university of groningen / faculty of economics and business

ASSIGNMENT 1

Churn Prediction

Course:

Data Science Methods for MADS

(EBM216A05.2022-2023.1)

Redacted for privacy reasons

TABLE OF CONTENTS

Managerial Summary 3

I. Business problem 4

II. Research design 4

2.1 Literature review 4

2.1.1 Relationship length 4

2.1.3 E-mail list 5

2.1.4 Starting Channel 5

2.1.5 Energy usage & Household Energy Label 5

2.2 Hypotheses Formulation 6

III. Data preparation 6

3.1 Data Overview 6

IV. Explorative analysis 9

V. Modeling 12

5.1 The baseline model 12

5.2 Stepwise logistic regression 13

5.3 Decision trees Forward 14

5.4 Bagging and Boosting 15

5.5 Random Forests 15

5.6 Support Vector Machines 16

VI. Managerial Conclusion & Model Recommendation 17

VII. References 18

Managerial Summary

When the core product of a business is identical for all of the market players, acquisition is often more expensive than retention. Having relevant insights into which customers are churning and what the reason for churning is, is therefore vital for companies. This research aims to develop a churn prediction model for a Dutch energy supplier, of which product is identical for all market players. Using such a model, the company will be able to more specifically target customers that are likely to churn, using retention strategies.

The *base model* of this research is logistic regression model, of which the independent variables are based on a literature review. We control this base model through six different models - among other things; stepwise logistic regression, decision trees, bagging, boosting, random forests and support vector machines- using machine learning techniques. All of the models are estimated based on a training set and validated on a validation set. The included independent variables of the base model are relationship length, contract length, e-mail list dummy, starting channel dummy, energy usage, and household label. This model performed pretty well with a hit rate of 70%, a TDL of 1.75, and a Gini coefficient of 0.54.

The *stepwise regression*, that uses forward classification and an AIC threshold, performed better with a hit rate of 75%, a TDL of 1.589, and a Gini coefficient of 0.658. The insights about the variables this stepwise regression model uses could help increase the accuracy of the base model. Furthermore, the *CART tree*, which is very easy to interpret, did slightly better than the base model in terms of the hit rate: 71.37%, having a TDL of 1.5 and a Gini coefficient of 0.5. The *bagging* and *boosting* models, which are quite hard to interpret, showed different results. The bagging model has a hit rate of 69.88%, a TDL of 1.46, and a Gini coefficient of 0.52, but is less sensible for variations in the data sample. The boosting model is very accurate, with a hit rate of 75.65%, a TDL of 1.57, and a Gini coefficient of 0.67. Also the Random Forest model is very accurate with a hit rate of 75.15%, a TDL of 1.57, and a Gini coefficient of 0.65. Finally the SVM RBF kernel model was also highly accurate with a hit rate of 75.47%, a TDL of 1.57, and a Gini coefficient of 0.64.

After controlling our base model, we conclude that although some machine learning techniques score higher in terms of accuracy, the base model is already performing pretty well. Also, not all of these control models are as easy to interpret. Therefore, the base model contributes already a lot to the understanding of which customers of this energy supplier are more likely to churn. This explanation could be complemented using the insights of the control models. Using these insights we found that income is an important variable, which adds accuracy to the churn predictions of our models. Therefore, we recommend adding that variable to the baseline model.

Finally, our estimation of the SVM model provided a very interesting insight. According to the SVM plot, people who have high electricity usage and a low-medium income could be classified as churners.

I. Business problem

In this assignment we are dealing with churn prediction. Hereby we are looking from the company's perspective. The problem at hand is that of a Dutch energy supplier whose core issue is customer churn, being concerned about customers leaving their business which can result in revenue loss. Research indicates this is particularly relevant, as customer acquisition is more expensive than retention in industries where the core product is identical for all market players, such as in the case of energy (Coussement & Van den Poel, 2008; Lu, 2002; Verbeke et al., 2011).

The objective behind the business problem is to develop the best model for churn prediction. Knowing which customers are likely to churn enables a company to target those customers with strategies to prevent them from churning. Using different predictive models, we acquire insights about which are the factors that influence churn the most and therefore, possible solutions to improve customer retention.

II. Research design

2.1 Literature review

2.1.1 Relationship length

The rate of customer churn is an important metric for managers because it is a critical input to assess customer lifetime value. To effectively plan interventions to reduce customer churn rates, managers need to obtain both an accurate prediction of the customers' churn rate and empirical estimates of the impact of different drivers of customer churn.

Since the cost associated with customer acquisition is much greater than the cost of customer retention, putting effort into satisfying and keeping customers over the long term can increase profitability. Moreover, long-term customers generate higher profits, tend to be less sensitive to competitive marketing activities, become less costly to serve, and may provide new referrals through positive word-of-mouth, while dissatisfied customers might spread negative word-of-mouth (Verbeke et al., 2011). Thus, the longer the relationship between the customer and the company, the more likely it is that the customer is satisfied and the lower the probability of churning.

2.1.2 Contract length

In the research paper of Lemmens and Croux (2006) the accuracy of a churn prediction model for a U.S. wireless telecommunications company has been improved using bagging and boosting classification techniques. Among other things, they found that the most important predictor of churn in this particular case is the amount of days people own their cellular phone, which could be explained by combined one-year subscriptions and free cellular phone packages. For this reason, it might also be interesting to investigate the effect of contract length on churn in our case as many energy suppliers offer 1 or 2-year contracts in combination with some entry

bonuses. Therefore, customers could be incentivized to switch energy providers after their contract ends.

2.1.3 E-mail list

Using churn management campaigns, customers could be prevented from churning, which can lead to enormous increases in revenue (Neslin, Gupta, Kamakura & Mason, 2006). It could be more costly and difficult to reach the customers of which the e-mail addresses are unknown, causing managers not to choose to target these customers for the churn management campaigns. Therefore, it might be interesting to investigate the effect of knowing customers' email addresses on churn.

2.1.4 Starting Channel

According to a stepwise logistic regression done by Nogueira (2022), the channel from which a customer is acquired has merit in churn prediction. Additionally, the researchers confirmed their logistic regression result by running Stat Explore analysis, which showed the acquisition channel to have the highest variable worth, indicating it is suitable to add to their churn prediction model. Furthermore, the research of Cristian (2016) also showcases the relevance of adding the acquisition channel in a churn prediction model, especially one for energy suppliers, which is appropriate. Therefore, it could also be interesting for us to add this to our prediction model.

2.1.5 Energy usage & Household Energy Label

Drawing upon the similarities between the telecommunication subscriptions and energy markets, in that both industries offer an identical core product, we turn our attention to research by Lemmens and Croux (2006). They found that customer churn risk varies as consumption habits change. Namely, when consumption decreases, the likelihood of churn increases. If the consumption is constant the customer is less likely to churn. Therefore it might be applicable for us to include electricity and gas usage in our model.

Next, we focus on home energy labels. The household energy label is defined by the energy efficiency of the home. Thereby homes with a higher energy label consume less energy. According to our initial descriptive statistics, namely examining mean electricity and gas consumption per energy label, we have found that homes with label “G” in our sample, use on average 41% more electricity and 243% more gas than ones with label “A”. Additionally, energy labels are also negatively correlated with churn. Based on these descriptive statistics and the finding of Lemmens and Croux (2006), it may be impactful to include electricity and gas usage as well as, household energy label as part of our model.

2.2 Hypotheses Formulation

Drawing from our literature review, we form the following directional and non-directional hypotheses between churn (DV) and the six above-proposed variables (IVs):

- **H_1: Customer relationship length is negatively associated with churn**. In other words, the higher the relationship length is, the lower the churn is and vice versa.
- **H_2: Contract length has a significant effect on churn**.
- **H_3: E-mail list is negatively associated with churn**. In other words, if a customer has an email listed with the company, the churn is lower.[1]
- **H_4: Starting channel has a significant effect on churn.**
- **H_5: Electricity usage is negatively associated with churn**: the lower electricity usage is, the higher the churn is and vice versa.
- **H_5: Gas usage is negatively associated with churn**: the lower the gas usage is, the higher the churn is and vice versa.
- **H_6: Home energy label has a significant effect on churn.**

III. Data preparation

3.1 Data Overview

To get the first insights in the data, we started by calculating summary statistics for all of the variables. Using the 'summary' function in R we immediately found that there are no NA values in our dataset. Using the 'dplyr' package, we computed mean statistics of relevant variables for customers that churned and for customers who did not churn. These statistics already gave some first insights as it seems like that customers who churned have higher gas- and electricity consumptions. Also, as seen the table 1, they seem to have a worse energy label, shorter relationship, and their contract expiration is closer on average then for the non-churners.

Churn	AVG_age	AVG_relation	AVG_contract	AVG_label	AVG_electricity	AVG_GAS	Total records
no	48.8 yr	66.6 months	11.5 months to go	3.81	2227 kWh	1217 M3	10 198
yes	46.6 yr	52.2 months	6.3 months to go	4.57	2577 kWh	1452 M3	9 802

TABLE 1: CHURN

[1] With the assumption that the company intends to implement churn management campaigns targeting customers through email.

Furthermore, we investigated what influence does the energy label of the home of the customer have on churn rate, yearly electricity usage (kWh) and yearly gas usage -cubic meters-.
Overall, it can be seen from the below figure that the usage per home for electricity and gas has an upward trend and people churn, they tend to have a worse energy label.

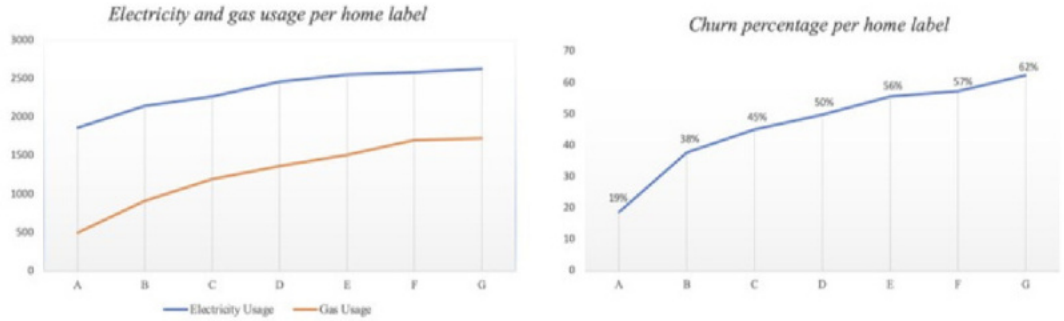


FIGURE 1: CHURN (left one A; right one B)

Variable distributions

In order to find potential outliers and inconsistencies in the dataset, we created boxplots and histograms for the relevant variables. In this analysis, we have observed a few customers that were below the age of 18 and we assumed this is a mistake in the data as it is not possible to have a contract before turning 18. Thus, those observations have been removed.

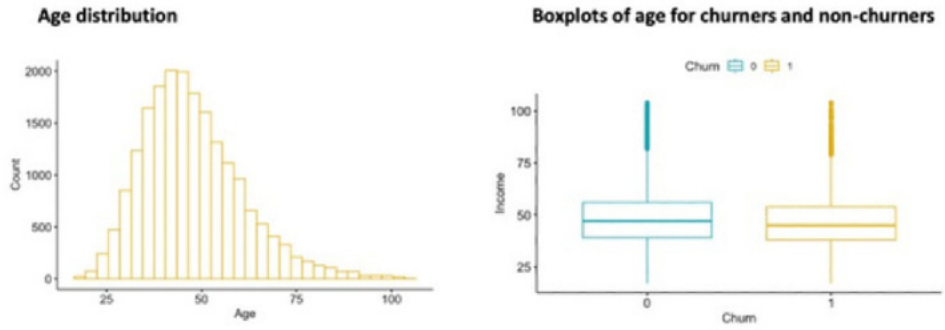


FIGURE 2: Age distribution (A) and Boxplot (B)

Furthermore, we found that some of the records contain age values higher than 100 as shown in the boxplot and histogram above. According to CBS there are about 2,600 people with the age of 100 years or higher in the Netherlands, which is about 0.015% of the population. Our dataset contains 30 customers with the age of 100 years or higher, which is about 0.15%. Given the fact

that these people also consume energy and the possibility that this energy supplier might be attractive for older people, we accept the high representation of 100 plussers.
We also found that there are quite a few records with gas consumption being 10 times higher than the average and electricity consumption being 20 times higher than the average as can be seen in the histograms below.

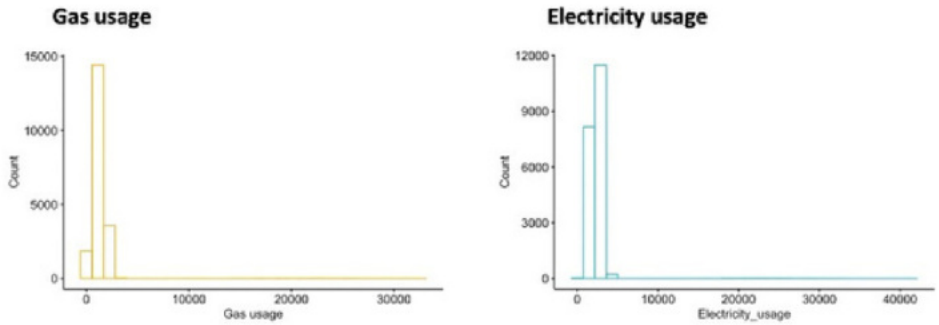


FIGURE 3: Gas usage (A) and Electricity usage (B)

We investigated the individual records that contain outliers in gas- and electricity usage, but the records all seem normal. However, by looking at the boxplots we see that there is an unexpected gap in electricity usage and gas usage as can be seen below. Because of the gap, we expect that there is a measurement error in which this cluster of records is being measured in years instead of months. As we are not able to contact the data provider we make the assumption that this is the case. We divided all the records above the gaps by 12 in order to transform them into a monthly-level measurement.

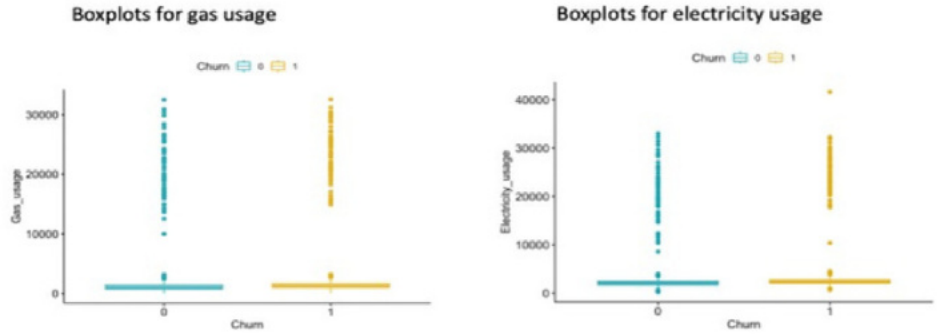


FIGURE 4: Boxplots for gas usage (A) and Boxplot for electricity usage (B)

Furthermore, there are records that have no gas consumption at all according to the data. However, this is explained by the fact that there are homes that do not use any gas, but do use electricity.
Also, for income we found records that are very high compared to the mean, like hundreds of thousands per month. We found the same gap in income by looking at the boxplot. Therefore, we

make the same assumption as with the energy usages and divide the records above by 12. Relationship length is very nicely distributed, but there are 101 records having a 0 value. This might be explained by the possibility that these customers just signed a contract with the energy supplier.

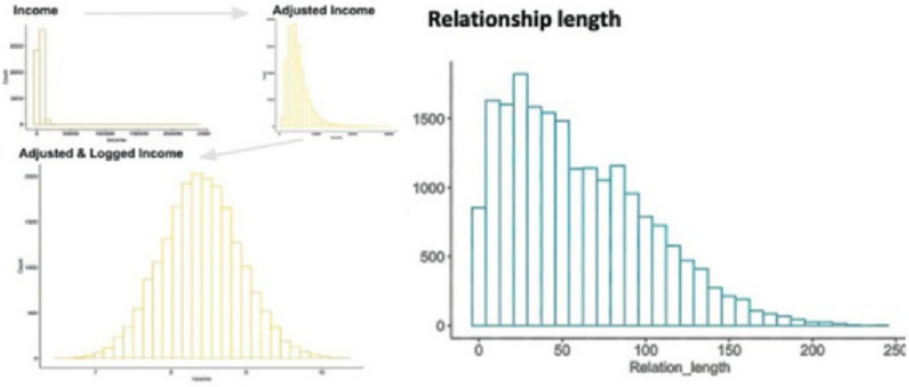


FIGURE 5: Income (A) and Relationship length (B)

IV. Explorative analysis

We created correlation matrices in order to find out to what extent the variables in the dataset are correlated with each other. In the first matrix it becomes clear that contract length and relation length are both negatively correlated with churn, just like we expected. Whereby, e-mail list, electricity- and gas usage, channel and home label are all positively correlated with churn, which is confirmed by our findings in the literature (FIGURE 6A). In FIGURE 6B the matrix shows that, except for the ID variable, all correlations with churn are statistically significant.

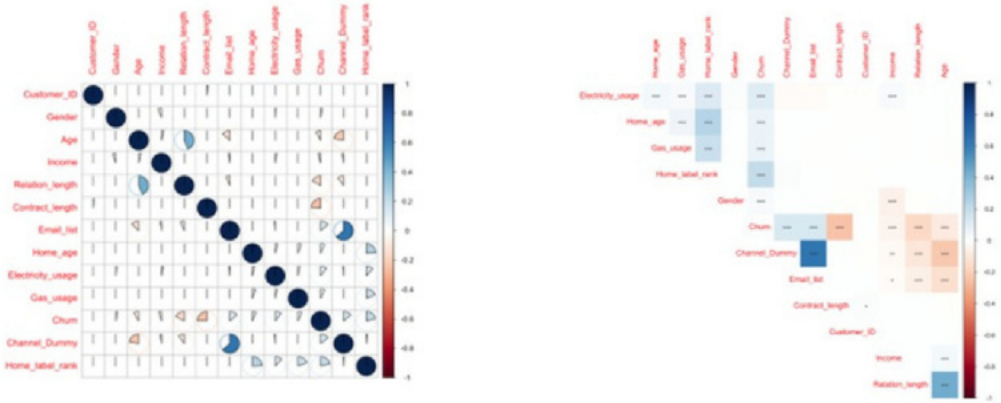


FIGURE 6: Boxplots for gas usage (A) and Boxplot for electricity usage (B)

Model-free evidence for hypothesis

Using the correlation matrices it already became clear that the hypothesis, which based on the literature, are also to some extent visible in the data. However, it is useful to get a better understanding of those relationships using some visualizations. According to our *H1*, relationship length is negatively associated with churn.

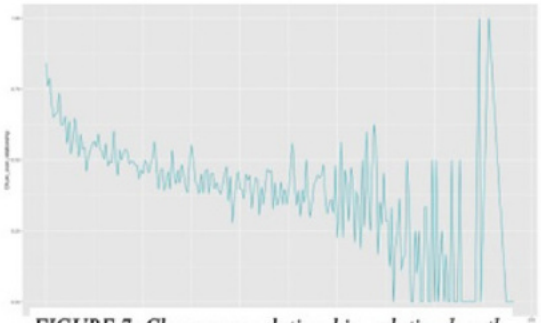


FIGURE 7: Churn over relationship, relation length

However, it is useful to get a better understanding of those relationships using some visualizations. According to our *H1*, relationship length is negatively associated with churn. We visualized it by taking the average of churn for all possible relationship lengths, having relationship length on the x-axis and churn on the y-axis. According to the attached graph, there is indeed a negative effect of relationship length on churn. This means that the longer people are in a relationship with the firm, the less likely they are to churn. This is also being confirmed by a simple regression test, which shows that if the relationship length increases with 1 month the churn decreases with 0.002.

According to our *H2*, contract length is negatively associated with churn. We visualized it by taking the average of churn for all possible contract lengths, having contract length on the x-axis and churn on the y-axis. In the following graph, the relationship between the two variables can be observed. Customers which have the contract length equal to 0 are allowed to switch their energy supplier without any penalties. Thus, it is more likely for this category of customers to churn.

A simple regression test gives a significant negative effect of 1 month increase in contract length results in an 0.11 decrease in churn. However, regression does not take into account that especially the contract length being 0 leads to this effect. Therefore, we also performed a simple t-test using a dummy variable for contract length being 0 or greater than 0. This simple t-test gives a significant difference between the 2 groups and the customers with the higher contract lengths are having about 50% less churn than the customers with 0 contract length.

According to our *H3* and *H4* hypothesis, being on the e-mail list and which start-channel you use is associated with churn. As we do not have time series data and churn is a binary value, it is difficult to plot the results for these variables nicely without a third variable. Therefore, we use home label rank as a third variable in order to show the differences among these variables and to show its effect on churn. In the following graphs it is shown that there are differences in churn rates among these groups. Churn seems to be higher among people of which the e-mail address is known by the company, which is the opposite of our hypothesis. Furthermore, customers that filled out the contract on the website seem to have higher churn rates then people who did it through phone. Also, it is clearly visible that the higher the home label rank, thus the worse the energy label is, the higher customers are likely to churn. This is also confirmed by the simple regression test which indicates an increase of 0.05 in churn when home label goes up by 1. To

test the difference between the 2 binary variables we used simple t-tests. These tests indicate significant differences in churn between the start channels 'online' and 'phone' and between being on the email list and not being on the email list.

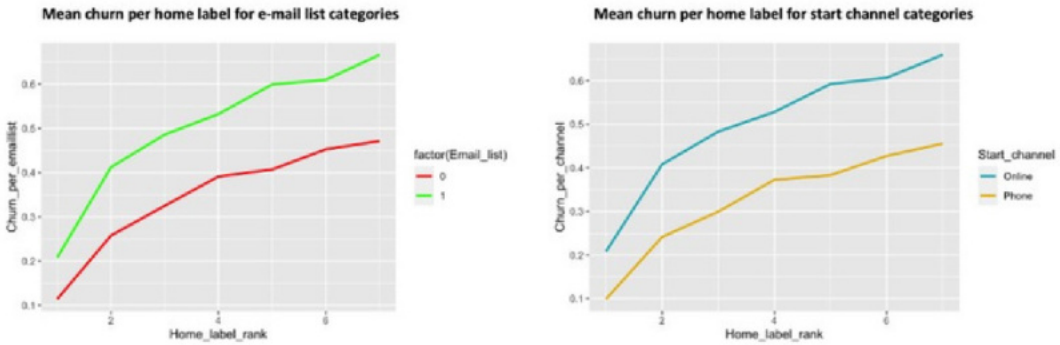


FIGURE 8: Mean churn "email"(A) and Mean churn "channel" (B)

According to our *H5* and *H6* hypothesis, gas- and energy usage are negatively related to churn. We visualize these 2 hypotheses by plotting churn over energy usage and plotting churn over gas usage. Therefore we divided the customers into 100 groups based on the amount of gas or electricity they use. The higher the number of the group, the higher the amount of gas or electricity the customers in these groups use. For every group we calculated the mean number of churn, resulting in the following graphs.

For both cases it is clear that the more energy consumed, the higher chances are that people churn. These relationships are also confirmed by simple regression tests that indicate the following effects; an 1000 kWh increase in electricity consumption leads to an increase in churn of about 0.03 and a 1000 M^3 increase in gas consumption leads to an increase in churn of about 0.02.

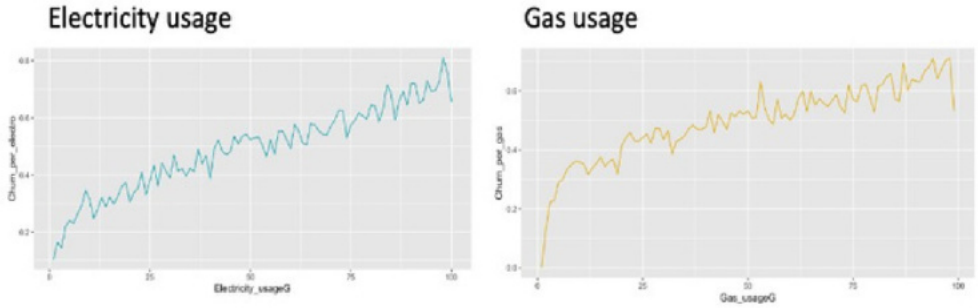


FIGURE 9: Electricity usage (A) and Gas usage (B)

V. Modeling

5.1 The baseline model

We use a logistic regression as our baseline model as this is quite a simple model which allows us to specify the independent variables beforehand. A logistic regression is a supervised machine learning technique that is used for classification of binary variables, in our case whether a customer churns or not.
The relevant variables for the baseline model are based on the literature, hypotheses and the explorative analysis. Thus, the following regressors have been included: *Relation_length_yrs*, *Contract_length_yrs*,*email_list*,*Startchannel_dummy*,*Home_label_rank*,*AdjElectricity_usage* and *AdjGas_usage*. After running the logistic regression, we see that almost all the variables are highly significant (*p-value = <2e-16*). The overall *Home_label_rank* is not significant, but the highest label (A) is significant, suggesting that eco-friendly houses do differ in their churn likelihood (~40%).
Due to the limitations of the pages, we will only interpret a few variables. Therefore, we see that *relationship length* has a coefficient equal to -1.160e-01, which means that with every one year increase in the relationship of the customer with the firm, the odds of churning decrease by 10.96%. In contrast, the value for *adjusted electricity usage* is equal to 1.199e-03, suggesting that with every unit increase in the electricity usage (kWh), the odds of churning increase as well by 1.001%. Regarding the *adjusted gas usage,* the direction of the effect is positive, meaning that when the gas usage goes up, the likelihood of a customer to churn increases.

To be able to see how well our model predicts, we can use different measures such as: the ***Hit ratio***, ***Top Decile Lift***, and the ***Gini coefficient***. In order to prevent overfitting, we split the dataset into estimation (75%) and validation sample (25%) and we use the aforementioned measures on the validation sample.

We use The ***Hit rate*** to see how well the model correctly and wrongly predicts churners against the observed churners and non-churners. Our sample is balanced, therefore a Hit ratio above 50% was expected, which means that our model does better than random guessing. The Hit rate is given by the percentage in the bottom right corner (70%) and represents the correct churn predictions made by the model.

Hit rate for the baseline model

	Observed		
Predicted	**Positive**	**Negative**	
Positive	1698	737	70%
Negative	731	1784	71%
	70%	71%	70%
	Sensitivity	**Specificty**	**Hit ratio**

TABLE 2: Hit rate

Next, we make use of the ***Top Decile Lift***. We use the trained model to predict churn and divide this group into 10 different groups, with group 1 having the highest chance to churn and group 10 the lowest. Afterwards we control our trained model with the overall actual churn rate, for which 1 means that our model is as good as a random selection model and 2 means twice as good. We get a value of ***1.75*** for the TDL which represents a fairly good score.

In contrast to the Top Decile Lift, the **Gini coefficient** focuses on the overall performance in which it compares the churn classification of our model with the churn classification of a random selection model. The closer the value is to 1 the better the model. For the logistic regression, we have obtained a value of *0.54* for the Gini coefficient, meaning that our model does better than random guessing.

Using raw data to validate the model

We did some data transformations due to the outliers we observed in the income, gas, and electricity variables. To make sure this did not affect our estimation negatively we estimate a new model using the raw data.

Control models

We control our base model by estimating 6 different models to see whether our baseline model is significantly worse than the more complete models. The techniques that we use for these 6 models are stepwise logistic regression, decision trees, bagging, boosting, random forest and support vector machines.

5.2 Stepwise logistic regression

For our baseline model we needed to decide manually, based on literature and the hypotheses we made, which variables we want to include in our model. However, in a stepwise regression, you can let the model decide which variables to include based on the contribution of the variable to the model performance. This enables us to potentially find important control variables for our base model.

A stepwise logistic regression can be based on forward, backward, and both direction selection. It adds or takes away variables until a certain threshold is reached. The thresholds we use are the *Bayesian Information Criterion (BIC)* which takes the sample size into account and the *Akaike Information Criterion (AIC)* which takes the degrees of freedom into account. BIC is more restrictive than AIC and is only recommended for larger sample sizes (>100 per variable) (Heinze et al., 2018). We estimated the 3 selection techniques for both BIC and AIC methods and then selected the model with the best results in terms of TDL, Gini coefficients and Hit rate. As visible in table 3, the forward AIC selection technique has the highest score. For interpretation purposes, we use the forward selection technique using the AIC method. The variables that the forward AIC selection model includes are: flexible contract, home label rank 1 to 7, electricity usage, relationship length, gas usage, income, email list, start channel dummy, high-income dummy, age, and age group 23 to 44 years.

Model	Hit ratio	TDL	Gini coeff.
Backward AIC	75%	1.5770	0.6590
Forward AIC	**75%**	**1.5893**	**0.6580**
Both AIC	75%	1.5770	0.6590
Backward BIC	75%	1.5728	0.6593
Forward BIC	75%	1.5728	0.6575
Both BIC	75%	1.5728	0.6593

TABLE 3: Fit criteria

5.3 Decision trees Forward

Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Due to their easy implementation and interpretation, decision trees are widely used in marketing. The idea of a regular decision tree is that it splits the data (root node), based on a variable, into subgroups. The model then keeps splitting the data until further splits are not useful according to the threshold. This threshold could be based on different rules, such as the GINI index, entropy, or the chi-square test. We use all of the techniques and compare these models to see which models compute the best results.
The CHAID model splits the data using the Chi-square and splits into groups with the highest scores until there are no significant splits possible. However, in order to run this model the independent variables should be factors, while our data set contains mostly numeric and integer variables. In order to save us from transforming the data, we use a CART model. This model splits the data using an impurity measure and splits into groups with the highest decreases of impurity until there is no decrease possible. This idea is similar to the idea of a C4.5 model.

We calculated 2 CART models, one using the default threshold and one using a more restricted threshold. The accuracy results of both models are presented in table 4 The cart model, which is visualized in appendix 1.1, first splits the data based on whether customers have a flexible contract. Then these resulting nodes are both split based on the electricity consumption of the customers. For the customers that do not have a flexible contract, the data is also split on relationship length, while both sides of the tree are split based on gas consumption. Looking at the tree it seems that people with a flexible contract are more likely to churn, that people with high electricity and gas usage are more likely to churn, and that people with a long relationship length will be less likely to churn.

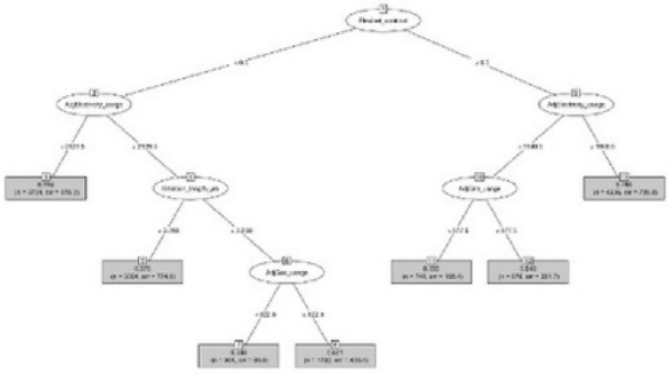


Figure 10: CART tree

Model	Hit ratio	TDL	Gini coeff.
CART-Tree (default)	71.37%	1.50	0.50
CART-Tree (restricted)	70.48%	1.46	0.49

TABLE 4: Fit criteria

5.4 Bagging and Boosting

Bagging is an ensemble learning technique that calculates many different decision trees based on different training sets. After these calculations, it takes the mean of all the trees which results in one final decision tree that is often significantly better than predicting just one tree. Using different samples for all trees and taking the average might be useful as slight changes in the training set might lead to entirely different models. Boosting is an ensemble learning technique that computes weights for all data points and updates these weights based on the accuracy of the computed decision trees. With bagging all trees are independently calculated, while with boosting every tree calculation is based on the previous tree. Due to the learning processes within these models, very good and accurate predictions can be estimated. However, it is very difficult to tell how these models came to the particular predictions, which creates difficulties for interpretation. We estimated both a bagging and boosting model of which the accuracy results can be found in table 5.

Furthermore, as both models do not result in 1 final tree, we created a table containing the importance of each predictor for the bagging model and the relative influence of each predictor for the boosting model. The most important predictor for the bagging model is electricity usage, followed by gas usage which are also the two predictors with the highest influence according to the boosting model. By looking at the importance and influence of the predictors of the bagging and boosting model, it can be concluded that many predictors have a high position in both lists, although the sequence differs.

Bagging model		**Boosting model**	
Predictors	***Relative importance***	***Predictors***	***Relative Influence***
AdjElectricity_usage	100,0	AdjElectricity_usage	24,419
AdjGas_usage	93,2	AdjGas_usage	17,058
AdjLogIncome	79,4	Contract_length_yrs	16,584
Relation_length_yrs	78,6	Flexible_contract	16,304
Home_age	58,6	Relation_length_yrs	11,299
Age	57,2	AdjLogIncome	6,761
Contract_length_yrs	55,7	Startchannel_dummy	3,623
Flexible_contract	24,7	Email_list	2,998
Email_list	18,3	Home_label_rank	0,466
Startchannel_dummy	17,2	Age	0,209
Gender	11,4	Home_age	0,179
Home_label_rank7	10,9	HighIncome_dummy	0,078
AgeGroup4568	7,9	AgeGroup_4568	0,009
AgeGroup2343	7,6	Gender	0,007
Home_label_rank4	7,0	AgeGroup_2344	0,005
Home_label_rank5	6,8	AgeGroup_022	0,000
Home_label_rank3	6,3	AgeGroup_68up	0,000
Home_label_rank6	5,4		
Home_label_rank2	4,6		
AgeGroup_68up	3,0		

Bagging Fit Criteria			**Boosting Fit Criteria**		
Hit Ratio	**TDL**	**GINI Coeff.**	**Hit Ratio**	**TDL**	**GINI Coeff.**
74.26%	1.55	0.63	75.66%	1.58	0.67

Table 5: Bagging and Boosting predictors

5.5 Random Forests

Random Forest is another ensemble learning technique that calculates decision trees based on a random set of variables using several training datasets. One advantage of this technique is that it gives every variable a chance to be in the tree, giving us a good indication of which variables are important. For our research we created multiple random forest models, testing out (1) with no special settings, (2) with 1000 trees and (3) with 10000 trees. We found the results overall do not change between (2) and (3), so we compared the 1000 trees model to the non-special settings model.

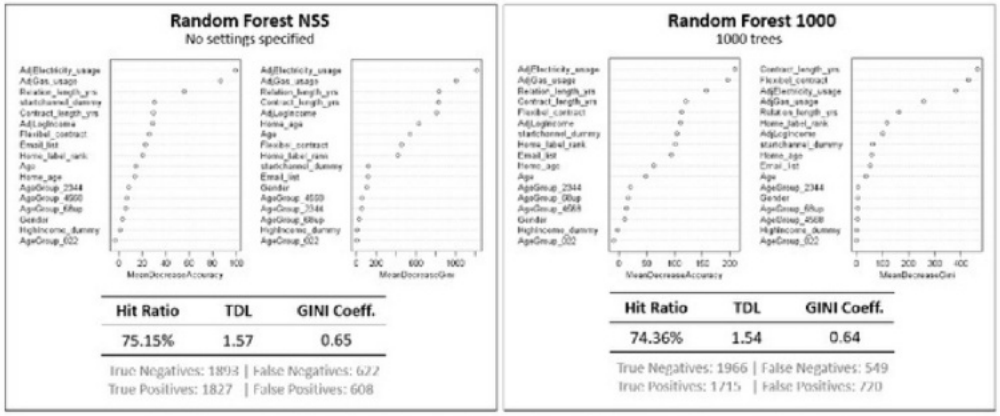
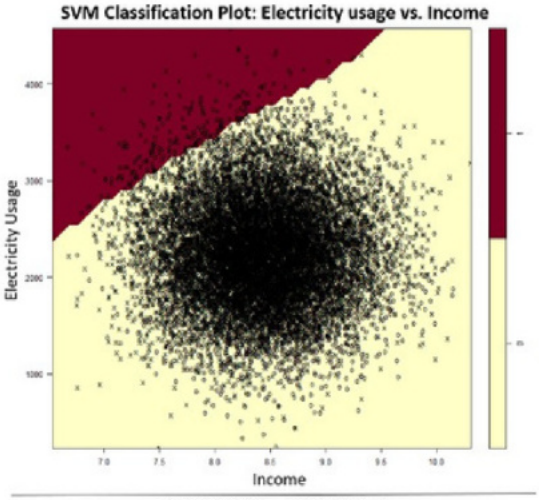


FIGURE 11: Random Forest NSS (A) and Random Forest 1000 (B)

We found both to score similarly on their fit criteria, with the 1000-tree model scoring slightly worse. The main difference between the two is in the true/false positives/negatives, with the 1000 tree model being better at predicting true negatives and has fewer false negatives, while the default model is better at predicting true positives. Considering it is more costly to predict a churner as a non-churner (false negative), it may be more profitable to choose the 1000-tree model. For the purposes of this assignment, we consider the variable recommendations of both and notice that the top 6 variables for both models are roughly the same for both the hit rate-based graph, as well as the GINI-based one. Generally, the top 6 variables are electricity usage, gas usage, relationship length, contract length, flexible contract and income.

5.6 Support Vector Machines

Support Vector Machines (SVM) is a machine learning technique with the main goal of finding a classifier with the lowest expected generalization error. With this method, it is important to already have a good idea of what variables to use. Based on all of our findings so far, we chose: electricity and gas usage, relationship length, income, start channel, flexible contract and contract length. When fitting the SVM model, we tested four kernel functions: Linear, Gaussian Radial Basis Function (RBF), Polynomial and Sigmoid. Next, we assessed which to focus on by examining the fit criteria of each. The Sigmoid kernel performed the worst with a 10% lower hit rate than all the other kernels. RBF performed the best, followed by the Linear and Polynomial kernels. Based on this we focused on RBF. After going through all the possible one-to-one plots, we found a meaningful relationship between electricity usage and income. According to the SVM plot, people who have high electricity usage and a low-medium income are classified as churners.



SVM RBF Kernel Fit Criteria		
Hit Ratio	**TDL**	**GINI Coeff.**
74.47%	1.57	0.64

Figure 12: Random Forest

VI. Managerial Conclusion & Model Recommendation

The baseline model, that we developed based on the literature review, was already quite accurate as it has a hit rate of 70%, a TDL of 1.75, and a Gini coefficient of 0.54. The relationships we found in the literature between churn and relationship length, contract length, e-mail list dummy, starting channel dummy, and household label, were confirmed by our baseline model. However, our hypothesis that a decrease in consumption of electricity or gas could lead to an increase in churn has been rejected by our baseline model. It is actually the other way around, which could be explained by the reason that customers might start to look more quickly for cheaper alternatives when the costs are more significant.

We built upon this by fitting several other machine-learning models which do not require variable specification with the goal of affirming our initially-chosen variables, as well as potentially finding new ones. Out of the 6 machine-learning models we built all but the CART tree model scored higher on their fit criteria than the baseline model by generally 5% higher on hit rate, and similarly or slightly higher on TDL and GINI. These results reaffirm the accuracy of our baseline model. The variables we included in our baseline model are also included in the stepwise regression model and are important predictors in the other machine learning models.

However, a variable we did not include in our baseline model is the income variable. Which, the stepwise regression did include it and according to the bagging, boosting, and random forest models, this variable has a relatively high importance and influence. Therefore, we recommend adding an income variable to the baseline model, which will increase its accuracy. Additionally, the SVM model revealed an insight that customers with high electricity consumption and low-medium income may be classified as churners.

Finally, we conclude that it is very important for managers to have access to information about their customers. Using this information, a churn prediction model could be developed. Being able to predict whether some customer is going to churn or not is very important as retention is cheaper than acquisition. Therefore, retention strategies could be developed and targeted specifically on customers with a high probability to churn.

VII. References

Ahn, J.-H., Han, S.-P., & Lee, Y.-S. (2006). Customer churn analysis: churn determinants and mediation effects of partial defection in the korean mobile telecommunications service industry. Telecommunications Policy, 30(10), 552–568. https://doi.org/10.1016/j.telpol.2006.09.006

Baker, S. R., Baugh, B., & Kueng, L. (2019). Income Fluctuations and Firm Choice. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3533766

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert systems with applications, 34(1), 313-327.

Cristian R. (2016). Churn Prediction for the Dutch Energy Market (Doctoral dissertation, Vrije Universiteit Amsterdam).

Lu, J. (2002). Predicting customer churn in the telecommunications industry—An application of survival analysis modeling using SAS. SAS User Group International (SUGI27) Online Proceedings, 114.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of marketing research, 43(2), 204- 211.

Nogueira, T. S. (2022). Churn prediction modeling comparison in the retail energy market (Doctoral dissertation).

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications, 38(3), 2354-2364.

Heinze, G., Wallisch, C. & Dunkler, D. (2018). Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. https://doi.org/10.1002/bimj.201700067