models at all. Also, individual external premium information was not included in the models reviewed, nor were any conclusions on premium sensitivity published. We therefore believe that this thesis can contribute to explore the potential of churn models from a new perspective.

- In the next chapter, we introduce the *theoretical concepts* underlying the thesis. These are the mathematical churn models and their corresponding loss functions as well as the measures of model performance and premium sensitivity.
  We introduce three logistic regression models with different loss functions: one standard loss function, one to reduce model complexity and one to account for the impact on the actuarial premium using different weights. Then, we fit two classification trees, of which one overcomes the issue of imbalanced data. Finally, a Gradient Boosting Machine is also fitted. We also describe the most popular metrics used to assess model performance and develop two measures to represent the impact on the premium, the absolute and net pricing losses. To measure premium sensitivity, we will calculate the derivative of the churn probability in function of the premium change.

- Finally, the *results* are provided. We compare the different models with respect to the classical AUC and the pricing losses. We find that the Gradient Boosting Machine is more accurate with respect to AUC, but the logistic regression generally leads to smaller pricing losses, especially for the model with a weighted loss function. It also turns out that calculating premium sensitivity is problemtic for the Gradient Boosted Machine, while it is intuitive and straight forward for the logistic regression.

The following paragraph summarizes the data available and the processing steps that were necessary to achieve the implementation of the theoretical concepts.

Data

A quite extensive data set was provided to us by a large Swiss health insurer, consisting of 6 years (2014-2020) of observations. For each year of observation, all policyholders having a mandatory health insurance contract are listed and for each of these a set of features is provided. These attributes range from biometric information, contract information, premium and claims information to CRM-related information. In particular, a categorical variable is provided, which indicates whether or not the insured has churned at the end of the year of observation. It will serve as the dependent variable in our models.
The sheer size and complexity of the data set led to a significant effort in order to prepare the data to make it ready for use. Two major challenges had to be overcome: The first was to introduce new features (competitors' premiums) in order to represent the insurance market situation, the second was to reduce the high dimension of some categorical variables. To deal with the first, the premium of the top 5 competitors was matched for every individual in the data set for every year of observation. To achieve this, internal information was combined with external public information on premiums and the corresponding premiums and their difference to the current insurer's premium were then introduced as new features. Concerning the second challenge, some categorical variables had thousands of levels, which was impractical to use for model fitting. To tackle this, the number of variables was reduced and some categorical levels were merged.

Chapter 2

Literature Overview

We start with an overview of the existing literature that adresses our research questions and related topics. Because the situation of mandatory health insurance is strongly regulated and therefore very specific, we will first look at studies in a similar setting, i.e. mandatory health insurance in Switzerland and comparable ones in other countries. We will then expand the focus to other industries in order to get a complete overview of all approaches that have been applied to model churn behaviour more generally.
When reviewing the scientific work, the following dimensions were of particular interest:

- What models were used?
- How was price sensitivity considered?
- How was model performance measured?

To find the relevant literature, keywords such as Price Elasticity, Premium sensitivity, Retention Models, Attrition Models, Churn Models, Portfolio Change, Switching Behaviour, and (in German) Wechselverhalten, Kündigungsverhalten, Preiselastizität/Sensitivität in combination with the corresponding country name and insurance keywords like Insurance, Health Insurance or Mandatory/Compulsory Health Insurance were used. As we focus on quantitative models, pure segmentation or qualitative analysises of churners were generally not considered.

2.1 Models in Mandatory Health Insurance

We first look at studies in mandatory health insurance in Switzerland and in comparable settings, that is, managed competition as we find it in Germany and the Netherlands.

An overview of some of the empirical studies mentioned hereafter is given in [33].

Switzerland

[2] modelled churn probabilities based on individual data of CSS insurance, a large Swiss health insurer. He used logistic regression and accounted for price sensitivity by using the ratio of CSS-premium and the average premium of the 10 largest competitors as independent variable. Even though the significance of the variables was assessed using the t-statistic, no assessment of the model as a whole was provided. However, a diagram showing the effective versus the expected churn rate-age-curve allows some visual assessment of errors.
Due to a lack of individual data, an aggregated approach was chosen by [7]. Here, the change in market share of mandatory health insurers was used as a proxy for the switching behaviour. As a model, they used univariate and multivariate linear regression, while the latter was modified using the assumption of log-normal distribution of the residuals. Among others, the relative difference between an insurer's premium and the market premium were used as variables as well as the difference between the insurers and the market's relative change of premium. The different models were assessed using the AIC and for some of the models, the adjusted R^2 was also provided.

[46] were interested in the possibilities of modeling switching rates and their effects on the risk structure in the context of the Swiss risk compensation scheme. To this end, they provide an overview of existing studies, which have also explicitly or implicitly been considered in this thesis' literature review, and of the suitability of the existing data. The authors conclude that the data available in the context of the risk compensation scheme is not sufficient to model switching behaviour, but some other sources might have potential for it. They mention the aggegrated data available to the Federal Office of Public Health (FOPH),which only allows to model net changes of number of insureds, and the results of the "Schweizerische Gesundheitsbefragung" (a survey on individual health status), which lacks information on the insurer. They find that data on individual level would be necessary to model switching probabilities, while aggegated data is suitable for determining price elasticity.

[9] aim to investigate the impact of complementary insurance on switching behaviour. They rely on data from a survey on health plan choice from the Federal Office of Social Insurance (OFAS), which contains information on individual level such as subjective health status, switch in the last 4 years, current insurance, but lack info on the previous insurer and hence the previous premium. A bivariate probit model is their first approach, but since the residuals are not correlated, they then concentrate on a simple probit model with several different specifications. Since the premium difference, which they call "monetary gain", is not known, the "expected monetary gain" is used as a proxy. However, it turns out that it cannot be used in the model due to simultaneity issues. Finally, the assessment of the models focuses on whether model assumptions are met and on the significance of the variables. The log-likelihood and (for the bivariate probit model) the p-value of the likelihood ratio test are also given.
[14] use the same logit approach and the same OFAS survey data to assess the impact of the number of health plans to choose from on switching behaviour. Hence, they encounter the same issue with the missing premium information of the previous insurer, so they use the standard deviation as expected gap to the mean plan instead. As before, only the significance of the variables seems to be of interest.

Germany

For Germany, [40] modelled the switching probability using a binomial probit model with a linear regression on the absolute premium, which is given as a percentage of the salary ("Beitragssatz"), and other features. This was done using panel data that provided information on individuals over several years, including their insurance company and socio-economic features. No particular quality assessment of the different proposed models was done, only the log likelihood of each model was provided.
[36] use the same data and follow the same probit model. However, other explanatory variables are used, such as a dummy variable for premium increase, the premium increase in absolute values and the premium increase as percentage. Furthermore, they introduce a time-component by distinguishing between different (time-dependent) regimes. In addition to modeling individual switching probabilities, they also include a model on aggregated data by using market share and the absolute number of insured as a dependent variable. As before, the focus lies on assessing significance of the variables, hence no comparison or quality assessment of the models is being performed. [39] use panel data of sickness funds in

Germany, on which they apply McFaddens random utility model, i.e. each fund's utility is assumed to be a linear regression on contribution rate and other factors, yielding a ln-linear function of the market share of each fund relative to other funds. This means, that the whole market is modelled, while the utility function is assumed to be the same for every individual. However, some groups of insureds are distinguished (e.g. self-employed) by introducing dummy variables in the market-share-function and time effects were also considered. As before, only significance of variables was tested while no specific model assessment was performed.

[45] use a similar model appraoch, but augment it by a lagged endogenous variable (the market share) in order to introduce market share persistency. First-differenced versions of this model are also considered, along with the basic static models. Besides testing the significance of each variable, these models are also compared via appropriate statistics (Sargan-Statistic for validaty of restrictions of GMM-estimated dynamic models, F-Tests and t-statistics) as well as the R^2.

[32] aim to detect risk selection effects and rely datawise on the same socio-economic panel data (i.e. of individuals over several years) as previously mentioned. In order to separate health effects from other impacts on the switching probability, they use a recursive two-equation model, which consists of an ordered probit model for a continous health index, and a multinomial logit-model for the switching probability. The health index is one of the independent variables of the latter model. However, premium or premium differences are not among the other explanatory variables considered and hence price-effects are not included in the model. A particularity of this model is the fact that it distinguishes between two types of insuerers ("BKK" and "non-BKK") and therefore there are three different outcomes for switching decisions: stay, switch to another "BKK" and switch to a "non-BKK". Some indicators for model performance (log likelihood, pseudo R^2 and LR chi squared) are provided for each model, which could serve for a comparison of the models.

Netherlands

Also in the Netherlands, estimating price elasticities of the mandatory health funds has been an area of research. [38] are interested in finding explanatory factors for premium and market share differences between Dutch sickness funds. To test the significance of the explanatory variables, such as premium, they fit a ln-regression model for the market share as dependent variable. As longitunal data is used, they use pooled Ordinary Least Squares (OLS) and Fixed Effects (FE) estimators to fit the model and compare these models via the

F-statistic.

[10] use changes of market shares instead of the absolute level of market share as dependent variable, following the suggestions of [45], while the premium is still taken as the level variable and the ranking of the premium in the market is also included. They suggest a model that considers the whole market, i.e. estimate bilateral price elasticities between insurers. For comparison, different linear regression models (including conditional logit and one with lagged market shares) are set up as well. As a criterion for model selection, the R^2 and the log-likelihood are provided.

[6] estimate price sensitivity using panel data, where they measure the loss in market share as a function of an unilateral increase of the premium. To do this, the bilateral flows of insureds between the insurance companies are modelled, this is done for different subgroups of the population. Among other variables, the premium differences between competing insurers is explicitly considered. The resulting models are assessed by the F-Statistic and the adjusted R^2 is also provided. Even though they take market competition into account to some extent, it needs to be noted that both studies only consider market shares, hence only consider net flows of insureds.

In [35] the balanced data of a Dutch health insurer is used to study the staying power of logit and classification tree models, that is, the validation of those models over time. Various variables were used, but premium was not among them. The top decile lift and the Gini coefficient were used for assessing the model performance over time.

More recently, customer churn in the Dutch health insurance has been the subject of two Master theses. [23] models churning behaviour of insured of a Dutch health insurer in a comprehensive way. In this regard, her setting and research question is very similar to ours. The thesis is based on extensive data about the insured and churners of the company, including information on their premium, but not those of the competition. As models, Log -Regression, Decision Trees, Artificial Neural Networks and Support Vector Machines are applied. In addition, profiling techniques such as k-means and self-organizing maps are explored. It is also to be mentioned, that the issue of imbalanced data, which is to be expected in all churning classification problems similar to ours, is explicitily adressed and treated by balancing the data sets. To compare the models, AUK, AUC, precision and sensitivity are used as measures. Futhermore, some robustness checks are also performed, like cross validation over time. Finally, a theoretical cost-benefit analysis of the suggested models is provided, which we consider an important component when assessing the practical use of the models.

In the second thesis, [25] categorizes the variables of a churn model for a Dutch health insurer according to the factors used for human migration modeling. Nevertheless, the variables finally used in the model are similar to the ones seen in the other studies, including premium. The models applied were logistic regression and latent class regression analysis, which consists of a segmentation of the portfolio. As a measure for model performance, the hit rate, the top decile lift and the Gini coefficient, together with the lift curve, and the mean squared error were provided.

Conclusion: Models in Mandatory Health Insurance

When reviewing scientific publications treating the modeling of switching behaviour of the insureds in the context of mandatory health insurance, it turns out that the main field of interest seems to be from a macroeconomic perspective. In particular, research focuses on the question of whether or not market competition works effectively in mandatory health insurance markets. This is due to the fact that in all these countries, reforms were undertaken in the past to induce the shift from a state-governed system to a managed competition market, where various insurers could compete. Assessing whether or not these reforms were effective is therefore an important question. In most papers that we found, market dynamics were measured by market shares and price elasticities of demand. Assessing the impact of other factors than the premium on switching decisions is also often a point of interest. In these studies, regression, linear and non-linear, seems to be the standard approach, sometimes different regression models are combined (e.g. two equation models, bilateral flow models). Usually, these are then used for hypothesis testing and often, model assessment is limited to providing classical indicators such as some test statistics (e.g. t-test for variables, F-Test for models, likelihood ratios), the R^2, the log-likelihood and, in one case, the AIC.
Most of these studies rely on aggregated data (which is sufficient for their research purpose), some rely on individual panel data. It is noteable that all of these models have an aggregated view on switching behaviour, meaning that only net changes of insureds in the portfolio are modelled. However, as seen in the introduction, it is important for an insurance company to distinguish between new and remaining costumers from a financial perspective. Hence, we consider these models to not give sufficient insight on the churning behaviour for a practical use.
Research and publications of churn models based on individual data which distinguish between churners and new business seem to be rare. In Switzerland, only one study was found [2], which uses logistic regression on an insurers portfolio. In the Netherlands, three studies

were using data from insurers on policy level, one using logistic regression, one using logistic Regression and classification Trees and one using logistic Regression, classification trees, artifical neural networks and Support Vector Machines. This latter study comes closest to what we are trying to achieve in this thesis. However, in none of these studies, the premium of the competition and, in particular, the premium after churning, was taken into account as explanatory variable. To us, considering this variable promises to improve the (prediction) power of the model and to gain valuable insight in the effect of premium setting on the churn behaviour of the insured in a competititive market.

In contrary to the econometric models, these churn models do solve a classification problem. When assessing model performance, metrics based on confusion matrices are therefore suitable. In the models under review, the top decile lift and the Gini coefficient were most used, but also the hit rate, the lift curve in general, the MSE and metrics like the AUK, AUC, precision and sensitivity were considered. As some of these models were more advanced than regression, cross-validation over time was also performed.

Note that this second stream of research was set in the context of CRM. It is therefore remarkable that none of the literature found treats the churning topic from an actuarial perspective, where the risk structure of the portfolio would be of interest.

Having said all this, these results do not mean that such models do not exist. It is possible that insurance companies develop them internally and don't publish the results because they are a competitive advantage.

However, to our knowledge, the information available to Swiss health insurers, concerning the past and future insurer of individual switchers is quite unique. These companies would therefore be predestined to setting up a model with detailed information on premium of the competition as we are trying to do, but it does not seem that any such model has been developed in this market yet [31].

2.2 Models in General Insurance and Other Industries

While the publications on churn modeling of mandatory health insurance seem to be quite limited, a look at the broader insurance industry promises more results. Within the insurance industry, car insurance is the most important field of application, so we will have a dedicated section for it. We will then expand to applications of churn models in the rest of the insurance industry. Finally, an overview of churn modeling applied in other industries (namely telecommunications) will be given in the last section.

Models in Car Insurance

[50] gives an overview of the literature on churn modeling in the financial industry until 2002. However, it doesn't mention any application in the insurance industry and our own literature research didn't reveal anything else, which is why we concluded that no substantial research was published until this date. One exception seems to exist though: Around 2001, a series of papers were published, that investigated data science techniques for churn modeling in the context of motor insurance and even proposesd a framework for including these models in the pricing exercise. More precisely, in [42] churn behaviour is modelled via logistic regression, decision trees and neural networks. Premium and premium difference are included as variables. Here, the lift chart and the accuracy are used as measures for the performance of the models. In [53], a k-means algorithm is applied on the same data to find clusters of claims (risk groups), on which then a neural network is applied for churn prediction. Here, premium level, absolute and relative premium change as well as the premium relative to the insured sum are included as variables; however, the competition's premium is not. Notably, as performance on the whole portfolio is deemed more relevant than for each individual predicition, the measure used is the average predicted churn rate per cluster versus the true rate per cluster. The resulting model is then used to optimize the premium to achieve maximal profit under risk minimization constraints. [52] gives more details on the neural network model applied in the previously mentioned study. For instance, it gives the resulting accuracy of the model in function of the premium change, showing that accuracy is particularly low in a certain range of premium changes.

Two very interesting aspects of applying churn models in the insurance context were adressed in [18] and [19]. In the first paper, it was suggested that the ultimate purpose of such a churn model was to identify those customers that were most likely to respond positively to CRM measures instead of only identifying those most likely to switch. As a consequence, an uplift model was proposed, which measures the change in probability of churn following a CRM intervention. To achieve this, a Random Forest model was used, were (among others) premium and premium change were included as variables. The prediction error was measured, in accordance with the model's purpose, as the difference of model uplifts for different top p percents. In [19], another important issue was pointed out: the fact, that actuarial (risk) premiums are deterministic functions of the covariates which are usually considered in the churn models. This means, that conclusions about the effect of premium changes on the churn rate are not reliable, because they get confused with the effect of the covariates. In order to tackle this issue und allow inference of causality between churn rates and premium

(changes), the authors apply Rubin's causal inference model in a car insurance context. As a side product, price elasticities on individual levels are also provided. Eventhough this topic seems generelly noteworthy, it is not deemed relevant in our setting. As mentioned in the introduction of this thesis, solidarities can be strong within a portfolio, and in our case we have, simply put, a unitary premium. So, in this sense, they are not actuarial risk premiums and are linked to the other covariates only in a very limited way.

Another study, similar to the ones presented previously, is [20], where data from a non-life insurer with different lines of business (Car, Home, Health) is used. Before fitting it to a logit churn model, the authors fit the data to a GAM in order to identify non-linear structures and transform the variables (which include premium and time-dynamic variables) accordingly. Prediction performance is assessed by out of sample and out of time ROC curves. In addition, the True Negative Rate of the top 1000 probabilities is also assessed.

In [29], the authors propose a model and an algorithm to optimize the expected value of the insurance portfolio. To achieve this, the portfolio is divided in clusters with respect to their future value, which is defined as the expected discounted loss ratio, taking into account a survival function. On these clusters, a GAM-logit approach is chosen to model the demand function, taking into account the premium, among other factors. No performance assessment of the resulting model is provided.

[43] applies various different models on churn data in car insurance: logistic regression, decision trees, naive Bayes, Random Forests, Support Vector Machines and survival analysis. Premium is considered as one of the features. These models are compared by their accuracy; for the winning neural network, the confusion matrix is also provided.

Similarly, [4] use logistic regression, decision trees, artificial neural networks and support vector machines to model churn probabilities in car insurance. Interestingly, three different measures are used to select the best model: the sum of sensitivity and specificity, the accuracy and the AUC and it is concluded that the selection of models depends on the preferences of the insurance company. In this sense, that study's topic is related to the focus of our study. It is to be mentioned that premium, change of premium and discounts are included as explanatory variables. In 2019, two Master theses were also concerned with churn prediction for car insurance. [3] fitted a GLM, Random Forest and artifical neural network for a car insurance, using premium of the current, previous and following year. As performance metric, the accuracy, AUS, sensitivity, precision, F-Score and Kappa were used. Finally, also the average error between predicted probability und churn rates was provided.

[44] applied logistic regression, Random Forest, Support Vector Machine, AdaBoost with

decision trees , k-nearest neighbors and neural networks to traffic and personal insurance. For each technique the best model was chosen by comparing AUC, F-score and accuracy. The ROC curves were also analysed. The winning models were then compared by their MSE. One of the considered features was the price difference to the market, however no further details on how this was determined were provided.
[26] estimates a demand function for car insurance within the framework of price optimization . For this, he uses a GLM/logistic regression approach. As price elasticity plays a major role here, a so-called price test is conducted in order to generate sufficient data for the model fitting. Here, premiums that differ (i.e. as increase or decrease) to the standard premium are proposed to a proportion of insured and their churn decisions are observed. However, this only reflects the effect of premium changes within the same insurer; the model does not include the competitors' premiums. The goodness of fit is assessed visually, by displaying the actual versus the expected premium (including confidence intervals).

Models in Other Insurance

[30] focuses on feature selection and dimension reduction in the context of time-stamped life insurance data. To predict lapses, decision trees, Support Vector Machines as well as the Apriori Algorithm and naive Bayes estimators were used. Model performance was assessed via the accuracy, precision, recall and the F-measure.
A very interesting approach is presented in [54]. Here, a logistic regression model (called 'shallow model') and an artifical neural network (called 'deep model') are combined. The two constituting models are trained simultanously via a common loss function applied to the sum of their outputs. One of the many factors considered is premium. As a metric for model assessment, precision, accuracy, recall, F_1 and AUC are considered.
[17] adress the issue of right-censoring, i.e. incomplete observations in longterm insurance data, by weightening the observations appropriately. To do this, survival functions need to be estimated, for which three different approaches (Kaplan-Meier, Cox-Model, Random Survival Trees) are used. These weighted observations are then used to fit a Random Forest model, which estimates the expected future value of the policyholders. As assessment criterion, accuracy can't be used due to right-censoring, so an 'adapted' (weighted) MSE is applied. Here, again, the weights are determined in accordance with the estimation approaches used for the the respective survival functions. For information, the C-index (i.e. the proportion of ordered pairs of the test set well-ordered by the model) is also provided. Note that the mentioned techniques are tested on synthetic data as well as on real sup-

plementary health insurance data. Premium information does not appear to be taken into account for the estimation of the churn rates, however it is considered when calculating the prospective value function.

[22] were interested in the effect on churning behaviour of first cancellations in a multi-product (life and non-life) environment. To model this, a survival function was estimated via a Cox-Model (Weibull Regression) approach. Subscription dummy variables of the 11 products were among the covariates, while their respective premium were not. The model performance was assessed by a likelihood ratio test.

[37] focus on the issue of missing data in the context of customer lifetime value calculation, where, for instance, acquisition and retention costs need to be considered. After the appropriate data treatment, quite a few methods were tested: Bernoulli and multinomial naive Bayes, Support Vector Machines, decision trees and artificial neural networks. For these models, accuracy was determined on a test set. Finally, the best models were compared using the following metrics: Accuracy, Precision, Recall, Specificity, and F_1-Score.

[8] focus on the problem of inbalanced data and the optimization of sampling techniques. The application is set in a life and non-life insurance context, but as premium are not considered, no conclusions on price elasticity can be drawn. The sampling method has consequences on the likelihood function used for estimating the logit-models in use (so called pseudo-likelihood) and hence on the confidence of the model estimates. As sampling methods are the focus, an efficiency measure for sampling methods is used for performance assessment.

Models in Other Industries

Churn prediction is also an important topic in other industries, notably telecommunications and online gaming. This is, because these are very dynamic and competitive markets where customer retention is crucial in order to create revenue. Consequently, an extensive literature is available, which will be summarized in this section.

[48] gives an overview on recent developments of churn modeling. It has an emphasis on telecommonuications, but also mentions the banking, energy and online social networks sector. Besides the most popular techniques such as decision trees, Support Vector Machines and logistic regression, it presents the Logit Lead Method (LLM), Bayesian Networks, algorithms based on rough set theory and genetic programming as other approaches. However, the review does not provide much forther detail than that.

[24] focuses on the main industry of application, telecommunications, and gives a broad overview of the models and methods applied, as well as on features and performance mea-

sures. It completes the list of models that we have encountered in the previous sections by two areas of research: hybrid models and classification. As hyprid models, the combination of neural networks (in this paper, they are often referred to as Perception Models, which is a subset of feed forward artificial networks) and a logistic regression are mentioned, as well as, again, the LLM. For classification, more details on algorithms and estimators are provided, which are: C5.0, K2, Fuzzy Classifiers, genetic algorithms, covering algorithms and LEM2 algorithm. Concerning random forests, different boosting methods were also listed. As metrics to measure model performance, no new indicators other than the ones already presented in this literature overview were found. For instance, indicators deducted from the conversion matrix were quite popular. In terms of features, attributes related to activity are most widely used, but call prices, (total) fees paid or charged were also considered. Note that these values correspond to the consumption of the service, not the tariff itself. Eventhough comparison is complicated, they would, in an insurance context, be more similar to claims costs than to premium. In that sense, they are also rather an indicator for activity, and not adequate to model price elasticity or sensitivity.

Conclusion: Models in General Insurance and Other Industries

Within the insurance industry, car insurance has been the most popular field of application for churn models. This is presumably because of the availability of suitable data bases and the relevance for business due to strong competition and high switching rates.
Besides that, the modelling churn (or: lapse) rates has also been a topic for life insurance or insurance similar to life insurance (like health insurance with long-term contracts). Here, right-censoring due to incomplete data is an issue that needs to be adressed.
Opposite to the previously considered mandatory health insurance, the premiums of these 'supplementary' insurances are generally allowed to be risk-adequate in the sense that they reflect the expected claims, which depend on characteristics of the insured. Also, competition is strong in these markets and insurers have more freedom when setting the market premiums, which is why premium optimization seems to play a bigger role. Several papers dealt with the embedding of the churn model in a pricing framework. Via demand functions and information on premium and claims, values such as the expected lifetime value (for a long term cover) and premium income could be modelled and optimized with respect to the premium. However, the risk-adequacy leads to an issue with interpretation of the effect of premium change. Since the premium (changes) are a deterministic function of the covariates of the churn rate, causal inference of premium changes and churn rate changes is not easy.

This problem was adressed in two papers, one solved it by a hypothetical premium change survey (price test), one applied Rubin's causal inference model.
In one case, premium sensitivity was assessed by setting up an uplift model such that changes in distribution could be measures.
Most of the models' performances were assessed by the metrics related to confusion tables as we have seen before and in one case, the accuracy was given as a function of premium change. But the true vs. expected churn rates (per 'cluster', which can be considered a tarif class), and the actual vs. expected premium were also considered. However, as premiums are considered to be risk-adequate, there is no pricing risk associated with structure changes of the portfolio, which explains why no corresponding measures were developed in this context.
It should also be noted, that no external premium information was considered as direct comparison in general insurance is not easy.
Also, there have been approaches to use the churn model to classify the insured in terms of expected value. To the set of the methods encountered in the previous chapter on mandatory health insurance add a few more, both supervised and unsupervised: Naive Bayes, AdaBoost, survival analysis and k-nearest Neighbor.
The variety in models expands even more, when considering other industries such as telecommunications. Algorithms based on rough set theory, genetic programming and hybrid models (a combination of neural networks and logistic regression), different specifications of algorithms like C5.0, K2, Fuzzy Classifiers and LLM are among them.

Chapter 3

Theoretical Concepts

3.1 Logistic Regression

We begin with a model that is widely used for standard decision problems like ours, the logistic regression. We start with presenting the model and then fit its coefficients with respect to different loss functions: the 'standard' Binomial Deviance loss, which yields the MLE-estimator, weighted loss functions, which yield versions of the Weighted-MLE-estimator and a penalized loss function, yielding the Penalized MLE-estimator.

3.1.1 Logit Model

In our first model we assume that the observation of the event "churn" of individual i is the realization of a random variable C_i, more concretely, we assume that C_i is Bernoulli distributed with success probability p_i:

$$\mathbb{P}[C_i = 1] = p_i \text{ and } \mathbb{P}[C_i = 0] = 1 - p_i \text{ for } p_i \in (0, 1) \tag{3.1}$$

This can be rewritten as:

$$\mathbb{P}[C_i = c_i] = p_i^{c_i}(1 - p_i)^{1-c_i} \text{ for } c_i \in \{0; 1\} \tag{3.2}$$

For N mutually independent C_i, the probability to observe the realizations c_i is then given by

$$P[C_1 = c_1, ..., C_N = c_N] = \prod_{i=1}^{N} p_i^{c_i}(1 - p_i)^{1-c_i} \tag{3.3}$$

Let X_j be the set of all possible expressions of a feature j and X $\subseteq X_1 \times \ldots \times X_q$ the set of all possible combinations of q features. We call X the *feature space* and an element $\mathbf{x} =$ $(1,x_1,...,x_q) \in 1\times$ X an *expression of features.*

Let $\mathbf{x}_i = (1,x_1^i,...,x_q^i)$ be the expression of features for the individual policyholder i.

We now assume that the individual churn probability p_i can be expressed as a function of the linear combination of the individual's features.

We define the *regression function*

$$\eta_\beta : X \to \mathbb{R}, \mathrm{x} \longmapsto \beta_0 + \beta_1 x_1 + ... + \beta_q x_q(x) := \langle \beta, x \rangle \text{ for } \beta = (\beta_0, .., \beta_q) \in \mathbb{R}^{q+1} \tag{3.4}$$

and set the *response function* as the logistic function:

$$F : \mathbb{R} \to (0,1), \eta_\beta \longmapsto \frac{\exp(\eta_\beta)}{1+\exp(\eta_\beta)} \tag{3.5}$$

Combining these then gives our logistic regression model

$$p_i := F(\eta_\beta(\mathbf{x}_i)) \text{ for a fixed } \beta \in \mathbb{R}^{q+1} \tag{3.6}$$

3.1.2 MLE-estimator

Model

We now need to find an estimator for β to specify the model, based on the observations O= $(O_i)_{i=1,..,N}$= $(\mathbf{x}_i, c_i)_{i=1,..,N} \subseteq 1 \times$ X $\times\{0,1\}$. A standard approach to do this is by maximizing the probability to observe O.

From equations(3.3) and (3.6) we get:

$$\mathbb{P}[C_1 = c_1, ...C_N = c_n \mid \beta, \mathbf{x}_1, .., \mathbf{x}_1] = \prod_{i=1}^{N} (\frac{e^{\langle \beta, x \rangle}}{1+e^{\langle \beta, x \rangle}})^{c_i} (1 - \frac{e\langle \beta, x \rangle}{1+e^{\langle \beta, x \rangle}})^{1-c_i} \tag{3.7}$$

Note that argmax(f)= argmax (log f). So, maximizing (3.7) with respect to β is equivalent to maximizing the log-likelihood:

$$\begin{aligned} l_O(\beta) :&= \log\left(\prod_{i=1}^{N}\left(\frac{e^{\langle\beta,x_i\rangle}}{1+e^{\langle\beta,x_i\rangle}}\right)^{c_i}\left(1-\frac{e\langle\beta,x_i\rangle}{1+e^{\langle\beta,x_i\rangle}}\right)^{1-c_i}\right) \\ &= \sum_{i=1}^{N} c_i \log\left(\frac{e^{\langle\beta,x_i\rangle}}{1+e^{\langle\beta,x_i\rangle}}\right) + (1-c_i)\log\left(1-\frac{e^{\langle\beta,x_i\rangle}}{1+e^{\langle\beta,x_i\rangle}}\right) \\ &= \sum_{i=1}^{N} c_i \log\left(e^{\langle\beta,x_i\rangle}\right) - c_i\log\left(1+e^{\beta,x_i\rangle}\right) + (1-c_i)\log\left(\frac{1}{1+e^{\langle\beta,x_i\rangle}}\right) \\ &= \sum_{i=1}^{N} c_i\langle\beta,x_i\rangle - c_i\log(1+e^{\langle\beta,x_i\rangle}) - (1-c_i)\log(1+e^{\langle\beta,x_i\rangle}) \end{aligned}$$

So, we have

$$l_O(\beta) = \sum_{i=1}^{N} c_i\langle\beta,x_i\rangle - \log(1+e^{\langle\beta,x_i\rangle}) \tag{3.8}$$

We now set the estimator for β:

$$\hat{\beta} := \hat{\beta}_{MLE} = \arg\max_{\beta} l_O(\beta) \tag{3.9}$$

and call $\hat{\beta}$ the *Maximum-Likelihood-Estimator*.

As l_O is differentiable in β, we can determine the maximum by calculating the partial derivatatives and setting: $\frac{\partial}{\partial\beta_l} l_O(\beta)\big|_{\hat{\beta}} = 0$ for all l=0,..,q. This gives:

$$\sum_{i=1}^{N}\left(c_i - \frac{e^{\langle\hat{\beta},x_i\rangle}}{1+e^{\langle\hat{\beta},x_i\rangle}}\right)x_l^i = 0 \text{ for l=0,..,q} \tag{3.10}$$

Remark: For the homogenous case, that is, if $\beta_1 = \ = \beta_q = 0$, we have $p_i = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ for all i = 1,..,N . This means that the churn probability is independent of the features of the individuals, and we have:

$$\frac{\partial}{\partial\beta_0} l_O(\beta) = \sum_{i=1}^{N}\left(c_i - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right) = 0 \Leftrightarrow \beta_0 = \log\left(\frac{-\sum_{i=1}^{N} c_i}{\sum_{i=1}^{N} c_i - N}\right)$$

However, in general, equation (3.10) does not have a closed form solution, but it does have a unique solution if the design matrix $(X_{i,j}) = (x_j^i)_{i=1,..,n,j=0,..,q}$ has full rank).
Finally, we observe that our MLE- estimator is the result of the minimization of the (nega-

tive) Binomial Deviance for:

Definition 1. The *Binomial Deviance Loss Function* is given by:

$$L_{Bin}(p) = -\sum_{i=1}^{N} c_i \log p_i + (1 - c_i) \log(1 - p_i),$$

and p is of the form defined in equation (4.6).

L_{Bin} is defined for a specific data set O, but we will omit this index for simplicity. Depending on the context, L_{Bin} will in thesis be calculated on the training data set for fitting purposes or on the test data set for validation purposes. The same applies for all other loss functions that will be introduced.

Result

Before we can fit the logistic regression on our sample, we need to reduce the number of categories for some of the variables. This is due to the fact that categorical variables are coded by dummy variables in the model, meaning that each categorical level is modelled as a separeate (dichotomous) variable. Since we have many categorical features in our data set with up to 6'880 levels each, we need to reduce the dimension of the features space in order to be able to fit the model in reasonable time. A detailed description of how this is done, is provided in the Appendix.
We then first naively run the logit regression model on all remaining input variables using the R-package `glm`. We call the resulting model Logit_Naive. In order to reduce complexity of this model, we perform a variable selection on the basis of significance for continuous variables. We also consult the ANOVA to consider categorical variables and analyse their importance for loss reduction.
It turns out, that the premium differences (the variables that were constructed especially for this analysis using external public data), were among the most important variables, they both were highly significant and had a high expected impact on the expected churn probability.
The logit model fitted on these manually selected variables is called Logit_Selected.
More concrete results are provided in Sections 5.2.1 and 5.2.2 of the Appendix.

3.1.3 Weighted MLE-estimator

Model

In a next step, we want to take into consideration the effect of deviance of modelled and effective probabilities on the technical result. To this end, we assume that the technical result per insured in one year of observation remains the same in the next year. The technical result is given as a feature in our data set, it is a number calculated for each insured for each year of observation. Of course, number and size of claims (among other components of the technical result) are also random variables, so if one wanted to predict next year's result, the result per person would also need to be modelled. However, here, we are interested in isolating the effect of the churn model. We do this by introducing weights to the (negative) binomial deviance loss functions that reflect the effect on the (expected) technical result.

Definition 2. The *Weighted Binomial Deviance Loss Function* is given by:

$$L_{Bin,w}(p) = -\sum_{i=1}^{N} w_i c_i \log p_i + w_i(1-c_i)\log(1-p_i),$$

where the weights are defined as

$$\begin{aligned} a)\ w_i^a &= \frac{|R_i|}{\sum_i |R_i|}, \\ b)\ w_i^b &= \frac{R_i - R_{i,min}}{\sum_i (R_i - R_{i,min})}, \\ c)\ w_i^c &= \frac{R_{i,max} - R_i}{\sum_i (R_{i,max} - R_i)} \end{aligned}$$

for R_i, the technical result of an observation of policy i, and

$$R_{i,min} = \min_i R_i,$$

$$R_{i,max} = \max_i R_i,$$

and p is of the form defined in equation (3.6).

The first weight w_i^a looks at the absolute result. More extreme losses or profits are weighted more strongly. On the other hand, policies that create no loss nor profit (R_i=0) will not be considered in the loss function at all. This follows the intuition, that we are indifferent whether or not we predict their churn probability correctly as their churn decision has no financial consequences for the insurer.

The two other definitions of the weights are centered and normalized" versions of the technical result. Here, we do distinguish between positive and negative contributions (i.e. proftable and unprofitable individuals).Centering the variables is done in order to ensure positivity of the weights, which is a requirement of easy implementation in R. The normalization reflects the proportion of each result on the total (centered) result; however the minimum of the loss function and hence the resulting estimator is not affected by this since it just corresponds to a (positive) factor on the loss function.

Remark: As the total technical result is additive, i.e. the R_i affect it in a linear way, it seems that a definition of the weights that is linear in R_i is adequate. However, if we are interested in considering other effects, such as, for example, the *value* of the result to the company or the full financial consequences of the churning decision of an individual, other (possibly non-linear) definitions of the weights could be more suitable. For example, utility functions could be applied on R_i or the costs in terms of required risk capital (via, for instance, a solvency measure) for an individual i could be calculated.

It is easy to verify that the equation characterizing the solution β to the optimization problem of maximizing $L_{Bin,w}(p)$, which is equivalent to equation (3.10) is given by:

$$\sum_{i=1}^{N} \left(c_i - \frac{e^{\langle \hat{\beta}, x_i \rangle}}{1 + e^{\langle \hat{\beta}, x_i \rangle}} \right) w_i x_l^i = 0 \text{ for l=0,..,q} \tag{3.11}$$

More details on the technical result can be found in Section 3.4.2.

Results

We fit the training data to a logit regression model with the same selected variables as for the model Logit_Selected, but this time with respect to a weighted binomial loss function for different weights. The resulting models are called Logit_Selected_Weighted_a, Logit_Selected_Weighted_b and Logit_Selected_Weighted_c.

3.1.4 Penalized MLE-estimator

Model

As described in Section 3.1.2, the naive logit model used all input variables, including features with a high number of levels. In order to reduce its complexity, variables for the model Logit_Selcted were manually selected. However, the selection was performed

on a variable level, not on a categorical level. Theoretically, this could have lead to the exclusion of variables that might have contained important explanatory information in *some* of their categorical levels. It is therefore reasonable to ask if there is a better way to reduce the number of input variables, including dummy variables (more precisely: the number of parameters to be fitted in the model) than manual selection.
The idea in this section is to manage the parameter-selection directly via the loss function. This will be done by introducing a penalty on the parameters, which is controlled by a *tuning parameter* λ.
Let $\|\beta\|_1 = \sum_{i=0}^{q} \mid \beta_i \mid$ be the L_1-Norm of the coefficient vector β.
We then modify the loss function the following way:

Definition 3. For a tuning parameter $\lambda > 0$, the L_1*-penalized Binomial Deviance Loss* is defined as

$$L_{Lasso,\lambda}(p) = -\left[\sum_{i=1}^{N} c_i \log p_i + (1 - c_i) \log((1 - p_i)\right] + \lambda \cdot \|\beta\|_1$$

with p defined as in equation (3.6).

As before, the coefficients β will be fitted such that $L_{Lasso,\lambda}(p)$ is minimized. But this time, a (positive) penalty term $\lambda \cdot \|\beta\|_1$ is added to the (negative) Binomial Deviance Loss. With increasing λ, this will shrink the coefficient estimates towards zero and eventually, when λ is sufficiently large, set them to zero. Since every dummy variable has its own coefficient, the selection happens on a categorical level.

Results

We fit the logistic regression model with respect to the L_1-penalized Binomial Loss function, using the same data as we used for fitting Logit_Naive. We choose the tuning parameter such, that the resulting number of coefficients (or, respectively, variables) is as close to the one of Logit_Selected.
Again, premiums of the current insurer and the competition are among the most important variables. In this model, selected levels appear as (dummy) variables, but not all levels. So, the penalized loss function indeed selected different (dummy) level variables than Logit_Select and we call the resulting model Logit_Lasso.
More details on this model can be found in the Appendix, Section 5.2.3.

3.2 Classification Tree

3.2.1 Classification Tree Model

We now introduce the classification tree model, which is different from the logit model in two key aspects: First, we do not assume a fixed structure for the churn probabilities (unlike the one given in equation (3.6)), and second, we do not directly model the churn *probability*, but the churn *decision*. The churn probability is then deduced from our model.
We begin by presenting the structure and building of the classification tree, where we transfer the considerations of [51] to our specific application. The foundations and thourough explanations of classification trees (including all aspects adressed in the following subsections of this section) are set out in [5]. Where applicable, we use the same notation as introduced in Section 3.1.
We call {0;1} the classes of the event "churn", where 1 stands for the occurence of a churn, and c=0 for no occurance of a churn. This is in line with (3.1), where C_i is the random churn variable of observation i.
We define the Classifier $\mathcal{C}$ on a subset of the feature space $\mathcal{X}_t \subseteq$ X :

$$\mathcal{C} : \mathcal{X}_t \to \{0;1\}, \mathbf{x} \longmapsto \mathcal{C}(\mathbf{x}) \tag{3.12}$$

$\mathcal{C}$ assigns a churn decision to each expression of features. More precisely, we assume that the estimated class of $\mathbf{x}$, given an estimator $\hat{p}$ of its churn probability, is assigned the following way.

$$\hat{\mathcal{C}}_b(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{p}(\mathbf{x}) \geq b \\ 0 & \text{if } \hat{p} < b \end{cases} \tag{3.13}$$

So, if the estimated churn probability in a subset is larger than a certain threshold b, we assign the event "churn" (i.e. 1) to all observations in that subset.
In order to calculate $\hat{p}$, we will proceed the following way. We start by defining the *empirical probabilities*, which are deduced from a sample of N pairs of observations (x_i, c_i) and a defined subset X' of the feature space X.
The probability, that a randomly chosen observation will be in subset $X' \subset$ X and have churn decision c, is:

$$p^{emp}(X', c) = \frac{n(X', c)}{N}, \tag{3.14}$$

for $n(c, X') = \sum_{i=1}^{N} \mathbb{1}_{\{X' \times \{c\}\}}(\mathrm{x}_i, c_i)$.

Analogously, we have:

$$p^{emp}(X') = \frac{n(X')}{N} \tag{3.15}$$

for $n(X') = \sum_i^N \mathbb{1}_{\{X' \times \{0;1\}\}}(\mathrm{x}_i, c_i)$.

And

$$p^{emp}(c \mid X') = \frac{p^{emp}(X', c)}{p^{emp}(X')} = \frac{n(c, X')}{n(X')} \tag{3.16}$$

So, if we are given subsets X' of X and estimates of the churn decisions $\hat{c}_i$, we can calculate estimates of the churn probabilities for the elements in the subset by calculating the empirical probabilities. For a given expression of features, we can assign a unique probability, if alls subsets X' represent a (disjoint) partition of the feature space.

We define this partition as $X'_{t\in\mathcal{T}}$ for a finite index set $\mathcal{T}$, which we construct by the Binary Tree Growing Algorithm using standartized binary splits. The following pseudo-code is based on the algorithms presented in chapter 6.1.1 and 6.1.2 of [51], but has been simplified and slightly adapted to our setting.

Algorithm 1 Binary Tree Growing Algorithm

Step 1.
Initialize $\mathcal{X}_0 = \mathcal{X}$ and T = $\mathcal{T}$={0}.

Remark: $\mathcal{X}_0$ is called the root

Step 2. Repeat while {stopping criterion = TRUE}:

- (a) Select t $\in \mathcal{T}$,$l_t \in$ {1,..,q} and
  { $b_t \in X_l$ for continuous components x_l or $S_t \subset X_l$ for nominal components x_l } such that
  for $\mathcal{X}_{t0}$ = {x $\in \mathcal{X}_t \mid x_l \leq b_t$ } or, respectively, $\mathcal{X}_{t0} = S_t$, and $\mathcal{X}_{t1} = \mathcal{X}_t \setminus \mathcal{X}_{t0}$:
  (t,b_t,S_t,l_t)= $\arg\max_{(t,b_t,S_t,l_t)} \Delta$L ($\mathcal{X}_{t0}$,$\mathcal{X}_{t1}$)

  Remark:$\mathcal{X}_{t0}$ and $\mathcal{X}_{t1}$ are called nodes
- (b)Set the new binary tree $(\mathcal{X}_t)_{t\in T}$ with indexes
  T $\leftarrow$ T $\cup$ {t0, t1}.,
  $\mathcal{T} \leftarrow$ (T\{t})$\cup$ {t0,t1}}

Step 3.
Return the final binary tree $(\mathcal{X}_t)_{t\in T}$.

Remark: $(\mathcal{X}_t)_{t\in\mathcal{T}}$ are called leaves

---

The leaves of this tree represent a disjoint partition of the feature space, which is optimal with respect to a loss function L that still needs to be specified. Furthermore, the algorithm stops when a stopping criterion is fulfilled, which also needs to be specified.

Based on this partition of the features space, we define a classification rule that assigns

a churn decision to each element of the feature space.

We start by looking at an arbitrary subset $\mathcal{X}_t$ for some t $\in \mathcal{T}$. We define for the elements in this subset:

c_t= $\arg\max_c$p(c | X_t),

and estimate this:

$\hat{c}_t$= $\arg\max_c \hat{p}(c \mid X_t)$.

It is easy to see that this is equivalent to the assignment rule in (3.13) for a threshold b =0.5.

As estimators, we use the empirical probabilities; from (3.16) we get:

$$\begin{aligned}\hat{c}_t &= \arg\max_c \frac{n(c, X_t)}{n(X_t)} \\ &= \arg\max_c n(c, X_t)\end{aligned}$$

It can easily be seen taht this assignment rule is equivalent to the majority vote, meaning that we assign the event that most of the observations belong to. Note that, since the churn rate in our application case is rather small, we can expect that the majority vote often leads to an assigment of "no churn" as the empirical churn probabilities would be below 0.5.

And we combine these subset classifiers to one classifier:

$$\hat{\mathcal{C}}(x) = \sum_{t\in\mathcal{T}} \hat{c}_t 1_{\{x\in\mathcal{X}_t\}}$$

So, basically our model consists of a partition $(\mathcal{X}_t)_{t\in\mathcal{T}}$ and a class assignment rule $\hat{\mathcal{C}}$. The inputs are, as in the logistic regression models, expressions of features $\mathbf{x}$ and the outputs are churn decisions $\hat{\mathcal{C}}(\mathbf{x})$.

The parameters of this classification tree model are the index set T (defining the structure of the tree), the component indexes $(l_t)_{t\in T}$ and the thresholds $(b_t)_{t\in T}$, and $(S_t)_{t\in T}$. As it was done for the logistic regression models, the parameters will be determined with respect to an optimization criterion. In the case of the classification tree, this is the loss function L that appears in Step 2 of the Binary Tree Growing Algorithm.

As before, we will look at different loss functions and the resulting models in the following sections.

3.2.2 Gini Index- Estimator

Model

The most straigh forward approach to define the objective function in a classification setting is to count the average cases of wrong assignments, i.e. calculate the *missclassification rate*, which we define here on a subset $\mathcal{X}_t$:

$$L_{miss}(\hat{\mathcal{C}}_b) = \frac{1}{\mid \mathcal{X}_t \mid} \sum_{x_i \in \mathcal{X}_t} 1_{\{c_i \neq \hat{\mathcal{C}}_b(x_i)\}}$$

However, this function is not continuous and calculating gradients in order to find the optimum is problematic. Therefore, often uses other measures are used, such as the Gini Index. As this is the default optimization criterion in the Tree Growing Algorithm Implementation in R (package `rpart`), we introduce it here.

Definition 4. The *Gini Index Loss Function* is given by:

$$L_{GINI}(\hat{p}) = \sum_{\{t \in \mathcal{T}\}} \hat{p}(\mathcal{X}_t) 2 \hat{p}_t (1 - \hat{p}_t),$$

where $\hat{p}_t = p^{emp}(c \mid \mathcal{X}_t)$ and $\hat{p}(\mathcal{X}_t) = p^{emp}(\mathcal{X}_t)$

It correponds to the empirical probability of classifying a random observation incorrectly. It is similar to the missclassification error, in particular, it has the same maximum. Another candidate to measure impurity worth mentioning is the entropy. For more information on these measures, refer to [21].

Set L(($\mathcal{T}$) = $L_{GINI}((\mathcal{X}_t)_{t \in \mathcal{T}})$ to emphasize the dependence on the partition and let $\Delta L(\mathcal{T}, \mathcal{T}') = L((\mathcal{X}_t)_{t \in \mathcal{T}}) - L((\mathcal{X}_t)_{t \in \mathcal{T}'})$ give the difference of Gini Index Loss between two trees. Let $\Delta L(\mathcal{X}_{t0}, \mathcal{X}_{t1}) := \Delta L(\mathcal{T}, \mathcal{T}') \geq 0$ be the improvement of the Gini Index Loss in Step 2(b) of the Binary Tree Growing Algorithm. Here, T' = ($\mathcal{T}$ \{t}) $\cup$ {t0,t1}} indicates the new tree after a binary split in Step 2(b) of the Algorithm. Finally, one can determeine the stopping criterion as a function of L, for example, a minimum improvement of the loss Function after an optimal split. It can also be defined in terms of minimum number of observations per leaf.

Plugging this definition of L into Step 2(a) of the Binary Tree Growing Algorithm then leads to the construction of the classification tree.

In order to model the churn decision, we also need to specify the threshold b of the class assignment rule in equation (3.13). We choose it such that the missclassification rate is

minimized, which is the case for b=0.5. This follows directly from the reasoning described in Section 3.2.3., where we look at a more general version of the missclassification rate.

Results

We implement naively the described algorithm on the original data set and obtain a model called Class_Tree_Naive. We notice, that the results of this naive classification tree are a bit disappointing. This is due to the fact that we have an imbalanced data set, combined with an optimization criterion (the Gini Index Loss Function) that is linked to accuracy. In other words: Because the overall empirical churn probability is small, the tree constructing algorithm favors classifications as "non churns". We will deal with this problem in the next section.

3.2.3 Imbalance Weighted Gini Index Estimator

Model

There are several ways to tackle the issue of imbalanced data, such as modifying the underlying data sample by undersampling, oversampling or synthetic data generation. We will, however, turn to the approach of modifying the loss function, as this is more consistent with the methods used to construct the previous models and allows directly for comparison.
Before, when constructing the naive classification tree, all missclassification errors were considered equal. The idea now is to punish wrong "no-churn" classifications more than wrong "churn" classifications in order to correct the incentive of the model to classify an observation as "no-churn" due to the data imbalance issue.
These weights are represented by a *Loss Matrix*:

$$L = \begin{pmatrix} 0 & L_1 \\ L_2 & 0 \end{pmatrix}, \tag{3.17}$$

where $L_{i,j}$ is the loss weight assigned to the observed decision i classified as decision j. (i,j =1 corresponds to "no churn" and i,j= 2 corresponds to "churn").

Naturally, correct classifications are not punished, hence $L_{i,i} = L_{j,j} = 0$. Without loss of generality, we can set $L_1 = 1$ as only the ratio of L_2 and L_1 matters. So, we are left to specify L_2, the weight for the error of not correctly detecting a churn.

Following the idea presented in [28] , we use the accuracy ratio

$$L_2 = \rho := \frac{N - n(X)}{n(X)},$$

where n(X) is the number of churns in analogy to the notation in definitions (3.14-3.16). The implementation of this loss matrix in the model is then done via the altered priors method, as this is the only option available in the `rpart` package. This method is outlined in the R documentation [47] and will be detailed for our specific case hereafter.

In Definition 4, we defined $\hat{p}_t$ as the empirical churn probability. We now rewrite this the following way:

$$\hat{p}_t = \frac{n(c, \mathcal{X}_t)}{n(\mathcal{X}_t)} \cdot \frac{N}{N} \cdot \frac{n(c)}{n(c)},$$

where n(c) = n(c,X) is the number of observed churns on the whole data sample.
We define the *prior churn probability* as

$$\pi := \frac{n(c)}{N} = \hat{p}_0.$$

Then

$$\hat{p}_t = \pi \cdot \frac{n(c, \mathcal{X}_t)}{n(c)} \cdot \frac{N}{n(\mathcal{X}_t)}$$

We now alter the prior by including the loss weights the following way:

$$\pi \longmapsto \tilde{\pi}_\rho := \frac{\pi \cdot \rho}{\rho\pi + (1 - \pi)} = \frac{\pi \cdot \rho}{\rho\pi + (1 - \pi)}$$

For example, if the overall empirical churn probability is 10 %, we have
$\rho = \frac{N - 0.1N}{0.1N} = 9$. Then $\tilde{\pi}_\rho = \frac{0.1 \cdot 9}{9 \cdot 0.1 + 0.9} = 50$ %.
Plugging the altered priors into the empirical probabilities then leads to

$$\hat{p}_t \longmapsto \tilde{p}_t = \tilde{\pi}_\rho \cdot \frac{n(c, \mathcal{X}_t)}{n(c)} \cdot \frac{N}{n(\mathcal{X}_t)}$$

To emphasize the dependence of the altered priors on the loss weight ρ we also write $\tilde{p}_t = \tilde{p}_{t,\rho}$. This leads to the following definition:

Definition 5. The *Imbalance Weighted Gini Index Loss Function* is defined as

$$L_{GINI,\rho}(\hat{p}) = \sum_{t \in \mathcal{T}} \hat{p}(\mathcal{X}_t) \cdot 2 \cdot \tilde{p}_{t,\rho}(1 - \tilde{p}_{t,\rho}),$$

where $\tilde{p}_{t,\rho} = \frac{\pi \cdot \rho}{\rho\pi + (1-\pi)} \cdot \frac{n(c, \mathcal{X}_t)}{n(c)} \cdot \frac{N}{n(\mathcal{X}_t)}$ for the prior churn probability π.

The classification tree is then constructed according to the Binary Tree Growing Algorithm explained in Section 5.1 with the loss reduction explained in Section 5.2.
Again, we need to specify the class assignment threshold b in order to make class predictions. Note that in this case, the missspecification rate is weighted by the loss matrix: We penalize false positives differently than false negatives.
The *L-weighted missspecification rate* looks like this:

$$L_{miss,L}(\hat{\mathcal{C}}_b) = \frac{1}{\mid \mathcal{X}_t \mid} \sum_{x_i \in \mathcal{X}_t} [L_1 \cdot \mathbb{1}_{\{c_i=0 \wedge \hat{\mathcal{C}}_b(x_i)=1\}} + L_2 \cdot \mathbb{1}_{\{c_i=1 \wedge \hat{\mathcal{C}}_b(x_i)=0\}}] \tag{3.18}$$

We want to choose b such that $L_{miss,L}(\hat{\mathcal{C}}_b)$ is minimized.
Note that, if $\hat{\mathcal{C}}_b(x_i) = 1 \ \forall x_i \in \mathcal{X}_t$, we have
$\{c_i = 1 \wedge \hat{\mathcal{C}}_b(x_i) = 0\} = \emptyset$ and $\{c_i = 0 \wedge \hat{\mathcal{C}}_b(x_i) = 1\} = \{c_i = 0\}$.
In this case,

$$L_{miss,L}(\hat{\mathcal{C}}_b) = \frac{1}{\mid \mathcal{X}_t \mid} \textstyle\sum_{x_i \in \mathcal{X}_t} L_1 \cdot \mathbb{1}_{\{c_i=0\}}$$

We proceed analogously for the case of an assignment $\hat{\mathcal{C}}_b(x_i) = 0 \ \forall x_i \in \mathcal{X}_t$ and use the notation introduced in definition (5.3) and onwards. The resulting misspecification rates are:

$$\mathrm{L}_{miss,L}(\hat{\mathcal{C}}_b) = \begin{cases} \frac{n(0, \mathcal{X}_t)}{n(\mathcal{X}_t)} \cdot L_1 & \text{if } \hat{\mathcal{C}}_b(x_i) = 1 \text{ for all } x_i \in \mathcal{X}_t \\ \frac{n(1, \mathcal{X}_t)}{n(\mathcal{X}_t)} \cdot L_2 & \text{if } \hat{\mathcal{C}}_b(x_i) = 0 \text{ for all } x_i \in \mathcal{X}_t \end{cases} \tag{3.19}$$

We want to choose b such that $L_{miss,L}(\hat{\mathcal{C}}_b)$ is minimized.
Let $\hat{\mathcal{C}}*$ be the related optimal assignment rule. Then

$$\begin{aligned} *(\mathrm{x}) = 1 \forall x_i \in \mathcal{X}_t \iff L_{miss,L}(1) < L_{miss,L}(0) \\ \iff \frac{n(0, \mathcal{X}_t)}{n(\mathcal{X}_t)} \cdot L_1 < \frac{n(1, \mathcal{X}_t)}{n(\mathcal{X}_t)} \cdot L_2 \iff L_1 \cdot (1 - \hat{p}_t(x)) < L_2 \cdot \hat{p}_t(x). \end{aligned}$$

So, we precise the following:

$$\text{The threshold in equation (3.13) is set as } b := \frac{L_1}{L_1 + L_2} \tag{3.20}$$

From this, it immediately follows, that if $L_1 = L_2$, we have b= 0.5. This is the case for default model introduced in section 3.2.2.
One final remark: In this section, we integrated a loss function to the original objective function (impurity measure), the Gini Index. A similar approach is also possible for the entropy impurity, which is called the focal loss. For details on this, refer to [27].

Results

When performing the Binary Tree Growing Algorithm using the modified Imbalance Weighted Gini Index Loss Function, we obtain a much more complex tree in the sense that we have more leaf nodes and that these are more balanced. We can also observe the change of the threshold b, which leads to more observations being classified as "churn" than previously. We name this model Class_Tree_Imbalance_Weighted.

3.3 Gradient Boosting Machine Model

3.3.1 Gradient Boosting Machine Model

In this section, we will combine many 'simple' classification trees like those constructed in Section 3.2 in order to improve prediction power.
To achieve this, we start with a simple classification tree and incrementally improve it by adding more simple classification trees. Here, 'simple' means that only a small number of leaf nodes is allowed, and 'improvement' means that it reduces a defined loss function.
We will begin by formally introducing the general approach, closely following Section 7.4.1 in [51] and combining it with the information on the implemantation in the R package `gbm` descibed in [34]
In line with the previous notation, let $\mathrm{L}(\hat{p})$ be a loss function, $\mathbf{x}_i = (1,x_1^i,...,x_q^i)$ the expression of features of observation i, c_i the churn decision of observation i and $(\mathbf{x}_i, c_i)_{i=1,..,N}$ the observations of the data set. Let $\hat{p}(\mathbf{x}_i)$ be an estimator (produced by a model) for the churn probability of an individual with features $\mathbf{x}$.
Now let

$$\hat{p}_{m-1} = \underset{\hat{p}\in CT(K))}{\arg\min} \frac{1}{N}\sum_{i=1}^{N} L(c_i, \hat{p}(\mathbf{x}_i)) \tag{3.21}$$

where CT(K)= $\{\hat{p} \mid \hat{p}(x) = \sum_{t\in\mathcal{T}} p^{emp}(c \mid X_t)\mathbb{1}_{\{x\in\mathcal{X}_t\}}$ with $\mathcal{X}_t$ the output of the Binary Classification Tree Algorithm and $\mathcal{T} \leq K$ $\}$ is the set of Binary Classification Trees with maximum K leaf nodes.

Note that N stands for the number of observations in the data set. If we calculate this on a training set with, say, N' of observations, and have K=N', we could set $\hat{p}(\mathbf{x}_i) = c_i$ and obtain zero loss. Even though this saturated model would be optimal, this result would not expand to external test data because of overfitting. It is therefore advisable to set K $\ll$ N.

In the next step, we try to improve $\hat{p}_{m-1}$ by finding

$$\hat{g} := \underset{g\in CT(K))}{\arg\min} \frac{1}{N}\sum_{i=1}^{N} L(c_i, \hat{p}_{m-1}(\mathbf{x}_i) + g(\mathbf{x}_i)) \tag{3.22}$$

To determine $\hat{g}$, we will use the gradient descent approach.

We write $\hat{p}_{m-1}(O) = (\hat{p}_{m-1}(\mathbf{x}_1), ..., \mathbf{x}_N)^T = (\hat{p}_{m-1,1}, ..., \hat{p}_{m-1,N})^T$ the vector of the (modelled) churn predictions of the observations O, and $L(\hat{p}_{m-1}) = \frac{1}{N}\sum_{i=1}^{N} L(c_i, \hat{p}_{m-1,i})$ for the corresponding loss.

Now we deviate from $\hat{p}_{m-1}$ pointwise in such a way that it leads to a maximal decrease of $L(\hat{p})$. This is exactly achieved by the partial derivates of L, which are summarized by the gradient

We have

$$\nabla L(\hat{p}_{m-1}) = \frac{1}{N}\left(\frac{\partial L(c_i,\hat{p})}{\partial\hat{p}}\bigg|_{\hat{p}=\hat{p}_{m-1,1}}, ..., \frac{\partial L(c_i,\hat{p})}{\partial\hat{p}}\bigg|_{\hat{p}=\hat{p}_{m-1,N}}\right)^T. \tag{3.23}$$

and define the model update by:

$$\hat{p}_{m-1} \longmapsto \hat{p}_{m-1} - \rho_m \cdot \nabla L(\hat{p}_{m-1}), \tag{3.24}$$

where ρ_m is the optimal step size:

$$\rho_m = \underset{\rho>0}{\arg\min}\, L(c_i, \hat{p}_{m-1} - \rho\nabla L(\hat{p}_{m-1})). \tag{3.25}$$

More precisely, ρ_m takes the values $\rho_{m,1}, .., \rho_{m,K}$ in the leaf nodes:

$$\rho_{m,t} = \underset{\rho>0}{\arg\min} \sum_{\mathbf{x}_i\in\mathcal{X}_t} L(c_i, \hat{p}_{m-1} - \rho). \tag{3.26}$$

Iterating these steps gives an algorithm to build a boosted tree prediction. In practice, two additional implementations are to be noted:

- The gradient $\nabla L(\hat{p})$ is approximated by a regression tree with the same constraints as

for $\hat{g}$, that is, maximum number of leaf nodes K.

- A shrinkage parameter λ is introduced to control the step width and learning rate of the algorithm.

After fixing the number of iterations T, the maximum number of leaves K, the shrinkage parameter λ and the subsampling rate p, we can define the Gradient Boosting Algorithm, which is outlined below.

Algorithm 2 Gradient Boosting Algorithm

Step 1. Initialize $\hat{p}_0(\mathbf{x}) = \arg\min_{\rho>0} \sum_{i_1}^{N} L(c_i, \rho)$ and t=0.
Remark: This is a constant
Step 2. Repeat while {t>T}:

- (a) Compute the negative gradient as the working response:

$$z_i = -\frac{\partial L(c_i, \hat{p})}{\partial \hat{p}}\bigg|_{\hat{p}=\hat{p}_{t,i}}$$

- (b) Randomly select p · N observations from the data set.
- (c) Fit a regression tree with K terminal nodes, $g_t(\mathbf{x}) = E(z \mid \mathbf{x})$ based on these selected observations.
- (d) Compute the optimal step width $\rho_t = \arg\min_{\rho>0} \frac{1}{N} \sum_{i=1}^{N} L(c_i, \hat{p}_t - \rho g_t(\mathbf{x}))$
- (e) Update the model:
  $\hat{p}_{t+1}(\mathbf{x}) \mapsto \hat{p}_t + \lambda \rho_t g(\mathbf{x})$
  and
  t $\mapsto$ t+1

Step 3. Return the final estimator $\hat{p}_T(x)$.

Step 2(d) and 2(e) slightly differ from the ones indicated in the R documentation. However, we believe our notation to be more consistent with the R documentation itself (such as the description of the shrinkage parameter) and assume that the documentation's description of the algorithm is not quite complete. Also note, that our notation is consistent with equation (7.16) in [51].

3.3.2 Bernoulli GBM Estimator

Model

In order to apply the Gradient Boosting Algorithm, an adequate loss function needs to be defined. We take the one from the GBM algorithm explained in [34].

Definition 6. The *Bernoulli GBM Loss* is given by:

$$L_{Ber}(f) = -\sum_{i=1}^{N}[c_i f(\mathbf{x}) - \log(1 + e^{f(\mathbf{x})})],$$

where the probabilities p are given by $f(x) = \log\left(\frac{p(x)}{1-p(x)}\right)$.

Plugging this loss function into the definition of the Gradient Boosting Algorithm allows to calculate the following specifications.

Step 1. We calculate f_0.

$f_0 = \arg\min_{\rho>0} -\sum_{i=1}^{N}[c_i\rho + \log(1 + e^{\rho})]$

We solve this by calculating the derivative:

$$\begin{aligned}
\frac{dL(\rho)}{d\rho} &= -\sum_{i=1}^{N}[c_i - \frac{e^{\rho}}{1+e^{\rho}}] = 0 \\
&\iff -\sum_{i=1}^{N} c_i = \frac{Ne^{\rho}}{1+e^{\rho}} \\
&\iff -\sum_{i=1}^{N} c_i - \sum_{i=1}^{N} c_i e^{\rho} = Ne^{\rho} \\
&\iff -\sum_{i=1}^{N} c_i = e^{\rho}(N + \sum_{i=1}^{N} c_i) \\
&\iff \rho = \log(\frac{-\sum_{i=1}^{N} c_i}{N + \sum_{i=1}^{N} c_i})
\end{aligned}$$

Step 2. We calculate z_i of interation t:

$$\begin{aligned}
z_i &= -\frac{\partial}{\partial f_i}(-1)\sum_{i=1}^{N} c_i f_i - \log(1 + e^{f_i})\Big|_{f_i=f_{t,i}} \\
&= (c_i - \frac{e^{f_i}}{1+e^{f_i}})\Big|_{f_i=f_{t,i}} \\
&= c_i - \frac{e^{f_{t,i}}}{1+e^{f_{t,i}}}
\end{aligned}$$

Remark: From $p(x) = \frac{1}{1+e^{-f(\mathbf{x})}}$ we have $f(\mathbf{x}) = -\log(1-p) + \log(p)$

Plugging this in the Bernoulli GBM Loss yields

$L_{Ber}(p) = -\sum_{i=1}^{N} c_i[-log(1-p) + \log(p)] - \log(1 + e^{-\log(1-p)+\log(p)})$

$=-\sum_{i=1}^{N} c_i[-\log(1-p) + \log(p)] - \log(1 + \frac{p}{1-p})$

$=-\sum_{i=1}^{N} c_i[-\log(1-p) + \log(p)] + \log(1-p)$

$= -\sum_{i=1}^{N} c_i \log(p) + (1 - c_i) \log(1 - p)$.

Results

We apply the Gradien Boosting Algorithm on our data set and call the resulting model GBM_Bernoulli. A thourough analysis of the results is provided in Appendix, Section 5.4. For example, as in the previous models, the premium/premium difference to the competition are among the most important variables according to this model. Partial Dependence Plots with respect to these premium differences show the form of an increasing step function around zero. This illustrates the non-linear dependence of churn rate and premium difference, which is caught by this model. Also the effects of other variables on the churn rate and interactions are analysed.

3.4 Model Performance Measures

3.4.1 Classification Metrics

Introduction

The loss functions that were used in the previous sections to fit the correspoding models were required to be continuous in order to be of use the respective optimization algorithms. However, when evaluating the prediction power of an existent fitted model, we can revert to discrete metrics.
The most straight-forward metric is the *accuracy*, which was introduced by [5]. It denotes the probability of missclassifying a point from the data set and depends on the specific classifier.

Definition 7. Let

$$\mathcal{R}(\mathcal{C}) := P[\mathcal{C}(\mathbf{x}) = c] \tag{3.27}$$

be the *accuracy* of the classifier $\mathcal{C}$, where ($\mathbf{x}$,c) $\in$ O is a pair of observations.

For a given data set with N observations, the accuracy can be estimated by the rate of correct predictions,

$$\hat{\mathcal{R}}(\mathcal{C}) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\mathcal{C}(x_i)=c_i\}}. \tag{3.28}$$

Note that this is the counterpart of the missclassification rate defined in Section 3.2.2.

Unfortunately, for unbalanced data sets like ours, using the accuracy to assess the model performance can be quite misleading. This is due to the fact that $\hat{\mathcal{R}}(\mathcal{C})$ gives the same weight to all (miss)classifications, without distinguishing between incorrectly predicted no-churns (which correspond to the majority class) and incorrectly predicted churns (which correpond to the minority class). As the large majority of the data points is expected to belong to the majority class, they will dominate the accuracy, which will biase the measure towards this class. Since we are actually more interested in predicting the minority class correctly, this can be misleading. For example, a classifier that assigns the majority class to all data points will automatically lead to a quite high accuracy, but will be useless in terms of predicting churns (which is our goal).
In order to define other indicators that can be used to assess the classification model perfomance, the model results are often summarized in a *confusion matrix*. It has the following structure.

$\hat{c}$ \c	1	0
1	TP	FP
0	FN	TN

It compares the predicted classes with the true classes, counting the elements. More precisely, we have:

$$\begin{aligned} TP &:=\mid \{c = 1 \wedge \mathcal{C}(x) = 1\} \mid \\ TN &:=\mid \{c = 0 \wedge \mathcal{C}(x) = 0\} \mid \\ FP &:=\mid \{c = 0 \wedge \mathcal{C}(x) = 1\} \mid \\ FN &:=\mid \{c = 1 \wedge \mathcal{C}(x) = 0\} \mid \end{aligned} \tag{3.29}$$

Based on this information, we can calculate the following popular indicators.

Definition 8. We define:

$$Sensitivity := \frac{TP}{TP + FN},$$

estimating the probability that a true churner will be predicted correctly (also called *recall*),

$$Specificity := \frac{TN}{FP + TN},$$

estimating the probability that a true non-churner will be predicted correctly,

$$Precision := \frac{TP}{TP + FP},$$

estimating the probability that a predicted churner is actually a true churner.

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall},$$

estimating the probability that a predicted churner is actually a true churner.

The last indicator could be an interesting measure for the model performance as it combines the information of the confusion matrix in a condensed way. Unfortunately, it also gives equal weight to precision and recall and therefore has the same issues evaluating the model on unbalanced data sets as the accuracy metric.
Of course, many other measures are available. We choose the one which is deemed both very popular and suitable for unbalanced date: The ROC and the AUC.

ROC and AUC

The Receiver Operation Characeteristics (ROC) graph is a popular way to visualize the capability of the model to distinguish between classes. The following introduction is based on [11].
First, we introduce another indicator, based on the confusion matrix.

Definition 9. Let the *false positive rate* be defined as

$$fp - rate := \frac{FP}{FP + TN}$$

In the same spirit, we call the previously introduced *Sensitivity* the *true positive rate (tp-rate)*.
The ROC is a plot that shows the true positive rate in function of the false positive rate and looks as follows (taken from [11]):

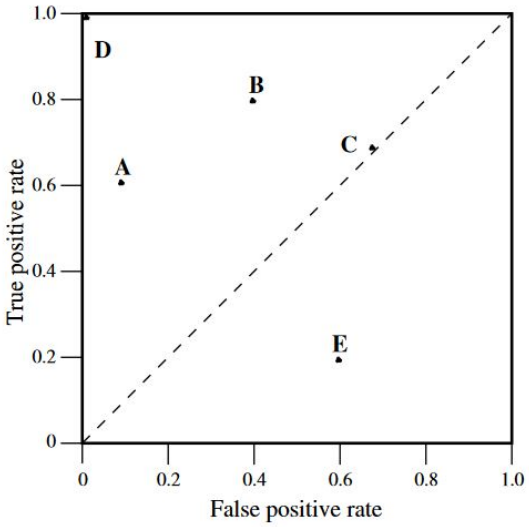

Figure 3.1: The ROC

Each point in this diagram corresponds to a specific classifier. To understand this better, first consider a classifier which classifies all data points as $\mathcal{C}(x) = 1$ (the "positive" prediction). This would for example be the case for our classifier defined in equation (3.13) with a threshold b set to 0. We would then expect to have:

$$TP = N \cdot p,$$
$$TN = FN = 0,$$
$$FP = N \cdot (1 - p).$$

Here, p is the true churn rate, which can be estimated by the empirical churn rate. Consequently, we get:

$$\textit{fp- rate} = \frac{FP}{FP + TN} = \frac{N(1-p)}{N(1-p) + 0} = 1$$
$$\textit{tp- rate} = \frac{TP}{TP + FN} = \frac{Np}{Np + 0} = 1$$

So, this classifier would be represented as point (1,1) in the diagram. Analogously, a classifier with $\mathcal{C}(x) = 0$ would lead to TP=FP=0 and the corresponding point in the ROC would be (0,0). Note that the classifiers of these examples are not very sophisticated. In fact, they don't include any feature information of the data points and just assign one value to all. Also, the structure we assumed in (3.13) is quite specific to our model. We can generalize this and assume that we have a classifier that purely randomly assigns a fixed proportion a

of the data points to the class 1. We then have:

$$\begin{aligned} E[TP] &= N \cdot P[\{c = 1 \wedge \mathcal{C}(x) = 1\}] \\ &= P[c = 1] \cdot P[\mathcal{C}(x) = 1], \text{ because } \mathcal{C} \text{ is purely random} \\ &= N \cdot p \cdot a \text{ (by definition)} \end{aligned}$$

And analogously,

$$E[TN] = (1 - a) \cdot (1 - p)$$
$$E[FP] = a \cdot (1 - p)$$
$$E[FN] = p \cdot (1 - a)$$

Plugging this in definition 8 and 9 yields
$fp - rate = \frac{FP}{FP + TN} = \frac{a(1-p)}{a(1-p) + (1-a)(1-p)} = a,$
$tp - rate = \frac{pa}{pa + p(1-a)} = a.$
So, all purely random classifiers correspond to points (a,a) for a ∈ [0,1], which is exactly the diagonal of the ROC (including example point C).
We can therefore consider the diagonal as a benchmark for random guessing and all classifiers lying above above this line (like points A and B) to correspond to models that are better than this. Point D corresponds to the perfect model, which produces no missclassifications at all (implying FN=FP=0).
All models lying below the benchmark line are 'worse' than random guessing, as they produce even more missclassifications than random guessing would.
Note that every point in the ROC correponds to exactly one classifier. If, like in our logit or classification tree models, the classification rule $\mathcal{C}_b$ is applied on estimated probabilities and is induced by a threshold b, then we can vary the threshold b and receive a distinct classifier for each choice of b. For all b ∈ [0,1] this results in a curve (the ROC curve) with endpoints (0,0) and (1,1), that correspon to b=0 and b=1, as we have outlined at the beginning of this section.

Finally, we need to summarize the information of the ROC-curve in one (numerical) measure in order to induce an ordering of different models' performances and allow for comparison. From what we have seen so far, we can conclude,that, the better the model, the further its ROC-curve will lie in the upper left part of the ROC.
This is reflected by the *Area under the ROC Curve (AUC)* measure, which is illustrated in

the figure below (also taken from [11]).

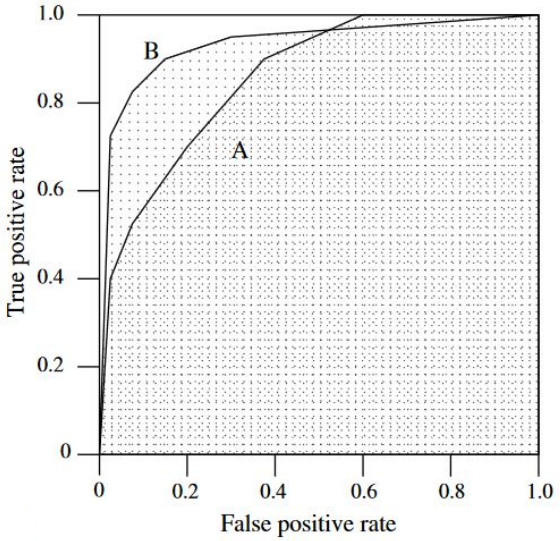

Figure 3.2: The Area under the ROC Curve (AUC) of two classification models

The diagonal has an AUC of 0.5. So, in order to be better than random guessing, the model should have an AUC higher than this value.
As a final remark, note that the false positive rate is equivalent to 1 - specificity, which is why sometimes the x-axis of the ROC shows the "inverted" scale of the specificity.

3.4.2 Pricing Loss

Since we are interested in using the churn model in an insurance pricing context, it seems legitimate to ask whether the classification metrics (namely the AUC) presented in the previous section are suitable criterions to assess the performance of a churn model.
Building on the reasoning of section 3.1.3., where we introduced weights to incorporate the effect on the technical result, we want to build a measure that assesses the financial impact of the prediction error of the model on the insurer.
As the initial application purpose of the developed churn model is to use it in pricing, we will deduce this 'pricing loss' measure from a pricing model.
Here, we will use the technical result instead of the expected claims. Why? In Swiss mandatory health insurance, the premium must cover all costs: the expected claims, reserves and payments from or to the risk compensation scheme. The technical result covers all these costs, and in our data set, it is approximated as:

Definition 10. Let the *Technical Result* of individual i be given by

$$R_i(t) = P_t(i) - RC_t(i) - C_t(i) * 135\%,$$

where $P_t(i)$ is the net premium of individual i in year t, $RC_t(i)$ is the Risk Compensation payment (can be positive or negative) of individual i in year t, and $C_t(i)$ are the claims paid in the current year t.

Note that this is an approximation as for recent years, the Risk Compensation is not known yet, and $C_t(i) * 135\%$ is an estimate for the ultimate i.e. includes the estimated development reserves for the observed claims.
Theoretically in Swiss mandatory health insurance, within the pricing process, the premium is set such that the expected result is zero. More generally, the expected result is often positive as it corresponds to the expected profits of the product. Interpreting risk as a deviation from expectation, we are interested in the pricing risk coming from the portfolio risk. This means that we want to quantify the difference of expected and effective technical result that comes from a difference of expected and effective churns. In our case, the premium is differentiated by a few factors, like canton and age group (children/young adults/adults), but within these categories, we have a unitary premium. We call this partition of the feature space the *pricing grid* and will therefore look at the deviations from expected result per *grid cell*. Note that we implicitly consider the (profit) margin of the premium here, so we actually do not only quantify the risk due to claims (actuarial/ risk premium), but the risk on the expected result (market premium), which is actually financially more relevant.
Then, we assume the average over the technical results of the observations is an unbiased estimator for the expected technical result (i.e. corresponds to the expected result) and that it does not change from one year to the next (stationarity assumption). This is important for two reasons: First, from a practical point of view, it would not be possible to take the next years (effective) technical results per person for all individuals in the preceding year's portfolio, simply because this information is not available for churners. Second, we are interested in isolating the effect of the churn model, which would be hard if effects get mixed up with random effects that come from the random claims.

To introduce this model, let $(G_k)_{k=1,..,K}$ be the pricing grid, that is a decomposition of the feature space into subsets. Then $|G_k|$ is the number of individuals per grid unit and G= $\bigcup G_k$ is the set of all individuals in the insurance portfolio. For each policy i $\in \{1, ..., |G_k|\}$ of grid cell G_k let the corresponding technical result be given by $R_{k,i}$.

Then the average technical result per grid cell k, $\bar{R}_k$, is given by

$$\bar{R}_k = \sum_{i \in G_k} \frac{R_{k,i}}{|G_k|}$$

and analogously the average churn rate $\bar{p}_k$ per grid cell k is given by

$$\bar{p}_k = \sum_{i \in G_k} \frac{c_{k,i}}{|G_k|} \tag{3.30}$$

Let $R(t)$ be the total technical result at time t:

$$R(t) = \sum_{k=1}^{K} \sum_{i=1}^{|G_k|} R_{k,i}(t) = \sum_{k=1}^{K} \bar{R}_k(t) \cdot |G_k(t)|$$

We then have for the expected total technical result at time t+1:

$$E[R(t+1)] = \sum_{k=1}^{K} E[\bar{R}_k(t+1)] \cdot E[|G_k(t+1)|] \tag{3.31}$$

By the stationarity assumption, we have

$$E[\bar{R}_k(t+1)] = \bar{R}_k(t), \tag{3.32}$$

which can be calculated based on observations O, available at t.

The next year's volume per grid depends on the churn decision of the individuals in the grid:

$$|G_k(t+1)| = \sum_{i=1}^{|G_k|} (1 - c_{k,i}) = |G_k| \cdot \sum_{i=1}^{|G_k|} \frac{(1 - c_{k,i})}{|G_k|} = |G_k|(1 - \bar{p}_k)$$

where we associated with $c_{k,i}$ the churning decision for individual i in grid cell k.

Plugging this and equation (7.5) into (7.4) yields:

$$E[R(t+1)] = \bar{R}_k(t) \cdot |G_k|(1 - E[\bar{p}_k]) \tag{3.33}$$

Under stationarity assumption, the critical object that needs to be estimated in order to predict the total technical result is the average churn probability per grid. So, in order to be of practical use in the pricing process, where the premium is calculated per grid cell, it is the expected churn probability per grid cell, that needs to be provided. Hence, we need to get to the expected average churn probability per grid cell $E[\bar{p}_k]$ (i.e. the Pricing Model Input) from the individual probabilities estimated by the model $\hat{p}_i = \hat{p}(\mathbf{x}_i)$(i.e. the Churn

Model Output).

From equation (7.4) we have

$$E[\bar{p}_k] = \sum_{i=1}^{|G_k|} \frac{E[c_{k,i}]}{|G_k|} = \sum_{i=1}^{|G_k|} \frac{\hat{p}_{k,i}}{|G_k|} \tag{3.34}$$

where $\hat{p}_{k,i} = \hat{p}(\mathbf{x}_{k,i})$ and we index the individuals as explained in this section's introduction. Now we want to calculate the pricing loss as the difference of the premium and the effective costs. More precisely, the pricing loss is the loss that occurs due to differences of expected and effective churn rates per grid (all other parameters equal). The pricing loss is the difference of expected and effective technical result:

$$\begin{aligned} E[R(t+1)] - R(t+1) &= \sum_{k=1}^{K} [\bar{R}_k(t) \cdot |G_k|(1 - E[\bar{p}_k]) - \bar{R}_k(t) \cdot |G_k(t)|(1 - \bar{p}_k)] \\ &= \sum_{k=1}^{K} \bar{R}_k(t)|G_k|(\bar{p}_k - E[\bar{p}_k]) \end{aligned}$$

The definition of the pricing grid, i.e. the choice of the decomposition G is arbitrary and irrelevant for the total pricing loss, as a transfer of losses and profits between risk cells is allowed. But for example, if we want to look at the aggregated result over different products to calculate a loss based on the insured's "value", we actually would need to adapt that function. In fact, the approach via the pricing grid was to understand how the pricing loss is composed and to illustrate how the churn model output is intergrated in a pricing model. So, wlog we can assume N= |G| and $|G_k| = 1 \; \forall$ k. Then $\bar{p}_k = \hat{p}_k$,$\bar{p}_k = p_k$ and $\bar{R}_k = R_k$ and we have

$$E[R(t+1)] - R(t+1) = \sum_{k=1}^{N} R_k(p_k - \hat{p}_k)$$

We write $R_{k,i} = R_k$ and then get:

$D(R, \hat{R}) := \sum_k R_k \cdot (\hat{p}_k - c_k)$

The difference of observed and estimated result is the sum of the differences of the observed churns and estimated churn probabilities, weighted by the technical result. A priori, we will treat positive and negative deviations equally, even though from a practical perspective, unexpected losses in profit are more problematic than unexpected gains. This motivates the following definitions.

Definition 11. The *absolute pricing loss function* is defined as

$L_{p,abs}(p) = \sum_i^N |R_i| \cdot |\hat{p}_i - c_i|$

The *net pricing loss function* is defined as
$L_{p,net}(p) = \sum_i^N R_i \cdot (\hat{p}_i - c_i)$

Initially, we compared the predicted churn rates with the effective churn rates (per grid cell), which then lead to a comparison of predicted churn probabilities with effective churn decisions. Since the inputs of the pricing model are the outputs of the churn model, it makes sense to assess the loss in terms of probabilities. Of course, theoretically, one could also compare classification outputs of the churn models (i.e. predicted churn decisions), but in a pricing context, this would not seem to be the a reasonable input.

3.5 Premium Sensitivity

The goal of this section is to understand how the predicted churn probability changes when premiums are adapted. This is particularly interesting, because premium (change) are the only feature that can actually be controlled by the insurer. We will do this by approximating the change by a linear function with the slope being defined as the *premium sensitivity.*
We will look at the overall churn probability of the whole portfolio, but of course, the following concepts can also be applied on subsets of the feature space.
The derivation of the premium sensitivity depends on the features/ explanatory variables. For example, the information on premiums or premium changes should somehow be contained in the observations. In our particular setting, several variables are concerned by premium changes: First, the variable that represents exactly the premium change from one year to the next. Second, the variables that represent the difference of the competitor's future premium to the insurer's future premium. This difference changes, if the insurer's future premium changes due to premium adaptations.
We will look at two applications: The Logistic Regression model and the Gradient Boosting Machine model. As the Logistic Regression Model provides a closed-form formula for the predicted churn probability, we will be able to analytically derive the premium sensitivity. This is not the case for the GBM, where we only have an empirical function for the churn probability, which comes from a "Black Box". Here, we will define a numerical approach to solve the problem.

3.5.1 Premium Sensitivity for Logistic Regression

In Chapter 3.1 , we introduced the Logit Models, which provide a direct formula of the churn probability, which is specified by coefficients, that are estimated on a given data set. Let $\hat{\beta} = (\hat{\beta}_0, .., \hat{\beta}_q)'$ be the estimated coefficients. According to the model structure of equation (3.6), we define the estimator of the expected churn rate based on N observations by:

$$\hat{E}[\hat{p}] = \frac{1}{N} \sum_{i=1}^{N} \hat{p}(\mathbf{x}_i) \tag{3.35}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \frac{e^{\langle \hat{\beta}, \mathbf{x}_i \rangle}}{1 + e^{\langle \hat{\beta}, \mathbf{x}_i \rangle}} \tag{3.36}$$

Here, we consider the expected value over the whole feature space for simplicity. Of course, we could also condition on subsets, such as a specific canton, if we are intersted in the

(isolated) effects of premium changes per canton.

We will now only vary the change in premium, keeping all other features fixed. Without loss of generality, let the last feature x_q give the premium change from one year to the next of the current insurer, and the previous m features $x_{q-m}, ..., x_{q-1}$ represent the difference of the current insurer's future premium to the competitor's future premiums. More precisely, we write for the current insurers premium change

$$x_q = P(t+1, \mathbf{x}) - P(t, \mathbf{x}) := \Delta P_t(\mathbf{x}) \tag{3.37}$$

where $P(t, \mathbf{x}) = P(t, x_1, .., x_{q-m-1})$ is the premium of the current insurer for an individual with features $x_1, .., x_{q-m-1}$ at time t. Of course, it could be that the premium does not depend on all features up to $q - m - 1$, but only on a few criterions such as age group and canton.
We write for the competitor k's premium difference

$$\begin{aligned} x_{q-k} &= P_k(t+1, \mathbf{x}) - P(t+1, \mathbf{x}) \text{ for } k = 1, .., m. \\ &= P_k(t+1, \mathbf{x}) - P(t, \mathbf{x}) - \Delta P_t(\mathbf{x}) \end{aligned}$$

where $P_k(t, \mathbf{x}) = P_k(t, x_1, .., x_{q-m-1})$ is the premium of insurer k for an individual with features $x_1, .., x_{q-m-1}$. The time component should reflect the decision process of the insured: Usually, switching happens once a year, after the publication of all insurers' premiums for the following year and the duration of the contract then is for one year. So, mid-year churns do not happen and we believe that the *current* year's premiums do not have an effect on the churn decision, as it is the *next* premium that will be the effective premium paid. In what follows we omit the time component since the features P_k, P and ΔP are unique for every entry in the data set, which corresponds to a specific year of observation. So, we have

$$\begin{aligned} \langle \hat{\beta}, \mathbf{x} \rangle = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_{q-m-1} x_{q-m-1} + \hat{\beta}_{q-m}(P_m - P - \Delta P) + ... \\ + \hat{\beta}_{q-1}(P_1 - P - \Delta P) + \hat{\beta}_q \Delta P \end{aligned}$$

Using the following notation

$$\begin{aligned} &\tilde{\mathbf{x}} := (1, x_1, ..., x_{q-m-1}, P_m - P, ..., P_1 - P, 0) \text{ and} \\ &\gamma := -\hat{\beta}_{q-m} - ... - \hat{\beta}_{q-1} + \hat{\beta}_q \end{aligned}$$

we can rephrase equation (3.36) as

$$\hat{E}[\hat{p}] = \frac{1}{N}\sum_{i=1}^{N} \frac{e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}}{1 + e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}} := \hat{E}[\hat{p}](\Delta P) \tag{3.38}$$

This gives us the estimated expected churn rate in function of the premium change ΔP. Since we assumed a logistic regression model, it also has the form a of a logistic curve. The following diagram depicts this shape.

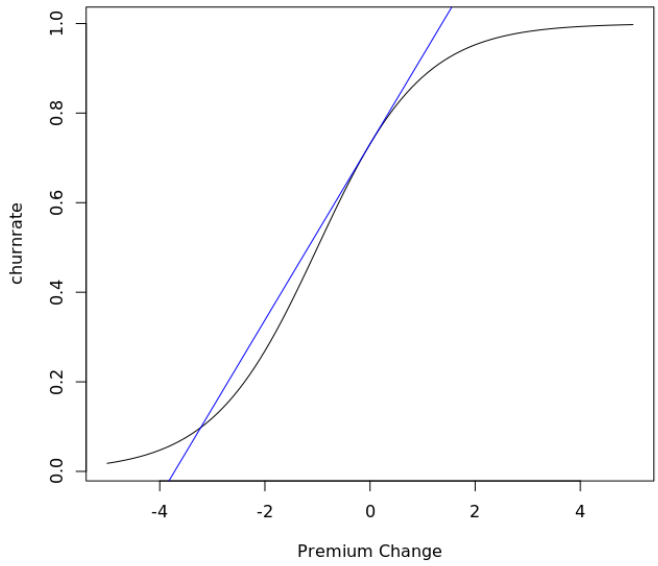

Figure 3.3: Illustration of $\hat{E}[\hat{p}](\Delta P)$ with $\langle \hat{\beta}, \tilde{x}_i \rangle = \gamma = 1$. The blue lines shows the slope of the function around 0.

We are interested in the effect of the premium change around zero, which we approximate by a linear relation using Taylor's Expansion:

$$E[\hat{p}](\Delta P) \approx E[\hat{p}](0) + \left.\frac{dE[\hat{p}](\Delta P)}{d(\Delta P)}\right|_{\Delta P = 0} \cdot \Delta P \tag{3.39}$$

We need to calculate the derivative:

$$\begin{aligned}
\frac{dE[\hat{p}](\Delta P)}{d(\Delta P)} &= \frac{d}{d(\Delta P)} \frac{1}{N} \sum_{i=1}^{N} \frac{e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}}{1 + e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}} \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{\gamma e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}}{1 + e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}} - \frac{e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}}{(1 + e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P})^2} \cdot \gamma e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P} \\
&= \frac{1}{N} \sum_{i=1}^{N} \gamma \cdot \left(\frac{e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}}{1 + e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}} \right) \cdot \left(1 - \frac{e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}}{1 + e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle + \gamma \Delta P}} \right) \\
\frac{dE[\hat{p}](\Delta P)}{d(\Delta P)}\Big|_{\Delta P = 0} &= \frac{1}{N} \sum_{i=1}^{N} \gamma \cdot \left(\frac{e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle}}{1 + e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle}} \right) \cdot \left(1 - \frac{e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle}}{1 + e^{\langle \hat{\beta}, \tilde{\mathbf{x}}_i \rangle}} \right)
\end{aligned}$$

This resembles the Gini Index of the predicted churn probability, where the observations were modified "as if" no premium change would occur. According to (8.6), the predicted churn rate changes by a factor on the premium change, which we define as follows.

Definition 12. The *premium sensitivity* for the logistic regression model is given by:

$$\epsilon := \gamma \frac{1}{N} \sum_{i=1}^{N} \hat{p}(\tilde{x}) \cdot (1 - \hat{p}(\tilde{x})) \tag{3.40}$$

So, we we have for the change in the expected churn rate:

$$\Delta E[\hat{p}] \approx \epsilon \cdot \Delta P. \tag{3.41}$$

Since γ can be calculated based on the estimated coefficients of the model and $\hat{p}(\tilde{x}$ by using the estimted model for prediction on modified inputs. We will apply this to our model Logit_Lasso in the Appendix 11.5.

3.5.2 Premium Sensitvity for GBM

Next, we want to derive the premium sensitivity for the GBM model. The challenge here is that we do not have a formula for the predicted churn probability and hence calculating the derivative is not directly possible.

However, we can approximate this term numerically.

For a function $f(x)$ the *symmetric difference quotient* is defined as

$$D_h(x) := \frac{f(x+h) - f(x-h)}{2h}$$

Applying this to the GBM, using the notation of the previous chapter, we have

$\hat{p}(\mathbf{x})$, the churn probability according the GBM-model for an individual with feature expres-

sion $\mathbf{x}$,

We set for the features of premium change ΔP:

$$\tilde{\mathbf{x}}(\Delta P) := (1, x_1, ..., x_{q-m-1}, P_m - P - \Delta P, ..., P_1 - P - \Delta P, \Delta P) \tag{3.42}$$

We then have the following estimate for the expected churn probability (estimated on the whole portfolio):

$$\hat{E}[\hat{p}] = \frac{1}{N} \sum_{i=0}^{N} \hat{p}(\mathbf{x}_i)$$

And, as a function of the premium change:

$$\hat{E}[\hat{p}](\Delta P) = \frac{1}{N} \sum_{i=0}^{N} \hat{p}(\tilde{\mathbf{x}}_i(\Delta P)) \tag{3.43}$$

We now approximate $\hat{E}[\hat{p}]'(\Delta P) = \dfrac{d\hat{E}[\hat{p}](\Delta P)}{d(\Delta P)}$ at zero by the symmetric difference quotient and get:

$$\hat{E}[\hat{p}]'(0) \approx \frac{1}{2h} \left(\sum_{i=0}^{N} \hat{p}(\tilde{\mathbf{x}}_i(h)) - \sum_{i=0}^{N} \hat{p}(\tilde{\mathbf{x}}_i(-h)) \right) \tag{3.44}$$

As this is identical with the premium sensitivity, we define the following.

Definition 13. The *premium sensitivity* for the GBM-Model is given by:

$$\epsilon := \frac{1}{2h} \sum_{i=0}^{N} (\hat{p}(\tilde{\mathbf{x}}_i(h)) - \hat{p}(\tilde{\mathbf{x}}_i(-h))) \tag{3.45}$$

Using this, we can numerically calculate the premium sensitivity and estimate the effect of premium changes. An example of this is provided in the Appendix in Sections 5.2.3 and 5.4

.

Chapter 4

Results

4.1 Comparison of Model Performance

We will now compare the models that resulted from implementing the concepts of Chapters 3.1, 3.2 and 3.3 with respect to the measures introduced in the previous section.
The overview below summarizes the models used, the respective (negative) loss functions that were maximized in the fitting process, and the model complexity measured by the number of parameters.

Model	Loss Function	# param.
Logit_Naive	Binomial Deviance	295
Logit_Selected	Binomial Deviance	68
Logit_Selected_Weighted_a	Weighted Binomial Deviance (w_i^a)	68
Logit_Selected_Weighted_b	Weighted Binomial Deviance (w_i^b)	68
Logit_Selected_Weighted_c	Weighted Binomial Deviance (w_i^c)	68
Logit_Lasso	L1-penalized Binomial Deviance	70
Class_Tree_Naive	Gini Index	4
Class_Tree_Imbalance_Weighted	Gini Index (Adjusted Priors)	11
GBM_Bernoulli	GBM-Bernoulli Loss Function	400

Table 4.1: Overview of the implemented models. Parameters set to "NA" were not counted. For GBM_Bernoulli, the number of parameters was approximated by the number of trees times the maximum number of leaves per tree.

The model Logit_Naive was quite complex, so a manual selection of variables was performed, resulting in Logit_Selected. The weighted Logit_Selected models rely on the same input variables. For Logit_Lasso, the tuning parameter λ=0.001611 was chosen such that

roughly the same number of parameters is achieved as in the Logit_Selected models and so that they are comparable in terms of model complexity. The complexity of the classification trees was not directly controlled, for GBM the maximum number of leaves was set to 4 and the number of iterations set to 100.

In order to assess the bias-variance tradeoff of the models and detect overfitting, we randomly split off a training and a test set of our large data set and evaluated the model performance measures on both. We begin with the most important performance measure, the AUC, and the optimization criterion Binomial Deviance and get the following results.

Model	AUC train	AUC test	Binomial Deviance train	Binomial Deviance test
Logit_Naive	0.794	0.791	-0.173	-0.173
Logit_Selected	0.777	0.776	-0.173	-0.177
Logit_Selected_Weighted_a	0.776	0.776	-0.189	-0.178
Logit_Selected_Weighted_b	0.778	0.776	-0.178	-0.177
Logit_Selected_Weighted_c	0.778	0.776	-0.178	-0.177
Logit_Lasso	0.790	0.789	-0.180	-0.179
Class_Tree_Naive	0.566	0.568	n.a	n.a
Class_Tree_Imbalance_Weighted	0.730	0.729	-0.190	-0.188
GBM_Bernoulli	0.837	0.835	-0.156	-0.155

Table 4.2: Overview of selected model performance measures. Total Binomial Deviance Loss was normalized by sample size.

Since the larger the AUC, the better the model performance, we immediately see, that GBM_Bernoulli is the winning model in comparison to the others. This does not come as a surprise since it incorporates the non-linear structures of the premium-differences' effect on churning, that is missed by the logit models. Contrary to a single tree, it systematically improves the prediction over a large number of interations and that way, combines many weak predictors (small trees) into one very strong predictor.
We observe that the Naive Logit model does also quite well, but this comes at the cost of high model complexity. However, the Logit Lasso model, which selected the parameters based on a penalty function, reaches almost the same results with considerably less parameters. So it seems that it is more efficient by filtering the relevant categorical levels.
The weighted logistic regression models perform a little bit worse, which makes sense, because by modifying their loss function, they were set up to optimize the Pricing Loss and not the AUC.

Next, the classification trees perform rather poorly. The naive tree reaches an AUC of 0.56, which is barely better than random guessing. This is because of the unbalanced data due to the low churn rate, which leads to the classification tree basically ignoring the minority class. We consider this result an instructive example of the pitfalls of naively fitting classification trees on imbalanced data.
The weighted tree alleviates this issue as it considers a loss function adapted by the imbalance ratio. As a result, it performs significantly better than the naive tree, but it still can't compete with the logistic regression models.
As a final observation, we see in none of the models a noteable difference between train and test AUC. So, overfitting does not seem to be an issue, which is due to the fact that we used a very large data set also for training the models.

In addition to the AUC, we listed the (negative) Binomial Deviance Loss in order to assess the impact of the modifications made to this loss function. Note that the goal was to maximize the negative Binomial Deviance Loss, so larger values correspond to better models. As outlined in the Remark in Section 6.2.1, the plain logit models (Logit_Naive, Logit_Selected) and the GBM_Bernoulli minimize the same loss function, but under different constraints for the input variable. Naturally then, these show the highest Binomial Deviance, while again, the GBM clearly is the best (as previously mentioned, it improves the Binomial Deviance over many interations). As for the modifications, the weights (b) and (c) only seem to only have a slight impact on the Deviance, which means that the optimal solution under their loss functions is very close to the optimal solution of the unweighted loss function.
The penalty of the Logit_Lasso model and the weights (a), which consider the absolute technical results, seem to have a more important effect. In fact, Logit_Selected_Weighted_a shows a Binomial Deviance Loss that is close to the one for
Classification_Tree_Imbalance_Weighted, which is a model yielded by optimizing the Gini Index and therefore a completely different loss function. For Classification_Tree_Naive, no Binomial Deviance Loss could be calculated, because it predicted probabilites of 0 and 1 (degenerate cases), for which the logarithm employed in the loss function is not defined.

The following ROC-curves give more detailed information on the performance of the models.

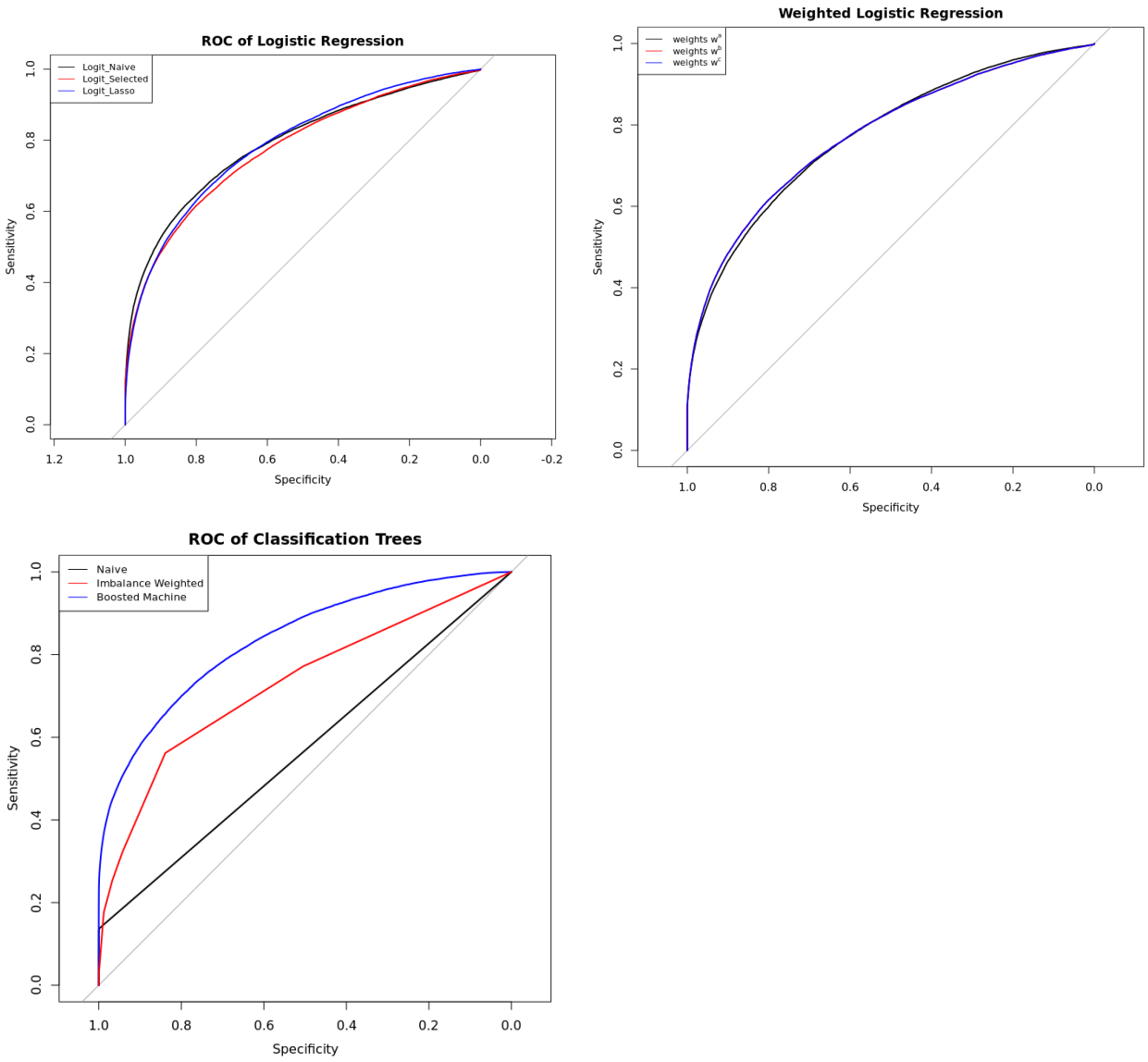

Figure 4.1: ROC-Curves of the implemented models. The blue and the red lines of the weighted logistic regression appear as identical.

While the ROC-Curves of the logit regression models are visually very close, a clear distinction can be made for those of the classification trees.

The naive classification tree model has the worst discriminatory power, its curve is a line very close to the diagonal, which corresponds to a purely random classification model. Like all other curves, it starts with a threshold of 1, yielding a sensitivity of 0 and a specificity of 1. Note that this classification tree has four leaves (see Figure 4.1), so the change of the threshold b only leads to a change in sensitivity and specificity, if it leads to a change in the

assignment rule in one of the leaves. For example, if we consider a threshold around b= 0.98 and consult the plot as well as the details in the R output 4.15, we can approximate a TP-rate of 0.0073 and a FP-rate of 0.0495, yielding a sensitivity of 0.128. This is exactly, where we see an "edge" of the black line. In fact, because we actually only have 4 observations points, this smooth line is a bit misleading. It is due to the linear interpolation between points, which is set as the plot type. Jumps or step functions would be a more adequate visualization.
The analogous observations can be made for the imbalance weighted classification tree, which has more leaves and therefore more points on the grid curve.
Finally, the curve of the GBM ("Boosted Machine") is approximately smooth and continuous, since small changes in the threshold indeed lead to smalle changes in the defined metrics. It is also clearly visible that this blue curve yields the highest AUC (regardless of how the other curves are exactly defined), which can be verified by consulting Table 3.2.

Now we assess the perfomance of the models with regard to the absolute and net pricing loss. In contrary to the previous measures, these have a very practical interpretation. The net pricing loss is the difference in the overall technical result that occurs due to the difference in effective and predicted churns. It is calculated on the assumption that the technical result for each person would remain the same in the following year. Of course, in practice, this is not the case, as claims are also random variables. But this assumption allows to isolate the effects of the churn model.
By looking at net differences, negative values represent a true loss (or profit reduction) and positive values represent a profit increase. For example, a positive loss can result from underestimating the churn rate of unprofitable insured. Or it can be a result from overestimating the churn rate of profitable clients. Also, this measure allows compensation between individuals. For example, if a model performs equally bad for very unprofitable and very profitable clients, this might compensate in such a way that the overall net result will not be much affected. The average technical result error per person, without allowing for compensation, is represented by the absolute pricing loss, which is therefore naturally much higher than the net pricing loss. In terms of absolute pricing loss, again, the classification trees perform worst and they also yield the highest net pricing loss. This means, that they underestimate the result per person by 1.88 CHF and 1.52 CHF, respectively. Generally, also a positive deviation from the expected result is a risk, eventhough management might be even more sensitive to negative losses. Also the Logit_Lasso model shows high absolute losses. We do

Model	Absolute Pricing Loss train	Absolute Pricing Loss(CHF) test	Net Pricing Loss (CHF) train	Net Pricing Loss (CHF) test
Logit_Naive	16.1978	16.177	0.000	0.149
Logit_Selected	16.669	16.636	0.000	0.156
Logit_Lasso	18.167	18.147	-0.109	-0.163
Logit_Selected_Weighted_a	16.445	16.451	0.337	0.466
Logit_Selected_Weighted_b	16.698	16.662	-0.041	0.117
Logit_Selected_Weighted_c	16.526	16.505	0.256	0.402
Classification_Tree_Naive	19.666	19.561	1.804	1.886
Class_Tree_Imbalance_Weighted	19.499	19.504	1.426	1.523
GBM_Bernoulli	16.159	16.21	0.287	0.406

Table 4.3: Model performance measured by Pricing Loss

see a compensation effect, as the net pricing loss error is comparably low. However, it represents a true loss and might be therefore more problematic than the larger positive losses of the classification trees. The plain regression models Logit_Naive and Logit_Selected perform quite well, but when expanding to the test set, their net pricing loss increases.
Finally, the results for the weighted logit models that were specifically adapted to minimize the pricing losses, are rather disappointing. Logit_Selected_Weighted_a considered the absolute technical results in the loss function, and indeed yields an absolute pricing loss that is lower than its "base" model, the Logit_Selected model. However, this does not expand to the net pricing loss, where we observe comparably high values. The model Logit_Selected_Weighted_b is again very close to Logit_Selected model in terms of performance, so it seems that the weighting in the loss function only has a minimal effect. The weighted models (b) and (c) were not expected to differ much, as the weights were similar and only differed by the way they were centered. (c) seems a bit more pessimistic in terms of technical result, as its losses are a bit higher than for (b), where we observe a negative loss on the training set. When comparing training and test set losses, it seems that the differences of the technical results of training and test set observations also affect these results. But there is also an indicator for overfitting: The weighted models that consider the technical results in their loss functions have a significantly higher increase in the net pricing loss when comparing training and test set. This could be explained the following way: Some observations and their respective features are weighted higher in the fitting process of the model than others and therefore have a more exact churn rate prediction. However, in the test set, there might be other observations with particularly high or low technical results, for which the model is then not as exact.

4.2 Premium Sensitivity

We implemented the ideas of Chapter 3.5 for two selected models and got the following key results.

Model	ϵ	$\hat{E}[\hat{p}](\bar{\Delta P})$	$\bar{\hat{p}}$
Logit_Lasso	0.08 %	5.5%	5.7 %
GBM_Bernoulli	0.025%	5.2%	5.7%

Table 4.4: Overview of estimated premium sensitivity. $\hat{E}[\hat{p}](\bar{\Delta P})$ corresponds to the expected churn rate approximated via the premium sensitivity for the average premium change $\bar{\Delta P}$ and $\bar{\hat{p}}$ correponds to the effective churn rate predicted by the model with given premium changes that average to $(\bar{\Delta P})$. These values were calculated on the test set.

The calculated premium sensitivities ϵ do vary, but are within the same range. In order to understand how "good" this linear approximation is, we applied it to the average premium change of the data set and compared the estimated churn rate with the effective predicted churn rate. Here, one major question arised: What is a reasonable way to define ΔP? Generally, using the average premium change as an input of these models will not lead to the average predicted churn rate (the same applies for all other variables). This is due to the non-linear nature of the functions, for example the logistic function used in the Logit-models, and has nothing to do with the linear approximation by the sensitivity-factor. Another way to illustrate this, is by noting that two different distributions of premium changes in the portfolio might lead to the same average premium change, but to different average predicted churn rates.

It would be necessary to build a 'weighted' average premium change in order to get more exact predictions for the resulting churn rate. However, looking at the concrete numbers for Logit_Lasso, the difference doesn't seem too big and seems suitable as a rule of thumb for pricing and management decisions.

GBM_Bernoulli produces a larger difference between average estimated churn rate and its linear approximation. In order to completely understand the reasons for it, it would be needed to further analyse the role of the step width h that is used for the approximation of the derivative and, more generally, the structure and shape of the curve of $\hat{E}[\hat{p}](\Delta P)$ as well as the difference produced by using the average of (ΔP) to predict the average churn rate. Due to the 'black box' nature of the GBM-model, this can not easily be seen nor be analytically derived, but for example the partial dependence (plots) could be used as a basis.

There, we observed asharp drop of the churn probability around a premium change of

0, which indicates that the choice for h would have a big impact on the calculated ϵ. To summarize, it seems that the premium sensitivity estimated via the logistic regression model provides more reliable and interpretable results, but with the drawback that the model itself is less reliable, as seen in the Comparison of the Model Performance. On the other hand, GBM_Bernoulli is the best model in terms of prediction power, but seems less suitable for assessing premium sensitivity in a robust and simple way.

4.3 Conclusion

In this thesis, we developed churn models for a Swiss health insurer, aiming to take into consideration an actuarial pricing perspective. The application in this specific context had two important aspects: First, due to the regulated standartized insurance cover, competition and churning behaviour was presumably mainly driven by premiums. We reflected this important factor by including information on the competitors's premiums as explanatory variables in the model. We also investigated the possiblity to develop a measure for premium sensitivity that could easily be used in practice when setting the (market) premiums.
The second particularity is the risk sharing between the insured in the portfolio becaus of a unitary premium, which is prescribed by law. This meant that actuarial premiums had implicit assumptions on portfolio structure and therefore a churn model needed to be integrated in the pricing model. This raised the question of requirements to a potential churn model from an actuarial pricing perspective. We adressed it in this thesis by developing an alternative measure to the standard model perfomance metrics: the pricing loss. In addition, we introduced weights to a loss function of one of the models in order to incorporate the impact on the pricing loss.

These ideas were applied on two families of models: Logistic regression models and classification trees (including an ensemble learning meathod, the Gradient Boosting Machine). In the first place, we were not interested in developing the most accurate model, which is why, for example, variable transoformations, cross-validation and (hyper) parameter selection was not given much attention. Instead, we were interested in the effects of different loss functions that were optimized in order to fit the respective models and comparing them with respect to the actuarial pricing perspective.

Besides overcoming practical challenges resulting from the presence of many categorical

variables in the data, the imbalance of the data and the effort of matching external data sets, the main findings were the following:

All models and analyses made clear that premium differences to the competitor were the most important variables. So, no matter what churn model used, if it will be applied in a "managed competition" context similar to Swiss health mandatory health insurance, this key information must be contained in the explanatory variables.
Theoretically, after integrating this information, the premium sensitivity can be derived, which can be used as a rule of thumb to understand premium change effects. In practice, the results seem sensible, however, they are easier obtained and understood in the logistic regression model than in the GBM.

When comparing different models and loss functions with respect to their performance, we found that the GBM achieved the highest AUC. It was the only model that was able to incorporate the non-linear relation of the premium differences and the churn probability. Eventhough the model itself is quite complex, its implementation was relatively easy. However, its pricing (net) loss was at best average. This could indicate that the resulting churn probability is biased, but could also be a random effect, as the technical result seems to have strong outliers (or might have a heavy-tailed distribution).
The results for the classification trees were disappointing and showed that unbalanced data sets pose a severe problem for classification trees, if they are not adequately incorporated in the model.
The logistic regression models yielded interesting results:The Naive Logit Regression model, which included all variables, was the most accurate in terms of AUC, and the Lasso Logit Regression model, which had a loss function modified by a penalty term, yielded a similar result with much less parameters.
In terms of net pricing loss, allowing the compensation of positive and negative technical results seem to make this measure more tolerant towards differences in the churn rate. In fact, the best result is achieved by the Logistic Regression where the loss function was modified by weights. Since the other two weighted Logistic Regressions show particularly high losses, we conclude that: a) The right definition of the weights can indeed improve the pricing loss result, but b) one should be cautious when defining these weights as the "wrong" definition can lead to the opposite effect.
Having said this, all other regression models do achieve better pricing net losses than the

GBM.

Bringing together all these findings, we can conclude that application in a pricing context can lead to favor other churn models than in a CRM-context:

- GBM is the most accurate one, but logistic regression models generally lead to smaller pricing losses
- GBM reflects the churn probability with respect to premium changes more accurately, but this makes it a priori more difficult to quantify and understand premium sensitivity in practice.

4.4 Outlook

The conclusions drawn in the previous section indicate which paths could be particularly interesting to be followed in subsequent research. Possible extensions and expansions concern all aspects of modeling adressed in this thesis:
As we have seen that premiums of the competition play a major role, a refinement of the current model would be achieved if the simplifications made when matching the categories of the premiums were replaced by the exact matches. Also, by taking the 5 top competitors, we only modelled 50 % of the market. So, there seems some room for improvement which would be expected to lead to improvement in the prediction accuracy.
In terms of models used, the results of the thesis suggest that it should be worthwhile to adapt the loss functions in such a way that the final performance measure (in our case: the pricing loss) is improved. This could be done by directly using the pricing loss as optimization criterion. Of course, the pricing loss itself is only one option for an 'actuarial' assessment of churn model performance. Depending on the context, using for example the expected claims, some definition of value of the insureds or a defined utility of the insureds could also make sense. Reflecting the assymetry in the perception of losses and gains could also be an improvement from a practical point of view.
Then, this thesis only focused on the most fundamental models, so in a next step other models should be explored, too. For example, it would be interesting to see whether combining a neural network with a loss function tailored to pricing purposes would be a candidate for an all-purpose model that can be employed in CRM and in a pricing context.
With regard to premium sensitivity, an application on a GBM or other non-linear models

requires to understand the properties of the function (such as differentiabilty) and the optimal numerical methods to approximate the derivative. For practical use, it would be useful to caracterize those groups of insured that are particularly sensitive to premium increases. Finally, from a broader perspective, the other component of the churn model, the new business model, shouldn't be forgotten either. Most of this thesis' conclusions and ideas can be transferred, and it is important from a pricing perspective to ultimately have one consistent optimiziation and evualuation scheme for both parts of the portfolio model.

Bibliography

[1] Athanassopoulos, A. (2000): *Customer satisfaction cues to support market segmentation and explain switching behavior.* Journal of Business Research, 47(3), p.191 - 207

[2] Beck, K. (2004): *Risiko Krankenversicherung: Risikomanagement in einem regulierten Krankenversicherungsmarkt.* University of Zurich. Wirtschaftswissenschaftliche Fakultät, p. 238

[3] Bellani, C. (2019): *Predictive Churn Models in Vehicle Insurance.* Master Thesis, Universidade Nova de Lisboa.

[4] Bolancé, C., Guillen, M., Padilla Bareto, A. (2016): *Predicting Probability of Customer Churn in Insurance.* International Conference on Modeling and Simulation in Engineering, Economics and Management, p. 82-91

[5] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C.J. (1984): *Classification and Regression Trees.* Chapman & Hall/CRC

[6] van Dijk, M., Pomp, M., Douven, R., et al. Wagner, J. (2008): *Consumer price sensitivity in Dutch health insurance.* International Journal of Health Care Finance and Economics, 8(4), p. 225-244

[7] Daily-Amir, D., Albrecher, H., Bladt, M., Wagner, J. (2019): *On Market Share Drivers in the Swiss Mandatory Health Insurance Sector.* Risks, 7(4)

[8] Donkers, B., Franses, P., Verhoef, P., (2003): *Selective Sampling for Binary Choice Models* Journal of Marketing Research, 40(4), p. 492-497

[9] Dormont, B., Geoffard, P.Y., Lamiraud, K., (2007): *The influence of supplemen-*

tary health insurance on switching behaviour: evidence on Swiss data. IEMS Working Paper No. 07(02), Universität Lausanne

[10] Douven, R., Lieverdink, H., Ligthart, M., Vermeulen, I., (2007): *Measuring annual price elasticities in Dutch health insurance: A new method.* CPB Discussion Papers, CPB Netherlands Bureau for Economic Policy Analysis.

[11] Fawcett, T. (2006): *An Introduction to ROC analysis.* Pattern Recognition Letters, 27(8), p. 861-874

[12] Federal Office of Public Health (2021): *Krankenversicherungsprämien* https://opendata.swiss/de/dataset/health-insurance-premiums, accessed in October 2021.

[13] Fleiss, J.L., Levin, B., Paik, M.C. (2003): *Statistical Methods for Rates and Proportions* 3rd Edition, Wiley Series in Probability and Statistics.

[14] Frank, R.G., Lamiraud, K., (2009): *Choice, price competition and complexity in markets for health insurance.* Journal of Economic Behavior & Organization, 71(2), p. 550-562

[15] Friedman, J.H. (2001): *Greedy function approximation: A gradient boosting machine.* The Annals of Applied Statistics, 29(5), p. 1189-1232

[16] Friedman, J.H., Popescu, B.E. (2008): *Predictive learning via rule ensembles.* The Annals of Applied Statistics, 2(3), p. 916-54

[17] Gerber, G., Le Faou, Y., Lopez, O., Trupin, M. (2020): *The Impact of Churn on Client Value in Health Insurance, Evaluation Using a Random Forest Under Various Censoring Mechanisms* Journal of the American Statistical Association, 116(536), p. 2053-2064

[18] Guelman, L., Guillén, M., Pérez-Marín, A.M. (2012): *Random Forests for Uplift Modeling: An Insurance Customer Retention Case.* Modeling and Simulation in Engineering, Economics and Management. Lecture Notes in Business Information Processing, 115, p. 123-133

[19] Guelman, L., Guillén, M., (2014): *A causal inference approach to measure price*

elasticity in Automobile Insurance. Expert Systems with Applications, 41(2), p. 387-396

[20] Günther, C., Tvete, I. F., Aas, K., Sandnes, G.I., Borgan, Ø., (2014): *Modelling and predicting customer churn from an insurance company.* Scandinavian Actuarial Journal, 2014(1), p.58-71

[21] Hastie, T., Tibshirani, R., Friedman, J. (2017): *The Elements of Statistical Learning: data mining, inference and prediction* Springer, Second Edition, 12th Printing.

[22] Henao Madrigal, M., Restrepo Tobón, D., Laniado, H. (2020): *Customer Churn Prediction In Insurance Industries: A Multiproduct Aproach* Doctoral dissertation, Universidad EAFIT

[23] Huigevoort, C. (2015): *Customer churn prediction for an insurance company.* Master Thesis, Eindhoven University of Technology

[24] Jain, H., Khunteta, A., Srivastava, S. (2021): *Telecom churn prediction and used techniques, datasets and performance measures: a review.* Telecommunication Systems: Modelling, Analysis, Design and Management, 76(4), p. 613-630

[25] Koornstra, M. (2021): *Churning behavior in the liberalizing Dutch health care market.* Master Thesis, Rijksunversiteit Groningen

[26] Krikler, S., Dolberger, D., Eckel, J., (2004): *Method and tools for insurance price and revenue optimisation* Journal of Financial Services Marketing, 9(1), p. 68-79

[27] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., (2020): *Focal Loss for Dense Object Detection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), p. 318-327

[28] Liu, H., Zhou, M., Lu, X.S. , Yao, C. (2004): *Weighted Gini index feature selection method for imbalanced data.* 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), p. 1-6

[29] Mayorga, W., Torres, D. (2017): *A practical model for pricing optimization in car insurance* ASTIN and AFIR/ERM Colloquia, 20-24 August 2017, Panama

[30] Morik, K., Köpcke, H. (2004): *Analysing Customer Churn in Insurance Data A*

Case Study Knowledge Discovery in Databases: PKDD 2004. Lecture Notes in Computer Science, 3202, p. 325-336

[31] Niklowitz, M. (2021): *Krankenversicherer: Erkennen, wenn ein Kunde auf dem Absprung ist* https://www.handelszeitung.ch/insurance/digitalisierung-krankenversicherer-erkennen-wenn-ein-kunde-auf-dem-absprung-ist (accessed 19.08.2021)

[32] Nuscheler, R., Knaus, T. (2005): *Risk selection in the German public health insurance system.* Health Economics, 14(12), p. 1253-1271

[33] Pendzialek, J.B., Simic, D., Stock, S. (2016): *Differences in price elasticities of demand for health insurance: a systematic review.* The European Journal of Health Economics, 17(1), p. 5-21

[34] Ridgeway, G., (2020): *Generalized Boosted Models: A guide to the gbm package* https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf (accessed on January 10, 2022)

[35] Risselada, H., Verhoef, P.C., Bijmolt, T.H. (2010): *Staying power of churn prediction models.* Journal of Interactive Marketing, 24(3), p. 198-208.

[36] Schmitz, H., Ziebarth, N.R. (2011): *In absolute or relative terms? How framing prices affects the consumer price sensitivity of health plan choice.* Ruhr economic papers, 304

[37] Scriney, M., Nie, D., Roantree, M. (2020): *Predicting Customer Churn for Insurance Data* 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 1417, 2020, Proceedings, p. 256-265.

[38] Schut, F.T., Hassink, W.H. (2002): *Managed competition and consumer price sensitivity in social health insurance.*Journal of Health Economics, 21(6), p. 1009-29

[39] Schut, F.T., Gress, S., Wasem, J. (2003): *Consumer price sensitivity and social health insurer choice in Germany and the Netherlands.* International Journal of Health Care Finance and Economics, 3(2), p. 117-138

[40] Schwarze, J., Andersen, H.H. (2001): *Kassenwechsel in der Gesetzlichen Krankenversicherung: Welche Rolle spielt der Beitragssatz?* Diskussionspapier Nr. 267.

[41] Siemes, T. (2016): *Churn prediction models tested and evaluated in the Dutch indemnity industry* Master Thesis, Open University of the Netherlands

[42] Smith, K.A., Willis, R.J., Brooks, M. (2000): *An analysis of customer retention and insurance claim patterns using data mining: a case study* Journal of the Operational Research Society, 51(5), p. 532 - 541

[43] Spiteri, M., Azzopardi, G. (2018): *Customer Churn Prediction for a Motor Insurance Company* Thirteenth International Conference on Digital Information Management, 2018

[44] Sucki, O. (2019): *Predicting the customer churn with machine learning methods - CASE: private insurance customer data.* Master's Thesis, Lappeenranta-Lahti University of Technology LUT

[45] Tamm, M., Tauchmann, H., Wasem, J., Gress, S. (2007): *Elasticities of Market Shares and Social Health Insurance Choice in Germany: a dynamic panel data approach.* Health Economics, 16(3), p. 243-256

[46] Telser, H., Becker, K., (2008): *Entwicklung von Methoden zur Analyse der Daten des Risikoausgleichs.* Schweizerisches Gesundheitsobservatorium, Forschungsprotokoll 6

[47] Therneau, T.M., Atkinson, E.J. (2019): *An Introduction to Recursive Partitioning Using the RPART Routines* https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf (accessed 3 January 2022)

[48] Tianyuan, Z., Moro, S. (2021): *Research Trends in Customer Churn Prediction: A Data Mining Approach.* Trends and Applications in Information Systems and Technologies, 1365, p. 227-237

[49] Tutz, G., Gertheiss, J. (2016): *Regularized regression for categorical data.* Statistical Modelling, 16(3), p. 161-200

[50] Van den Poel, D., Larivière, B. (2004): *Customer attrition analysis for financial services using proportional hazard models.* European Journal of Operational Research, 157(1), p.196-217

[51] Wüthrich, M.V., Buser, C. (2021): *Data Analytics for Non-Life Insurance Pricing.* Lecture Notes. Version October 27, 2021, https://ssrn.com/abstract=3951109

[52] Yeo, A.C., Smith, K.A., Willis, R.J., Brooks, M. (2001): *Modeling the Effect of Premium Changes on Motor Insurance Customer Retention Rates Using Neural Networks.* Computational Science - ICCS 2001, Lecture Notes in Computer Science, 2074

[53] Yeo, A.C., Smith, K.A., Willis, R.J., Brooks, M. (2002): *A mathematical programming approach to optimise insurance premium pricing within a data mining framework.* Journal of the Operational Research Society, 53(11), p. 1197-1203

[54] Zhang, R., Li, W., Mo, T., Tan, W. (2017): *A Deep and Shallow Model for Insurance Churn Prediction Service.* 2017 IEEE 14th International Conference on Services Computing