Lappeenranta-Lahti University of Technology LUT

School of Business and Management

Master's Degree Program in Strategic Finance and Business Analytics

Jeremias Marttinen

Modelling customer churn with private electricity customer data

Master's Thesis

2021

1st Supervisor Pasi Luukka

2nd Supervisor Jan Stoklasa

ABSTRACT

Author:	Jeremias Marttinen
Title:	Modelling customer churn with private electricity customer data
Faculty:	School of Business
Major:	Strategic Finance and Business Analytics
Year:	2021
Master's Thesis:	Lappeenranta-Lahti University of Technology LUT
	57 pages, 13 figures, 10 tables, 1 formula and 3 appendices
Examiners:	Professor Pasi Luukka
	Post-Doctoral Researcher Jan Stoklasa

Keywords: Customer churn, logistic regression, decision trees, machine learning, supervised learning, electricity

The objective of this thesis is to study customer churn problem in a Finnish electricity company. First, the theory of customer churn, logistic regression, and decision tree methodology is studied and then that information is utilized in building models that could predict customer churn. The methods used to build the models are logistic regression and decision trees. The study is based on the data from a Finnish electricity company. The research questions are what are the features that can be used to identify churning customers and can we predict which customers are likely to leave next. The modelling is done by dividing the data into training and testing datasets with ten times cross validation. Imbalances in the data are treated by removing customers who had been with the company for over five years. The findings suggest that the product category was a common feature that can be used to identify churning customers. It was found that the decision tree model gives better results than the logistic regression. The accuracy of the models is measured with ROC-curve, confusion matrix and F-score.

TIIVISTELMÄ

Tekijä:	Jeremias Marttinen
Tutkielman nimi:	Asiakaspoistuman mallintaminen yksityishenkilöiden sähköasiakasdatan kanssa.
Tiedekunta:	Kauppatieteellinen tiedekunta
Pääaine:	Strategic Finance and Business Analytics
Vuosi:	2021
Pro gradu tutkielma:	Lappeenrannan–Lahden teknillinen yliopisto LUT
	57 sivua, 13 kuvaa, 10 taulukkoa, 1 kaava ja 3 liitettä
Ohjaajat:	Professori Pasi Luukka
	Tutkijatohtori Jan Stoklasa

Hakusanat: Asiakaspoistuma, logistinen regressio, päätöspuut, koneoppiminen, ohjattu oppiminen, sähkö

Tämän pro gradu – tutkielman tavoite on tutkia tutkimaan asiakaspoistumaa suomalaisessa sähköyrityksessä. Ensiksi käydään läpi asiakaspoistuman teoriaa, logistisen regression ja päätöspuumallien metodologiaa ja sitten sitä tietoa hyödynnetään mallien rakentamisessa, jotka voisivat ennustaa asiakaspoistumaa. Mallien rakentamiseen käytetty metodologia on logistinen regressio ja päätöspuut. Tutkimus perustuu dataan suomalaisesta sähköalan yrityksestä. Tutkimuskysymykset ovat mitä ominaisuuksia voidaan käyttää poistuvien asiakkaiden tunnistamiseen ja voimmeko ennustaa mitkä asiakkaat lähtevät todennäköisesti seuraavaksi. Mallinnus tehtiin jakamalla data harjoitus- ja testidataan kymmenenkertaisella ristiinvalidoinnilla. Dataluokkien epätasapainoa vähennettiin poistamalla asiakkaat, jotka olivat olleet asiakkaina yli viisi vuotta. Löydöt osoittavat, että tuotekategorian havaittiin olevan yhteinen ominaisuus, jolla voidaan tunnistaa lähteviä asiakkaita. Löydettiin että päätöspuumalli antoi parempia tuloksia kuin logistinen regressio. Mallien tarkkuutta mitattiin ROC-käyrän, sekaannusmatriisin ja F-pisteiden avulla.

ACKNOWLEDGEMENTS

It has been a memorable journey to get here and be finally graduating from the LUT. The writing process of this thesis started last year, when I thought about possible topics for a thesis where I could utilize the knowledge I have gotten from my studies. I got an opportunity to study customer churn, and supervised learning models were fit for that task.

I would like to thank everyone I had the privilege of knowing while studying at LUT, all the great students and teachers. I would like to thank my supervisor Jan Stoklasa for helping me during the thesis writing process. I ran into many problems during the writing process and the support of the supervisors were helpful for finishing this thesis. I would also like to thank the case company for providing the data for the thesis, it allowed me to study a real-world example and make the thesis writing more interesting.

Even under unusual circumstances caused by the Covid-19 virus, the university handled the thesis process well and I can really recommend LUT as a place to study.

In Rauma, Finland on 14.6.2021
Jeremias Marttinen

Table of contents

List of Figures

List of Tables

List of Formulas

List of Abbreviations

AUC	Area under the ROC curve
CCP	Customer churn prediction
CLV	Customer lifetime value
CRM	Customer relationship management
DT	Decision tree
FN	False negatives
FP	False positives
LR	Logistic regression
ML	Machine learning
NN	Neural networks
RF	Random forest
ROC	Receiver operating characteristic curve
SVM	Support vector machines
TN	True negatives
TP	True positive

1 Introduction

1.1 Motivation and background

The purpose of this thesis is to study a customer data from an electricity company which has been suffering from customer churn. The case company has lost private customers and would like to get more insight into the matter. The customer churn has been substantial during the first and second quarter of the 2020 and all the factors behind it all not entirely clear for the case company. Electricity contracts are relatively similar products, and the competition is fierce in the market. There have been also some new companies entering the Finnish electricity market during last years.

Customer churn has been studied in many different industries, including in banking, telecommunication services, insurance, and online gambling. Many different methodologies have been used, including linear regression, logistic regression (LR), neural networks (NN), decision trees (DT), random forest (RF), support vector machines (SVM) and others. Multiple studies like Risselada et al. (2010) and Coussement et al. (2016) have found that DT and LR algorithms resulted in robust results while studying customer churn, and they were methodology included in this study. Table 1 shows examples of customer churn prediction (CCP) in literature.

Examples of churn prediction in the literature.			
Article	**Industry**	**Sizes of datasets**	**Methodology**
Coussement, K. et al.	Online Gambling	3729	DT, RF
Pribil, J. et al.	Electricity	30 000	LR, DT
Wang, Y. et al.	Telecom	4000	DT
Xie, Y. et al.	Banking	1524	RF
Lima, E. et al.	Telecom	15 000	LR, DT
Bolance, C. et al.	Insurance	14 000	LR, DT, NN, SVM

Table 1. Examples of churn prediction in the literature.

Multiple studies have noticed the higher cost of acquiring new customers than keeping existing ones. Olle (2014) found that the cost of acquiring a new customer is five to ten times higher than retaining an existing customer. The future churn of customers is of a paramount importance since knowledge of the churn allows companies to target potentially churning customers and prevent them from leaving. Ideally companies would have probabilities of all customers churning based on accurate model using customer data, and these probabilities could be ranked and customers with most risk of churning could be contacted. Robust feedback system would allow companies to possibly notice reasons for churn and allow effective churn prevention.

There have been some studies made about CCP in electricity industry using machine learning (ML). An article by Pribil and Polejova (2017) used LR and DTs on real electricity data and found that both models are accurate at predicting customer churn in the electricity industry. The data from the article came from Czech Republic. Magdalena et. al. (2017) studied the customer churn of electricity companies in Austria using regression models. Apart from these articles using electricity data, the research on the topic is scarce.

There are not many similar articles available, but electricity business shares some similarities in general with telecom industry in Finland. They both are former monopolies that have been opened to competition due to deregulation. Private customers in both industries generally have low barrier to switch service provides since the competitors offer relatively similar products. The lack of available articles makes it interesting to study whether it is possible to model the customer behaviour.

Customer churn causes huge annual losses to telecommunications companies who struggle to keep their existing customers. (Wei and Chiu, 2002) Customer churn harms telecommunications companies since they lose price premiums, face decreased profits and possible loss of referrals from the churning customers. (Ahn et al. 2006)

Coussement et. al. (2016) studied customer churn in telecommunications industry, and they emphasized the importance of customer centric approach to reduce churn. This means focusing on preventing customers from churning instead of only focusing on acquiring new customers. Potentially churning customers should be targeted with marketing campaign to prevent them from leaving. This type of preventive actions are commonly used by telecommunications companies.

1.2 Focus of the study

This thesis focuses on CCP with DT and LR modelling to build a model that predicts customer churn with the electricity dataset.

The study is a quantitative study and is based on a data set of private electricity customer data. The case company wanted to only focus on private customers and not business customers, since the churn primarily happened in private customers.

This thesis is based on empirical research and building a predictive model with the data from the case company. After that the reliability of the model can be assessed to see whether it can predict the customer churn accurately or not.

Customer churn has not been studied much with ML methodology and electricity data, so this study also works as a case study whether we can build a model which gives better than random predictions of customer churn with electricity customer data with this type of methodology.

1.3 Objective, research questions and hypothesis

The objective of this thesis is to study customer churn problem in a Finnish electricity company. Based on this objective, the following research questions can be formed:

Research questions:

1. What are the features that can be used to identify churning customers?
   Hypothesis: In/out network variable can be used to predict customer churn.

2. Can we predict which customers are likely to leave next?

The existing literature on using ML methodology in studying electricity customer data or finding the features that can be used to identify churning electricity customers is scarce. The existing articles by Pribil and Polejova (2017) and Magdalena et. al. (2017) suggest that the methodology chosen for this study can accurately predict customer churn with electricity data. The results of this study are based

on single dataset received from the electricity company. The limitations of the data make it difficult to generalize the results of the study, but the company's data collection put limits to the complexity of the available data.

The hypothesis on the first research question is formed because the case company's sales strategy is based on defending their market share aggressively around their own network area. This generally means lower prices and better service for the in-network customers. This could be one of the features that can be used to identify churning customers.

To answer to these research questions, literature review into customer churn and ML is done. Then LR and DT models are built with private electricity customer data. These models are then used to classify and predict customer churn. The performance of the models will be evaluated with accuracy, ROC-curve, F-score, and confusion matrix.

1.4 Limitations of the study

The existing research literature provides insight into customer churn in telecommunications, insurance, online gambling, and banking industry, but there is a limited amount of research done studying customer churn in electricity companies. We do not know if all methodology and theory from those studies are applicable to electricity industry.

This study is a quantitative study based on a single dataset and focused on a single company providing the data. It is done from the perspective of the case company and the results are not meant to be generalized to the whole industry.

The data used in this study had 18020 customers and 5 variables, one dependent (churn or not churn) and 4 independent variables. The amount of data seemed reasonable compared to other customer churn studies which can be seen from the table 4 in the page 30. It is possible that better and more accurate results would have been received with larger and more complex datasets. We also do not know whether there was some non-voluntary churn, in other words if some customers agreements were terminated by the case company. If these cases existed in the data, they were rare.

The time span of the data was two quarters because that was the period with most churn and the case company wanted to focus on. Ballings and Van den Poel (2012) found that minimum time span should be one quarter, but generally longer time horizons result in better functioning models.

The data was divided with 80/20, 70/30, 60/40 hold-out and 2, 5 and 10 times cross-validation in terms of the training and testing data. The dataset was treated for imbalance issues with under sampling the data. Under sampling can decrease the accuracy of the model. Different data cleaning and data dividing could have led to a better functioning model.

The results of the models are averages, thus limitation for the study is lack of minimum and maximum values for the models. We do not know the true range of possible results from the worst to the best outcome.

Overfitting is always a risk when using DT type of models. Different type of trees was tested, and the final model used 4 branch tree with ten times cross-validation.

Obvious limitation is also time and resource constraint, this thesis needed to be finished according to the schedule agreed with the supervisor, and this limited the amount of research and models that could have been tested. Researchers have used other models in testing customer churn, for example SVM and NN, but DT and LR were used in this study since these methods were common in the literature and were discussed with the case company.

Features that can be used to identify churning customers were tested by modelling only one predictor variable at time. This method does not tell us about the results of combining multiple different predictor variables together for models.

1.5 Structure of the thesis

The chapter two explains the literature review of customer churn literature and supervised learning models which are used in this study. It also explains which other models were used in relevant studies. There is also a short section for literature selection process, some information about customer churn in electricity industry and a literature framework for evaluation criteria of the models.

In chapter three the ML methodologies used in this study are used in building a model for analysing and predicting customer churn for this study. The chapter starts with the data part which explains the data in more detail and the data pre-processing part. Data tools used in this study are also shortly described and then the algorithms are implemented.

In chapter four the results of the model are being discussed and the accuracy of the models is discussed in the light of the evaluation criteria. The research questions and hypothesis posed earlier will be answered. Finally, there will be some lessons to the industry and future research ideas are suggested based on this study.

2 Literature Review

This part of the thesis deals with the literature review. It includes the literature search process, the customer churn and ML theory, information about the methodology, evaluation criteria and the pre-processing of the dataset.

2.1 Searching for relevant literature

Literature search was done with LUT Primo-service, which searches for peer-reviewed articles from different leading journals. Keywords used were customer churn in electricity industry, customer churn modelling and ML methodology. After defining keywords which were used, total of 1039 articles were found. Relevant articles were filtered by reading the abstracts to determinate the relevance of the articles for the thesis. After filtering non-relevant articles, 41 articles were left which were used for references. Table 2 shows the relevant articles used in this study. Different keywords were used to search from database to find as many articles as possible. The search was limited to English language peer-to-peer reviewed articles, between 2000 and 2020. The literature search process is described by the figure 1.

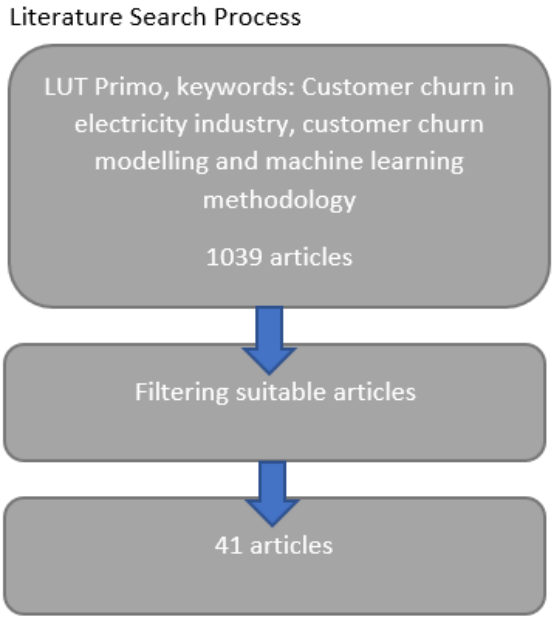


Figure 1. Literature search process

List of relevant articles used in this study		
Authors	Article	Year
Ahn, J. et al.	Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry	2006
Athanassopoulos, A. D.	Customer satisfaction cues to support market segmentation and explain switching behavior	2002
Bishop, C. M.	Pattern Recognition and Machine Learning	2006
Bolance, C. and Guillen, M.	Predicting Probability of Customer Churn in Insurance	2016
Burez, J. and Van den Poel. D.	Handling class imbalance in customer churn prediction	2009
Burez, J. and Van den Poel. D.	CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services	2007
Cawley, G. C. et. al.	On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation	2010
Chakrabarti et. al.	Data Mining Curriculum: A proposal	2006
Chen, C. et al	Using random forests to learn imbalanced data	2004
Coussement, K., De Bock, K	Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning	2013
Coussement, K., De Bock, K	A comparative analysis of data preparation algorithms for customer churn prediction	2016
Cramer J. S.	The Origins of Logistic Regression	2003
Dankers, F. J. W. M et al.	Prediction Modeling Methodology	2018
De ville, B.	Decision trees	2013
Drummond, C. and Holte, R. C.	Class imbalance, and cost sensitivity	2003
Fawcett, T.	An introduction to ROC analysis	2006
Gur Ali, O. and Arıtürk, U.	Dynamic churn prediction framework with more effective use of rare event data	2014
Hosmer, D. W. and Lemeshow, S.	Applied Logistic Regression	2000
Huigevoort, C.	Customer churn prediction for an insurance company	2015
Idris, A. et al.	Intelligent churn prediction in telecom	2013
Keaveney, S. M. and Parthasarathy, M.	Customer switching behavior in online services	2001
Kim, H. and C. Yoon.	Determinants of Subscriber Churn and Customer Loyalty in the Korean Mobile Telephony Market	2004

Lazarov, V. and Capot, M.	Churn Prediction	2007
Lima, E. et. al.	Domain knowledge integration in data mining using decision tables	2009
Magdalena et. al.	Information as potential key determinant in switching electricity suppliers	2017
Narasimha, M. M. and Susheela, D. V.	Introduction to Pattern Recognition and Machine Learning	2015
Neslin, S, A. et al.	Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models	2006
Olle, G.	A Hybrid Churn Prediction Model in Mobile Telecommunication Industry	2014
Peng, J.	An Introduction to Logistic Regression Analysis and Reporting	2002
Powers, D. M. W.	Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation	2008
Pribil, J. and Polejova M.	A Churn Analysis Using Data Mining Techniques: Case of Electricity Distribution Company	2017
Risselada, H. et al.	Staying power of churn prediction models	2010
Sabbeh, F. S.	Machine-Learning Techniques for Customer Retention: A Comparative Study	2018
Vafeiadis, T. et al	A comparison of machine learning techniques for customer churn prediction	2015
Verbeke, W. et al.	Building comprehensible customer churn prediction models with advanced rule induction techniques	2011
Wang, Y. et al	A recommender system to avoid customer churn	2009
Weiss, G. M.	Mining with rarity: A unifying framework	2004
Wei, C. P. and Chiu, I.	Turning telecommunications call details to churn prediction	2002
Witten, I. H. et al.	Data Mining: Practical Machine Learning Tools and Techniques	2016
Xie, Y. et al	Customer churn prediction using improved balanced random forests	2009
Zimek, A. and Filzmoser, P.	Outlier detection between statistical reasoning and data mining algorithms	2018

Table 2. List of relevant articles used in this study.

2.2 Customer churn prediction

Customer relationship management (CRM) refers to building and managing long-term customer relationships. Part of this process is customer churn (Vafeiadis et al. 2015).

Customer churn happens when customers move to a competitor company. This can happen because customers want better quality of service, offers or benefits. The best measurement for this is a churn rate, which all companies should minimize. This had led CCP to become a central part of proactive customer retention planning (Sabbeh, 2018).

Churning customers can be described as either voluntary or non-voluntary churners. Non-voluntary churning customers are customers whose customer relationship has been terminated by the company, for example because they did not pay their bills or violated the terms of service. Voluntary churn refers to customers who have voluntary terminated their contract with the company. The reasons behind this are more difficult to find. Voluntary churn can be either incidental or deliberate. Incidental churn happens when there is a change in the customers financial situation that makes it difficult for them to stay as a customer even if they want to. This type of churn is difficult for the companies to control. Deliberate churn happens when customer makes a deliberate decision to terminate the contract with the company. This type of churn is the problem companies try to solve (Kim and Yoon, 2004).

CCP models try to find customers with a high likelihood of churning. The three most important characteristics of CCP model are accuracy, comprehensibility, and justifiability. Customer segmentation is needed to find and target potential churning customers with marketing to retain these customers. This increases the efficiency of the resources used in marketing. Customer retention is profitable because: a) new clients are more expensive than retaining existing ones, b) old clients are less sensitive to competition, c) dissatisfied customers generate negative word of mouth and d) churning customers have an opportunity cost because of lost sales (Verbeke et al. 2011).

Customer lifetime value (CLV) means the present value of the customer minus the cost of acquiring, keeping, and developing the customer. Revenue, cost, discount, and time horizon are important parts of this process. CLV should be compared to the probability of the churn so appropriate marketing measures can be targeted at the customer (Lima et al. 2009).

Customer retention is supported as a hard but at the same time gradually developing and important business strategy in services. It is argued that there is a higher cost for acquiring new customers than retaining existing ones (Athanassopoulos, 2000).

Reactive retention in the CRM is when customer has already terminated the contract and the customer will be approached after that. The customer can be offered better prices or services. The advantage of this system is that it allows the company to know which new supplier the customer wants to change, and this makes it possible to give the customer a better offer. The disadvantage of this is the complexity of the process, contacting the customer, keeping up with the deadlines, and responding quickly and efficiently. Proactive retention on the other hand tries to prevent customers from leaving before they have filed for termination of their contract. Ideally, company would know about the customers who are potentially churning beforehand and could target them with marketing to prevent them from leaving (Pribil and Polejova, 2017).

The CCP can be studied with a large enough dataset which has churning and non-churning customers. This dataset can then be used to build a classifier. This classifier classifies datapoints to categories based on the dataset. Examples of models used to build churn classifiers are NN or DT. (Lazarov and Capot, 2007).

CCP is a key part of a customer retention strategy. It is a process based on the data from the customers past behaviour which is used to build a model to whether customers churn or not. Once the likelihood has been established, the customers can be ranked based on that, and the most likely churners can be targeted with customer retention campaign (Keaveney and Parthasarathy 2001).

This CCP process raises many practical questions. How should we decide which customer we are trying to retain? How do we calculate customer lifetime value? When is the optimal time to contact customers? How should we reach the customers, and what should we offer them? The goal of this process is to reduce the number of leaving customers, but customers cannot be held at any cost. (Pribil and Polejova, 2017)

Coussement et al. (2016) found multiple logistic regressions to be effective at studying customer churn in the telecommunications industry. Also, Neslin et al. (2006) found that for companies doing predictive modelling logistic regression is a good technique to start with.

Coussement and De Bock (2013) used random forest methodology to study customer churn with gambling industry. They found that RF models have a higher than random likelihood in classifying churning customers.

There are multiple studies in which DTs have been successfully used to predict customer churn in the telecommunications sector. Gürsoy (2010) achieved very high accuracy on them model predicting customer churn in telecommunications sector. Also, Wang et al. (2009) found that DTs are suitable for studying customer churn at telecommunications industry. Using DTs for studying customer churn in insurance sector has been studied for example by Huigevoort (2015), although in her study the model did not give the most reliable results.

Bolance and Guillen (2016) used also NN and SVM to study customer churn in insurance industry. NN and SVM resulted in strong accuracy and Area under the ROC curve (AUC) figures in this study. The accuracy and AUC for NN were 88,72% and 92,71%, and for SVM they were 86,24% and 90,34%.

It is also important to notice that even though the performance of the models used in CCP is the most important criteria for model selection, there are other criteria as well. The expertise of the employees at the company, the running time of the algorithms and the access to necessary software for using the models. More complex models require specialized skills and training analyst for using them is costly (Coussement and De Bock 2013).

Burez and Van den Poel (2007) emphasised the early detection and targeting of churning customers to increase profits. The study used three different ways of targeting potentially churning customers, free incentives, special events and loyalty surveys. They found that CRM approach of first recognizing the churning customers, and then targeting them with customer loyalty surveys lead to doubling of the profits of the case company. The methodology used was logistic regression and random forests. The study shows the benefits of customer retention strategy in significantly increasing the business profits. The AUCs of the random forest and logistic regression models were both 77%.

2.2.1 Customer churn in the electricity Industry

There have been few articles studying customer churn in the electricity industry. An article by Pribil and Polejova (2017) from Czech Republic used LR and DT models on electricity data. Both models were found to be accurate at predicting customer churn in the electricity industry.

Magdalena et. al. (2017) studied the customer switching of electricity companies in Austria. Switching is defined as customer voluntary changing electricity supplier. The difference between switching and churn is that churning customer can just terminate the electricity contract and stop using electricity altogether, whereas switching customer always changes to another service provider. They found multiple reasons for switching customers between electricity suppliers.

1. Customers with better knowledge of electricity providers are more likely to switch.
2. The higher the self-reported popularity of the electricity provider is the lower the switch rate is.
3. Customers who have switched supplier once before are less likely to switch again (possibly due to thinking that they already have the best deal or because the found the process too difficult)

European Energy Markets Observatory (2010) writes about the change in electricity pricing in European Union. Increased competition puts pressure on decreasing electricity prices and thus margins of the service providers. To avoid customers from churning, companies must adapt their pricing. This highlights the importance of cost cutting measures. Cost savings potential can be found from:

1. Credit and default management: Credit defaults of customers have increased, and bad debt management has become important.
2. Customer retention: Electricity companies have followed telecommunications and insurance companies in retention strategies, offering multiple services to customers to increase the customer stickiness.
3. IT efficiency: Electricity companies have increased investment into CRM systems and data mining to collect more information about customer churn and to implement CCP strategies

Even uncompetitive markets can become competitive quickly due to deregulation. Once electricity market is open to competition, dramatic customer switching can happen. In many European electricity markets increased customer switching is caused my aggressive marketing, decreased prices, high levels of customer awareness, easy online switching of service providers and comparing electricity providers on social media (European Energy Markets Observatory, 2010).

Customer switching is one of the signs of competitive electricity market. The figure 2 shows the switching rates of different European countries. We can see the general trend of increased switching between the suppliers after the liberalization of the European electricity market. Finland also has relatively high switching rate, the average between 2011-2015 was 10.12 percent. (CEER report, 2017)

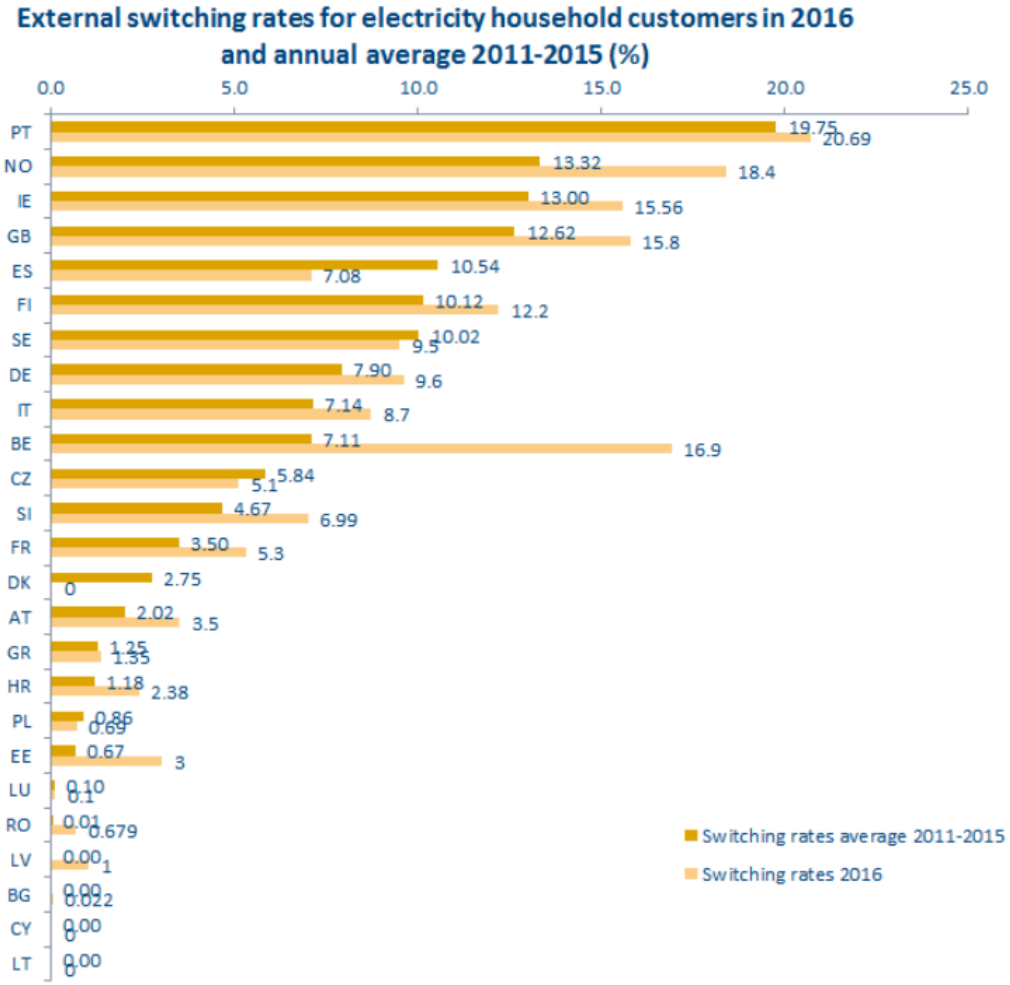


Figure 2. Annual electricity customer churn rate (%) in different European countries. (CEER report, 2017)

The 2017 CEER report offers some explanations for customer switching behaviour. Regulatory barriers can prevent customer's from easily switching between suppliers. Inadequate monetary benefits can deter customers from switching, since changing the contract only makes sense if the benefit is higher than the cost of switching. The complexity of the switching process, lack of trust in the supplier, and overall satisfaction and loyalty towards the supplier are mentioned as explanations for switching behaviour.

2.3 Machine learning

Computers cannot think abstractly in the same way as humans do. They work with the representations of the patters and not the patterns itself. For example, discriminating between two objects with computer, we must abstract these objects and use the representations of them on computer. The input and the output of the ML or pattern recognition system is abstraction of pattern (Narasimha and Susheela, 2015).

Pattern recognition is divided into two parts: classification and clustering. Classification can be explained as giving a class label to unlabelled classes and clustering as giving a class label to unlabelled patterns. Classification models are built based on learning process. There are two types of learning processes for classification, supervised learning, and unsupervised learning. Supervised learning uses set of variables which already have class labels on them, whereas unsupervised learning tries to find patterns of regularities and irregularities in a dataset. (Narasimha and Susheela, 2015)

For example, digit recognition cases in which the goal is to classify each input to one of a finite number of discrete categories, are classification problems. Churn modelling is another example of classification since churn or non-churn is a binary classification problem. Regression problem is supervised method when the output variable in continuous (Bishop, 2006).

Other ML tasks are regression, ranking and dimension reduction. In ranking instead of classifying we are trying to get a probability which we can use to give a pattern a class. In classification dimensionality reduction can be done by feature selection because of the space requirements for algorithms. (Narasimha and Susheela, 2015) Data mining is at the cross-point of ML and artificial

intelligence, and it tries to convert data into more easily understandable form. (Chakrabarti et. al. 2006).

The data used in this thesis did not require reducing dimensionality since the dataset and variables came from readymade reports from the company's database and were not too complex.

Most of the classifiers have a training and testing phase, so the data is divided first to train the model and then to test it. In this thesis both hold-out and cross validation were used to divide the data and test different models.

2.4 Methodology

2.4.1 Logistic Regression

Logistic regression outlines and tests the relationship between a categorical dependent variable and one or multiple categorical or continuous independent variables. An example is a linear regression for one continuous predictor X and one dichotomous outcome variable Y. The plot of this data is in two parallel lines, each corresponding to a value of the dichotomous outcome. This leads the plot to have a S-shape, which would be hard to show with linear equation since the extremes do not follow a linear trend. In practice, the logistic regression now predicts the logit of Y from X. Logit is the natural logarithm of odds of Y, and odds are ratios of probabilities of Y happening (Peng et al. 2002). Formula 1 shows a simple logistic model form, where α is the Y intercept, β is the regression coefficient.

$$\text{logit}(Y) = \textit{natural}\log(\textit{odds}) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X.$$

Formula 1. Simple logistic model form. (Peng et al. 2002).

In customer churn modelling logistic regression is often used since it uses binary prediction of a categorical variable (for example churn) which is modelled based on one or more predictor variables. The model is used to classify observations to two classes, churn, and no-churn. The model is trained

with training data and then the trained model is used to guess the probability of observations belonging to one group or another. Logistic regression has been used with good results in customer churn studies, sometimes with as good results as decision trees (Vafeiadis et al. 2015).

Example of difference between linear and logistic regression is seen from the figure 3. This figure shows results for a continuous outcome using linear regression and for a binary outcome using logistic regression. The logistic regression predictions are rounded to class A or B based on a threshold. (Dankers 2018).

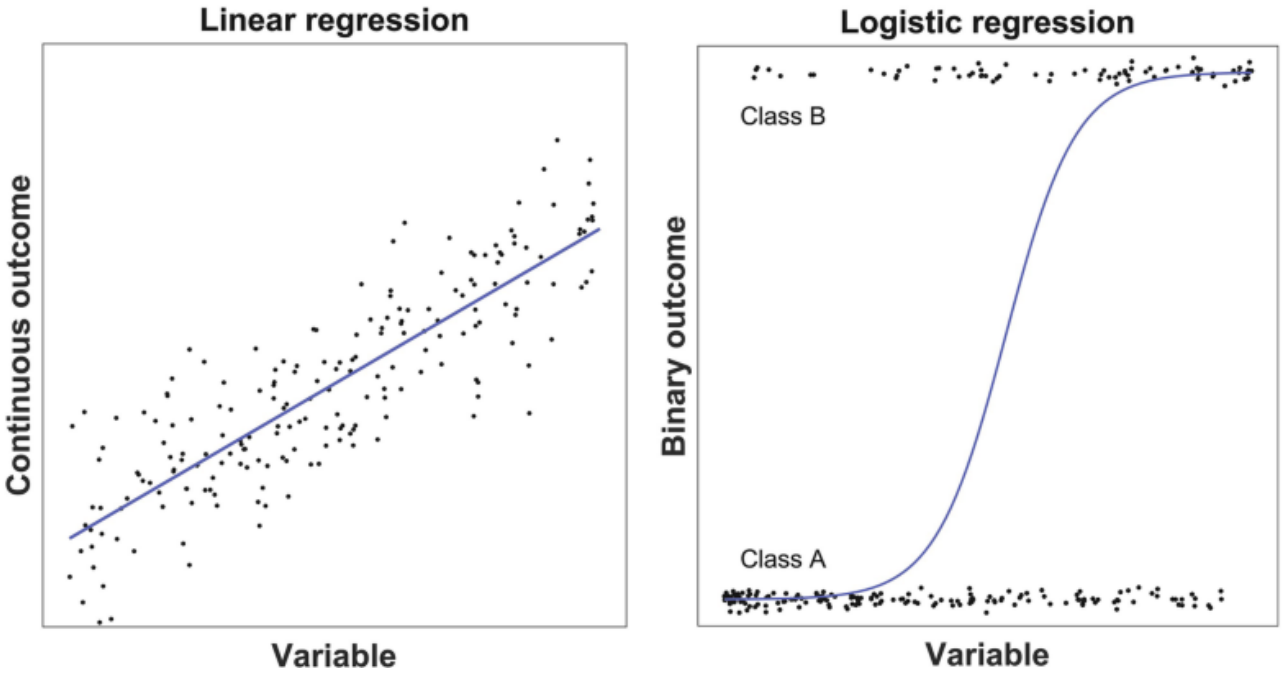


Figure 3. Comparison of linear regression to logistic regression plot. Dankers (2018).

2.4.2 Decision tree

Decision trees are exceptionally adaptable methods for statistical analysis, adaptable for different data qualities. They offer robust performance even when data is missing and offer different ways of including missing data into models. DTs are also flexible and simple to operate. They offer robust performance with few assumptions. (Barry de Ville, 2013). Also, Risselada et al. (2010) found that

DT algorithms performed best in studying customer churn. Figure 4 shows a simple example of a decision tree.

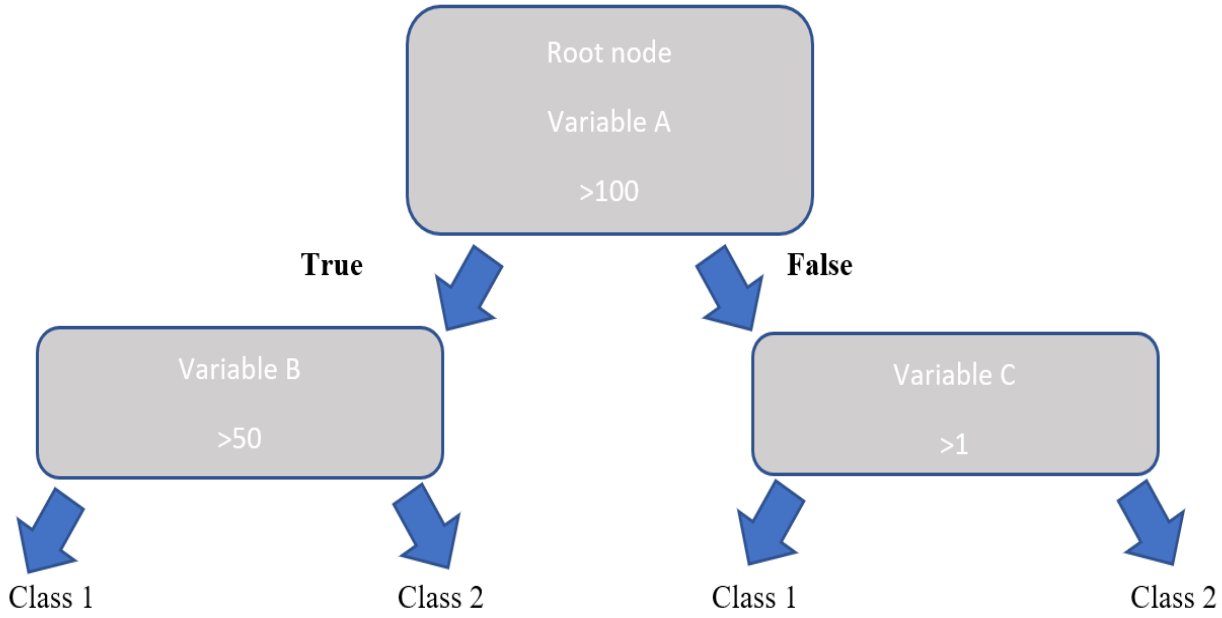


Figure 4. Example of a simple decision tree

DTs are tree like structures which generate classification rules for datasets. They are also called classification trees or regression trees. In these trees, leaves mean class labels and branches mean combination of features that lead to those labels. They can be very accurate at studying customer churn, depending on the data used (Vafeiadis et al. 2015).

DT is made of a set of rules that divide the variable into smaller, more similar sets. With prediction models, to find out the probability of the churn the customer enters the tree from the root node. After that each node is a test that moves the customer to the lower-level nodes until terminal node is reached. This path from the root node to the terminal node makes up the rules used to calculate the probability. All customers with the same path receive the same probability. (Coussement and De Bock 2013).

DT is an adaptable model that supports both categorical and continuous data. Due to their adaptability, they have become one the most commonly used methods in CCP (Sabbeh, 2018).

2.5 Data cleaning and pre-processing

Fortunately for companies, customer churn numbers are often not as large as the numbers of non-churning customers. (Burez and Van den Poel, 2009) However this does create a problem when studying customer churn with ML methods which work better without large class imbalances.

Weiss (2004) describes different problems often found in imbalanced datasets.

1. Inadequate evaluation metrics. Sometimes the best metrics are not used for evaluation the results of the methods.
2. Absolute rarity: The number of examples in one class is small in absolute sense, which makes it difficult to find out irregularities.
3. Relative rarity: The number of examples in one class is small in relative to other objects.
4. Data fragmentation: Many methods like DT models work by dividing the data in to smaller and smaller subsets.
5. Inductive bias. Model findings should not be generalized.
6. Noisy data. Noisy data can affect the way the models work.

The first problem is dealt with by introducing multiple different evaluation metrics which will be described in the next subchapter.

The absolute rarity is generally not an issue with churn studies, since usually there are a lot of data available. However, the relative rarity can be an issue, the differences in the amount of data can cause severe data imbalance issues, which can be helped by under or over-sampling of the data or treating the outliers (Burez and Van den Poel, 2009).

The most common method to deal with relative rarity (imbalances in dataset) is to under or over sample the data. Under sampling removes the most common cases from the class, whereas over sampling multiples the rarest cases in the class. They both decrease the class imbalance, but also have

shortcomings. Under sampling ignores cases that could be useful and thus decrease the accuracy of the model. Over sampling can increase the time used in building the model, and possibly lead to overfitting the model. (Weiss, 2004)

Studies like Drummond and Holte, (2003) have shown over sampling to be ineffective at building the models. Studies like Chen et al. (2004) have shown that under sampling is more effective than over sampling. These findings combined with the risk of overfitting in over sampling led under sampling to be used in this thesis to deal with class imbalances.

Another issue with the data is outliers. Outlier is a data point that significantly differs from the other observations. According to Zimek and Filzmoser, (2018), there is not and objective mathematical test for determining what is an outlier and thus it is subjective.

2.6 Evaluation Criteria

Objective and strong performance evaluation is the foundation of ML research. Without performance indicators for different algorithms, it is impossible to assess the strength of the model. (Cawley et al. 2010)

Cawley et al. (2010) write about the undesirable optimistic bias that can happen with models due to over-fitting in model selection. Overfitting in choosing the model is likely to be worst when data is insufficient, and the parameters are large. To achieve less biased performance, more robust methods are needed such as cross-validation. Cross-validation is an easy and robust method for model selection and performance evaluation. The k-fold cross-validation splits the data randomly to form k disjoint subsets of roughly equal size. These sets are used to evaluate the performance of the model. The average of the performance of all k folds gives a slightly pessimistic estimation of the performance of the model.

2.6.1 Confusion Matrix

Confusion Matrix, or error matrix, allows the visualisation of the performance of the classification models by showing the amount of correctly and falsely predicted classes. This can be expressed either

in absolute or relative terms. (Powers, 2011) Table 3 shows a confusion matrix with correctly and falsely predicted classes.

		Predicted class	Predicted class
		Churning	Non-churning
Actual class	Churning	True positive (TP)	False negative (FN)
Actual class	Non-churning	False positive (FP)	True negative (TN)

Table 3. The confusion matrix.

Vafeiadis et al. (2015) described evaluation measures as follows:

Precision is the proportion of predicted positive cases that are correct.

Precision = $TP / (TP+FP)$ (1)

Sensitivity, or recall is the proportion of positive cases that are correctly classified.

Sensitivity = $TP / (TP+FN)$ (2)

Specificity is the proportion of negative cases that are correctly classified.

Specificity = $TN / (TN+FP)$ (3)

Accuracy is the total amount of predictions that were correct.

Accuracy = $(TP+TN) / (TP+FP+TN+FN)$ (4)

F-score is the harmonic mean of precision and recall.

F-score = $(2 * Precision * Recall) / (Precision + Recall)$ (5)

2.6.2 Area under the ROC curve

Receiver operating characteristic (ROC) is a graphical plot that shows the performance of binary classifier model. AUC is used to measure the quality of the classifier. The AUC number should be

between 0.5 and 1, closer to 1 means higher quality classifier. ROC analysis has become the standard evaluation metric (Powers, 2011).

Burez and Van den Poel (2009) also noted that AUC is good evaluation metric, especially because it does not depend on any threshold, unlike for example accuracy.

In terms of interpretating the results of AUC, the score of 0.5 suggest no discrimination, score of 0.7 to 0.8 can be viewed as acceptable, 0.8 to 0.9 is excellent and higher than 0.9 is outstanding (Hosmer and Lemeshow, 2000).

The confusion matrix is used to calculate the AUC. The ROC-graphs have two axes, and the TP rate is on the y-axis and the FP rate is on the x-axis. The graph shows the trade-off between benefits (TP) and costs (FP). The point (0.0) is where there is never a positive classification, the point (1.1.) only has positive classifications and the point (0.1) is the perfect classification. The AUC score tells us the probability that the classifier classifies a randomly chosen positive example higher than a randomly chosen negative one. ROC-curves can be used to show the connection between sensitivity and specificity in different cut-offs for a test or a combination of tests (Fawcett, 2006). Figure 5 shows an example of ROC curve.

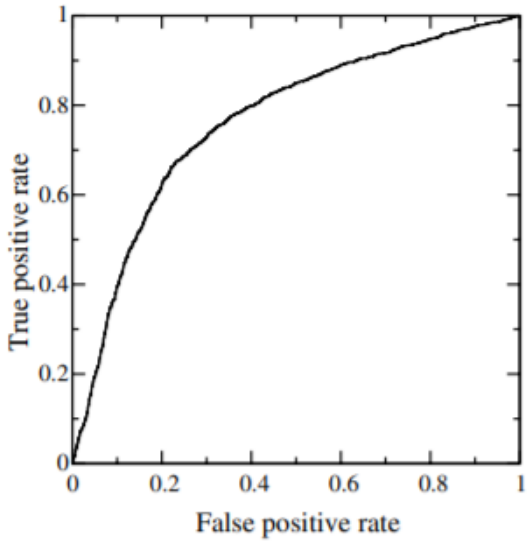


Figure 5. Example of ROC curve (Fawcett, 2006).

3 Developing the model to study and predict the customer churn

This section deals with exploring the data used in the thesis and building the model to study customer churn. It starts with describing the data and the pre-processing of the data. Then model tools are explained and finally models are built, and results are visualised in tables.

3.1 Data and description

The data used in this study was downloaded from customer data reports from the company's database. The reports had been already made using SQL-searches from the company's database. The dataset contains 18020 rows (customers) and 5 columns (variables). The dataset is from the first and second quartal of the 2020 and contains only customers from Finland. It contains data from the customers who have left the company during these quarters. There has not been any noticed seasonality in the churn. There is a total of 11644 non-churning customers and 6376 churning customers in the dataset. The ratio of non-churning customers to churning customers in this final cleaned dataset is 65:35. Figure 6 shows the data searching and model building process.

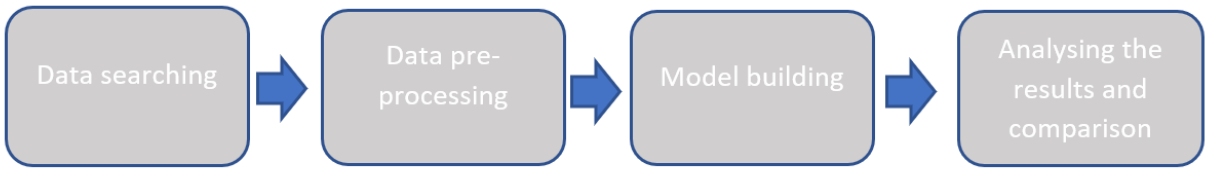


Figure 6. Data searching and model building

Most of the customer churn research studies the datasets are between a few thousand customers and tens of thousands of customers. The typical churn rates are between 20-30% in these studies. Few examples can be seen from the table 4. The dataset used in this thesis is like these datasets in terms of the amount of data used.

Examples of churn rates in the literature.		
Article	**Total customers**	**Churn Ratio**
Coussement, K. et al.	3729	31 %
Bolance, C. et al.	14 000	20 %
Pribil, J. et al.	30 000	20 %
Lima, E. et al.	15 000	14 %

Table 4. Examples of churn rates in literature.

Neslin, S, A. et al (2006) found that logistic regression and decision trees work well in customer churn modelling when the data period is at least three months, but they recommend preferably longer time horizons. The time horizon of two quarters was chosen because that was the time of most customer churn at the company, and longer time horizons would have added to more complexity to the thesis.

Variables of the dataset		
1.	Customer churn	(1=not churn, 0=churn)
2.	Annual electricity consumption	(kilowatt hours)
3.	Electricity product the customers are using	(5 different product classes)
4.	In company's electricity network	(1=in, 0=out)
5.	Length of customer relationship	(years)

Table 5. Variables used in the study.

Table 5 describes the variables used in this study. The variables included categorical variables (customer churn, product category and electricity network) and discrete data (electricity consumption and length of customer relationship). Customer churn and electricity network are binary 1 or 0, annual electricity consumption is in kilowatt hours between 0 and 100 000 kilowatt hours per year. Electricity

consumption is the estimation of the current year's consumption, which is calculated by the last 12 months average. There are five different product categories, which are not described in detail due to privacy concerns. They are only referred to with numbers from one to five. The length of the customer relationship varies between 0 and 5 years. Figure 7 visualizes the annual electricity consumption versus product categories and customer churn.

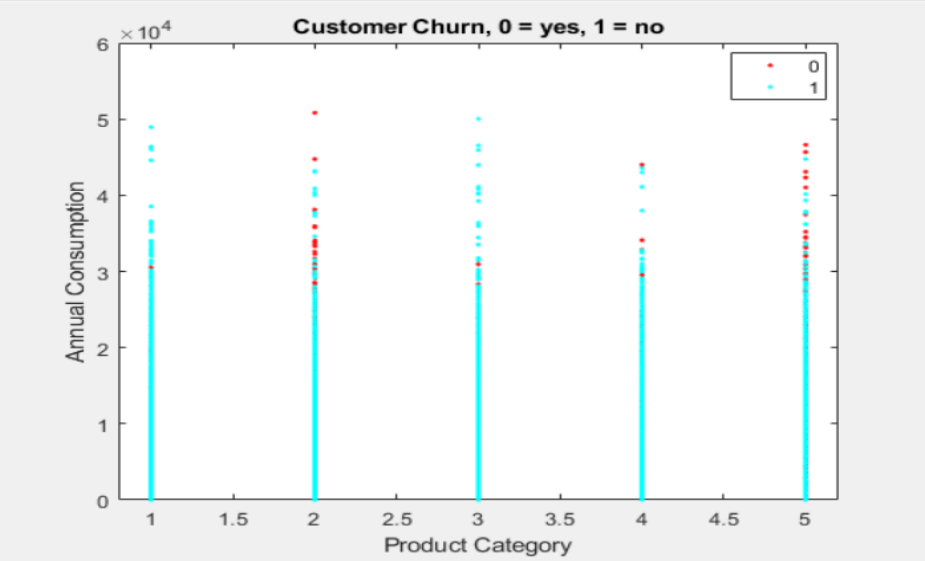


Figure 7. Annual electricity consumption versus product categories and customer churn.

3.2 Data cleaning and pre-processing

The dataset contained 30777 rows (customers) and 7 columns (variables). The dataset is from the first and second quartal of the 2020. The dataset was derived from readymade quarterly reports the company's database and converted into an excel file. After gathering the data, it was possible to exclude certain variables and customers that were not relevant for this study. The business customers were excluded from this study because the case company wanted to only focus on private customers. There were total of 1660 business customers which were excluded. Customer identification numbers and business customer labels were not necessary variables for the analysis and thus they were excluded. This left total of five variables for the final data used in this study.

In the data cleaning part, some customers who did not have necessary info were excluded. Total of 902 nulls, missing values, duplicated information, and mathematical symbols were also excluded.

As stated in the literature part, under sampling was decided to use in this thesis since studies like Chen et al. (2004) have shown that under sampling is more effective than over sampling. The literature review combined with the risk of overfitting in over sampling led under sampling to be used in this thesis to deal with class imbalances.

In this thesis the major class was non-churning customers. By under sampling non-churning customers, the class imbalance was reduced. Customers who have been with the company for over 5 years were excluded since the churn rates among the long-term customers were very low and the case company wanted to focus on newer customers. Also, the customers who had stayed with the company for more 5 years had lower annual electricity consumption figures, which made them less interesting to study. This under sampling excluded 10195 customers. This significantly decreased the class imbalance in the data and made the model more robust, since the classifiers are more accurate with more churning customers. However these older customers could also contain some important information. Figure 8 shows the total customers and churning customers before and after under sampling.

This data treatment left the dataset with total of 18020 customers, which 6376 were churning customers. This means that the churn rate is 35% after the data treatment. Figures 9, 10 and 11 show the distribution of the data within variables in/out of network, length of the customer relationship and product category.

Two new variables were created from the existing variables. The churning customers were derived from the ending contract dates, and the how long the customers have been with the company was derived from the start date of the contracts.

Before dividing the data with the hold out and cross validation methodology, the whole dataset was randomized by the Matlab classification learner application.

3.2.1 Class balances

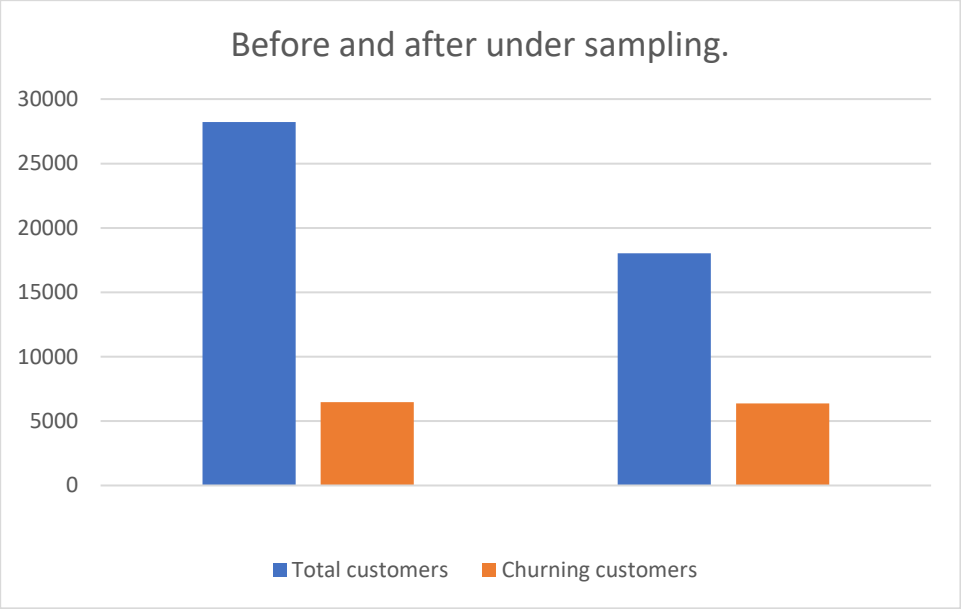


Figure 8. Total customers and churning customers before and after under sampling.

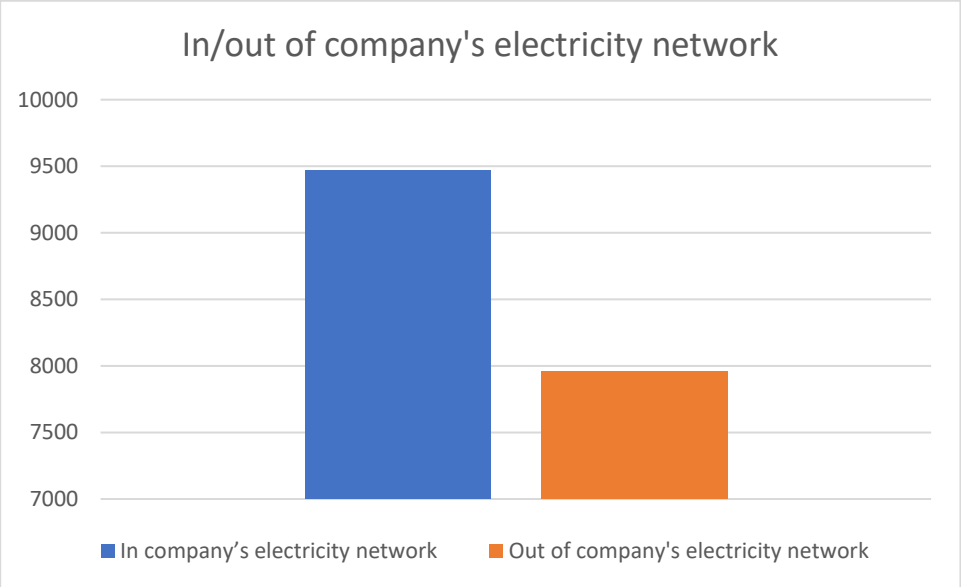


Figure 9. Number of customers in the company's electricity network.

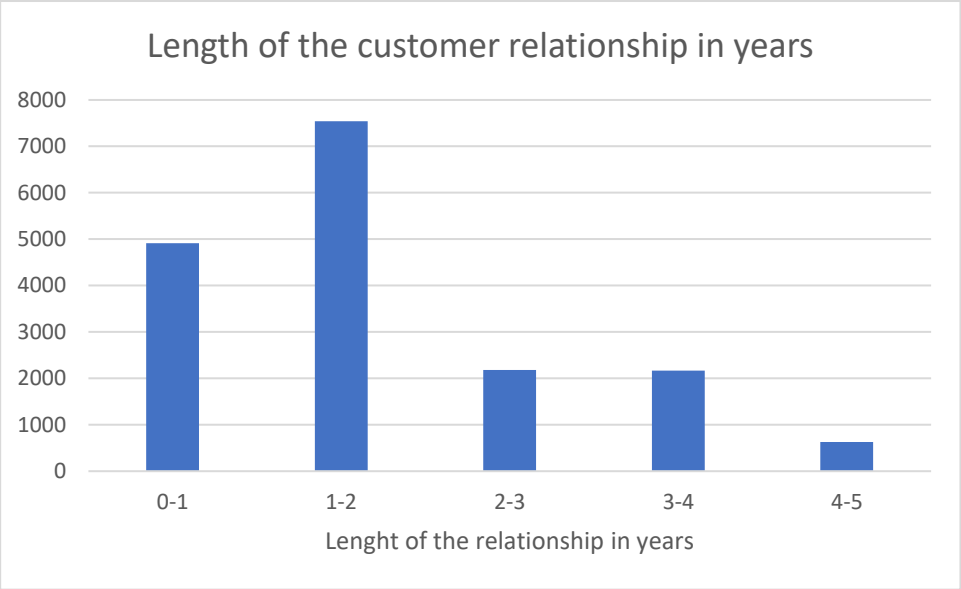


Figure 10. The length of the customer relationship in years.

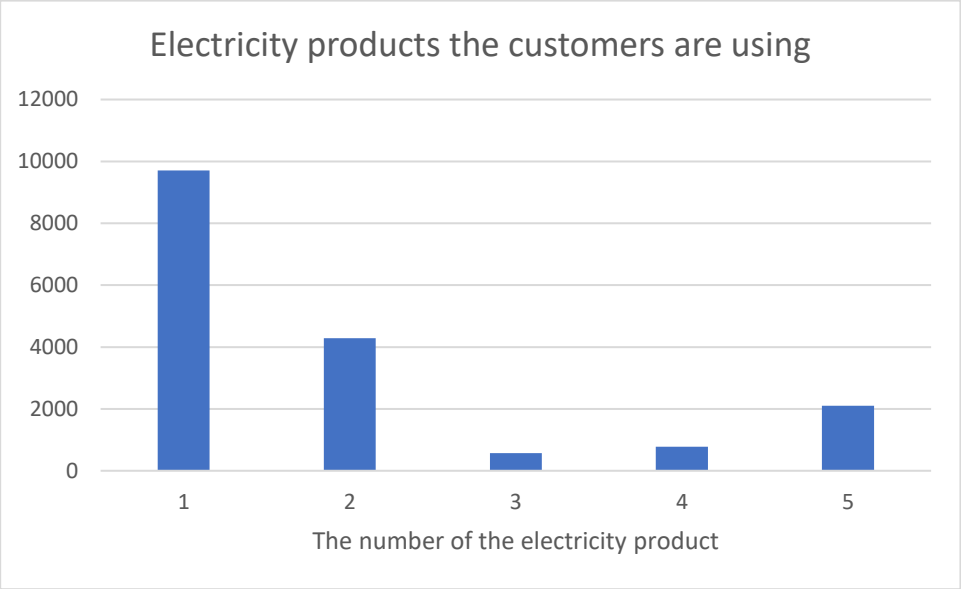


Figure 11. Electricity products the customers are using.

3.3 Data tools and libraries

The data used in this study was searched with a SQL-search form the company's database and converted into an excel file. The database had readymade reports which contained the customer data needed for the analysis. These reports could be exported to Microsoft Excel. The data preparation, cleaning, pre-processing, and organising was done with Matlab R2017b statistical software. The analysis and visualisation used in this thesis were also done with Matlab using the classification learner application. Matlab classification learner application was also used for the plots, and histograms used in this thesis.

3.4 Implementing the algorithms

First, we run the LR and DT models with all variables and with six different data dividing methods. The best results were achieved with the ten-time cross validation as the data dividing method. The 10-times cross validation was also suggested by the literature, such as Witten, 2016.

Later individual predictor variables are modelled against the response variable using to see which of the features can be used to identify churning customers.

3.4.1 Comparison of the different data diving methods

The results of the algorithms with these three different models can be seen from tables 6, 7, and 8. The data was divided with 80/20, 70/30, 60/40 training data split and with 2, 5- and 10-times cross validation split. Different data splits were tried to see how the results differ under different data separation. The two times cross validation was done as an experiment to see what kind of results we get even though two times cross validation is not enough for the final model. The cross-validation models gave almost the same results. As we can see, regardless of how the data is split the results are robust.

These tables show results of the whole model with all 5 variables.

	80/20		70/30		60/40	
(%)	LR	DT	LR	DT	LR	DT
Accuracy	79	87	79	87	79	86
AUC	82	87	83	88	83	88
F-score	43	74	44	74	44	74

Table 6. Comparison of the different hold-out data dividing methods.

	2 times		5 times		10 times	
(%)	LR	DT	LR	DT	LR	DT
Accuracy	79	87	79	86	79	87
AUC	83	87	83	87	83	87
F-score	44	74	44	74	45	74

Table 7. Comparison of the different cross-validation data diving methods.

Comparison of the models (with ten times cross validation)		
(%)	Logistic Regression	Decision Tre
Accuracy	0.79	0.87
AUC	0.83	0.87
F-Score	0.45	0.74

Table 8. Comparison of the logistic regression and decision tree models with 10 times cross-validation.

3.4.2 Comparison of the logistic regression and decision tree models

Accuracy of the Logistic Regression model (ten times cross validation)			
(%)	Predicted Churn	Predicted Non-churn	Class recall (%)
Actual churn	1423 (TP)	1120 (FN)	56
Actual Non-churn	2524 (FP)	12343 (TN)	
Class precision (%)	36		

Table 9. Accuracy of the logistic regression model (ten times cross validation).

Accuracy of the Decision Tree model (ten times cross validation)			
(%)	Predicted Churn	Predicted Non-churn	Class recall (%)
Actual churn	3362 (TP)	1758 (FN)	66
Actual Non-churn	603 (FP)	11705 (TN)	
Class precision (%)	85		

Table 10. Accuracy of the decision tree model (ten times cross validation).

3.4.3 Influence of individual predictor variables on the response variable

One of the research questions was to identify which variables have most influence over the dependent variable out of the variables taken from the company's database. After running the LR and DT models we can see the DT model performed better than the LR model. DT model with ten times cross validation was used to study just the individual variables since it was more effective model than LR. Matlab uses the mean ROC curve for the cross validations.

The feature that can be used to identify churning customers was the product category. The annual electricity consumption had little influence. The in/out network and how long the customer had been with the company had no influence. The figure 12 shows the ROC-curve of the decision tree with the

product category as predictor variable and customer churn as response variable. The figure 13 shows ROC-curve of the decision tree with the in/out of network as predictor variable and customer churn as response variable. The AUC with only product category as independent variable was 0.85. The same figure for in/out of network was only 0.51. The AUC for electricity consumption was 0.77, and for the length of the customer relationship it was 0.58.

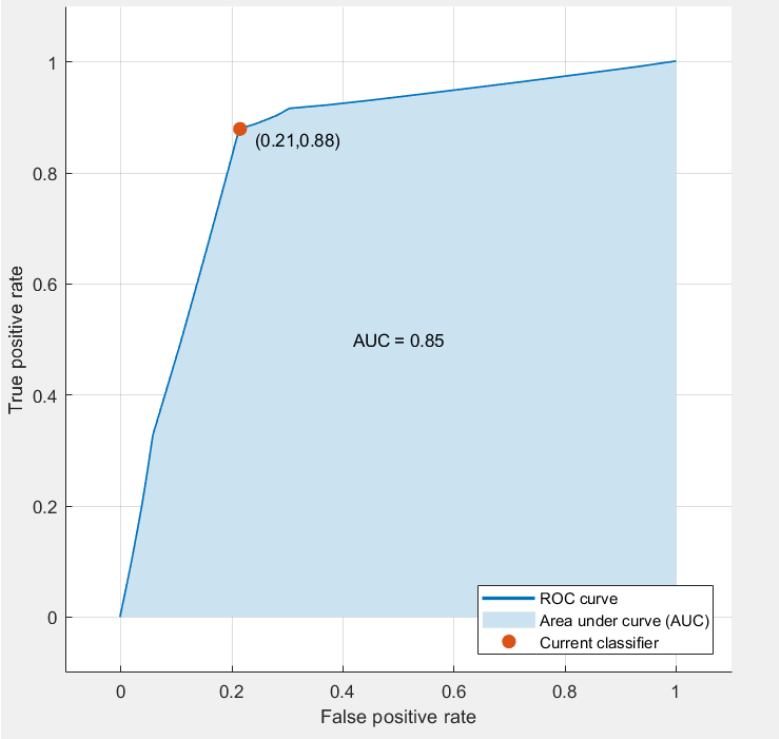


Figure 12. The ROC-curve of the decision tree with product category and customer churn with 10 times cross validation.

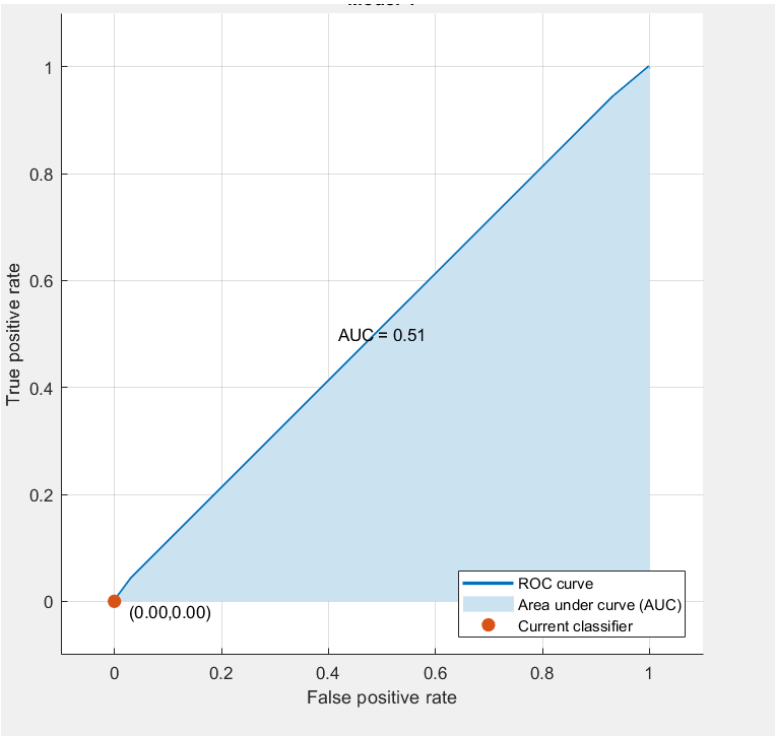


Figure 13. The ROC-curve of decision tree with in/out of network and customer churn with 10 times cross validation.

4 Summary and conclusions

All the models were better at predicting non-churning customers than churning customers. Class imbalances are usually with customer churn studies, there are usually fewer churning customers than non-churning customers and that can explain why the model works better this way.

These results of this thesis can be used in multiple ways. Since DTs were found to be accurate at predicting churn, the company's churn modelling can be based on DT. More accurate segmentation of customers is needed to recognize churn patters better. The company recognized that more accurate customer data is needed.

Secondly, since product category had a highest effect on churn, more focus is needed on product categories in customer development. Some of the products need to be revaluated and possibly discontinued.

Thirdly, the company's sales strategy was based on the fact whether the customer belongs to in/out of network, focusing sales on the in-network customers. There seems to be no difference in churn between those groups, thus this sales strategy can also be revaluated.

Bolance, C. and Guillen, M. (2016) also used DT and LR to study customer churn in insurance industry, and their dataset was similar size to the one used in this study. The results they had for the accuracy and AUC of LR were 86,19% and 90,23% and for DT model 85,28% and 90,28%. However, they cited risks of overfitting in their study.

Pribil and Polejova (2017) article which had similar dataset of electricity customer data and methodology got relatively similar results. They found that both the DT and LR model performed well. Their best DT model had an accuracy of 88% in predicting churning customers, AUC of 81% and f-score of 64,4%.

The results of this thesis are like theirs; the DT model is very effective at predicting customer churn with high accuracy.

From the reviewed papers, Coussement et al. (2016), Wang et al. (2009) and Neslin et al. (2006) among others had similar results in that the articles showed that DTS can be used to accurately

measure customer churn. Most of the reviewed articles agreed that DTS are more accurate, but in some papers both models got the same accuracy.

Neslin et al. (2006) used combination of several DTs with each two to eight tree nodes. Coussement et al. (2016) used Classification and Regression Trees, using two-way splits which split a node in two smaller nods until no more splits can be done since all customers are identical in terms of the target variable. Wang et al. (2006) did not specify the decision tree type they used.

The results clearly indicated that the DT model is better at predicting the churn with this dataset than the LR. The reason for this could be that the DTs can examine better the relatively non-linear nature of the division of the data. The data also has nominal variables. This gives better classification accuracy for the DT model. However, the DT model has a risk of overfitting, and this must be taken into consideration. The results of the models are averages, so we do not know the minimum and maximum values for the models. There was also a significant difference between the f-score of the DT and LR model.

After implementing the algorithms, the research questions and hypothesis can be answered.

4.1 Answers to research questions

What are the features that can be used to identify churning customers?

The AUC for the DT with 10 times cross validation with only product category as independent variable was 85%. The similar figure for in/out of network was only 51%. The AUC for electricity consumption was 77%, and for the length of the customer relationship it was 58%. Within the independent variables, the product category was the feature that had the highest effect on identifying churning customers. Also, annual electricity consumption had little effect. How long the customer had been with company and the whether the customer was in network or out network proved to be worse than random selection.

Hypothesis: In/out network variable can be used to predict customer churn.

The AUC of in/out network variable was only 51%. The hypothesis that in/out network variable can be used to predict churning customers can be rejected based on the performance of the classifier.

Can we predict which customers are likely to leave next?

Both the models had a high accuracy in the predictions and can be used to predict churning customers. Especially the DTs was robust and effective at predicting churning customers. LR had an 79% accuracy and DTs had 87% accuracy. The AUC for LR was 83% and for DTs 87%. The f-score LR was 45% and for DTs 74%.

4.2 Further research topics

Both LR and DT were found to be usable methods to study customer churn in the electricity industry. It would be interesting to do more broad study about customer churn with other ML methods as well. Xie et al. (2009) used balanced RF for studying customer churn, which reduced the noise and helped with the imbalance issues often found in customer churn studies. This type of more refined DT model could be interesting topic to study and then compare it to the simple decision tree model used in this thesis.

There are many other possible research opportunities in customer churn as well. More complex data with more variables and larger datasets could be used. It is important to remember that the results of this study are based on single dataset received from the electricity company. The limitations of the data make it difficult to generalize the results of the study, thus more research into the topic is needed.

One further research topic is to increase the time span of the data to years. Ballings and Van den Poel (2012) found that increasing the time span over 5 years in studying customer churn does not significantly increase the accuracy. But studying a data set of 1-5 years could be a topic for next research.

Finally, customer churn modelling could be used at electricity companies to increase customer retention by predicting the customers who are at danger of leaving the company and focusing retention efforts on them. Lima et al. (2009) suggested adding more domain knowledge to customer churn modelling, for example the number of calls to the customer helpdesk. As acquiring new customers is more expensive that retaining existing ones, these types of models have potential to increase profits for the companies.

References

Ahn, J. et al. (2006) 'Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry', *Telecommunications Policy,* 30, 10-11, pp. 552–568.

Athanassopoulos, A. D. (2000) 'Customer satisfaction cues to support market segmentation and explain switching behavior', *Journal of Business,* Research 47, pp.191–207.

Bishop, C. M. (2006), 'Pattern Recognition and Machine Learning', Springer.

Bolance, C. and Guillen, M. (2016) 'Predicting Probability of Customer Churn in Insurance ', *Research Gate.* DOI: 10.1007/978-3-319-40506-3_9

Burez, J. and Van den Poel, D. (2009) 'Handling class imbalance in customer churn prediction' *Expert Systems with Applications.* Volume 36, Issue 3, Part 1, pp. 4626-4636. Available at: https://doi.org/10.1016/j.eswa.2008.05.027

Burez, J. and Van den Poel, D. (2007) 'CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services' *Expert Systems with Applications.* Volume 32, pp. 277–288. Available at: https://doi.org/10.1016/j.eswa.2005.11.037

Cawley, G. C. et al. (2010) 'On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation', *Journal of Machine Learning Research* 11.

Chen, C. et al (2004). 'Using random forests to learn imbalanced data', *Technical Report* 666. Statistics Department, University of California at Berkeley.

Council of European Energy Regulators CEER (2017). Retail Markets Monitoring Report. [www.document] [Accessed 30 August 2020] Available at: https://www.ceer.eu/documents/104400/6122966/Retail+Market+Monitoring+Report/56216063-66c8-0469-7aa0-9f321b196f9f

Coussement, K. and De Bock, K. (2013) 'Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning', *Journal of Business Research* 66, pp. 1629–1636.

Coussement, K et al. (2016) 'A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry', *Decision Support Systems*, Volume 95, Pages 27-36

Cramer J. S. (2003) 'The Origins of Logistic Regression', *Tinbergen Institute Working Paper* No. 2002-119/4.

Dankers, F. J. W. M et al. (2018) 'Prediction Modeling Methodology', *Fundamentals of Clinical Data Science*. pp 101-120.

Chakrabarti et. al. (2006). 'Data Mining Curriculum: A proposal', ACM SIGKDD.

De ville, B. (2013). 'Decision trees', *WIREs Comput Stat* 2013, 5:448–455. doi: 10.1002/wics.1278

Drummond, C. and Holte, R. C. (2003). 'Class imbalance, and cost sensitivity: Why under-sampling beats over-sampling', In: *Workshop on learning from imbalanced data sets II, international conference on machine learning*.

Elitedatascience. (2017) Overfitting in machine learning: What it is and how to prevent it. [www document]. [Accessed 1 August 2020]. Available https://elitedatascience.com/overfitting-in-machine-learning.

European Energy Markets Observatory (2010): 'Energy, utilities and chemicals', 2009 and Winter 2009/2010 Data Set Twelfth Edition.

Fawcett, Tom. (2006) 'An introduction to ROC analysis', Pattern Recognition Letters Volume 27, Issue 8, June 2006, Pages 861-874.

Gur Ali, O. and Arıtürk, U. (2014) 'Dynamic churn prediction framework with more effective use of rare event data: The case of private banking', Expert Systems with Applications 41, pp. 7889–7903.

Hosmer, D. W. and Lemeshow, S. (2000). 'Applied Logistic Regression'. 2nd Ed. Chapter 5. New York, NY: John Wiley and Sons. pp. 160 –164.

Huigevoort, C. (2015) 'Customer churn prediction for an insurance company' Eindhoven University of Technology, research.tue.nl

Idris, A. et al. (2013) 'Intelligent churn prediction in telecom: Employing mRMR feature selection and rotboost based ensemble classification', Applied Intelligence 39, pp. 659–672.

Keaveney, S. M. and Parthasarathy, M. (2001). 'Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors', Journal of the Academy of Marketing Science, 29(4), 374–390.

Kim, H. and C. Yoon. (2004) 'Determinants of Subscriber Churn and Customer Loyalty in the Korean Mobile Telephony Market', *Telecommunications Policy*. 28: pp. 751-765.

Lazarov, V. and Capot, M. (2007) 'Churn Prediction', Bus. Anal. Course. TUM Comput. Sci, 2007 - Citeseer.

Lima, E., Mues, C., & Baesens, B. (2009). 'Domain knowledge integration in data mining using decision tables: Case studies in churn prediction.', *Journal of the Operational Research Society*. 60, pp. 1096–1106.

Magdalena et. al. (2017). 'Information as potential key determinant in switching electricity suppliers.' Zeitschrift für Betriebswirtschaft; Heidelberg Vol. 87, Iss. 2, pp. 263-290. DOI:10.1007/s11573-016-0821-9

Mathworks (2020), Available at: https://se.mathworks.com/?s_tid=gn_logo

Narasimha, M. M. and Susheela, D. V. (2015) 'Introduction to Pattern Recognition and Machine Learning', pp.1-5.

Neslin, S, A. et al (2006) 'Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models', *Journal of Marketing Research* 204 Vol. XLIII 204–211. Available at: https://doi-org.ezproxy.cc.lut.fi/10.1509/jmkr.43.2.204

Olle, G. (2014) 'A Hybrid Churn Prediction Model in Mobile Telecommunication Industry', *Int. J. E-Educ*. E-Bus. E-Manag. ELearn.

Peng, J. (2002) 'An Introduction to Logistic Regression Analysis and Reporting', *The Journal of Educational Research* 96(1):3-14 DOI: 10.1080/00220670209598786

Powers, David M W (2008). 'Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation'. *Journal of Machine Learning Technologies*. **2** (1): 37–63.

Pribil, J. and Polejova, M. (2017) 'A Churn Analysis Using Data Mining Techniques: Case of Electricity Distribution Company ', Proceedings of the World Congress on Engineering and Computer Science Vol I WCECS.

Risselada, H. et al. (2010) 'Staying power of churn prediction models', *Journal of Interactive Marketing* 24, pp. 198–208.

Sabbeh, F. S. (2018) 'Machine-Learning Techniques for Customer Retention: A Comparative Study', *International Journal of Advanced Computer Science and Applications*, 9(2). doi: 10.14569/ijacsa.2018.090238.

Vafeiadis, T. et al (2015) 'A comparison of machine learning techniques for customer churn prediction', *Simulation Modelling Practice and Theory* Volume 55, pp. 1-9. Available at: https://doi.org/10.1016/j.simpat.2015.03.003

Verbeke, W. et al. (2011) 'Building comprehensible customer churn prediction models with advanced rule induction techniques', *Expert Systems with Applications* 38, pp. 2354–2364

Wang, Y. et al (2009) 'A recommender system to avoid customer churn: A case study ', *Expert Systems with Applications* Volume 36, Issue 4, pp. 8071-8075. Available at: https://doi.org/10.1016/j.eswa.2008.10.089

Weiss, G. M. (2004). ‘Mining with rarity: A unifying framework.’ SIGKDD Explorations, 6(1), pp. 7–19.

Wei, C. P. and Chiu, I. (2002) ‘Turning telecommunications call details to churn prediction: a data mining approach. *’Expert Systems with Applications*, Volume 23, Issue 2, Pages 103–112

Witten, I. H. et al. (2016) ‘Data Mining: Practical Machine Learning Tools and Techniques. ’ Morgan Kaufmann.

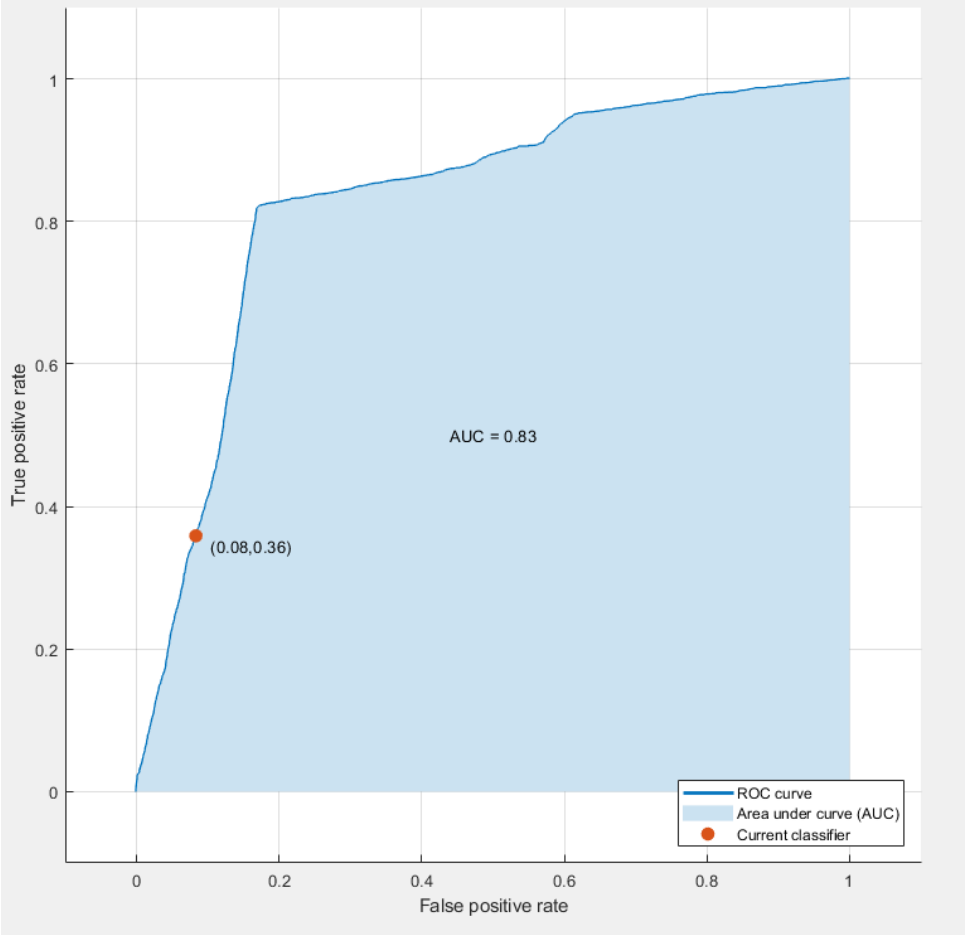
Xie, Y. et al. (2009) ‘Customer churn prediction using improved balanced random forests’, *Expert Systems with Applications* 36, pp. 5445–5449.

Zimek, A. and Filzmoser, P. (2018). ‘There and back again: Outlier detection between statistical reasoning and data mining algorithms’ *Wiley Interdisciplinary Reviews*: Data Mining and Knowledge Discovery.

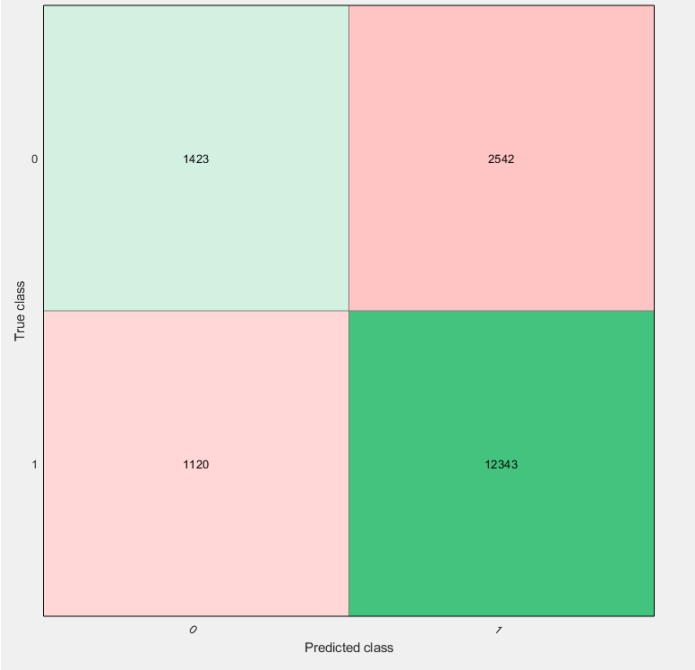
Appendices

Appendix 1. Logistic Regression Model

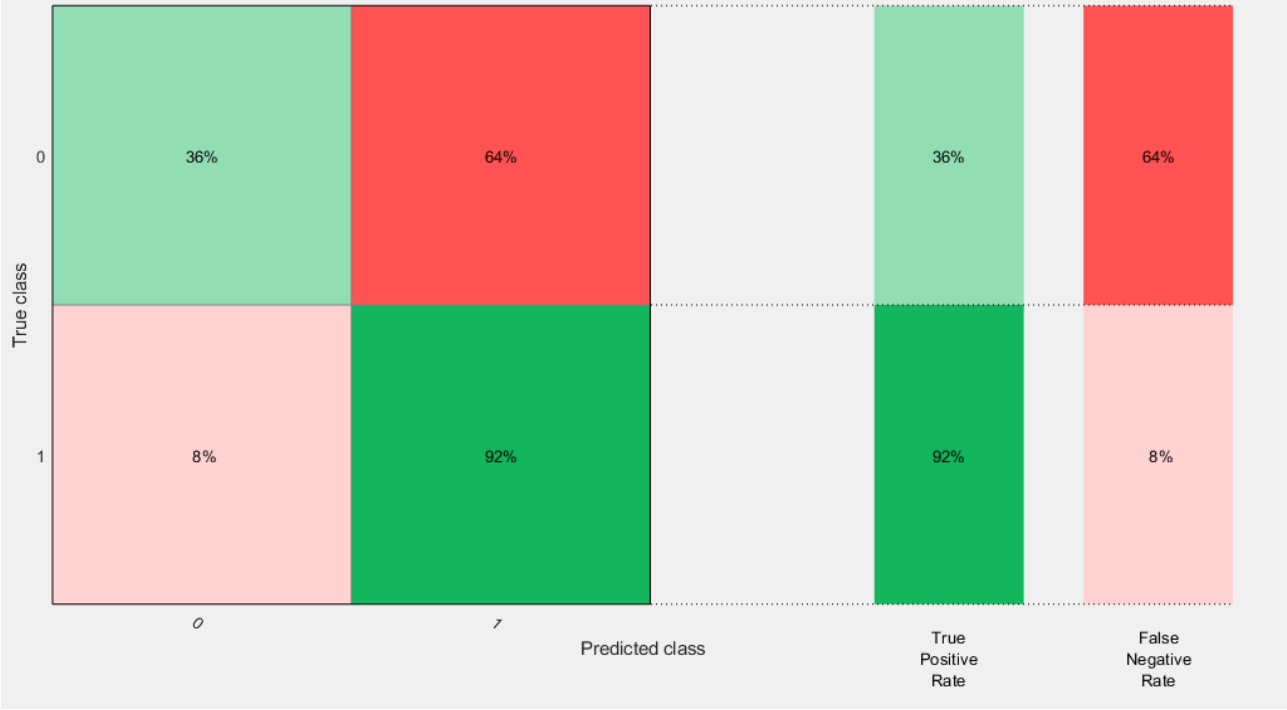
ROC-curve



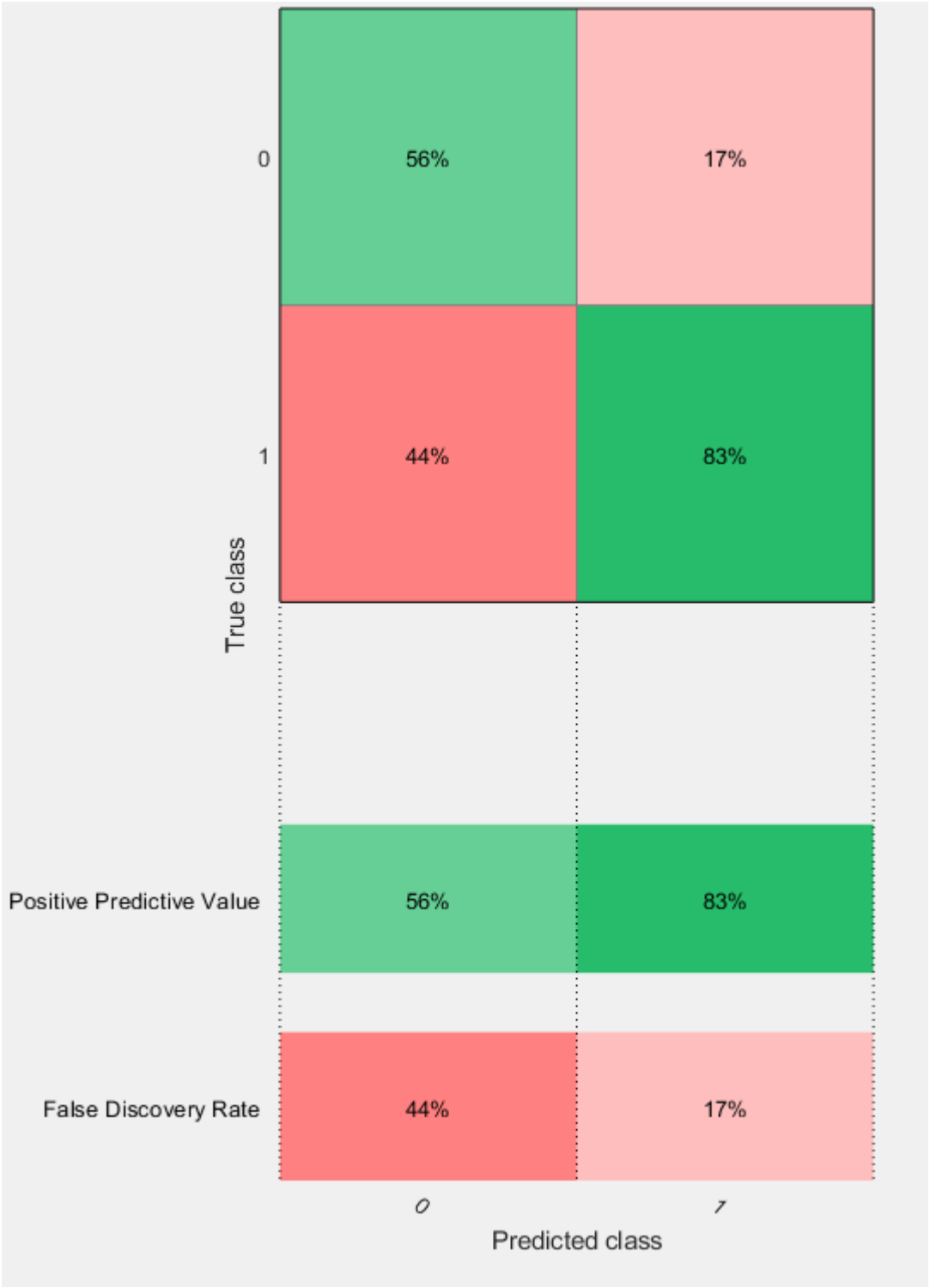
Number of observations



True positive and true negative rates

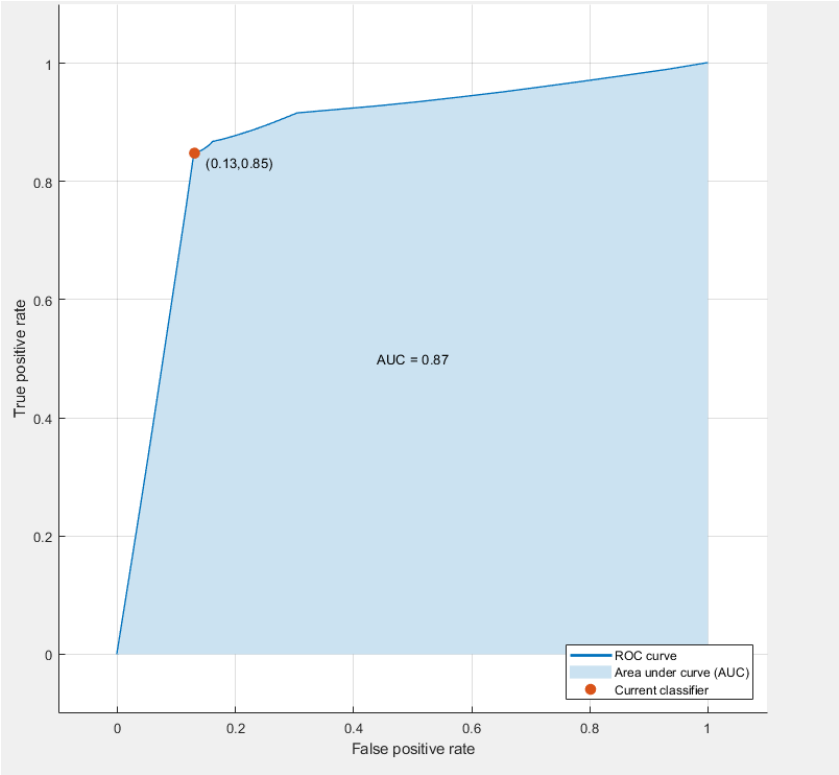


Positive predictive values and false discovery rates

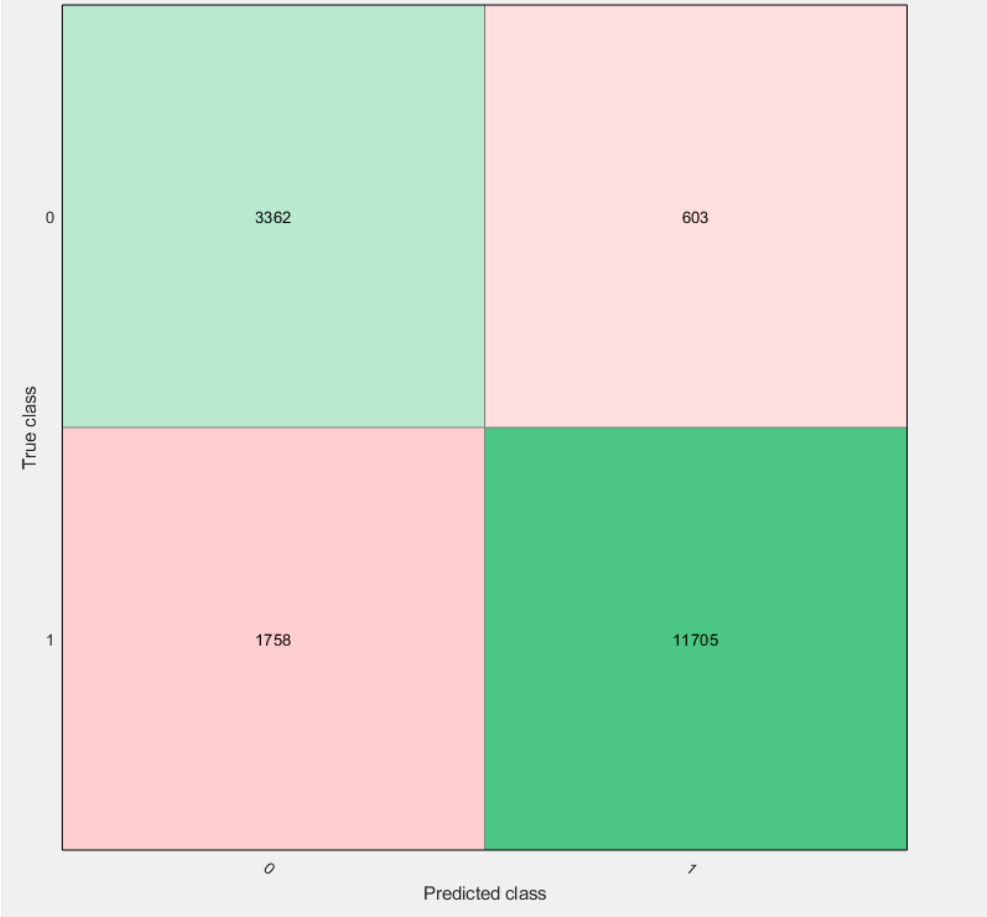


Appendix 2. Decision tree Model

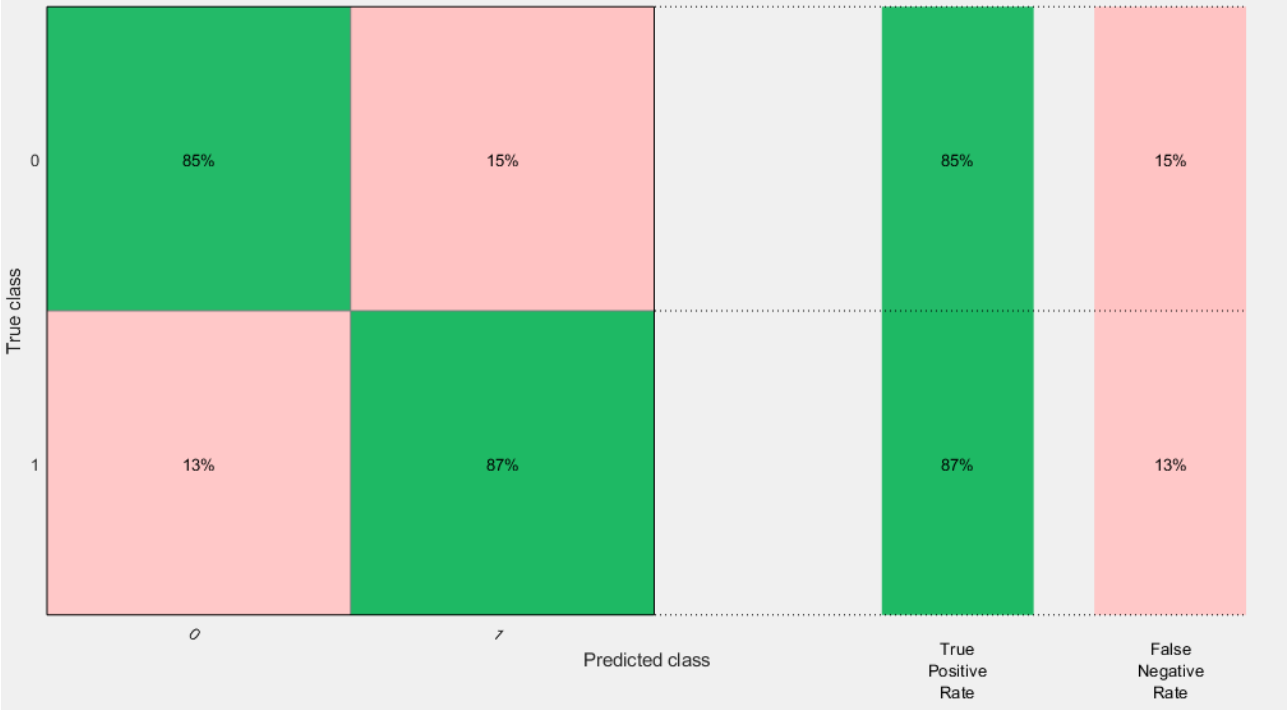
ROC-Curve



Number of observations



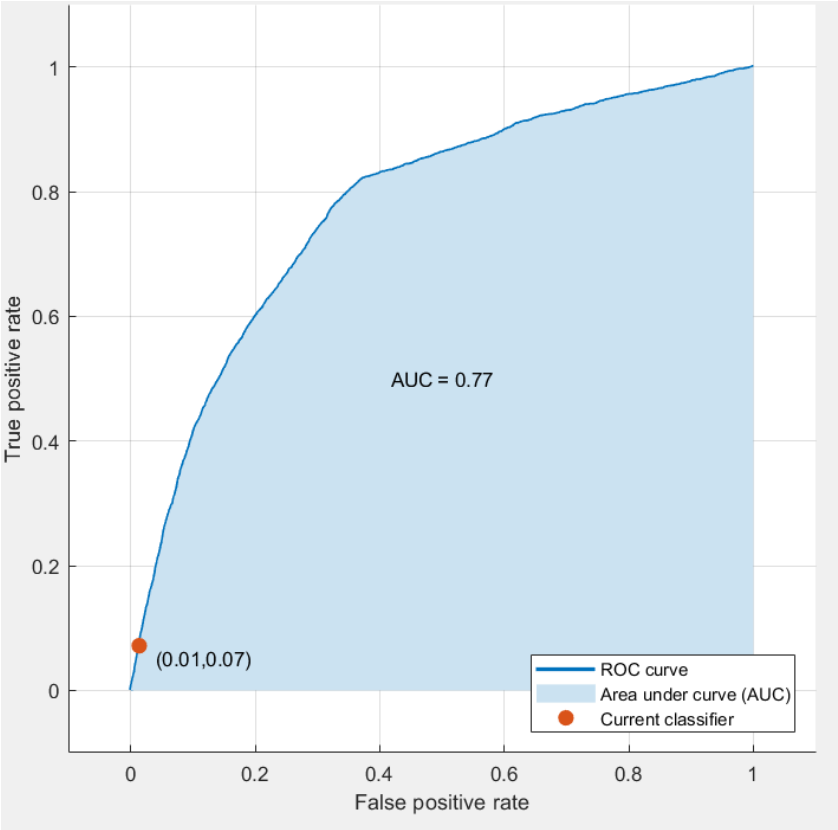
True positive rates and false negative rates



Positive predictive values and false discovery rates

Appendix 3. ROC-curves for electricity consumption and length of customer relationship as predictor variables

ROC-curves 10 times cross validation with only electricity consumption as predictor variable



ROC-curves 10 times cross validation with only length of customer relationship as predictor variable