Managerial Summary

This project aims to develop a robust predictive model to identify customer likeliness to churn for a Dutch energy supplier. European energy, especially in the retail sector, is undergoing significant transformations due to market liberalization and evolving consumer expectations. These changes have intensified competition among energy providers. Customer churn significantly impacts revenue and increases acquisition costs, making it a focal point for managers.

The analysis is based on a dataset with 20,000 customer records containing demographic, behavioral, and contractual information. The overall churn rate in this dataset is 49.01%. We analyzed one baseline model using a logistic regression based on theory, and subsequently built various models, selecting five to focus on for this paper. Key findings indicate that behavioral and contractual factors are crucial, while demographic variables like age and income play secondary roles. Surprisingly, email engagement has a counterintuitive effect. Based on our results, we recommend the boosting model to be implemented and used for churn prediction, as it scores the highest on a number of criteria such as hit rate, top decile lift, and gini coefficient.

Main Paper

Defining the Business Problem

The business problem for the report is to develop a predictive model to identify customers who are most likely to churn from a large Dutch energy supplier. Customer churn poses a significant challenge for the company, as it leads to revenue loss and increased acquisition costs to replace lost customers. By accurately predicting churn, the company aims to proactively implement retention strategies to reduce churn rates and improve customer loyalty, thereby enhancing profitability and market competitiveness.

This project will leverage a provided dataset containing customer demographic, behavioral, and contractual information to build and validate a churn prediction model that informs targeted interventions.

Designing the Research

Industries such as Utilities, Telecom, and Groceries typically maintain stable customer bases over time. Conversely, sectors like Hotels, Car Rentals, and Clothing retailers experience significantly more fluctuation in their customer bases from year to year (Baker, 2020). This indicates that long-term relationships, likely facilitated by contracts, play a significant role in reducing churn in industries like utilities.

The Dutch energy market is less competitive compared to the US and UK, with fewer market participants and lower customer turnover. A notable portion of Dutch consumers remain on older contracts that offer relatively high profit margins and exhibit low churn rates (Van Der Schrier, 2019). This indicates that not everything about customer churn behaviors in the energy sector can be generalized across countries, as market structures and regulatory environments significantly shape consumer behavior. In the Dutch market, churn is driven largely by price increases, contract endings, or people moving houses (Van Der Schrier, 2019). Conversely, other research has found that for households struggling with financial issues, churn rate does not differ significantly from those households without said struggles (He, 2017). This indicates that there is not an overall consensus on whether income has a direct influence on customer churn rates.

Furthermore, demographic factors such as age have been shown to play a critical role in influencing churn rates, as older customers tend to exhibit higher loyalty compared to younger ones, who are more inclined to switch providers based on perceived benefits (Nechet, 2024). This variability highlights the importance of localized studies to understand the unique predictors of churn in specific markets. It also underscores the need for targeted retention strategies that address both contractual and demographic variables to mitigate customer turnover effectively.
Another variable of interest with adequate theoretical implications is the presence of an email list for customers. Customers who engage with onboarding emails exhibit more positive behaviors in terms of repayment and retention (Marshan, 2019). This interaction indicates that well designed and strategically timed onboarding emails can directly influence customer churn rates.

General marketing theories can also be directly applied to the business challenge. Social identity theory, life course theory, Maslow's hierarchy of needs, and relationship marketing theory can all be linked to why consumers make decisions leading to churn. In combination with the specific theoretical implications above, these findings underline the importance of determining why customers churn so that managers can take appropriate actions to either reduce churn rate or focus on acquiring customers with a lower churn rate.

From this theoretical foundation, the following hypotheses are derived:
H1: Customers with longer relationships exhibit lower churn rates, as longer-term relationships facilitate stability in industries like utilities (Van Der Schrier, 2019; Baker, 2020).
H2: Income does not have a significant direct influence on churn rates, as financial struggles do not necessarily predict customer churn behaviors (He, 2017).
H3: Older customers have lower churn rates due to higher loyalty, while younger customers are more likely to churn because of their inclination to switch providers for perceived benefits (Nechet, 2024).
H4: Customers whose email addresses are on file and who receive onboarding emails have lower churn rates, as email engagement positively influences retention behaviors (Marshan, 2019).

Throughout this research R Studio will be used as a primary data analysis tool to address the challenge of churn prediction, as defined in the first section of this paper. The process will begin with thorough data cleaning and exploration, leveraging the provided dataset's demographic, behavioral, and contractual information. Building on insights from the literature review, such as the role of contractual stability in utilities, the demographic influence of age on churn, and the impact of customer engagement through email communication, a baseline logistic regression model to predict Churn and test the hypotheses 1-4 is developed. After acquiring and assessing the baseline model, a number of models will be employed and compared using various prediction assessment criteria. From the R code, 5 models will be selected and discussed thoroughly in comparison to the baseline model in the final section of the main paper.

Data Preparation

The data consist of 14 variables, with a total of 20,000 observations. At a first glance, we can see that the Churn rate for the overall data set is 49.01%. We created new columns for categorizing Age, Income, and Relation_length for descriptive/explorative analysis using graphs. These additional columns are solely used for the descriptive analysis and not in the models.

Additionally, the Customer_id column, which contains unique values for each row, was excluded from the dataset. When building a model, it is crucial to omit the Customer_id column. This column acts as a unique identifier for each customer but does not provide any meaningful information or predictive value. Including it could introduce noise and potentially lead to overfitting, as the model might learn to associate specific customer IDs with certain outcomes rather than identifying the underlying patterns in the data. Therefore, this column was excluded from the dataset beforehand to ensure the model focuses on the relevant features.

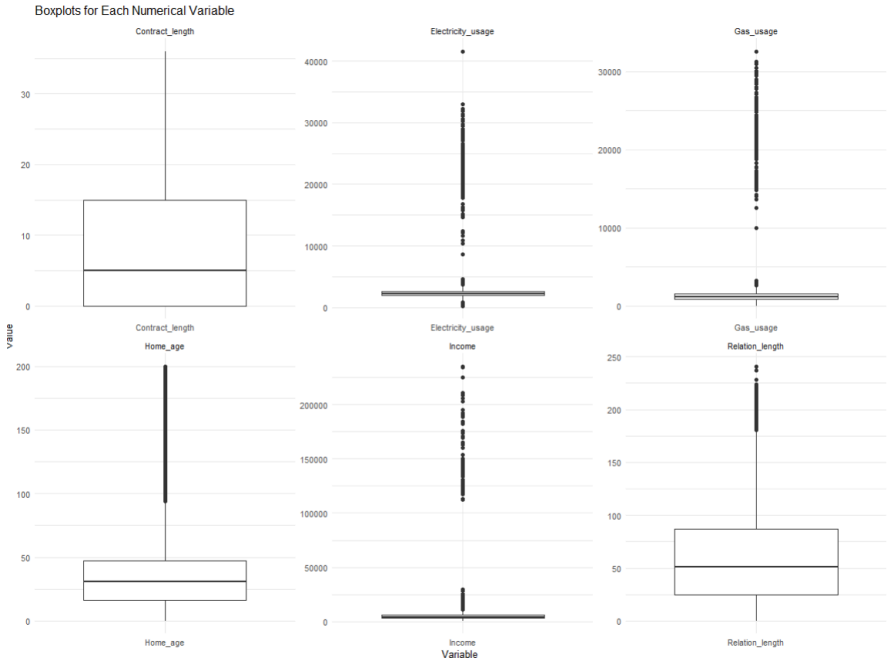
Outliers

During our analysis of the dataset, we observed several instances where Gas_usage was recorded as 0 while Electricity_usage was not. Upon further investigation, we discovered that all these cases were associated with homes labeled as Home_label A. This label indicates that these homes are highly environmentally friendly and possess excellent insulation.

Given this context, it is logical that these homes would have zero gas usage due to their efficient design, insulation, and electrical heating units, which minimizes the need for gas heating. Consequently, what initially appeared to be outliers are actually consistent with the characteristics of Home_label A homes. Therefore, we have decided to retain these data points in our dataset as they accurately reflect the energy usage patterns of these environmentally friendly homes.

Using boxplots we can detect possible outliers in the dataset. As can be seen in Figure 1, there is quite considerable variance in some variables. However, upon closer examination of this variance, we determined that these data points are logical and do not require further action.

Figure 1: Boxplot for Each Numerical Variable



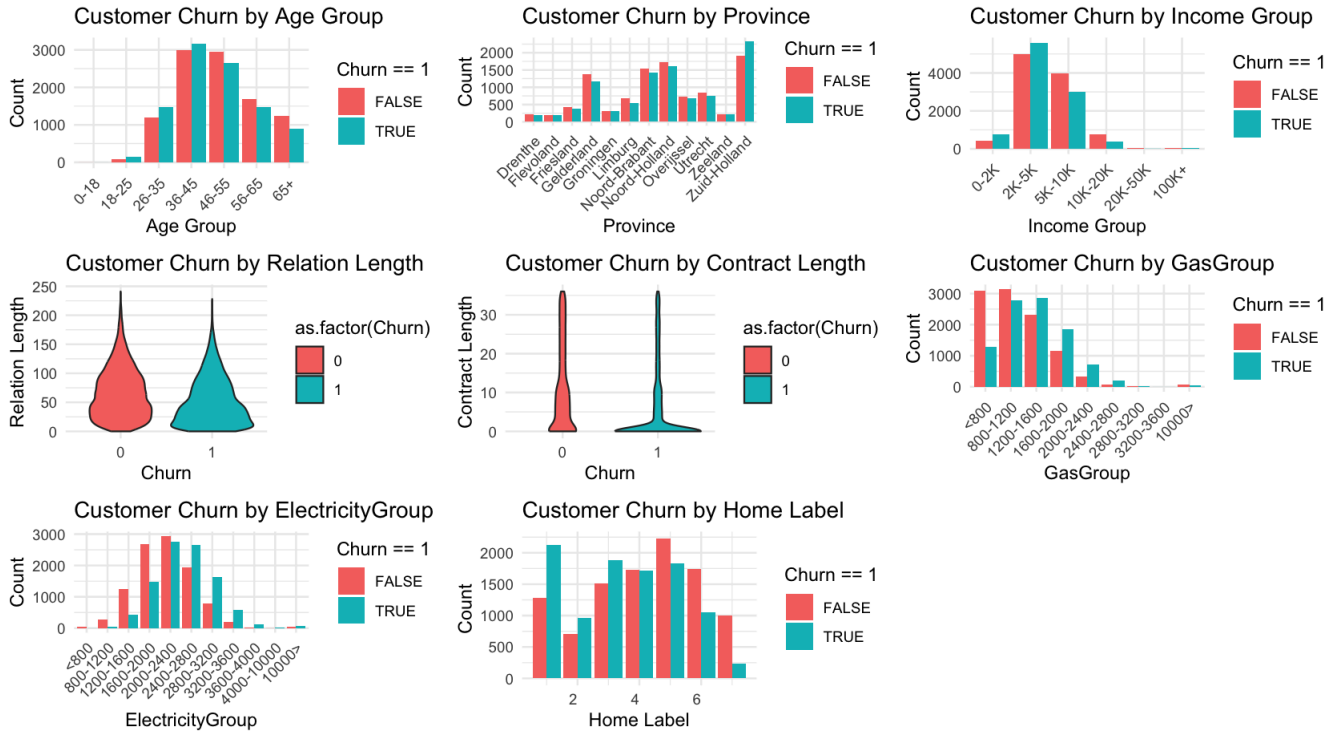
In figure 1 it can be seen that sometimes home age is 0. We found some cases where Home_age is 0 but Relation_length is not. After further investigation we came to the conclusion that in the Netherlands energy contracts are tied to individuals rather than properties (DutchReview, n.d.). This means that people can maintain their energy contracts, with contract adjustments, even when they move to a new home. As a result, it is possible for someone to have a new home (Home_age = 0) but still have an ongoing energy contract (Relation_length > 0) from their previous residence. Consequently, these data points do not require further action.

Data Transformation

For categorical variables, such as Gender, Email_list, Start_channel and Province, we transform them into factor variables. For Home_label we transform them into an ordinal factor as the variable follows an order from good (“A”) to bad (“G”).

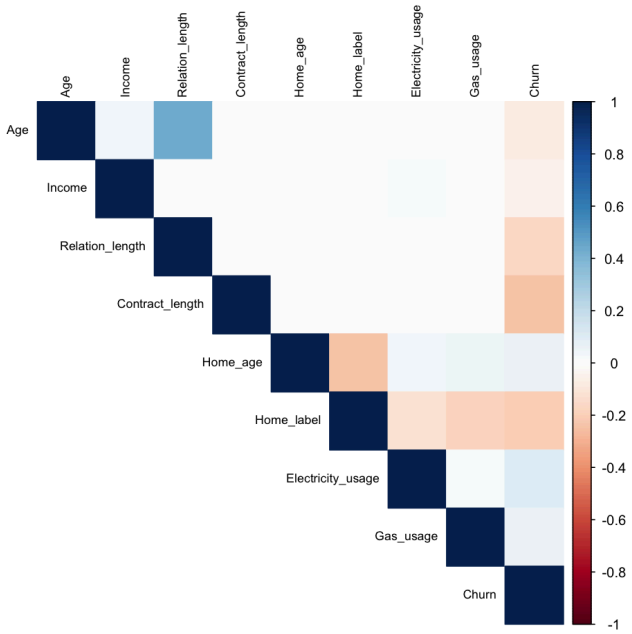
Explorative Analysis

Figure 2: Individual Variables Plot



Before we started to estimate the models, we visualized the distribution of customer churn in different variables in Figure 2. For each variable, the distribution of the data is separated by customer churn indicated by "TRUE" for churning customers and "FALSE" for non-churning customers. Figure 2 shows that the middle-young age group, starting from 18 to 45 of age, tend to have higher possibility of customer churn compared to the older age group. Customers in Zuid-Holland show the highest churn rates compared to other provinces, this province also has the largest number of customers. The density of the churned group peaks at shorter relationship lengths. This suggests that customers are more likely to churn within their first few months of joining the company. The density of retained customers is broader, with peaks at mid-to-long relationship lengths, indicating that customers who have longer relationships with the firm tend to stay with the company. Customers with flexible contracts (zero contract length) are significantly more prone to churn, as there are no financial consequences for leaving. Fixed-term contracts seem to reduce churn, especially those with longer durations. Customers with high gas consumption rate (>1200) and also customers with moderate to high electricity consumption (>2800) have higher churn rates. Lower consumption groups (<2000) have slightly lower churn rates, possibly due to lower service engagement or pricing sensitivity.

Figure 3: Correlation Plot



The correlation plot in Figure 3 reveals correlations for both positive and negative. A positive correlation is observed between Age and Relation_length, indicated by a light blue color. This suggests that older customers tend to maintain longer relationships with the energy company. Conversely, a negative correlation is found between Home_label and Home_age, shown by a light red color. This implies that higher Home_label values are associated with newer homes, which is intuitively reasonable, as newly built houses often have better energy efficiency. Regarding customer churn, weak negative correlations are observed with Contract_length, Home_label, and Relation_length, indicating that longer contracts, lower energy labels, and longer relationships might reduce the likelihood of churn. These correlations might introduce multicollinearity in a model. On the other hand, white color indicates that there is no correlation between the two variables.

Figure 4: Individual Statistical Test

type	variable	p_value	test
Numerical	Age	3.386403e-25	Wilcoxon Rank-Sum Test
Numerical	Income	8.148936e-95	Wilcoxon Rank-Sum Test
Numerical	Relation_length	1.315678e-135	Wilcoxon Rank-Sum Test
Numerical	Contract_length	0.000000e+00	Wilcoxon Rank-Sum Test
Numerical	Home_age	1.129764e-35	Wilcoxon Rank-Sum Test
Numerical	Home_label	5.656830e-188	Wilcoxon Rank-Sum Test
Numerical	Electricity_usage	0.000000e+00	Wilcoxon Rank-Sum Test
Numerical	Gas_usage	8.338043e-255	Wilcoxon Rank-Sum Test
Categorical	Gender	1.034562e-03	Chi-Square Test
Categorical	Start_channel	8.012030e-87	Chi-Square Test
Categorical	Email_list	3.829406e-83	Chi-Square Test
Categorical	Province	3.296045e-14	ANOVA Test

Statistical test is performed for each of the variables individually by using Churn as the dependent variable to get a better understanding of the dataset. By performing this we can assess the relationship significance of each variable towards Churn, identify which variables are the predictors for Churn, and ultimately select the relevant variables for the model. Figure 4 consists of four columns which provide information about type of variables, name of the variable, the significance level of the test (p_value column) and type of the statistical test used. The result of the statistical test shows that all of the variables are significant (p-value < 0.05).

Figure 5: Generalized Linear Model

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             9.687e-01  1.469e-01   6.596 4.21e-11 ***
Gender1                 7.412e-02  3.107e-02   2.386  0.01705 *
Age                     2.598e-03  1.334e-03   1.947  0.05150 .
Income                 -1.530e-05  2.537e-06  -6.031 1.63e-09 ***
Relation_length        -9.215e-03  4.160e-04 -22.150  < 2e-16 ***
Contract_length        -5.349e-02  1.540e-03 -34.730  < 2e-16 ***
Start_channelPhone     -4.776e-01  5.373e-02  -8.889  < 2e-16 ***
Email_list1             4.653e-01  4.814e-02   9.665  < 2e-16 ***
Home_age                9.301e-04  4.760e-04   1.954  0.05071 .
Home_label             -2.367e-01  9.338e-03 -25.350  < 2e-16 ***
Electricity_usage       1.854e-04  1.947e-05   9.524  < 2e-16 ***
Gas_usage               4.034e-05  1.002e-05   4.027 5.64e-05 ***
ProvinceFlevoland      -1.625e-01  1.530e-01  -1.062  0.28831
ProvinceFriesland       4.663e-03  1.311e-01   0.036  0.97164
ProvinceGelderland     -8.737e-02  1.152e-01  -0.759  0.44803
ProvinceGroningen       2.590e-03  1.371e-01   0.019  0.98493
ProvinceLimburg        -1.291e-01  1.236e-01  -1.044  0.29655
ProvinceNoord-Brabant  -2.954e-02  1.139e-01  -0.259  0.79535
ProvinceNoord-Holland  -2.656e-02  1.131e-01  -0.235  0.81436
ProvinceOverijssel     -6.357e-02  1.216e-01  -0.523  0.60097
ProvinceUtrecht        -1.122e-01  1.198e-01  -0.936  0.34908
ProvinceZeeland         1.483e-01  1.504e-01   0.986  0.32411
ProvinceZuid-Holland    2.903e-01  1.118e-01   2.596  0.00944 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We use a generalized linear model (GLM) to evaluate the combined effect of multiple predictors on the dependent variable (Churn). While individual statistics tests focus solely on the bivariate relationship between dependent and independent variables, GLM takes into account correlations and interactions between variables simultaneously. Thus, it can be used to build a predictive model. In Figure 4, we can see that Contract_length, Home_label, Relation_length, Start_channel, Income, Gas_usage, Electricity_usage, and Email_list are highly significant predictors of Churn. The findings from Figure 5 align with the pattern observed from the correlation plot from Figure 3, showing reliability of the result. Further, we can refine the model based on the result of Figure 5 and our hypotheses.

Figure 6: Multicollinearity Test

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.006343	1	1.003166
Age	1.242626	1	1.114731
Income	1.008784	1	1.004382
Relation_length	1.211325	1	1.100602
Contract_length	1.024903	1	1.012375
Start_channel	1.662052	1	1.289206
Email_list	1.619846	1	1.272732
Home_age	1.056711	1	1.027965
Home_label	1.157968	1	1.076089
Electricity_usage	1.052843	1	1.026082
Gas_usage	1.038602	1	1.019118
Province	1.007007	11	1.000317

We further test the multicollinearity in the dataset. The GVIF values indicate that multicollinearity is present between the above variables. Figure 6 shows that the GVIF values are close to 1, in this case we can neglect it as multicollinearity is present but it will not negatively affect the reliability of the model’s coefficient and prediction.

Baseline Model

In order to test our 4 main hypotheses, we will run a logistic regression model using the variables Age, Income, Email_list, and Relation_length to predict Churn.

Figure 7: Baseline Model Summary

```
Call:
glm(formula = Churn ~ Age + Income + Email_list + Relation_length,
    family = binomial, data = training_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      6.470e-03  7.516e-02   0.086    0.931
Age              5.431e-04  1.416e-03   0.383    0.701
Income          -1.055e-05  2.351e-06  -4.486 7.26e-06 ***
Email_list       6.086e-01  4.125e-02  14.754  < 2e-16 ***
Relation_length -8.356e-03  4.478e-04 -18.660  < 2e-16 ***
```

The results indicate that the baseline churn rate when all predictor variables are 0 is not

statistically significant. Additionally, age does not have a statistically significant effect on churn. Its near-zero coefficient implies little to no impact. Income negatively impacts churn, meaning that as income increases, the likelihood of churn decreases. However, the effect size is very small due to the low coefficient. Contrary to our theory research, the model shows that being on the email list significantly increases churn, as indicated by the positive coefficient. Finally, relation length has a strong negative relationship with churn. Longer relationships with the company reduce the likelihood of churn.

According to this first logistic regression, we can reject hypotheses 2, 3, and 4, while we fail to reject hypotheses 1.

We use a number of validation criteria to assess the accuracy and fit of the model by running predictions of the model on the validation dataset.

Figure 8: Hit Rate Table - Baseline Model

	Predicted		
Observed	0	1	Total
0	1488	1025	2513
1	1016	1422	2438
Total	2504	2447	4951

In the hit rate table of our first model it can be calculated that Type 1 error (1025/2513) and the type 2 error (1016/2504) are both quite high, indicating the model does not have a high accuracy. The model correctly identifies 59.2%(1488/2513*100%) of non churners, and 58.3%(1422/2438*100%) of churners, corresponding to an overall hit-rate of 58.8%.

Figure 9: Decile Lift Table - Baseline Model

	Decile									
Observed	1	2	3	4	5	6	7	8	9	10
0	138	205	228	226	243	258	273	280	315	347
1	358	290	267	269	252	237	222	215	180	148

Based on the decile lift table, it can be concluded that customers in decile 1 are 1.4657 times more likely to churn. This is calculated by dividing the actual churn rate of group 1 and dividing it by the overall churn rate, and multiplying the outcome by 100%.

Figure 10: Lift Curve - Baseline Model

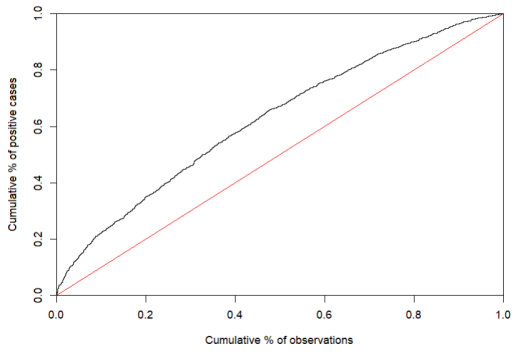


Figure 10 shows the lift curve for the baseline regression model. This graph is an indicator of the model's accuracy. It can be seen that the AUC is relatively small compared to a well performing model, indicating that the model could potentially be improved. The gini coefficient of this graph is equal to 0.2544766, meaning that the model performs better than random guessing but it is quite far away from a perfect discriminating model (which would have a gini coefficient of 1).

AIC Backward stepwise regression

Figure 11: Comparison of considered stepwise regression models

Method	Hit Rate table	Hit rate		TDL	Gini	Final AIC
Regression models						
AIC Forward	Predicted Observed 0 1 0 1600 913 1 700 1738	67,421%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 82 133 161 211 231 277 289 341 368 420 1 414 362 334 284 264 218 206 154 127 75	1,695032	0,4817845	AIC=18072.94
AIC Backward	Predicted Observed 0 1 0 1600 913 1 700 1738	67,421%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 82 133 161 211 231 277 289 341 368 420 1 414 362 334 284 264 218 206 154 127 75	1,695032	0,4817845	AIC=18072.94
AIC Both	Predicted Observed 0 1 0 1600 913 1 700 1738	67,421%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 82 133 161 211 231 277 289 341 368 420 1 414 362 334 284 264 218 206 154 127 75	1,695032	0,4817845	AIC=18072.94
BIC Backward	Predicted Observed 0 1 0 1656 857 1 762 1676	67,300%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 79 136 160 216 238 261 306 326 371 420 1 417 359 335 279 257 234 189 169 124 75	1,707315	0,4773965	AIC=18235.07
BIC Forward	Predicted Observed 0 1 0 1656 857 1 762 1676	67,300%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 79 136 160 216 238 261 306 326 371 420 1 417 359 335 279 257 234 189 169 124 75	1,707315	0,4773965	AIC=18235.07
BIC Both	Predicted Observed 0 1 0 1656 857 1 762 1676	67,300%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 79 136 160 216 238 261 306 326 371 420 1 417 359 335 279 257 234 189 169 124 75	1,707315	0,4773965	AIC=18235.07

In our primary analysis, we ran 6 stepwise regression models, including forward, backward, and both direction stepwise regressions, using AIC and BIC model selection criteria for each direction type. Interestingly, the 3 AIC models converged on the same model, having the same hit rate, top decile lift, and gini coefficient. The same applies to the 3 BIC models, which showed slightly lower performance in these measures as the BIC models punish harder for more complex models. Therefore, we decided to focus on an AIC backward stepwise regression model in this section, as it is the best performing in terms of prediction and CPU power.

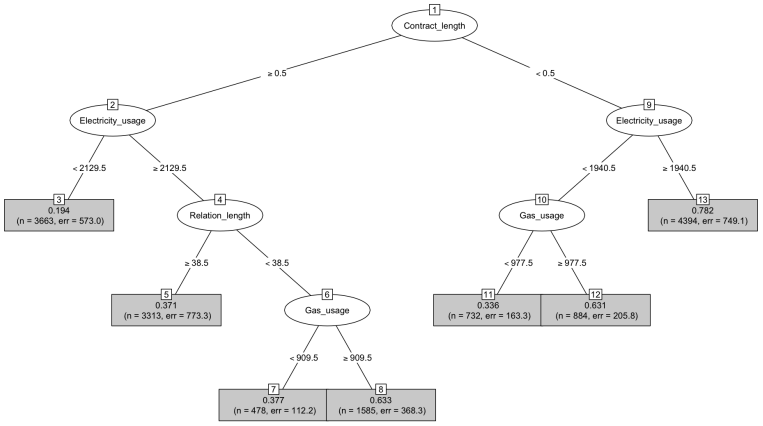
The AIC-based backward stepwise regression model outperforms the baseline model in predictive power, achieving a hit rate of 67.42%. Additionally, the model demonstrates a higher top decile lift (TDL = 1.695), indicating improved ability to distinguish between deciles with respect to the likelihood of churn. Finally, the gini coefficient is also higher, being equal to 0.481, meaning that this model performs better at discrimination than the baseline model.

While the validation criteria mentioned above clearly shows that this model outperforms the baseline model, it is important to note that while the baseline model only had 4 predictor variables, this stepwise regression model uses 11. The algorithm only excluded home_age from the final model, as this variable did not significantly contribute to the predictions of the model.

CART Decision Tree

We considered both CHAID and CART decision tree models, but the CART decision tree resulted in higher predictive power. By experimenting with some settings in R, 2 CHAID trees and 1 CART tree model were derived. The setting changes did not affect the number of splits of the CART tree model.

Figure 12- Cart Decision Tree



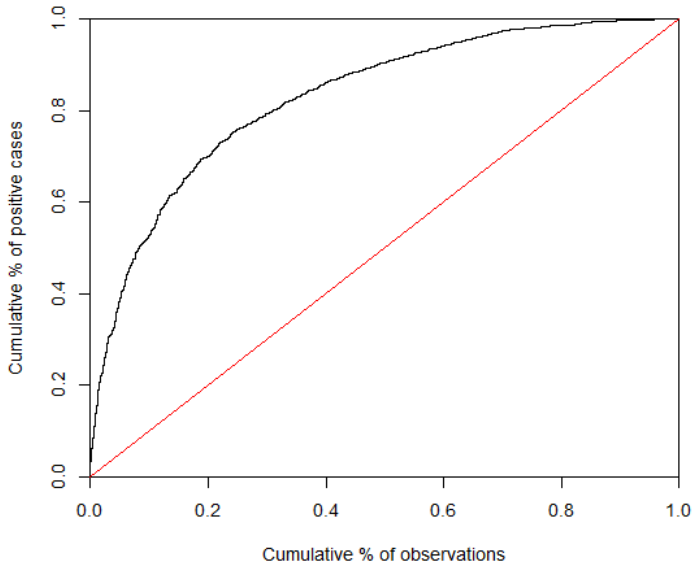
In figure 12, it can be seen that the root of the tree splits on Contract Length, indicating that it is the most important factor in predicting churn. Customers with shorter contracts (<0.5) are at significantly higher risk of churn compared to those with longer contracts (≥0.5). For customers with shorter contracts (<0.5), Electricity Usage further splits the population. Higher electricity usage is associated with a higher likelihood of churn, while lower usage reduces churn risk. The tree then gives further implications following the same logic regarding the other variables.

This CART decision tree achieves a hit rate of 0.715411, a TDL of 1.564015, and a gini coefficient of 0.5053637, indicating that it is both better at predicting and classifying churn when compared to the baseline model.

Boosting

In our R code, we have explored both bagging and boosting methods. Both offer high predictive power, with the boosting method outperforming the bagging method. This indicates the dataset requires a model with higher complexity and lower bias to uncover, because base decision trees used in bagging are too simple while their iterative combination improves overall accuracy. By focusing on correcting the hardest to predict instances by weighting misclassified examples more heavily, performance is increased. Furthermore, it can be concluded that the dataset is quite clean with little noise, as boosting tends to perform better in datasets like these. This is supported by earlier conclusions drawn from the stepwise regression model, where we saw that 11 out of the 12 predictor variables were significantly impactful in the model. This means that misclassified examples reflect true underlying patterns, rather than random noise. The boosting method achieved a hit rate of 75.38% and a TDL of 1.89. Additionally, the gini coefficient is 0.661, indicating this model also performs well in classification and discrimination of churn.

Figure 13- Boosting Lift Curve



In the lift curve it can be seen that the boosting model successfully concentrates a large proportion of positive cases in the top deciles. The curve remains well above the Diagonal and flattens towards the right, which is expected. Compared to the baseline model lift curve, this visual tool shows that the boosting model performs significantly better at ranking and identifying positive cases, which is important for managerial implications.

Random Forest

We implemented a Random Forest model with 10,000 trees to achieve high predictive and classification power and evaluate its ability to outperform the boosting method. The model achieved a hit rate of 75.04%, a top decile lift (TDL) of 1.875, and a Gini coefficient of 0.650. While these metrics indicate that the Random Forest is a robust model with excellent performance, it slightly underperforms compared to the boosting method in all three measures. A hit rate of 75.04% demonstrates that the Random Forest is effective in classifying both churners and non-churners with relatively high accuracy. This indicates strong overall predictive power and an ability to generalize well across different customer groups. A TDL of 1.875 indicates that the model is effective at ranking customers by their likelihood of churn, concentrating the highest-risk individuals in the top decile. However, the marginally lower TDL compared to boosting (1.89) suggests that Random Forest is slightly less effective at distinguishing between high-risk and low-risk customers. A Gini coefficient of 0.650 confirms that the Random Forest provides strong discriminatory power, far exceeding the baseline model's Gini coefficient of 0.254. However, it is still slightly lower than the boosting model's Gini coefficient of 0.661, reinforcing the conclusion that boosting better captures the relationships within the dataset. Random Forest models are inherently robust to noise due to their aggregation of many independent trees. This result confirms earlier findings that the dataset is relatively clean, with minimal noise affecting predictions. An upside Random Forests is that the trees are easily parallelizable, making them computationally efficient for large datasets despite using 10,000 trees.

Figure 14 - Random Forest variable importance

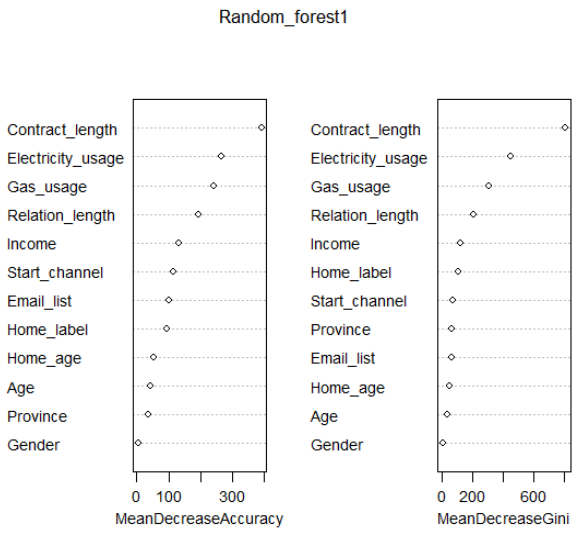


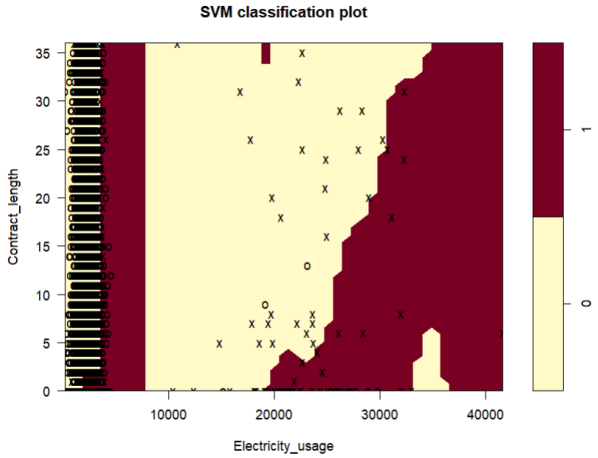
Figure 13 identifies Contract Length, Electricity Usage, and Gas Usage as the most important predictors of churn, indicating that contract length and electricity usage patterns significantly influence customer retention. Relation Length and Income also play key roles, with longer relationships and higher incomes reducing churn risk. Conversely, demographic features like Age and Gender have minimal impact, suggesting churn is driven more by behavioral factors than demographics.

Support Vector Machines

We explored various support vector machine (SVM) models using different kernels and degrees to classify churning and non-churning cases. The primary goal of SVMs is to create a decision boundary that separates these cases while maximizing the margin between the closest data points of each class, known as support vectors.

In our analysis, we evaluated five SVM methods with linear, polynomial, radial, and sigmoid kernels using the training data. We focused on two key predictors of churn, Electricity_usage and Contract_length, identified as the most significant by Random Forest and logistic regression models.

Figure 15 - SVM Radial



After developing and plotting several models, and predicting on the test data, we concluded that the radial kernel outperformed the other SVM models. This kernel provided the best classification performance for our churn prediction task with an overall hit rate of 68.55%. The linear kernel method, which ranked second, achieved a hit rate of 65.02%. Additionally, when comparing the Gini indexes and TDL, the radial kernel method also outperformed all of the other methods (Figure 16).

Figure 16 - Support Vector Machines method comparison

Support Vector Machines					
Method	**Hit Rate table**	**Hit rate**		**TDL**	**Gini**
Linear	Predicted Observed 0 1 0 1437 1076 1 656 1782	65,01717%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 195 202 180 182 190 206 353 321 344 340 1 301 293 315 313 305 289 142 174 151 155	1,232378	0,3027535
Radial	Predicted Observed 0 1 0 1860 653 1 904 1534	68,55181%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 152 144 144 157 258 350 334 311 343 320 1 344 351 351 338 237 145 161 184 152 175	1,408432	0,3693555
Polynomial	Predicted Observed 0 1 0 2507 6 1 2427 11	50,85841%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 269 244 262 257 235 242 253 269 238 244 1 227 251 233 238 260 253 242 226 257 251	0,9294016	0,00212431
Polynomial - Degree 2	Predicted Observed 0 1 0 414 2099 1 235 2203	52,85801%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 251 242 263 238 223 234 258 229 267 308 1 245 253 232 257 272 261 237 266 228 187	1,003099	0,06835285
Sigmoid	Predicted Observed 0 1 0 1448 1065 1 1115 1323	55,96849%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 232 230 199 232 225 282 281 277 289 266 1 264 265 296 263 270 213 214 218 206 229	1,08089	0,1188617

Managerial Conclusion

Recommendation

The analysis revealed critical insights into customer churn drivers and predictive power of different models. Boosting emerged as the most effective approach, achieving the highest hit rate (75.38%), top decile lift (1.89), and Gini coefficient (0.661). These metrics indicate its superior ability to classify churners and prioritize low and high risk customers for targeted retention strategies. When simplicity or computational power is of importance, the random forest and CART model also performed well, but we recommend the Boosting model to be used for implementation to achieve the best churn prediction.

Figure 17: Summary of 5 main models

Method	Hit Rate table	Hit rate		TDL	Gini	Final AIC
Baseline logistic regression	Predicted Observed 0 1 0 1488 1025 1 1016 1422	58,776%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 138 205 228 226 243 258 273 280 315 347 1 358 290 267 269 252 237 222 215 180 148	1,465752	0,2544766	AIC =20153
AIC Backward	Predicted Observed 0 1 0 1600 913 1 700 1738	67,421%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 82 133 161 211 231 277 289 341 368 420 1 414 362 334 284 264 218 206 154 127 75	1,695032	0,4817845	AIC=18072.94
CART	Predicted Observed 0 1 0 1915 598 1 811 1627	71,541%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 114 98 111 180 254 321 387 347 389 392 1 382 397 384 315 241 174 188 148 106 103	1,564015	0,5053637	
Boosting	Predicted Observed 0 1 0 1916 597 1 622 1816	75,379%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 34 79 118 167 228 296 337 369 422 463 1 462 416 377 328 267 199 158 126 73 32	1,891557	0,4773965	
Random Forest	Predicted Observed 0 1 0 1947 566 1 670 1768	75,035%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 38 73 128 166 225 289 343 370 421 460 1 458 422 367 329 270 206 152 125 74 35	1,87518	0,6496726	
Radial	Predicted Observed 0 1 0 1860 653 1 904 1534	68,552%	Decile Observed 1 2 3 4 5 6 7 8 9 10 0 152 144 144 157 258 350 334 311 343 320 1 344 351 351 338 237 145 161 184 152 175	1,408432	0,3693555	

Contract Length is the most critical predictor. Shorter contracts significantly increase churn risk, indicating the importance of promoting longer-term agreements. In other words, contract lengths represent an exit barrier for the customer, which directly influences churn rate. Usage patterns such as gas and electricity usage also provide opportunities for targeted marketing strategies, as these variables are also of high importance.

According to this first logistic regression, we can reject hypotheses 2, 3, and 4, while we fail to reject hypotheses 1. It is also interesting to add that contrary to our initial assumptions, random forest showed that the age variable is relatively less important in predicting churn compared to other variables.

Sources

Baker, S. R., Baugh, B., & Sammon, M. C. (2020). Measuring customer churn and interconnectedness (NBER Working Paper No. 27707). National Bureau of Economic Research. https://www.nber.org/papers/w27707

DutchReview. (n.d.). *Utilities in the Netherlands: How to set up water, gas, and electricity*. Retrieved December 4, 2024, from https://dutchreview.com/expat/utilities-netherlands/

He, X., & Reiner, D. (2017). Why consumers switch energy suppliers: the role of individual attitudes. *The Energy Journal*, *38*(6).

Marschan, C. (2019). Customer onboarding messages: the effects on customer repayment & immediate retention in the consumer loans business (Master's thesis, Hanken School of Economics).

Van Der Schrier, B., Miyasaka, S., White, J., Twartz, B., KPMG in the Netherlands, KPMG in Japan, KPMG in the US, KPMG in the UK, Global Strategy Group, & KPMG International. (2019).

The use of AI

We used AI to copy edit the assignment, i.e., we did use it to get a better understanding about the model concept , spelling check, grammar and rewriting. We also used AI to improve our R-code. For all of this, we used ChatGPT 4o. For this, we used ChatGPT-4o.

Signature:

Nils Depner (s4766946)

Qi Zhu(s5493838)

Ruben Meijer (s4543831)

Erlangga Roesdyoko (s2667983)