CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Bachelor's Thesis

2023 **Andrej Židek**

CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Application of the logit leaf algorithm for customer churn prediction in the energy distribution industry in the Czech Republic

Bachelor's Thesis

Author of the Thesis: Andrej Židek

Study programme: Economics and Finance

Supervisor: prof. Ing. Karel Janda, Dr., Ph.D., M.A.

Year of the defence: 2023

Declaration

1. I hereby declare that I have compiled this thesis using the listed literature and resources only.
2. I hereby declare that my thesis has not been used to gain any other academic title.
3. I fully agree to my work being used for study and scientific purposes.

In Prague on 24th July 2023 Andrej Židek

References

ŽIDEK, Andrej. *Application of the logit leaf algorithm for customer churn prediction in the energy distribution industry in the Czech Republic*. Praha, 2023. 44 pages. Bachelor’s thesis (Bc). Charles University, Faculty of Social Sciences, Institute of Economic Studies, Department of Social Sciences. Supervisor prof. Ing. Karel Janda, M.A., Dr., Ph.D.

Length of the Thesis: 58 180 characters

Abstract

The thesis investigates determinants of losing customer (customer churn) in the Czech energy sector. For this purpose, the data from MND Energie, a.s., one of the largest Czech energy suppliers, on average consumption, tariff, and sociodemographic characteristics about 9254 of their customers whose natural gas contracts terminated at the end of 2019 are used. The main goal of this thesis is to build a model capable of predicting probability of non-renewal of the individual customers' contracts. Before the contract termination date, some of the customers randomly selected from the dataset were directly notified of the possibility of a new fixed-price contract. The thesis, in compliance with its main goal, evaluates the influence of this treatment on the churn probability. The experiment has so far only been carried out in 2019. Thus, the thesis deals with supervised machine learning task performed on cross-sectional data. The logit leaf model (LLM) was chosen as the way of obtaining the desired predictions. The LLM algorithm used in this thesis was published in 2018 and it builds on previous research in this area. Its main contribution lies in combining the two generally accepted approaches, decision trees and logistic regression, in order to eliminate their disadvantages. LLM's performance was compared with the performance of its two building blocks used individually. The results were compared with relevant literature.

JEL Classification	C53, L97, Q49
Keywords	logit leaf algorithm, customer churn prediction, energy industry
Title	Application of the logit leaf algorithm for customer churn prediction in the energy distribution industry in the Czech Republic
Author's e-mail	13903645@fsv.cuni.cz
Supervisor's e-mail	karel-janda@seznam.cz

Abstrakt

Práce zkoumá determinanty migrace zákazníků v českém energetickém sektoru. Za tímto účelem poskytla společnost MND Energie, a.s., jeden z největších českých dodavatelů energií, databázi sestávající z informací (např. průměrná spotřeba, tarif, sociodemografické charakteristiky) o 9254 zákaznících, kterým ke konci roku 2019 skončila smlouva na odběr zemního plynu. Hlavním cílem práce je sestavit model schopný predikovat pravděpodobnost neprodloužení smluv jednotlivých zákazníků. Před datem ukončení smlouvy byla část zákazníků, náhodně vybraných z databáze, přímo informována o možnosti uzavření nové smlouvy se zafixovanou cenou. Práce v souladu se svým hlavním cílem vyhodnocuje vliv tohoto postupu na pravděpodobnost přechodu k jinému dodavateli. Experiment byl zatím proveden pouze v roce 2019. Práce se tedy zabývá úlohou strojového učení pod dohledem prováděnou na průřezových datech. Jako způsob získání požadovaných předpovědí byl zvolen logit leaf model (LLM). Algoritmus byl publikován v roce 2018 a navazuje na předchozí výzkum v této oblasti. Jeho hlavní přínos spočívá v kombinaci dvou obecně uznávaných přístupů, rozhodovacích stromů a logistické regrese, s cílem eliminovat jejich nevýhody. Výkonnost LLM byla hodnocena ve srovnání s výkonností jeho dvou stavebních prvků použitých samostatně. Výsledky byly porovnány s příslušnou literaturou.

JEL Klasifikace	C53, L97, Q49
Klíčová slova	logit leaf algoritmus, predikce migrace zákazníků, energetický sektor
Název práce	Předpověď migrace zákazníků v českém energetickém sektoru pomocí logit leaf algoritmu
E-mail autora	13903645@fsv.cuni.cz
E-mail vedoucího práce	karel-janda@seznam.cz

Acknowledgement

I would like to express my gratitude to the thesis supervisor, prof. Ing. Karel Janda, M.A., Dr., Ph.D., for his professional guidance.

I am also deeply thankful to Mgr. Adam Hloušek, the pricing manager at MND Energie, a.s., for introducing me to the aspects of the analysed database and smooth negotiation.

Furthermore, I extend my appreciation to my dear friend and colleague Bc. Cyril Škorvaga, who accompanied me throughout my studies.

This thesis is part of a project GEOCEP that has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 870245.

Table of Contents

Introduction

The migration of customers to competitors is a global phenomenon that is being addressed by companies in various industries.

This thesis aims to analyse the customer database of a Czech energy supplier in order to predict loss of customers (churn) and reveal the relationships between customer characteristics and churn propensity. In addition to customer characteristics, random assignment of customers to groups that differ in how they are treated by management in terms of retention policy is also used to refine the predictions.

The thesis is narrowly focused on natural gas customers. According to statistics from the Ministry of Industry and Trade of the Czech Republic, more than 2.6 million households in the Czech Republic consume natural gas. According to OTE (state owned operator of the Czech short-term energy market), currently the ten largest suppliers cover the majority of the Czech natural gas market. Compared to previous years, this is a relatively diversified and competitive market. As recently as 2010, one supplier controlled more than 60 % of the market. This gradual transformation has its origins first in the liberalisation and privatisation of the energy sector after the fall of the communist regime in 1989 and then in the antimonopoly regulations associated with the Czech Republic's accession to the European Union in 2004. From this point onwards, the Czech Republic's energy policy was influenced by EU directives (Vlček & Černoch (2013)). Along with this development, customer churn rates increased. According to data from OTE, more than 40 % of all natural gas customers in the Czech Republic have switched to new suppliers since the beginning of liberalisation. Between 2015 and 2022, the annual number of churning customers increased by more than 37 %. This trend is, among other things, reinforced by the simplification of the administrative process of supplier switching and the variable price of natural gas. However, compared to Western European countries, the Czech Republic has lower churn rates on average. For example, when comparing statistics published by OTE and the Department for Energy Security of the Government of the United Kingdom, it can be seen that over the past decade, the average annual churn rate in the Czech Republic has been around 7 %, while in the United Kingdom this value has typically exceeded 10 %.

The implications of increased churn rates prompt energy suppliers to focus on churn prediction and customer retention strategies. Predicting churn accurately can enable suppliers to take proactive measures to retain at-risk customers, address their concerns, and offer personalized incentives to encourage loyalty. Identifying the key characteristics associated with churn can also help suppliers tailor their marketing and customer service efforts more effectively to meet the specific needs and preferences of different customer segments.

Investing in reliable churn prediction models can lead to substantial cost savings (Wahul et al. (2023)) for energy companies by reducing the expenses incurred in acquiring new customers to replace those lost due to churn. Additionally, a robust churn prediction system can help suppliers allocate their resources more efficiently by prioritizing retention efforts for high-value customers who are more likely to churn.

Furthermore, effective churn prediction can enhance customer satisfaction levels as suppliers become more proactive in resolving issues and providing better services. Satisfied customers are more likely to remain loyal and even recommend the supplier to others, contributing to positive word-of-mouth marketing and organic customer growth.

In conclusion, the increasing churn rates within the Czech natural gas market present a significant challenge for energy suppliers. The evolving market dynamics, coupled with the competitive landscape, make churn prediction and customer retention strategies critical for the sustainable growth and success of energy companies. By delving into the customer database and analysing churn patterns, this thesis seeks to provide valuable insights that can assist suppliers in developing more effective churn mitigation strategies and ensuring long-term customer loyalty.

Literature review

Introduction

Customer churn is considered a crucial problem faced by companies in various industries. The energy sector is not an exception. Churn refers to the loss of customers over a period of time, which can have a significant impact on a company's revenue and profitability. To mitigate its effects, companies in the energy industry have explored different methods and models to predict customer churn and develop effective retention strategies.

Tackling attrition efficiently requires the management to understand the determinants of churn and identify customers with high probability of leaving. Serving loyal customer base helps the management to understand needs of individual customers and supports the good name of the company through positive word of mouth. These factors among others cause several times lower cost of retaining an existing customer compared to the cost of acquiring a new customer (Torkzadeh et al. (2006)).

Customer churn has been tackled from two different angles in previous research. On the one hand, researchers focus on improving customer churn prediction models in which more complex models are being developed and proposed in order to boost the predictive performance. On the other hand, researchers want to understand the cause of customer churn, which requires more comprehensible models. A trade-off occurs between these two goals.

Having structures suitable for this problem, decision trees and logistic regression have been considered popular methods in customer churn prediction. Recently, more complex methods such as random forests, support vector machines or neural networks were applied in studies concerned with datasets from various sectors. While achieving significantly better predictive results in comparison with decision trees and logistic regression, results of these models are more difficult to interpret, which is in line with the trade-off mentioned above. The management of MND Energie, a.s. indicated a preference for a more comprehensible model in order to not only identify customers with high probability of churning but also better understand the reasons for such decision. Understanding these may allow combating attrition efficiently. This literature review aims to summarize the research conducted in this area and highlight the different approaches taken by researchers to address this problem.

Review of Customer Churn Models

Relatively simple models

Several studies have explored different benchmark models to predict customer churn. Ballings & Van Den Poel (2012) employed logistic regression and decision trees in combination with bagging for their analysis of the entire customer database of a newspaper company. Rather than revealing predictors significantly correlated with the target variable and discussing possible reasons for churn, the authors focused on the relationship between the length of customer event history included in the examined dataset and the classification performance of individual models. Apart from showing superior performance of bagged trees compared to the remaining two methods mentioned, they concluded that the length of the predictors period is logarithmically related to the predictive performance (measured by the area under the receiver operating characteristic curve).

Logistic regression assumes that the expected value of a target variable with binomial distribution equals the value of the logistic function of the linear combination of the predictors and the unknown regression parameters. Coussement et al. (2010) relaxed the linearity assumption by employing generalized additive models in order to allow for different functional forms of the predictors. After their analysis of the data provided by the largest Belgian newspaper publishing company, the authors concluded that their approach reached higher predictive performance than logistic regression. Furthermore, they estimated increase in company's profits in case the management started using churn prediction model proposed by them. The most important contribution of this approach lies in the ability to visualize non-parametric relationships between predictors and the target variable.

Moeyersoms & Martens (2015) employed decision trees and logistic regression to include high-cardinality attributes – ZIP codes, family names and bank account numbers, in predicting customer churn in the energy sector. They analysed a database of a large energy supplier in Belgium which contained information (including high-cardinality data) on more than 1 million customers. In order to achieve inclusion of this type of data in the models, the authors applied dummy encoding, systematic grouping and creating affiliated continuous variable as methods of data transformation. Regression parameters and tree visualization were not disclosed. On the other hand, the authors focused on the relationship between the predictive performances of the models and the chosen data transformation method. The main conclusion is that inclusion of high-cardinality variables boosts predictive performance of the classification models.

More complex models

Ensemble algorithms are methods with high computational complexity, which are based on single algorithms such as decision trees. Such algorithms combine results of multiple models with the same specification but different training sets and selected features. Ensemble methods boost the predictive performance, often at the cost of worse interpretability.

De Bock & Van Den Poel (2012) built on the Coussement et al. (2010) study and proposed a similar approach based on generalized additive models. In their model, the logistic function argument contained, in addition to the sum of the unknown functional forms of the continuous predictors, a linear combination of the dummy-coded components of categorical variables and the unknown regression parameters. This logistic semi-parametric model specification was used as a core of an ensemble method. The authors applied their algorithm on data from several industries and shown its highly competitive performance in comparison with other benchmark methods by evaluating predictive performance metrics such as area under the receiver operating characteristic curve (AUC) and top-decile lift (TDL). Of the six datasets analysed, the resulting parameters and non-parametric relationships were published only for the financial services company dataset, as a part of the authors' case study. In this case study they created two instruments for their ensemble model interpretability – generalized feature importance scores (calculated as contributions of individual features to the chosen predictive performance metric), and bootstrap confidence intervals for smoothing splines (allowing for graphical display of the non-parametric relationships between continuous predictors and the target variable). The overall intention of the authors was to reconcile performance and interpretability in customer churn prediction.

Coussement & De Bock (2013) investigated the beneficial effect of applying ensemble models by comparing simple machine learning algorithms to their ensemble counterparts in terms of predictive performance. Analysing bwin's database of poker players, the study focused on customer churn prediction in the online gambling industry. In this study, decision trees and generalized additive models were benchmarked against affiliated ensemble methods – random forests and GAMens. Apart from showing better predictive performances of the ensemble methods, the authors revealed feature importance scores resulting from the ensemble methods.

Classification algorithms such as support vector machines or neural networks are ranked among the more advance methods. Application of these approaches allows for more precise analysis of linearly inseparable data. While support vector machine uses mapping function to assign a point in space to every observation from the training set so that the gap between points assigned to observations from different classes is maximized, neural networks aim to simulate biological brain responding to external signals. That is achieved through building a network of nodes (simulating neurons) characterized by an activation function which transforms sum of weighted signals entering a neuron into an output signal to be passed further, a network architecture which specifies the number of neurons and layers in the model together with the manner in which they are connected, and a specification of input signals' weights setting. Such advanced methods were employed in the churn literature. Chen et al. (2012) proposed a hierarchical multiple kernel support vector machines to predict churn using longitudinal behavioural data, while Sharma & Kumar Panigrahi (2011) used a neural network-based approach to predict customer churn in the cellular network services industry. Keramati et al. (2014) provided complex churn analysis of a telecommunication market customer database. Several advanced data mining techniques such as support vector machines or neural networks were employed along with simple techniques such as decision trees.

De Caigny et al. (2018) proposed a new hybrid classification algorithm for customer churn prediction based on combining logistic regression and decision trees based on the underlying idea of elimination of individual methods' weaknesses. The study demonstrated that the proposed algorithm outperforms its building blocks – simple machine learning algorithms, in predicting customer churn, while sustaining acceptable level of interpretability. The authors' goal was to present a model that is a possible compromise in terms of high predictive performance and sufficient interpretability. Their algorithm will be further discussed in the Methodology chapter.

In addition to predicting churn, some studies have focused on identifying the factors that influence gas consumption in the energy industry. Tang et al. (2014), for example, assessed the impact of derived behaviour information on customer attrition in the financial service industry. Moitra et al. (2020) used a long short-term memory (LSTM) approach to predict crude oil prices. Raju et al. (2022) proposed an approach for demand forecasting in the steel industry using advanced machine learning techniques, while Ofori-Ntow Jnr et al. (2021) focused on electricity demand forecasting by introducing a hybrid ensemble intelligent model based on artificial neural network combined with other more complicated data mining methods.

Review on the churn determinants in the energy sector

Despite the high number of studies addressing the churn issue, we encounter a significant lack of relevant literature discussing influential customer features in relation to attrition in the energy sector. This is due to two reasons: (i) low data availability, (ii) authors publish predictive performance metrics rather than parameters and relationships resulting from their models.

Moeyersoms & Martens (2015) conducted a case study in churn prediction in the energy sector. Apart from continuous (age, average amount of bill, contacts with company) and traditional nominal (gender, type of contract, package, payment method) variables, they also explored the inclusion of high-cardinality attributes in predictive model and found that it improved the performance of the model. Specifically, they compared the performance of a model that excluded high-cardinality attributes to one that included them. They found that the inclusion of high-cardinality attributes such as ZIP codes and bank account numbers as determinants of customer churn in the energy sector resulted in a significant improvement in the performance of the model. They concluded that high-cardinality attributes should be included in predictive models to improve churn prediction in the energy sector. Nevertheless, the results of the algorithms applied by them were not disclosed.

Studies analysing datasets from other industries have in most cases found customer behavioural data to be significant determinants of churn. For example, when analysing customer databases from the banking sector, De Bock & Van Den Poel (2012) revealed the significance of features related to account services usage, such as the checking account balance, the average amount of credit transactions or the total number of debit transactions. Chen et al. (2012) in turn identified volume and frequency as the most discriminative predictors of customer churn in the food industry. Authors of numerous studies concerned with predicting churn did not publish the parameters estimated by the model, but only measures of the predictive ability of the model. Studies focusing on understanding churn from a managerial perspective rather than maximizing predictive performance (Gustafsson et al. (2005); Hansen et al. (2013)) often identify factors such as customer satisfaction and calculative commitment as key drivers of churn.

It is important to note that features shown to be influential in terms of churn in one sector cannot automatically be considered significant in other sectors. The results may depend strongly on the sector and the country from which the database originates.

Gaps in existing literature

Despite the different models proposed to predict customer churn and the factors that influence gas consumption in the energy industry, some gaps remain in the existing literature. First, most of the studies have focused on the predictive modelling aspect of churn prediction, with limited attention given to the development of effective retention strategies. Second, there is a lack of research on the use of deep learning approaches in churn prediction in the energy industry, despite the success of these models in other industries. Third, there is a need to investigate the effect of customer segmentation on churn prediction, as different segments may exhibit different churn behaviours.

Overall, the studies mentioned in this literature review have used various models and techniques to predict customer churn, mainly focusing on predictive performance. Apart from aforementioned, these studies suffer from certain gaps in relation to the research object of this thesis. For instance, most of the studies reviewed in this paper focused on predicting customer churn based on customer behaviour data, while few studies explored the impact of contextual factors such as economic factors, weather conditions, and energy policy on customer churn. Additionally, some of the studies focused on forecasting demand rather than predicting churn, and the applicability of their findings to customer churn prediction in the energy industry is uncertain.

Summary of Main Issues in the Churn Modelling

Customer churn is a pressing issue faced by companies across industries, including the energy sector. Churn, the loss of customers over time, can have a significant impact on a company’s revenue and profitability. To tackle this problem, energy companies have explored various methods and models to predict customer churn and develop effective retention strategies.

Efficiently addressing attrition requires understanding the factors that drive churn and identifying customers with a high likelihood of leaving. Serving a loyal customer base not only helps understand individual customer needs but also promotes positive word-of-mouth, benefiting the company’s reputation. Moreover, retaining existing customers is generally more cost-effective than acquiring new ones.

Previous research has approached customer churn from two angles: improving churn prediction models and understanding the causes of churn. More complex models, such as random forests, support vector machines, and neural networks, have shown better predictive performance compared to traditional methods like decision trees and logistic regression. However, these advanced models may be less interpretable, creating a trade-off between predictive power and comprehensibility.

In conclusion, this literature review highlights the need for more research on customer churn prediction in the energy industry. Specifically, future studies should focus on interpretation of customer features relationships towards their churn behaviours and the development of models that consider both customer behaviour data and contextual factors to improve churn prediction accuracy. Additionally, more studies should be conducted on predicting customer churn specifically in the energy industry rather than relying on research from other industries.

The literature review reveals that various machine learning algorithms have been applied by Moeyersoms & Martens (2015) to predict customer churn in the energy industry. The study has focused on the development of ensemble learning methods to improve predictive performance. However, few studies have investigated the factors affecting customer churn in the energy industry, particularly in the Czech Republic.

Regarding the methodology used in the De Caigny et al. (2018) paper, the proposed hybrid classification algorithm based on logistic regression and decision trees achieved better performance than other benchmark methods in predicting customer churn in the telecom industry. However, this method has not been applied to the energy industry yet.

The literature review provides insights into the application of machine learning algorithms for customer churn prediction in the energy industry. Future studies could focus on investigating the factors affecting customer churn in the Czech energy industry and explore the potential application of the hybrid classification algorithm proposed by De Caigny et al. (2018).

Various models and techniques have been explored by the researchers to predict churn, a significant challenge faced by many industries, and to identify factors that influence gas consumption. However, further research is needed to address the gaps in the existing literature in order to develop more effective retention strategies suitable specifically for the natural gas supply industry.

Motivation for the research

General information

The first impulse which led to writing this thesis came from MND Energie, a.s. (Moravské naftové doly), whose data department decided, especially due to the current uncertainty on the world energy markets and the high cost of acquiring new customers, to focus on optimizing the retention policy and evaluating their current progress in this area. Therefore, the head of the data department provided access to such sample of their client database that customers belonging to this section had their gas contracts with MND Energie, a.s. terminating at the same time. This dataset is to be used to create an algorithm that is as accurate as possible and at the same time relatively easy to understand and interpret. The goal is to be capable of predicting which customers will want to switch to another supplier after their contract with MND Energie, a.s. expires. Thus, the primary motivation of the thesis is to optimize the retention policy of the database provider and subsequently stabilize its profits.

In line with this issue, the output of the thesis should present the reasons for customer churn within the Czech energy sector. There are studies investigating this phenomenon worldwide, e.g. on the Belgian energy market (Moeyersoms & Martens (2015)). However, as mentioned in the previous chapter, the literature of this type focused on the Czech market with its specific features (Vlček & Černoch (2013)) is not sufficient. Although this thesis analyses a database from only one supplier, conclusions based on this sample can be to some extent generalised to the Czech population. According to the OTE's statistic for May 2022, MND Energie, a.s. supplies approximately 4% of gas consumers and 2% of electricity consumers in the Czech Republic. Consequently, it is ranked among the ten largest energy suppliers in the Czech Republic. Furthermore, MND's customer portfolio is highly diversified and contains clients of from all regions of the Czech Republic.

Understanding churn may significantly affect company's earnings, as the cost of acquiring new customers exceeds the cost of retaining current customers (Torkzadeh et al. (2006)). This is due to, for example, the opportunity to understand and adapt to the needs of long-term customers or their recommendation of the company to new potential customers in case of satisfaction. Nevertheless, minimising customer churn does not only benefit the supplier in terms of profit. The Consumer Surplus Theorem (Syam & Hess

(2006)) states that in case of a company adopting a retention strategy which lowers the churn rate under competition, the company's customers are better off. Furthermore, the Czech energy market specifically is notorious for fraudulent offers. A well-established company being able to lower its churn rate automatically implies lower chance of a consumer switching supplier getting manipulated to sign an onerous contract.

Another motivation for writing this paper is to apply a hybrid algorithm for customer churn prediction (De Caigny et al. (2018)) on data from a different industry and verify its superiority to its individual parts used separately. It is presented by the authors as a tool with a high level of predictive ability, whose structure is at the same time relatively intuitive and easy to grasp, which are also parameters required by the data department which provided access to the client database.

Consumer Surplus Theorem

In their paper, Syam & Hess (2006) examined the consequences of different Customer Relationship Management (CRM) approaches chosen by competing companies for the companies' profits and their customers utilities.

CRM is a business practice that entails the identification and prioritization of valuable customers, followed by the provision of specialized products or services to them. The application of game theory tools can be employed to model this process. In this context, the authors viewed two companies as players who can choose between two distinct strategies: attracting and retaining customers (known as the acquisition strategy) or preventing customer defection (known as the retention strategy).

The model's outcome reveals following Nash equilibrium in a competitive environment: one firm adopts an acquisition strategy, while its rival (the first-mover) pursues a more profitable retention strategy. Consequently, employing a retention strategy leads to a comparatively smaller group of devoted customers.

A noticeable implication of the model emerges when examining the surplus of customers as a function of the churn rate, assuming a firm implements a retention strategy in a competitive equilibrium. The model's findings directly allow for the formulation of the following theorem:

'*If a firm adopts a retention strategy in a competitive equilibrium, both the firm itself and its club members experience enhanced well-being when the churn rate is minimized.*' (Syam & Hess (2006))

The conclusions of the described model support (considering certain assumptions) understanding lowering churn as a phenomenon generally beneficial for both, a company adopting a retention strategy and its customers. Thus, based on these conclusions, employing well performing churn prediction algorithms may lead to implementing churn-lowering strategies resulting in an increase in the profitability of a company and its customers surplus.

Data

General information about the dataset

The thesis investigates dataset consisting of information about 9254 MND Energie, a.s. customers whose natural gas contracts terminated at the end of 2019. Each customer either renewed their contract or left the company. This information is stored in the target variable *Churn*.

Before the contract termination date, the customers were exposed to a form of a simple retention policy measure. Two thirds (the treatment group of the experiment) of the customers, randomly selected from the dataset, were directly notified of the possibility of a new fixed-price contract setting 789 CZK/MWh including VAT, while the remaining third (the control group of the experiment) received only information about the transition to the new standard cost-plus contract, which at the time implied 889 CZK/MWh including VAT. Information about group membership for individual observations is stored in the variable *Experiment*.

The remaining explanatory variables are of three types. First, variables provided directly by MND Energie, a.s. containing information about the product usage such as *Pricelist*, *Length of contract*, *Consumption*. Second, variables provided directly by MND Energie, a.s. containing sociodemographic characteristics of individual customers – *Sex* and *ZIP code* (of the customer's consumption point). The last variable mentioned is of high cardinality. Due to the requirement of relatively easy interpretation of the results by MND Energie, a.s., this variable was not directly used for the analysis. However, its values were used to assign the third type of explanatory variables, which the original dataset provided by MND Energie, a.s. did not contain. These are the contextual factors – *Town pop*, *Avg temp in reg* and *Avg wage in reg*. The data consisting of the values of the third type of explanatory variables were obtained from the Czech Statistical Office (CSO) and the

Czech Hydrometeorological Institute (CHMI). The exact meanings and units of all variables are summarised in the following chapter.

Description of the variables

The provided dataset consists of variables of various types, units and distributions. Detail description of all variables included in the dataset is provided in Table 1.

Table 1 Overview of the variables included in the dataset

Name	Meaning	Units	Default type
ID	Unique number for each individual customer (1 to 9254)		Integer
Pricelist	Type of pricelist agreed on in the contract terminated at the end of 2019 (two types: Online and Offline – Online has lower price for the fixed component (fee for a consumption point independent of consumption) of *Advance payments*; same price for MWh)		Character, values: Online, Offline
Length of contract	Length of the time period between the start of the terminated contract and the end of 2019	Years	Numeric

Churn	Dummy variable based on the customer loyalty to the company (1 in case of churn, 0 in case of new contract); the target variable		Integer, values: 1, 0
Consumption	Yearly gas consumption of a customer	MWh	Numeric
Advance payments	Monthly advance payments for gas	CZK	Integer
Experiment	Group membership regarding the retention experiment		Character, values: TG, CG (treatment group, control group)
Sex	Self-explanatory		Character, values: M, F, L (male, female, legal entity)
No of cons points	Number of consumption points registered by a customer		Integer
Time w the comp	Length of the time period between the start of the first contract and the end of 2019	Years	Numeric
ZIP code	ZIP code specified in the contract		Character, high cardinality

Town	Municipality linked with the *ZIP code*		Character, high cardinality
Town pop	Population of *Town* as of 1st Jan 2020		Integer
Avg temp in reg	Average temperature in the region linked with the *ZIP code* in the 2019 (14 administrative regions of the Czech Republic)	°C	Numeric
Avg wage in reg	Average monthly wage in the region linked with the *ZIP code* in the 2019	CZK	Integer

As already partially evident from Table 1, the data must be pre-processed before proceeding to the actual analysis. The steps involved are described in the following chapter.

Data pre-processing

First, data types were modified accordingly. Character variables *Pricelist*, *Experiment* and *Sex* were turned to factors (the type of variables that statistical software treats as categorical). High cardinality attributes – *ID*, *ZIP code* and *Town* were used for construction of area specific variables (*Avg temp in reg* etc.)

Missing values were treated accordingly to methodology used in the De Caigny et al. (2018) paper. Each variable has less than 5 %, values missing. Thus, all data points containing missing values are removed from the dataset in order to avoid the impact of imputation procedures. After this procedure, 9055 data points remained in the dataset. The proportion of customers falling into the treatment group remained close to two thirds (exactly 65.84 %).

Two methods were used for outlier detection – identifying values further than three standard deviations from the mean value and generating boxplots. Outliers among all variables were inspected with respect to the values. Except for one variable, it was found that all outliers are very likely true values rather than errors. Therefore, the datapoints containing these outliers might carry valuable information and it was decided to keep them in the dataset.

Variables *Advance payments*, *No of cons points*, *Time w the comp*, *Town pop* and *Avg wage in reg* were log-transformed.

Methodology

Introduction

In compliance with the board's requirement of building a prediction model which is both comprehensible and relatively well performing in terms of prediction, the logit leaf algorithm, developed and presented in the paper titled *A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees* by A. De Caigny in 2018, became the method of choice. This algorithm combines two simple methods often used for customer churn prediction – decision trees and logistic regression. The proposed algorithm aims to solve for disadvantages of both of its parts applied separately. Apart from the logit leaf model itself, this thesis also employed its two building blocks separately in order to provide benchmark in terms of predictive performance. AUC and TDL were chosen as predictive performance metrics.

Churn prediction models (or binary choice models in general) are usually trained on a training set randomly generated from the whole dataset and eventually evaluated in terms of performance on a testing set consisting of the remaining data points of the whole dataset. The robustness of the results is to be tested by 5 x 2 folds cross-validation. This procedure consists in generating four other random training sets (splitting the whole dataset into two folds – training and testing), using them for training the algorithm from scratch and comparing the results with those originated from fitting the model on the first training set. If the results are similar, the first fit is considered robust.

The analysis is performed through code run in the R environment. R is an open-source programming language and statistics software. R is suitable for the purpose of this thesis due to basic statistical techniques implemented in its libraries and graphical visualization possibilities. User-created packages allow for extension of these uses.

The predictive performance metrics and the main ideas behind all three prediction methods are further described in the following subchapters.

Prediction methods

Logistic regression

Logistic regression is a way of predicting the values of the dependent binary variable y (in this case churn), which can take values of 0 and 1. This method is based on classical linear regression. Linear regression itself is not appropriate in this case, especially since it may predict values greater than 1 or less than 0. For this reason, the concept of an underlying latent variable (Wooldridge (2008)) y^*, which directly implies the values of the dependent variable, is introduced in logistic regression. In the context of the problem of this thesis, this latent variable can be viewed as the net utility of individual customers, which they potentially gain from churning. If this net utility is greater than 0, the customer chooses to churn.

The logistic regression assumes that the underlying latent variable follows the following linear model:

$$y_i^* = \boldsymbol{\beta}\mathbf{x}_i + u_i$$

Where y_i^* stands for the i-th customer's net utility gained from churning, $\boldsymbol{\beta}$ stands for the vector of regression parameters, $\mathbf{x}_i$ stands for the i-th customer's vector of values of the set of dependent variables and u_i stands for i-th disturbance. Logistic regression also assumes no perfect collinearity among the dependent variables and homoscedastic (having constant variance independent of the set of dependent variables) zero-mean disturbances independent of the set of the dependent variables.

Logistic regression aims to predict the expected value of the target variable y (the probability of y being equal to 1. Probability $p(y_i = 1|\mathbf{x}_i)$ can be expressed the following way:

$$p(y_i = 1|\mathbf{x}_i) = p(\boldsymbol{\beta}\mathbf{x}_i + u_i > 0|\mathbf{x}_i)$$
$$p(y_i = 1|\mathbf{x}_i) = p(u_i > -\boldsymbol{\beta}\mathbf{x}_i|\mathbf{x}_i)$$

Assuming u_i being symmetrically distributed, the following holds:

$$p(y_i = 1|\mathbf{x}_i) = p(u_i \leq \boldsymbol{\beta}\mathbf{x}_i|\mathbf{x}_i)$$

$p(u_i \leq \boldsymbol{\beta}\mathbf{x}_i|\mathbf{x}_i)$ is the definition of the value of the cumulative distribution function of u's distribution for $\boldsymbol{\beta}\mathbf{x}_i$.

$$p(y_i = 1|\mathbf{x}_i) = F(\boldsymbol{\beta}\mathbf{x}_i)$$

where F stands for the aforementioned cumulative distribution function. Logistic regression assumes u following standard logistic distribution. Logistic function is the cumulative distribution function of standard logistic distribution.

$$F(x) = \frac{1}{1 + e^{-x}}$$

Plugging into the equation describing the probability $p(y_i = 1|\mathbf{x}_i)$ yields following relationship:

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}}$$

The assumption of u following standard logistic distribution allows for tidy representation of the log-odds of the target variable.

$$\frac{p(y_i = 1|\mathbf{x}_i)}{1 - p(y_i = 1|\mathbf{x}_i)} = \frac{\frac{1}{1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}}}{1 - \frac{1}{1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}}}$$

$$\frac{p(y_i = 1|\mathbf{x}_i)}{1 - p(y_i = 1|\mathbf{x}_i)} = \frac{\frac{1}{1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}}}{\frac{e^{-\boldsymbol{\beta}\mathbf{x}_i}}{1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}}}$$

$$\frac{p(y_i = 1|\mathbf{x}_i)}{1 - p(y_i = 1|\mathbf{x}_i)} = e^{\boldsymbol{\beta}\mathbf{x}_i}$$

$$\log\left[\frac{p(y_i = 1|\mathbf{x}_i)}{1 - p(y_i = 1|\mathbf{x}_i)}\right] = \boldsymbol{\beta}\mathbf{x}_i$$

The vector of regression parameters $\boldsymbol{\beta}$ is unknown. It cannot be estimated using the least squares method because the values of the underlying latent variable y_i^* are not observed. Therefore, maximum likelihood estimator (MLE) is employed. The idea behind this estimation consists in finding such vector of estimates $\widehat{\boldsymbol{\beta}}$ which maximize the predicted probability of the y outcomes having the observed values. That is performed through expressing the joint probability of $y = y_i$ given $\mathbf{x}_i$ for all y_is and $\mathbf{x}_i$s. This joint probability is a function of $\widehat{\boldsymbol{\beta}}$. The optimal $\widehat{\boldsymbol{\beta}}$ is found using the first-order condition for finding extremes of a function.

MLE estimator is consistent, asymptotically normal and asymptotically efficient. Thus, it is possible to calculate the standard errors for individual estimates.

The marginal effect of an increase in the x_k variable on the estimated probability of the target variable (keeping the values of the remaining independent variables constant) being equal to 1 can be described in the following manner:

$$\frac{\partial p(y = 1|\mathbf{x}_i)}{\partial x_k} = \frac{\partial \left(\frac{1}{1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}}\right)}{\partial x_k}$$

The marginal effect does not only depend on the value of $\widehat{\beta_k}$ (the estimate of the regression parameter linked with the x_k variable), but also on the values of the remaining estimates and on the $\mathbf{x}_i$ values. Thus, the marginal effect differs among observations. Therefore, in order to interpret the estimated effect of the independent variable on the target variable, either the average marginal effect for all observations or the marginal effect for the average observation (all independent variables averaged) is presented.

For dummy independent variables, the marginal effect can be expressed as the difference between the estimated probabilities of the target variable being equal to 1 for individual levels of the dummy independent variable, assuming fixed values of the remaining independent variables.

Logistic regression is one of the basic churn prediction methods. Its main advantage lies in handling linear relationships among the observed variables. Thus, in cases where the values of the target variable are rather dependent on the interactions between the values of the independent variables, logistic regression tends to be unsuitable.

Decision trees

Decision trees refer to another basic classification algorithm. The idea consists in splitting the data points into groups, which are as homogenous as possible in terms of values of the target variable, by conditioning on the values of the independent variables. The tree structure starts with the root node (the whole dataset), which is split into two non-overlapping subsets (child nodes) based on the value of the selected independent variable. This process is repeated on the latest child nodes created until there is no benefit in further splitting regarding the predictive performance. The terminal nodes are referred to as the leaves. The tree structure assigns each data point into one of the leaves.

In case of customer churn prediction, the training set is used to build the tree structure by splitting the training set customers into groups homogenous in the value of the churn variable. For each group, the predicted value of churn is set as the average churn value for the training set customers belonging to this group. Subsequently, the tree structure is used to split the testing set and assign the predicted churn values to all testing set customers based on the group membership.

For each node, the decision tree algorithm creates as homogenous child nodes as possible. Thus, the splitting criterion is chosen with respect to the given node, rather than the overall tree structure.

For the purpose of this thesis, the *rpart* package is employed in order to perform the analysis in the R environment. The decision tree function included in this package calculates the Gini Index potential child nodes in order to select the optimal splitting criterion at a given node. The Gini Index, which measures impurity of a node with respect to the target variable, is defined in the following manner:

$$GI = 1 - \sum_{i=1}^{n_c} [P(i)]^2$$

where n_c stands for the number of classes of the target variable (two classes in the churn prediction case – 1 and 0) and $P(i)$ for the ratio of the node's observations belonging to the i class of the target variable. The function selects such splitting criterion that the weighted (by the number of observations in each child node) average Gini Index of the arising child nodes is minimized. In other words, the arising child nodes are as pure (homogenous) as possible regarding the values of the target variable.

In comparison with logistic regression, decision trees are more capable in capturing the influence of interactions between the values of the independent variables on the target variable. That is allowed due to generating the terminal nodes by consecutive conditions. Generating decision trees may lead to creating complex structures which perfectly fit the training set at the expense of general predictive ability (overfitting). The *rpart* package functions allow for controlling the tree structure complexity by setting the tree's hyperparameters – cost complexity, tree depth and minimal number of observations in a node required for further splitting. The decision tree is usually built for numerous combinations of hyperparameters' values and the optimal hyperparameters are selected with respect to the overall predictive performance of the tree. This procedure is called hyperparameter tuning.

Logit leaf model

The main idea of the LLM lies in combining two relatively simple models with good comprehensibility – logistic regression and decision trees, in order to eliminate disadvantages of individual methods and boost the predictive performance while preserving the comprehensibility. While decision trees used separately cannot properly capture linear relationships, logistic regression cannot properly handle interactions. The logit leaf model aims to eliminate these disadvantages.

The algorithm starts with growing a decision tree using the whole customer set. The decision tree splits the customers into groups based on their terminal node affiliation. Logistic regression is then applied on individual groups. It is important to note that the hyperparameters of the decision tree are tuned with respect to the overall predictive performance of the logit leaf model. The training phase of the algorithm consists of the following steps:

1. Generate decision tree using the whole training set

2. Define non-overlapping subsets of the training set based on the terminal node membership

3. For i in $(1{:}\,n_g)$, where n_g stands for the number of terminal nodes: Run logistic regression using the data points belonging to terminal node i

Performance metrics

Area under the receiver operating characteristic curve

Basic measures such as sensitivity and specificity depend on the choice of the classification threshold. In the churn prediction case, the classification threshold refers to the value that sets the border for classification – observations with predicted churn probabilities above this value are labelled as churners. Receiver operating curve is introduced in order to evaluate the predictive performance of the model among various thresholds. It is a curve depicting the values of true positive rate and false positive rate in space (true positive rate on the vertical axis and false positive rate on the horizontal axis) for various thresholds. True positive rate is the number of correctly predicted churners among all churners, while false positive rate is the number of falsely predicted churners among all non-churners. Both true positive rate and false positive rate increase with decreasing classification threshold. The area under the receiver operating characteristic curve serves as a measure of the overall predictive performance of the model. For a random classifier, this area is close to $\frac{1}{2}$. For a perfect classifier (in most cases unrealistic), the area equals 1.

Top-decile lift

Top-decile lift is a self-explanatory predictive performance metric, which compares the number of churners among the 10 % customers with highest predicted churn probabilities to the number of churners in the whole dataset. E.g., top-decile lift equal to 2 suggests the average churn rate of the 10 % customers with highest predicted churn probabilities being two times the average churn rate of the whole dataset.

Results and discussion

LLM

The first part of the logit leaf algorithm splits the training set into four groups. The first group consists of customers which have more than one consumption points registered. The *No of cons points* variable is positively correlated with the dummy variable which describes whether the customer is a legal entity. Thus, there is higher concentration of legal entity customers in the first group compared to the whole dataset. The second group consists of customers with one consumption point registered whose monthly advance payments are below 3670 CZK. Most customers belong to the second group. The remaining two groups consist of customers with one consumption point registered whose monthly advance payments are above or equal to 3670 CZK. Customers from the third group are from regions with average monthly wage in the region above or equal to 30 000 CZK. The last group consists of customers with average monthly wage in their region below 30 000 CZK. In 2019, these regions were Karlovarský kraj, Pardubický kraj, Olomoucký kraj, Zlínský kraj and Moravskoslezský kraj. The last two groups do not contain enough data points for reasonable degrees of freedom number. Therefore, logistic regression was performed only on the first two groups. The final testing set predicted churn probability for the remaining two groups was set as the average churn rates of the training set customers belonging to these groups. The fourth group has very high predicted churn probability compared to the average churn rate of the whole dataset. The average churn rate of the whole dataset equals 4.51 %, while the predicted churn probability of the testing set customers who belong to the fourth group exceeds 55 %. The predicted churn probability for the third group equals 7.69 %.

The average churn rate among the training set customers, who belong to the first group, equals 8.02 %. The logistic regression applied on the first group revealed three variables significant at 10 % level – *Sex.M* (the dummy variable which describes whether the customer is a male), *logTwC* (the natural logarithm of *Time w the comp*) and *logAW* (the natural logarithm of *Avg wage in reg*). The effects of these variables may be interpreted in the following manner:

- Male customers' odds of churning are 0.91 % higher compared to female and legal entity customers.
- 1% increase in the *Time w the comp* is expected to decrease the odds of churning by 0.46 %.
- 1% increase in the *Avg wage in reg* is expected to increase the odds of churning by 3.41 %.

The average churn rate among the training set customers, who belong to the second (the largest) group, equals 3.83 %. The logistic regression applied on the second group revealed two variables significant at 10 % level – *Length of contract* and *Experiment*. The effects of these variables may be interpreted in the following manner:

- Additional year of *Length of contract* is expected to decrease the odds of churning by 0.33 %.
- Treatment group customers' odds of churning are 0.31 % lower compared to control group customers.

The decision tree part of the model is visualized in Figure 1. The results of logistic regression applied on the first and second group are summarized in Figure 2 and Figure 3 respectively.

Figure 1 Decision tree part of the LLM algorithm

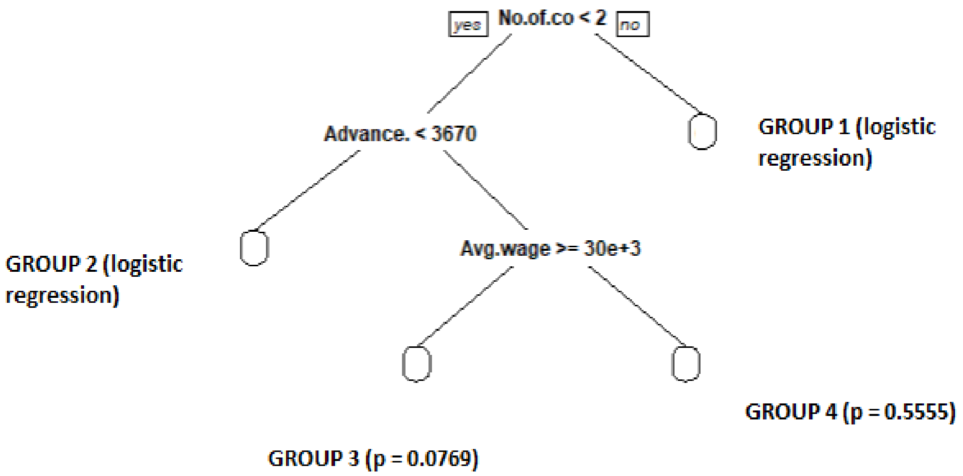
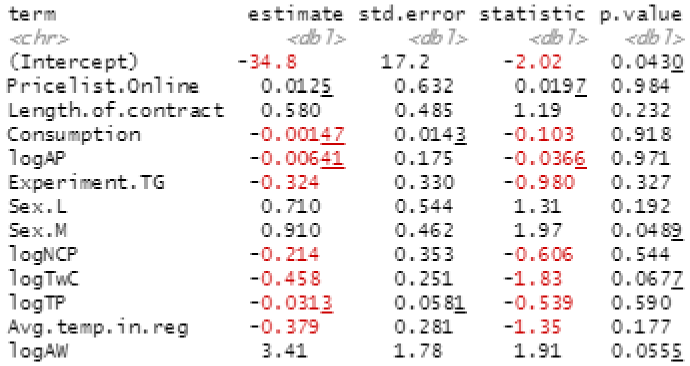
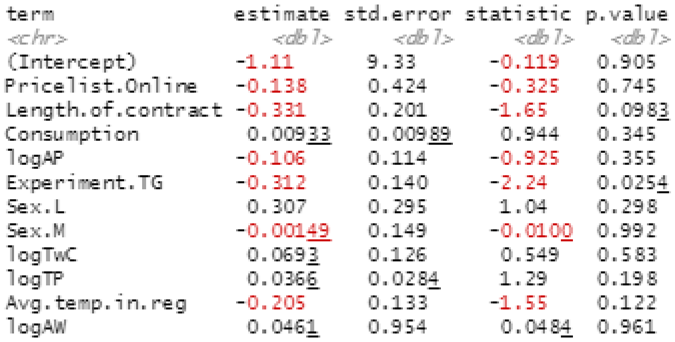


Figure 2 Results of the logistic regression performed on GROUP 1



```
term                 estimate std.error statistic p.value
<chr>                   <dbl>     <dbl>     <dbl>   <dbl>
(Intercept)          -34.8       17.2      -2.02    0.0430
Pricelist.Online       0.0125     0.632     0.0197  0.984
Length.of.contract     0.580      0.485     1.19    0.232
Consumption           -0.00147    0.0143   -0.103   0.918
logAP                 -0.00641    0.175    -0.0366  0.971
Experiment.TG         -0.324      0.330    -0.980   0.327
Sex.L                  0.710      0.544     1.31    0.192
Sex.M                  0.910      0.462     1.97    0.0489
logNCP                -0.214      0.353    -0.606   0.544
logTwC                -0.458      0.251    -1.83    0.0677
logTP                 -0.0313     0.0581   -0.539   0.590
Avg.temp.in.reg       -0.379      0.281    -1.35    0.177
logAW                  3.41       1.78      1.91    0.0555
```

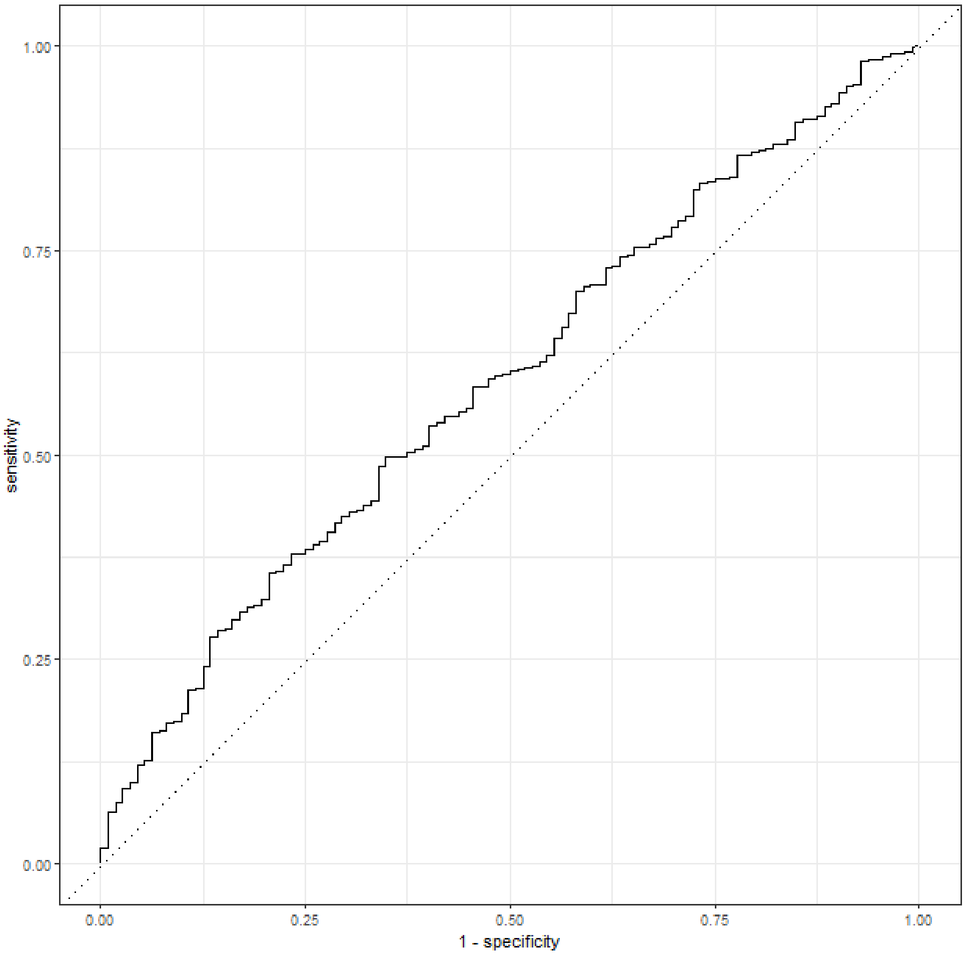
Figure 3 Results of the logistic regression performed on GROUP 2



```
term                estimate std.error statistic p.value
<chr>                  <dbl>     <dbl>     <dbl>   <dbl>
(Intercept)         -1.11      9.33     -0.119   0.905
Pricelist.Online    -0.138     0.424    -0.325   0.745
Length.of.contract  -0.331     0.201    -1.65    0.0983
Consumption          0.00933   0.00989   0.944   0.345
logAP               -0.106     0.114    -0.925   0.355
Experiment.TG       -0.312     0.140    -2.24    0.0254
Sex.L                0.307     0.295     1.04    0.298
Sex.M               -0.00149   0.149    -0.0100  0.992
logTwC               0.0693    0.126     0.549   0.583
logTP                0.0366    0.0284    1.29    0.198
Avg.temp.in.reg     -0.205     0.133    -1.55    0.122
logAW                0.0461    0.954     0.0484  0.961
```

The LLM scored 0.59 in the AUC performance measure and 1.521 in the TDL performance measure. The receiver operating curve is depicted in Figure 4.

Figure 4 Receiver operating curve for the LLM



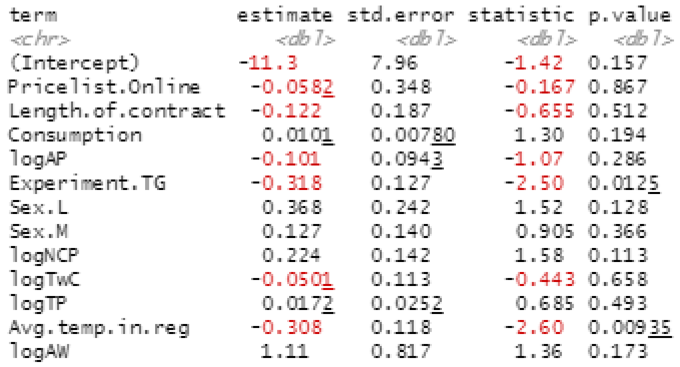
Logistic regression

Logistic regression performed on the whole training dataset revealed two variables significant at 10 % level – *Avg temp in reg* and *Experiment*. The effects of these variables may be interpreted in the following manner:

- Additional °C of *Avg temp in reg* is expected to decrease the odds of churning by 0.31 %.
- Treatment group customers' odds of churning are 0.32 % lower compared to control group customers.

The results of logistic regression applied on the whole training set are summarized in Figure 5.

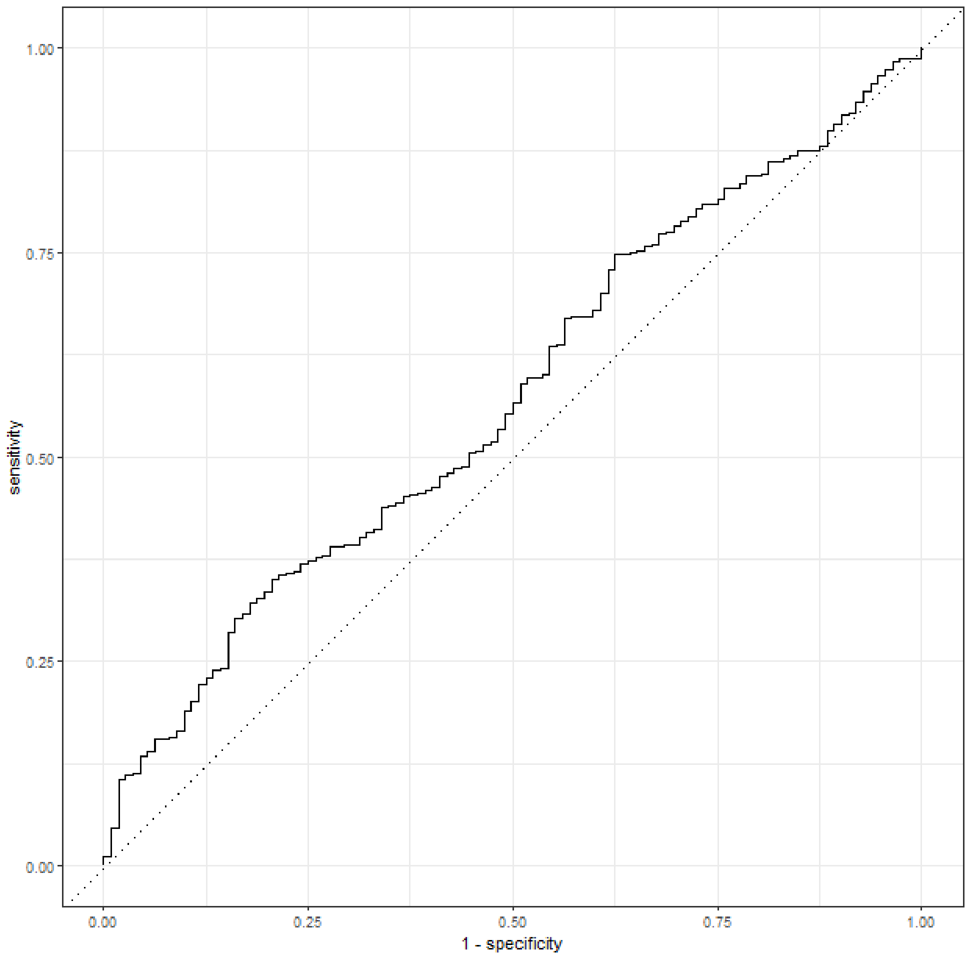
Figure 5 Results of the logistic regression performed on the whole training set



```
term                estimate  std.error statistic p.value
<chr>                  <dbl>      <dbl>     <dbl>   <dbl>
(Intercept)         -11.3      7.96        -1.42  0.157
Pricelist.Online     -0.0582   0.348       -0.167 0.867
Length.of.contract   -0.122    0.187       -0.655 0.512
Consumption           0.0101   0.00780      1.30  0.194
logAP                -0.101    0.0943      -1.07  0.286
Experiment.TG        -0.318    0.127       -2.50  0.0125
Sex.L                 0.368    0.242        1.52  0.128
Sex.M                 0.127    0.140        0.905 0.366
logNCP                0.224    0.142        1.58  0.113
logTwC               -0.0501   0.113       -0.443 0.658
logTP                 0.0172   0.0252       0.685 0.493
Avg.temp.in.reg      -0.308    0.118       -2.60  0.00935
logAW                 1.11     0.817        1.36  0.173
```

The logistic regression scored 0.571 in the AUC performance measure and 1.073 in the TDL performance measure. The receiver operating curve is depicted in Figure 6.

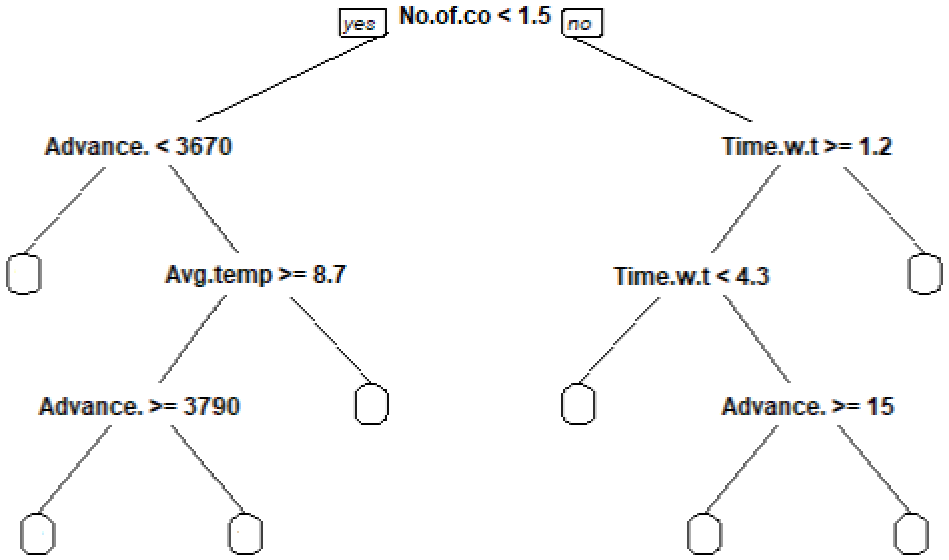
Figure 6 Receiver operating curve for the logistic regression



Decision tree

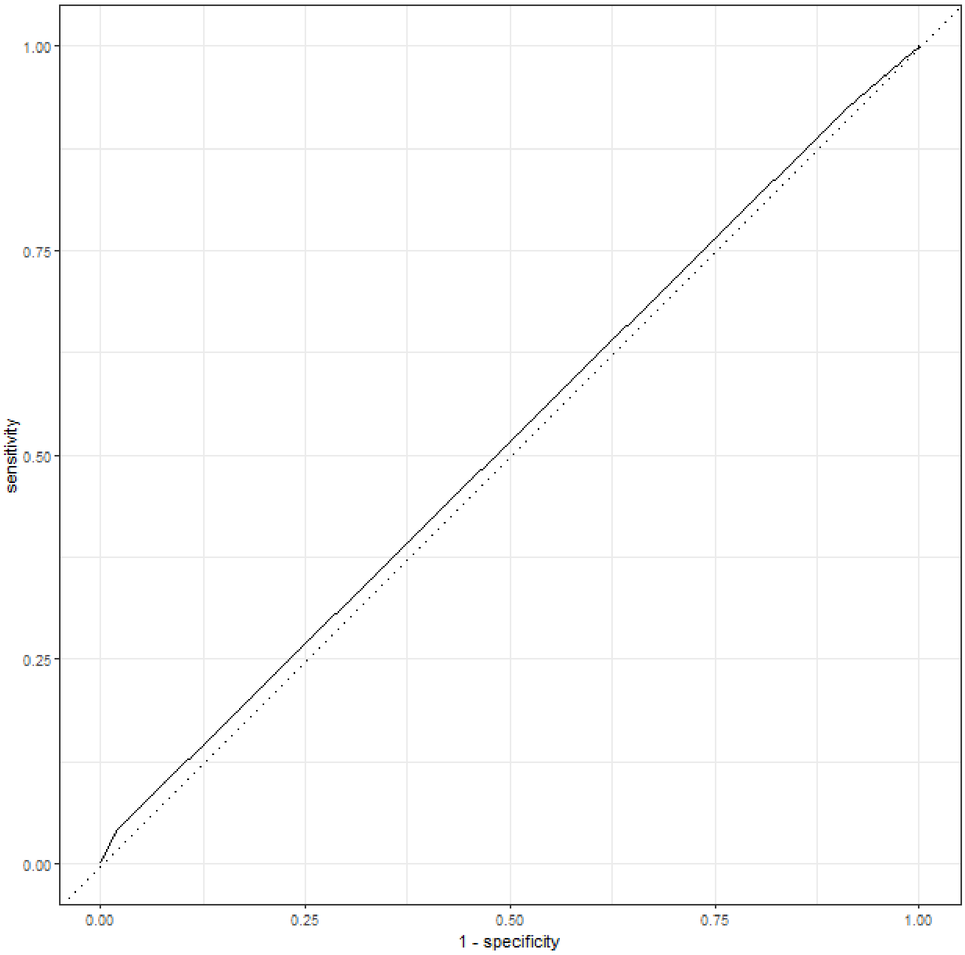
The first two nodes of the decision tree split the data into the same groups as the first two nodes of the first part of the LLM. The tree is then further split conditioning mainly on product usage variables. The decision tree part of the model is visualized in Figure 7.

Figure 7 Decision tree visualization



The decision tree scored 0.517 in the AUC performance measure and 1.342 in the TDL performance measure. The receiver operating curve is depicted in Figure 8.

Figure 8 Receiver operating curve for the decision tree



Discussion and cross-validation

The LLM algorithm defined four groups, which are characterized by different churn behaviour. The first group consists of customers with more than one consumption points registered. This group has large share (30.16 %) of legal entity customers compared to the whole dataset (8.69 %). The second group consists of customers with one consumption point registered who pay less than 3670 CZK in monthly advance payments. Most customers belong to this group and the results of logistic regression performed on training set customers belonging to this group are similar to the results of logistic regression performed on the whole training set. The most important difference between the first two groups is connected to the *Experiment* variable. For the first group, there is no significant difference between the treatment group and the control group members regarding the churn. In the second group on the other hand, the treatment group members are expected to have significantly lower log-odds ratio of churning compared to the control group members. The remaining customers are further split into two small groups conditioning on the average monthly wage in their region. Both of these groups, especially the one consisting of customers from regions with lower average monthly wages, have higher expected probabilities of churning compared to the first two groups.

The LLM algorithm achieved better predictive performance than both benchmark methods, logistic regression and decision trees, considering the selected performance metrics – AUC and TDL. The robustness of these findings was tested via 5 x 2 folds cross-validation. Summary is provided in Table 2.

Table 2 Cross-validation results

Method	**Average AUC**	**Average TDL**
LLM	0.579	1.585
Logistic regression	0.575	1.456
Decision trees	0.515	1.127

The retention experiment i.e., directly notifying some customers of the possibility of a new fixed-price contract, may only had the desired effect (lower churn probability) on customers with only one consumption point registered.

Overall, according to the results, the product usage information and retention experiment group membership turned out to be the variables predicted to affect churn most significantly.

The model also suggests possible relationship between churn and contextual factors. Nevertheless, it is in most cases statistically insignificant, and it varies depending on the method chosen. This may be due to the facts that individual contextual factors are relatively highly correlated (Appendix 1) and that the data were collected mostly on regional level, rather than county level (there are only fourteen regions in the Czech Republic).

In this analysis, as one of the LLM groups consisted of majority of customers, the logistic regression (performed on this group) represented the main part of the computing process of the LLM. For two of the data splits used during cross-validation, the LLM algorithm even decided it is optimal not to split the training set during the decision tree part at all and perform logistic regression on the whole dataset. Nevertheless, the remaining data splits generated trees similar to the one generated with the original data split.

This may be caused by lack of data. Most of the studies mentioned in the Literature Review work with approximately ten times larger datasets.

Conclusions

In conclusion, this thesis addressed the prediction of churn and the identification of customer characteristics that have a significant effect on the predicted probability of churn, in the Czech energy sector. For this purpose, an analysis of the database of natural gas customers (provided by MND Energie, a.s.), whose contracts terminated in 2019 was conducted.

The thesis includes a cursory literature review that summarizes the approaches used in studies predicting churn in different sectors. The LLM algorithm has become the method of choice due to its easy interpretability as well as its relatively high predictive performance compared to baseline methods.

The Results chapter provides a comparison of the predictive performance of LLM with its constituent parts – logistic regression and decision trees. In addition, the results of the model are interpreted. Variables identified by the model as having a significant effect on the target variable and the estimated relationships with respect to the target variable are discussed. Furthermore, the model also evaluates the impact of management's retention strategy performed in 2019. The conclusions of the analysis suggest that LLM achieves superior predictive performance compared to logistic regression and decision trees. AUC and TDL metrics were used to evaluate the predictive performance. The results were verified using cross-validation. Conditioning mostly on product usage variables, the LLM algorithm split the customers into four groups that exhibit different behaviours in terms of churn. The management's retention strategy turned out to have significant effect on predicted churn probability only for some customers.

The analysis suffers from a relatively low number of data points (compared to other studies predicting churn) and correlations between contextual factors. Therefore, logistic regression formed the main part of the computational process of the LLM and the influence of contextual factors on the probability of churn is rather unclear.

Further research could focus on analysing customer databases of a larger number of natural gas suppliers. It could also include more time periods and contextual data collected at lower administrative levels. More advanced machine learning methods may be employed. However, in line with the trade-off mentioned in the previous chapters, this would come at the cost of more difficult interpretability.

List of References

Ballings, M., & Van Den Poel, D. (2012). Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications*, *39*(18), 13517–13522. https://doi.org/10.1016/j.eswa.2012.07.006

Chen, Z. Y., Fan, Z. P., & Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research*, *223*(2), 461–472. https://doi.org/10.1016/j.ejor.2012.06.040

Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications*, *37*(3), 2132–2143. https://doi.org/10.1016/j.eswa.2009.07.029

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, *66*(9), 1629–1636. https://doi.org/10.1016/j.jbusres.2012.12.008

De Bock, K. W., & Van Den Poel, D. (2012). Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Systems with Applications*, *39*(8), 6816–6826. https://doi.org/10.1016/J.ESWA.2012.01.014

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760–772. https://doi.org/10.1016/J.EJOR.2018.02.009

Gustafsson, A., Johnson, M. D., & Roos, I. (2005). The Effects of Customer Satisfaction, Relationship Commitment Dimensions, and Triggers on Customer Retention. *Journal of Marketing*, *69*(4), 210–218. https://doi.org/10.1509/jmkg.2005.69.4.210

Hansen, H., M. Samuelsen, B., & E. Sallis, J. (2013). The moderating effects of need for cognition on drivers of customer loyalty. *European Journal of Marketing*, *47*(8), 1157–1176. https://doi.org/10.1108/03090561311324264

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing Journal*, *24*, 994–1012. https://doi.org/10.1016/j.asoc.2014.08.041

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems*, *72*, 72–81. https://doi.org/10.1016/J.DSS.2015.02.007

Moitra, N., Raj, P., Saxena, S., & Kumar, R. (2020). Crude Oil Prediction Using Lstm. In *International Journal of Innovative Science and Research Technology: Vol. x* (Issue 2). www.ijisrt.com

Ofori-Ntow Jnr, E., Ziggah, Y. Y., & Relvas, S. (2021). Hybrid ensemble intelligent model based on wavelet transform, swarm intelligence and artificial neural network for electricity demand forecasting. *Sustainable Cities and Society*, *66*. https://doi.org/10.1016/j.scs.2020.102679

Raju, S. M. T. U., Sarker, A., Das, A., Islam, M. M., Al-Rakhami, M. S., Al-Amri, A. M., Mohiuddin, T., & Albogamy, F. R. (2022). An Approach for Demand Forecasting in Steel Industries Using Ensemble Learning. *Complexity*, *2022*. https://doi.org/10.1155/2022/9928836

Sharma, A., & Kumar Panigrahi, P. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. In *International Journal of Computer Applications* (Vol. 27, Issue 11).

Syam, N. B., & Hess, J. D. (2006). Acquisition versus retention: Competitive customer relationship management. *University of Houston*.

Tang, L., Thomas, L., Fletcher, M., Pan, J., & Marshall, A. (2014). Assessing the impact of derived behavior information on customer attrition in the financial service industry. *European Journal of Operational Research*, *236*(2), 624–633. https://doi.org/10.1016/j.ejor.2014.01.004

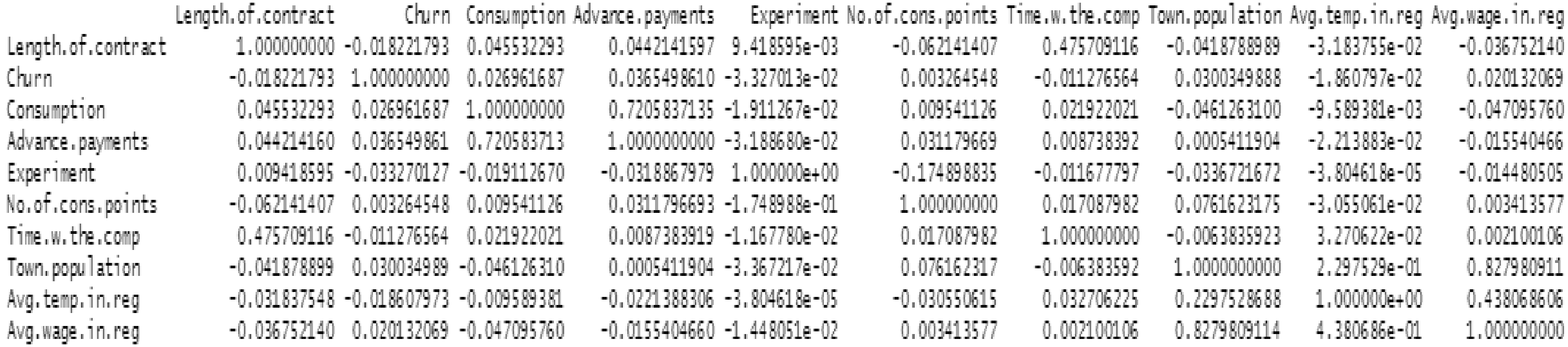
Torkzadeh, G., Chang, J. C. J., & Hansen, G. W. (2006). Identifying issues in customer relationship management at Merck-Medco. *Decision Support Systems*, *42*(2), 1116–1130. https://doi.org/10.1016/j.dss.2005.10.003

Vlček, T., & Černoch, F. (2013). The Energy Sector and Energy Policy of the Czech Republic. In *The Energy Sector and Energy Policy of the Czech Republic*. https://doi.org/10.5817/cz.muni.m210-6523-2013

Wahul, R. M., Kale, A. P., & Kota, P. N. (2023). An Ensemble Learning Approach to Enhance Customer Churn Prediction in Telecom Industry. *International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING*.

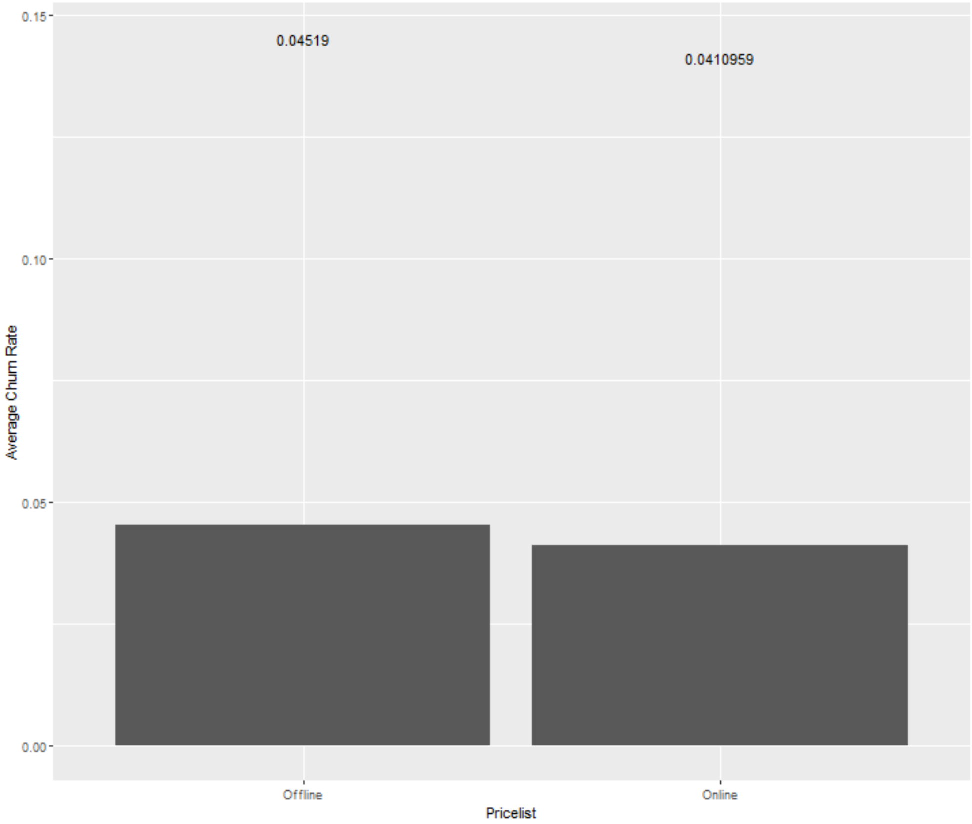
Wooldridge, J. M. (2008). *Introductory Econometrics: A Modern Approach*. Cengage Learning. https://books.google.ch/books?id=64vt5TDBNLwC

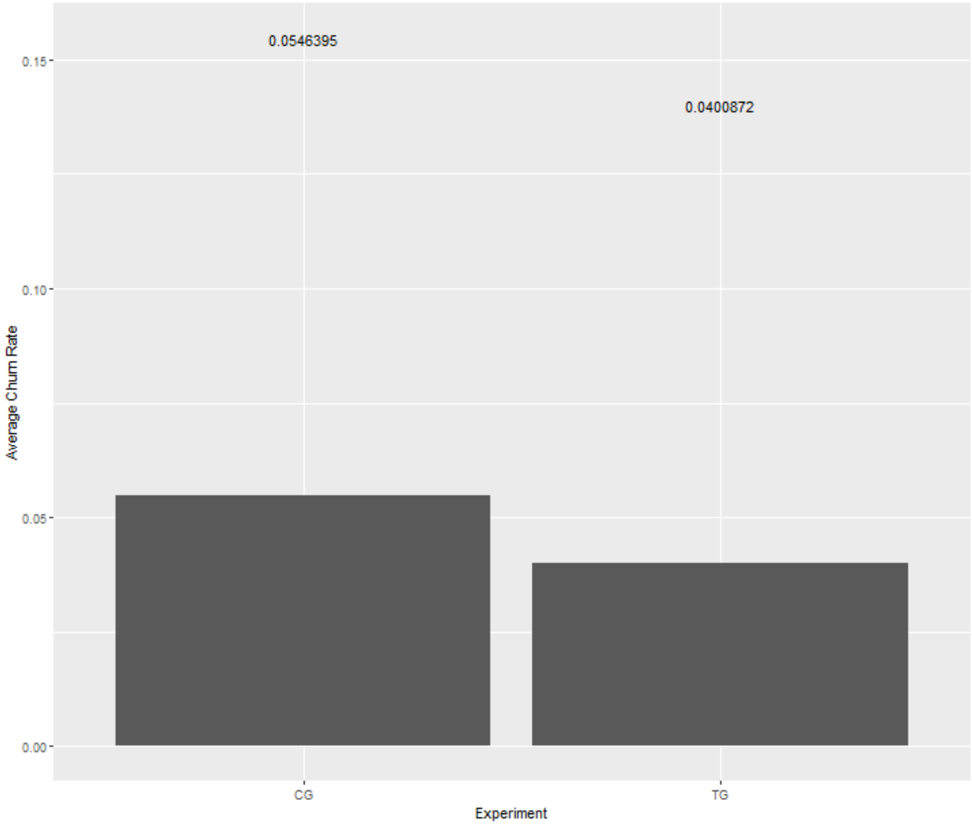
Appendix 1 – Numerical variables’ correlation matrix

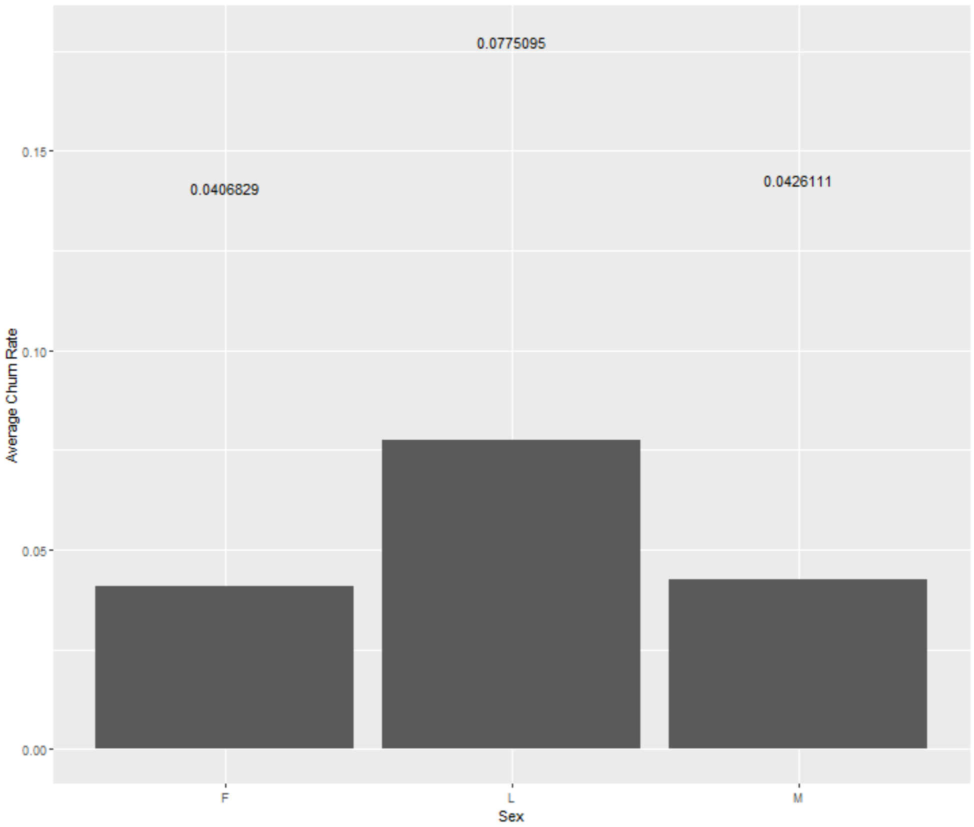


```
                   Length.of.contract        Churn  Consumption Advance.payments    Experiment No.of.cons.points Time.w.the.comp Town.population Avg.temp.in.reg Avg.wage.in.reg
Length.of.contract        1.000000000 -0.018221793  0.045532293     0.0442141597  9.418595e-03      -0.062141407     0.475709116   -0.0418788989   -3.183755e-02    -0.036752140
Churn                    -0.018221793  1.000000000  0.026961687     0.0365498610 -3.327013e-02       0.003264548    -0.011276564    0.0300349888   -1.860797e-02     0.020132069
Consumption               0.045532293  0.026961687  1.000000000     0.7205837135 -1.911267e-02       0.009541126     0.021922021   -0.0461263100   -9.589381e-03    -0.047095760
Advance.payments          0.044214160  0.036549861  0.720583713     1.0000000000 -3.188680e-02       0.031179669     0.008738392    0.0005411904   -2.213883e-02    -0.015540466
Experiment                0.009418595 -0.033270127 -0.019112670    -0.0318867979  1.000000e+00      -0.174898835    -0.011677797   -0.0336721672   -3.804618e-05    -0.014480505
No.of.cons.points        -0.062141407  0.003264548  0.009541126     0.0311796693 -1.748988e-01       1.000000000     0.017087982    0.0761623175   -3.055061e-02     0.003413577
Time.w.the.comp           0.475709116 -0.011276564  0.021922021     0.0087383919 -1.167780e-02       0.017087982     1.000000000   -0.0063835923    3.270622e-02     0.002100106
Town.population          -0.041878899  0.030034989 -0.046126310     0.0005411904 -3.367217e-02       0.076162317    -0.006383592    1.0000000000    2.297529e-01     0.827980911
Avg.temp.in.reg          -0.031837548 -0.018607973 -0.009589381    -0.0221388306 -3.804618e-05      -0.030550615     0.032706225    0.2297528688    1.000000e+00     0.438068606
Avg.wage.in.reg          -0.036752140  0.020132069 -0.047095760    -0.0155404660 -1.448051e-02       0.003413577     0.002100106    0.8279809114    4.380686e-01     1.000000000
```

Appendix 2 – Average churn rates for different levels of categorical variables






0.0775095
0.0406829
0.0426111
0.15
0.10
0.05
0.00
Average Churn Rate
F
L
M
Sex