university of groningen

# Data Science Methods for MADS

Assignment 1

**Group 2.6:**

Ny Ha Dao - S5677858

Nanyun Zhang - S5658292

Yichong Tao - S5065623

Tim van Noord - S3734978

**December 5th, 2024**

# Table

# 1. Managerial Summary

The company's objective is to predict which customers are most likely to churn in order to implement targeted retention strategies. To achieve this, selecting the best model for churn prediction is crucial, with an emphasis on accuracy and strong out-of-sample performance.

First, we built the baseline model from hypotheses derived from previous research, and the result of logistic regression confirms that factors such as younger age, lower income, higher home labels, and higher electricity usage significantly impact churn. These insights validate the hypotheses and provide a clear direction for targeted interventions.

Second, upon comparing multiple models, **Boosting** stands out as the top performer, with the highest hit rate (74.33%), Top Decile Lift (1.92), and GINI coefficient (0.698). These results demonstrate Boosting's superior predictive accuracy and consistent performance, making it the ideal model for the company's churn prediction needs.

Key findings from both the Boosting and baseline models suggest that **electricity usage** is the most significant factor driving churn. This is consistent with previous research and supported by all models. Similarly, **gas usage** is highlighted as an important predictor by Boosting, CART trees, and random trees. Based on these insights, the supplier should prioritize customers with high energy consumption and consider offering a **Tiered Loyalty Program** with personalized services to retain these valuable customers and reduce churn. The Boosting model also identifies **contract type** and **contract length** as important variables influencing churn. The company should leverage both **contract type** and **contract length** for customer segmentation, distinguishing between loyal customers and those more likely to switch to competitors.

Additional factors like **relationship length** and **income** also emerge as potentially useful variables. The **relationship length** reflects the customer's familiarity with the supplier, which may positively influence churn prediction. **Income** has a negative relationship with churn, suggesting that low-income customers are more likely to churn, particularly if they form a significant portion of the customer base. The company should consider implementing retention strategies focused on low-income customers to ensure they are not overlooked.

In conclusion, while the Boosting model offers a strong foundation for churn prediction, it should be complemented by further analysis to uncover the underlying causal relationships, ultimately enabling the company to develop more effective, data-driven retention strategies.

# 2. Defining the business problem

This study addresses the critical issue of customer churn for a Dutch energy supplier, an industry where product homogeneity amplifies the challenge of retaining customers. High churn rates pose

significant risks to revenue, as customer acquisition is notably more expensive than retention (Coussement & Van den Poel, 2006; Verbeke et al., 2011).

The research aims to develop an effective churn prediction model, enabling the company to identify and target at-risk customers with tailored retention strategies. By employing various predictive models, this study not only identifies key drivers of churn but also provides actionable insights to enhance customer retention efforts.

# 3. Designing the research

## 3.1 Dependent variable: Churn

Churn in an energy context is when customers are most likely to churn and thus switch to another energy supplier (Jullie et al., 2015). A "churner" is a customer who chooses to cancel their service provider subscription and switch to a new one, thereby reducing the financial loss incurred by the company (Umayaparvathi & Iyakutti, 2012). This switching intention is defined as a signal of termination of the relationship between the consumer or customer and the old energy provider (Haridasan et al., 2021). Churn can have a major financial impact, as there is not only one customer leaving, but there could also be a chain reaction as unhappy consumers tell others about you, causing more churn (Rahimullah et al., 2024). In summary, customer churn represents a critical challenge across industries, particularly in the energy sector, where its ripple effects can amplify financial losses and damage brand reputation through negative word-of-mouth.

## 3.2 Demographic variables

Demographics play an important role in the analysis of churning behaviour across customers in the energy sector. From a marketing perspective, it is key to understand wether specific age groups or income groups show different churn behaviour.

Customer churn is shown to vary significantly between different age groups. Where younger customers are more likely to churn than older customers. Younger customers tend to be more aware of available substitutes and are more prone to churn (Svendsen & Prebensen, 2013). Older customers, for instance, on the other hand show to be more resistant to churn due to established relationships with companies (Hübner et al., 2023).

***H1: Younger customers are more likely to churn compared to older customers.***

Previous research shows that lower-income groups tend to exhibit higher churn rates compared to their higher-income counterparts. This trend can be attributed to several factors, including financial instability and the ability to switch services more readily due to fewer financial commitments. For instance, Oetama (2023) highlights that income category is a crucial variable influencing customer churn, suggesting that customers in lower income brackets are more likely to discontinue their services. This aligns with findings from Basiri et al (2010), who suggest that customer retention strategies must consider income levels, as lower-income customers are more likely to churn when faced with better deals from competitors

***H2: Customers with lower household incomes are more likely to churn compared to customers with higher household incomes.***

## 3.3 Contract type

Flexible contracts, characterized by the absence of penalties for early termination, inherently lower transaction costs and reduce the psychological burden associated with switching decisions (Seo et al., 2008). This freedom enables customers to respond more swiftly to competitive market offerings, such as lower prices or better service quality (Lipowska et al., 2024). In energy markets, where deregulation has introduced significant competition, flexible contracts amplify customers' sensitivity to alternative suppliers, as noted by Giulietti et al. (2013). Additionally, behavioral theories, such as the Push-Pull-Mooring (PPM) framework, indicate that when the "mooring" effects of contractual obligations are removed, customers are more likely to explore "pull" factors such as promotions from competitors or enhanced service benefits (Keaveney, 1995). Flexible contracts, therefore, diminish customer inertia and create an environment conducive to churn. This argument is further reinforced by Lipowska et al. (2024), who found that customers on flexible agreements exhibit higher intention-to-switch rates due to the absence of penalties or procedural complexities.

***H3: Customers with flexible contracts are more likely to churn compared to those bound by fixed-term contracts, as the absence of financial penalties or obligations facilitates switching behavior.***

## 3.4 House label

According to Marmolejo-Duarte and Bravi (2017), a survey of 250 respondents in the Spanish residential market shows that educated and energy-conscious groups are more likely to choose a house with a high-energy label, which has a significant impact on the choice of house and willingness to pay (WTP). This group of consumers is not only concerned with energy efficiency performance but also sees choosing highly energy-labeled housing as a way of supporting sustainable development. Meanwhile, a finding from another study of 376 customers in the Canary Islands further suggests that sustainability-conscious customers tend to prioritize companies that offer renewable energy and have an environmental commitment when choosing an electricity supplier (Amador et al., 2013). These findings suggest that consumer preferences in choosing an energy supplier are related to consumers' own concerns about sustainability and their lifestyles. In this study, we believe that consumers who live in houses with higher energy labels are more likely to churn energy services to renewable energy sources and sustainable related energy supplies, therefore, this study investigates the effect of housing energy labels as a relevant variable on churn.

***H4: Customers residing in homes with higher label levels are more likely to churn.***

## 3.5 Electricity usage

Among the primary sources of household energy consumption, electricity plays a central role, with growing demand among households (Matsumoto et al., 2022). However, increased electricity consumption leads to higher costs, making electricity prices a significant factor influencing the consumption behavior of key customers (Ezzard de Lange, 2008). Monetary savings are one of the main motivators for households to search for and switch suppliers, as highlighted by research showing that financial incentives are key drivers of this behavior (Ek and Söderholm, 2008; Deller et al., 2007).

With a greater number of suppliers offering a variety of deals, households are more likely to switch when they perceive the benefits of switching outweigh the costs involved (He et al., 2017). Consequently, switching energy suppliers has become one of the most straightforward ways for consumers to reduce energy bills in recent years (He et al., 2017). This motivates customers to

switch from their current supplier and seek alternative energy providers offering more favorable deals. Based on these findings, we can hypothesize that:

***H5: Higher electricity usage can increase customers' churn intention from their current energy supplier.***

# 4. Data Preparation

## 4.1 Data Transformation

The purpose of data transformation is to ensure that the form of data will be suitable for further analysis. In our process, transformations are applied to the home label and contract length.

First, the original data type of the home label is a character, presented in letters from "A" to "G", where A is the best grade. We used mutate function transforming them to numerical type, assigning 7 to "A", and decreasing in order.

Secondly, we wanted to find out not only the influence of remaining contract length but also the influence of contract type (flexible/ long term) on the churn. We created this new variable based on the contract length, an "if-else" function was adopted to identify if the contract length is larger than 0. When the contract length is greater than or equal to 1, it is marked as 1 (indicating a long-term contract), otherwise it is 0.

## 4.2 Missing Data Cleaning

We use summary to check the data for missing values, and the result shows that there are no missing values in the data.

## 4.3 Outlier processing

To detect the outliers of data, we visualized the variables. Among all variables, we found that there is an extremely unusual pattern in the boxplots of income, electricity usage, and gas usage. From these three plots we can see that all three variables have a very large number of outliers and that these outliers are clustered together and have huge discontinuities from the main data (Fig.1, Fig.2, Fig.3).

We believe that this result is due to the fact that the respondents incorrectly included annual income, annual electricity usage, and annual gas usage in the data for monthly income, monthly electricity usage and monthly gas usage. Hence, we applied 50000 as a threshold for income, 5000 as a threshold for electricity and gas usage to divide the data into 2 groups, for incorrectly entered data, we divide it by 12.

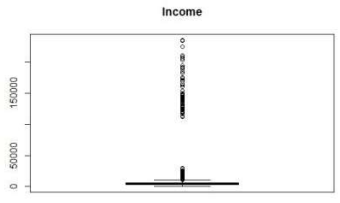



*Figure 1: Boxplot for income*

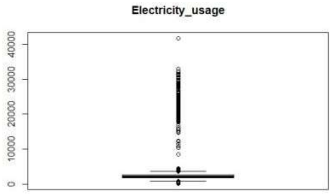



*Figure 2: Boxplot for Electricity_usage*

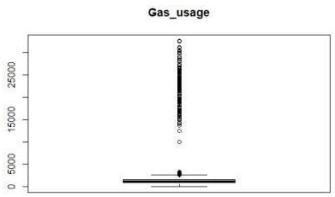



*Figure 3: Boxplot for Gas_usage*

Through these steps, we successfully cleaned the data and eliminated the effects of outliers, ensuring that the data could be used more accurately for subsequent analysis and modeling.

# 5. Explorative analysis

## 5.1 Descriptive analysis

***- Age:*** A histogram was generated to investigate the distribution of age in this dataset, we can see from Fig.4, that this distribution is close to normal, but with a skew to the right. The model is 40 years old.

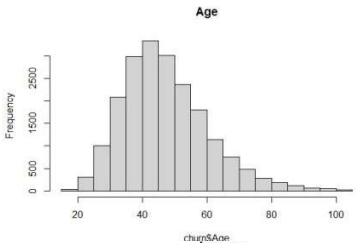



*Figure 4: Histogram for Age*

***- Income***: From the boxplot [Fig.5] of income, we can see that there are still a lot of outliers after our outlier treatment of income, but from the pattern, we can see that the distribution of these outliers is linear, so we think that these outliers are reasonable and that people's incomes do have significant variations. We can also see from the graph that the median income is around 5000, and the range from Q1 to Q3 is around 4000-6000.

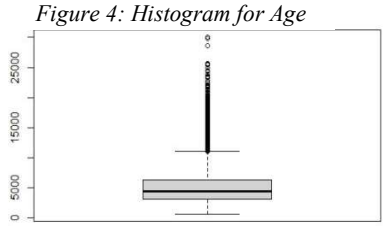


*Figure 5: Boxplot for income*

***- Contract type:*** Histograms are used to compare contract types. An x of 0 is a flexible contract and a 1 is a permanent contract. As we can see from Fig.6, the number of people using long-term contracts is significantly higher than the number of people using flexible contracts.

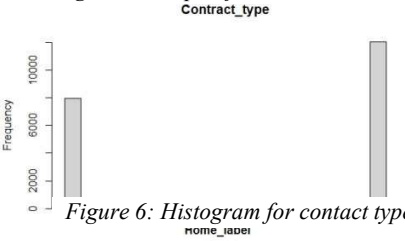



*Figure 6: Histogram for contact type*

***- Home label***: A histogram was used to study the distribution of home labels. A larger number in the x-axis indicates a more environmentally friendly house. As we can see from Fig.7, the houses of class 7(A) are the least frequented, followed by the houses of class 2(F). The most populated houses are classified as class 5(C).

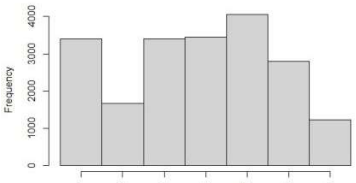



*Figure 7: Histogram for Home_label*

***- Electricity usage:*** From the boxplot [Fig.8] of electricity usage, we can see that there are still a lot of outliers after our outlier treatment, but from the pattern, we can see that the distribution of these outliers is linear, so we think that these outliers are reasonable and that people's electricity usage does have significant variations. We can also see from the graph that the median is around 2300, and the range from Q1 to Q3 is around 1900-2600.

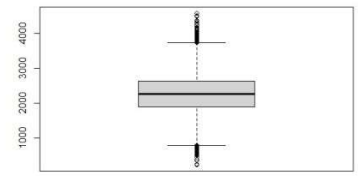



*Figure 8: Boxplot for Electricity_usage*

***- Churn***: We use a histogram [Fig.10] to look at the distribution of churns, which shows that the number of churns and the number of non-churns are roughly the same.

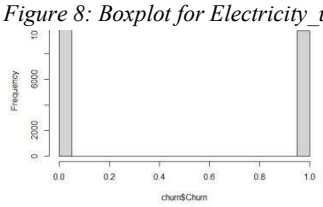



*Figure 9: Histogram for Churn*

## 5.2 Correlation analysis

To investigate the correlation between the main variables, we used the correlation test. First, we examined the correlation between age and churn, p-value < 2.2e-16, cor = -0.08358589, so we conclude that age and churn are significantly negatively correlated, but the correlation is small. We then examined the correlation between income and churn, p-value < 2.2e-16, cor = -0.1383183,so we concluded that income and churn are significantly negatively correlated. We then tested the correlation between contract type and churn, p-value < 2.2e-16, cor = -0.3541713, so we concluded that contract type and churn are significantly negatively correlated. Afterward, we tested the correlation between home labeling and churn, p-value < 2.2e-16, cor = -0.2077779, so we concluded that home labeling

and churn are significantly negatively correlated. Finally,we examined the correlation between electricity use and churn, p-value < 2.2e-16, cor = 0.2907818, so we conclude that electricity use and churn are significantly positively correlated. Besides, we adopted a corrplot to illustrate the relationship among variables. In Fig.11, red indicates a positive correlation, while blue is a negative correlation, and the color intensity and ellipse indicate the magnitude of the correlation.

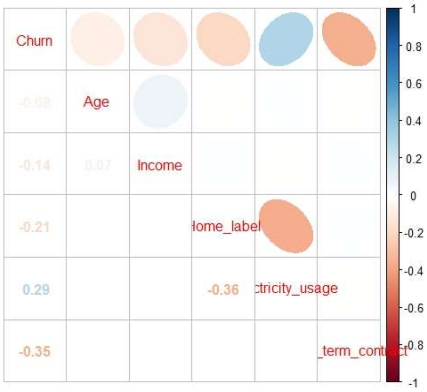



*Figure 10: Correlation heatmap*

## 5.3 Statistical tests

In this section, we performed several statistical tests to verify patterns and relationships in the data.

- ***Normality Test***

  First, to assess the normality of the data, a Lilliefors test (lillie.test) was performed for each variable, grouped according to Churn status, for Age, Income, Home_label, and Electricity_usage. The results showed that all p-values < 0.05, implying that the data for these variables were not normally distributed in both Churn and Un-churn groups.

- ***Wilcoxon Test***

  For non-normally distributed data, the Wilcoxon Test was used to compare the differences between Churn and Un-churn groups. The tests were applied Age, Income, Home_label and Electricity_usage were significantly different between Churn and Un-churn groups (P<0.05). There is a significant difference between the groups.

- ***Chi-Square Test***

  Since contract type is a categorical variable, we analyze the relationship between churn and contract type using the chi-square test. p-value < 2.2e-16, we reject the null hypothesis and show that there is a significant association between churn status and long-term contract type.

Furthermore, we used visualization to observe the distribution of different variables in different churning states.

- ***Age***

  In the Wilcoxon test p-value < 2.2e-16, 2 groups have different churn rates. We used a boxplot to study the age distribution of different churning states. As we can see from Fig.12, the median age of the un-churn is older than that of the churn, i.e., the older the age the less likely the churn is.

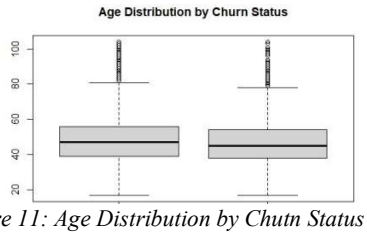



*Figure 11: Age Distribution by Chutn Status*

- ***Income***

  In the Wilcoxon test p-value < 2.2e-16, 2 groups have different churn rate. We used a scatter plot with jitter to study the distribution of income for different churn states. From Fig.13, we

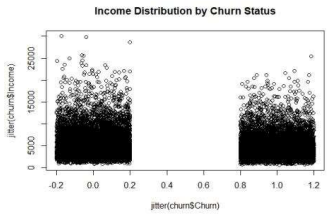



*Figure 12: Income Distribution by Chutn Status*

can see that extremely high-income people are less likely to churn.

- ***Contract Type***

  In the Chi-square test, p-value < 2.2e-16, 2 groups have different churn rate. In order to analyze the contract types, we first used random sampling on the data and used scatterplots with jitter to study the contract types under different churn states. As we can see from Fig.14, customers with long-term contracts are generally not churned, while customers with flexible contracts are more likely to churn.

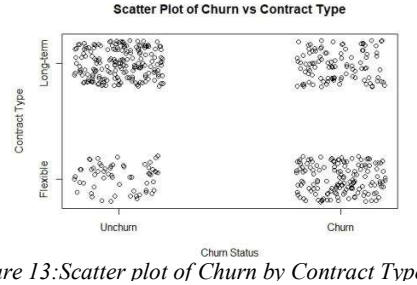



*Figure 13:Scatter plot of Churn by Contract Type*

- ***Home Label***

  In the Wilcoxon test p-value < 2.2e-16, 2 groups have different churn rate. We used boxplot to study the home label distribution of different churning states. As we can see from Fig.15, customers who choose not to churn will live in homes with higher labeling levels and are more environmentally friendly than those who churn.

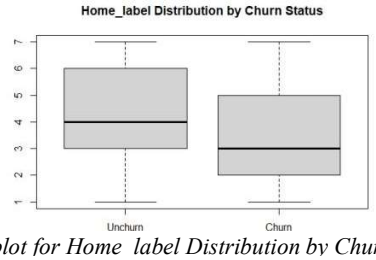



*Figure 14:Boxplot for Home label Distribution by Churn Status*

- ***Electricity Usage***

  In the Wilcoxon test p-value < 2.2e-16, 2 groups have different churn rate. We used box plots to study the distribution of electricity usage under different churn states. As can be seen from Fig.16, customers who choose churn usually use more electricity.

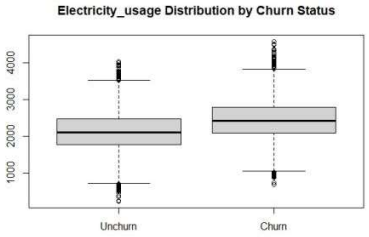



*Figure 15:Boxplot for Electricity usage Distribution by Churn*

# 6. Modelling:

## 6.1 Baseline Model

Drawing on hypotheses from prior research, we developed a baseline model as follows:

To assess this baseline model, we employ logistic regression. The logistic regression results indicate that:

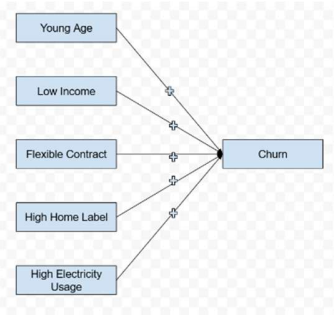



*Figure 16 Baseline Model*

Age, with an estimate of -0.016 (p-value < 2e-16, which is less than 0.05), has a significant negative effect on churn. Based on the odds ratio of 0.984, we accept H1 and conclude that each additional year of age reduces the odds of churn by 1.6%.

Income, with an estimate of -0.00013 and a p-value less than 0.05, also has a slight negative effect on churn. Based on the odds ratio of 0.9998, we accept H2 and conclude that an increase of 1 EUR in monthly income reduces the odds of churn by just 0.0125%.

Contract type, with an estimate of 1.787 and a p-value less than 0.05, shows the most significant negative effect on churn. Given the odds

ratio of 0.16741, we accept H3 and conclude that having a flexible contract increases the odds of churn by 497% compared to fixed contracts.

Home label, with an estimate of -0.167 and a p-value less than 0.05, also shows a negative effect on churn. The odds ratio of 0.84585 indicates that a one-unit increase in the home label reduces the odds of churn by 15.4%. Therefore, we accept H4.

Electricity usage, with an estimate of 0.001 and a p-value less than 2e-16, demonstrates a significant but very small positive effect on churn. The odds ratio of 1.00121 indicates that a 1 kWh increase in electricity usage raises the odds of churn by 0.1%. Therefore, we accept H5.

To evaluate the predictive power of the baseline model, we compute the hit rate, top decile lift, and GINI coefficient. The results are presented below:

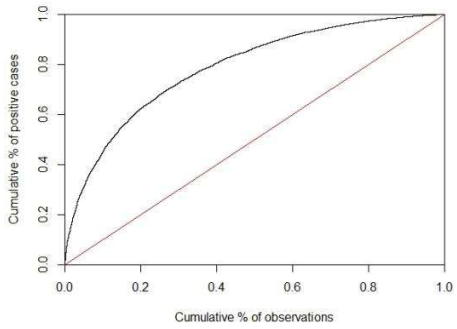



*Figure 17: Life curve*

The model's hit rate is 0.71505, indicating that 71.505% of churners were correctly predicted. Since the dataset is balanced (50% churners, 50% retainers), the model's hit rate exceeds the random guessing rate of 50%, demonstrating its good effectiveness.

The top decile lift results indicate a TDL of 1.827, meaning the model is 1.8 times more effective at predicting churn in the top decile of predicted probabilities. It also shows that the probability of churn increases from 10.2% in decile 1 to 89.6% in decile 10, demonstrating that the model ranks individuals well, with higher deciles having a higher probability of churn. The GINI coefficient is 0.5773489, which is greater than 0.5, and the plot (Figure 17) shows that the lift curve is closer to the top-left corner. This indicates that the model performs moderately well to good, outperforming random guessing.

To evaluate the model's ability to generalize to unseen data, we proceed with model validation by training it on the estimation sample (75% of the original dataset) and then testing its predictive performance on the validation sample (25% of the original dataset).

The estimates and p-values from the logistic regression yield similar results, with contract type having the largest effect on churn (estimate = -1.775) and electricity being the only variable with a small positive impact on churn (estimate = 0.001). The hit rate remained stable at 0.712, and the TDL value was 1.8219, with the probability of churn increasing from 11.5% in the first decile to 89.3% in the tenth decile, indicating a strong ranking. The GINI coefficient is 0.5760506, and the lift curve closely mirrors that of the original dataset. This suggests that the model is not overfitting and exhibits consistent performance.

## 6.2 Stepwise regression

Stepwise regression is a useful method that helps in explorative analysis to decide which variables to include in a model. It is based on the variable’s contribution to model performance. The stepwise regression method analyses whether the model should include variables in a, as the name suggests, stepwise manner, and this can be done either backward, forwards, or in both directions. Besides, the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (BIC) are both used, which are information criteria that reward model improvements by penalizing model

complexity (de Haan, 2024). Also, all three directions were done and compared. A lower score for both is preferred, and the BIC prefers simpler models in comparison with the AIC.

The common predictors that consistently appeared in the final models across all the stepwise regressions were: contract_type, electricity_usage and gas_usage, relation_length, income, age, and start_channel. Due to the fact that BIC models favor simpler generalizable models, the AIC-based models retained more variables, which included email_list, province, and home_label. The variables that were excluded from the models were customer_ID, gender, contract_length, and home_age. The final AICs for the AIC models were the same and were 15261.32, similar to the BIC models, which had a final AIC score of 15385.59.

| Predictor | Backward AIC | Forward AIC | Both AIC | Backward BIC | Forward BIC | Both BIC |
|---|---|---|---|---|---|---|
| contract_type | O | O | O | O | O | O |
| Electricity_usage | O | O | O | O | O | O |
| Gas_usage | O | O | O | O | O | O |
| Relation_length | O | O | O | O | O | O |
| Income | O | O | O | O | O | O |
| Start_channel | O | O | O | O | O | O |
| Email_list | O | O | O | O | O | O |
| Age | O | O | O | X | X | X |
| Province | X | O | O | X | X | X |
| Contract_length | X | X | X | X | X | X |
| Home_label | X | X | X | X | X | X |
| Home_age | X | X | X | X | X | X |
| Gender | X | X | X | X | X | X |
| Customer_ID | X | X | X | X | X | X |
| Final AIC Score | 15261.32 | 15261.32 | 15261.32 | 15385.59 | 15385.59 | 15385.59 |

*Table 1: Stepwise regression*

## 6.3 CART trees

CART (Classification and Regression Tree) is a non-linear classification technique that divides data into segments using binary splits based on different features. In this study, two CART models were developed for predicting customer churn: CART Custom and CART Optimized. These models help assess the performance of a customized model versus an optimized version, focusing on prediction accuracy and the ability to capture key variables.

**Model 1: CART Custom Model**

The CART Custom model was created with stricter parameters to serve as a baseline. The key parameters for this model are as follows: Maximum tree depth: 3; Minimum split sample size: 100; Minimum leaf sample size: 50; Pruning parameter (cp): 0.01.

Figure 1 shows the decision tree for the CART Custom model. These constraints were chosen to generate a simpler model that would likely underfit but serve as a baseline for comparison with the more complex CART Optimized model.

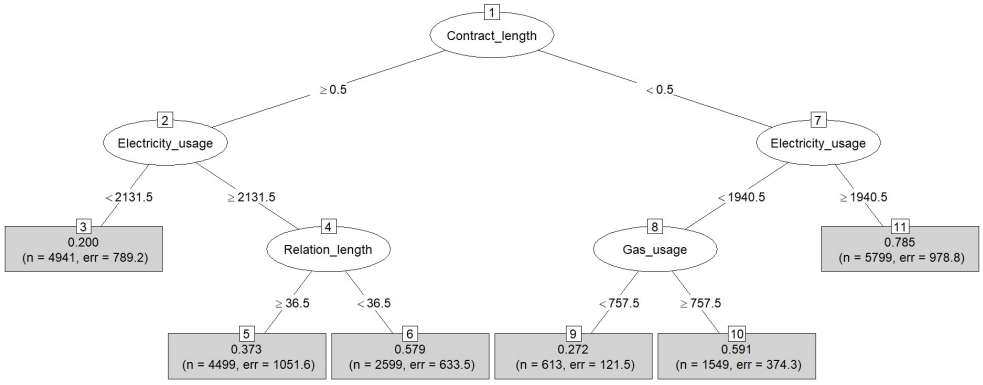



*Figure 18: Cart tree for CART Custom Model*

**Model 2: CART Optimized Model**

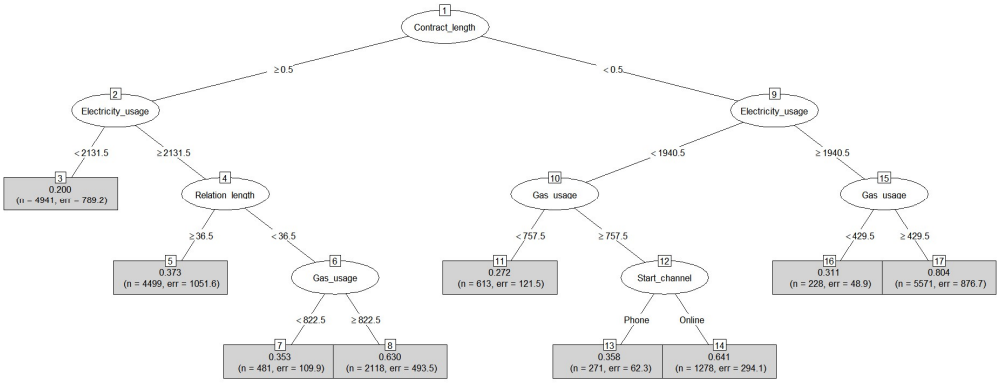



*Figure 19: Cart tree for CART Optimized Model*

The CART Optimized model was designed to improve predictive performance by relaxing some of the constraints from the Custom model. The parameter adjustments were made based on the size of the tree and the model's ability to fit the data. The optimized parameters are: Maximum Depth: 5 (increased from 3 to capture more interactions between variables); Minimum Split Sample Size: 50 (reduced to allow more splits in the tree); Minimum Leaf Sample Size: 25 (reduced to increase the number of leaves and refine predictions); Pruning Parameter (cp): 0.005.

Figure 3 shows the decision tree for the CART Optimized model. By increasing the maximum depth and reducing the minimum split and leaf sample sizes, we allowed the model to capture more

complex relationships between the features and churn. The pruning parameter was adjusted to 0.005 to further improve the model's flexibility without overfitting.

**Feature Importance**

During model estimation, feature importance analysis revealed that Electricity Usage was the most significant predictor in both models. Interestingly, in the CART-optimized model, contract_type and contract_length also gained importance, suggesting these variables provide additional predictive value when the model is optimized.

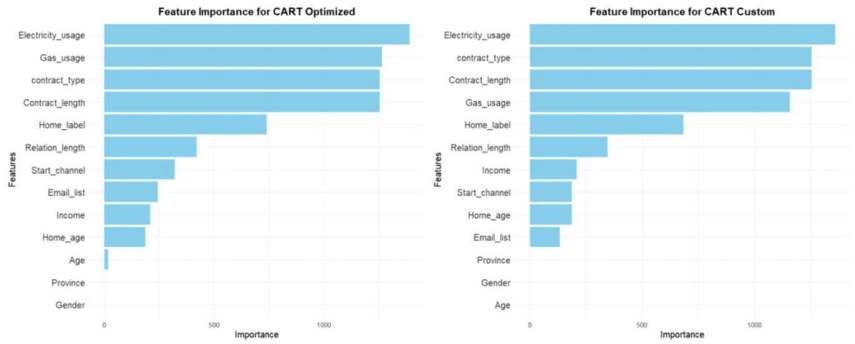



*Figure 20: Feature Importance for CART Custom&Optimized*

On the other hand, variables like Gender and Province showed negligible contributions in both models and were excluded in the CART Optimized version, which focused on the most influential features. By concentrating on these key variables, the CART Optimized model demonstrated better prediction efficiency and accuracy.

## 6.4 Bagging

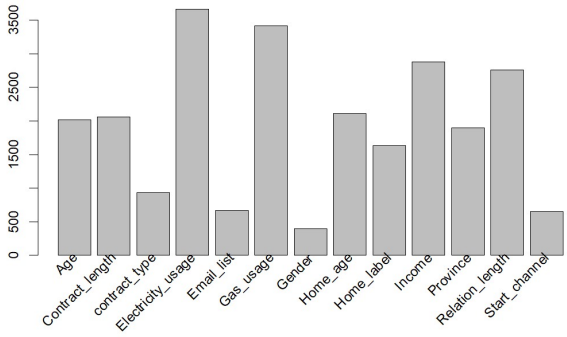



*Figure 21:Importance - Bagging*

Bagging is an ensemble method that uses bootstrap as a model training method. It randomly slices the dataset into different subsets, uses the subsets to train various models, and finally synthesizes all the trees to obtain the optimal model, which is more accurate than the use of only one decision tree. In this study, we combined 500 repetitions of bagging with decision trees. The Hit Rate of the final model is 75.28%, the Top Decile Lift is 1.85, and the Gini index is 0.64, indicating that the model is good at distinguishing between churn and non-churn customers. We also used the varImp() function to calculate the importance of the variables in the model, from the figure we can see that Electricity_usage, Gas_usage, Income, and Relation_length are very important and are the key variables, Email_list, Start_channel, Gender are the variables that have the least influence on the model.

## 6.5 Boosting

Boosting is also an ensemble learning technique that calculates the weights of all the data points, in iterative model training it identifies the data points that are misclassified in the decision tree and

gives more weights to these misclassified data points, which gradually improves the accuracy of the overall model after iterative model training. In this study, boosting combines 10,000 iterations of the decision tree. By using the gbm.perf() function, we obtained the optimal number of iterations for the model, which is 956. The Hit Rate of the model is 76.83%, the Top Decile Lift is 1.92, and the Gini coefficient is 0.698. In addition to this, we also obtained the Relative Influence of the variables, which is shown in the following graphs Electricity_usage, Contract_length, Gas_usage, and contract_type are the four most important variables, while Province, Email_list, and Start_channel have relative importance less than 5%.

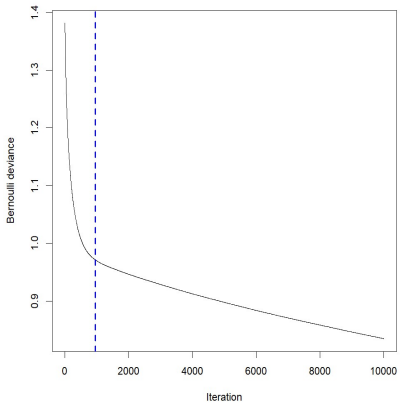



*Figure 22: Bernoulli deviance (error) line of boosting*

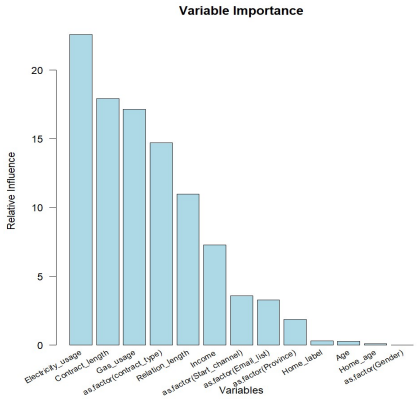



*Figure 23: Importance - Boosting*

## 6.6 Random Tree

In this study, we used Random Forest models to predict customer churn and compared two versions of the model: the baseline Random Forest model and the tuned Random Forest model. The evaluation of these two models sheds light on the impact of hyperparameter tuning on model performance.

**Model Performance Comparison**

- Hit Rate: This measures the proportion of correctly identified churn customers. Both models have nearly identical Hit Rates: 0.7549 for the baseline and 0.7530 for the tuned model, indicating little difference in overall accuracy.
- Top Decile Lift: This measures accuracy within the top decile of churn predictions. The tuned model (1.8831) shows a slight improvement over the baseline model (1.8544), indicating better performance in identifying high-risk churn customers.
- Gini Coefficient: This measures the model's ability to distinguish between churned and non-churned customers. The tuned model has a slightly higher Gini coefficient (0.6554) than the baseline model (0.65499), suggesting better discriminatory power.

While the Hit Rates are nearly the same, the tuned model performs slightly better in terms of Top Decile Lift and Gini coefficient. This suggests that hyperparameter tuning improves the model's focus on high-risk churn customers and enhances its ability to discriminate between churned and non-churned customers.

### Feature Importance Analysis

Feature importance analysis revealed that Electricity_usage and Gas_usage were the most significant predictors in both the baseline and tuned models. However, in the tuned model, the importance of Income and Start_channel increased, suggesting that these features provide additional predictive value when the model is optimized. This highlights how hyperparameter tuning allows the model to better utilize these features, improving the accuracy of predictions.

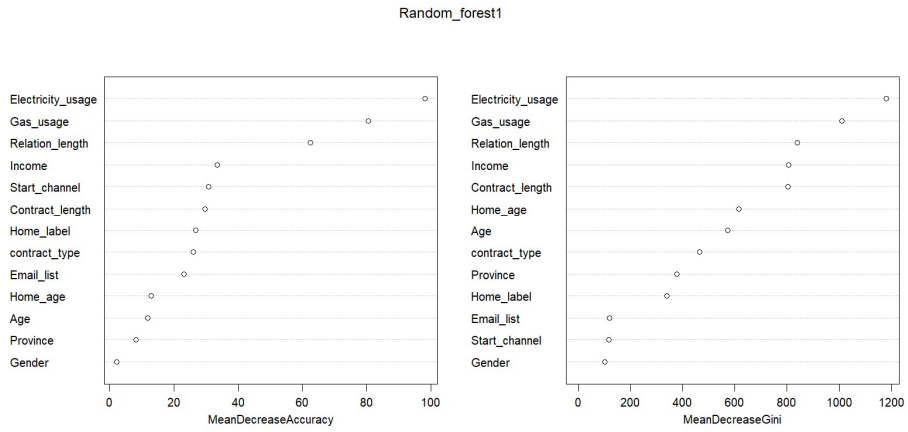



*Figure 24: Comparison of Model Performance for Random Forest and Tuned Random Forest*

## 6.7 Support Vector Machines

In the final analysis of the models, the SVM model was used. After comparing the different model predictions and fit criteria and based on the theory, the SVM model with a radial basis function (RBF) kernel was chosen. This model is able to effectively capture non-linear relationships in the data, as evidenced by the curved decision boundary in the classification plot of Electricity_usage and Gas_usage. Features such as Gas_usage and Electricity_usage play a significant role in explaining churn but show substantial overlap between the classes (Churn = 0 and Churn = 1). This overlap indicates that these two features alone are insufficient for a perfect separation, and thus the other importance of all the predictors such as Age, Income, and Relation_length. Due to

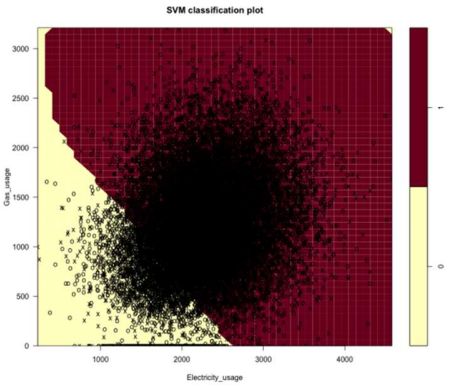



*Figure 25: SVM plot for Electricity_usage*

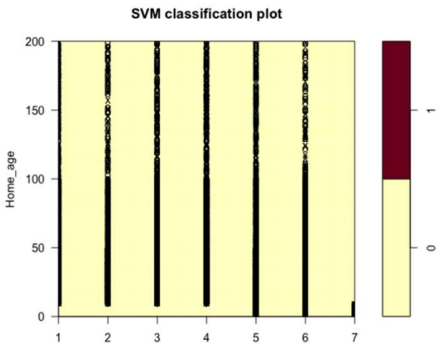



*Figure 26: SVM plot for Home_label*

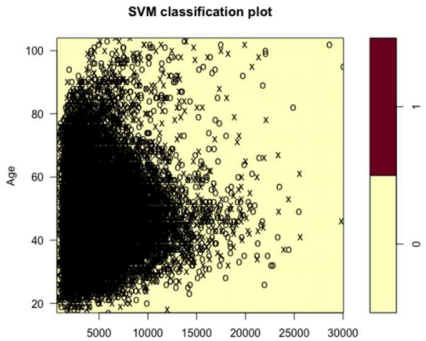



*Figure 27: SVM plot for Income*

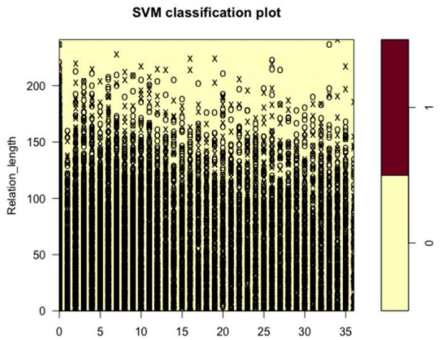



*Figure 28: SVM plot for Contract_length*

the difficulty of visualizing relationships in the models, multiple 2-dimensional relationships in regards to churn 4 classification plots are shown.

While the SVM model provides a robust non-linear approach, it requires careful evaluation of a validation dataset to confirm its predictive performance and generalizability. Additionally, the model's probabilistic outputs enable practical business applications like ranking customers by churn risk.

# 7. Conclusion

## 7.1 Model Evaluation

In the validation phase, all models were evaluated using three key metrics: classification accuracy (Hit-Rate), Top-Decile Lift (TDL), and Gini coefficient. The results are summarized in the table below:

| Model | Hit-Rate | Top-Decile Lift (TDL) | Gini Coefficient |
|---|---|---|---|
| Logistic Regression | 71.2% | 1.82 | 0.576 |
| Stepwise AIC | 74.87% | 1.87 | 0.66 |
| Stepwise BIC | 74,83 | 1,89 | 0.66 |
| CART Custom | 71.0% | 1.63 | 0.51 |
| CART Optimized | 72.7% | 1.67 | 0.53 |
| Bagging | 74.33% | 1.83 | 0.63 |
| Boosting | 76.83% | 1.92 | 0.698 |
| Random Tree | 75.3% | 1.86 | 0.66 |
| Random Forest Tuned | 75.0% | 1.87 | 0.65 |
| Support Vector Machines | 74.49% | 1.79 | 0.63 |

*Table 2: Validation Results Comparison By Three Key metrics*

## 7.2 Managerial conclusion

In line with the business objective, the company aims to predict which customers are most likely to churn, necessitating the selection of the best model for churn prediction. Given this goal, *accuracy and strong out-of-sample performance are crucial factors in choosing the model*.

When comparing the three key metrics across all models, Boosting stands out with the highest hit rate (74.33%), TDL (1.92), and GINI coefficient (0.698). As a result, Boosting emerges as the best model, offering the highest predictive accuracy and consistent out-of-sample performance, making it the most suitable solution for the company's needs.

The baseline model we developed, based on hypotheses from previous research, performs well on out-of-sample data, achieving a hit rate of 71.2%, a TDL of 1.82, and a GINI coefficient of 0.576. The logistic regression results confirm that the factors identified by the model significantly impact churn. We validated all the hypotheses, concluding that younger age, lower income, higher levels of home label, and higher electricity usage are associated with increased churn.

Based on the results from both the Boosting model and the baseline model, we can identify the key factors driving churn for the energy supplier and provide recommendations for the company's actions:

Electricity usage emerges as the most significant factor influencing consumers' churn intentions. This finding aligns with the hypothesis proposed by previous studies and is supported by all models. Similarly, gas usage is identified by Boosting, CART trees, and random trees as one of the strongest factors related to churn. Based on these insights, the supplier should prioritize these energy consumption variables, focusing on building strong relationships with high-consumption customers. One potential approach could be offering a Tiered Loyalty Program with personalized or special services to retain these customers and prevent them from switching to competitors.

Next, contract length and contract type are identified as the next most important variables by the Boosting model. While the baseline model explored the impact of contract type and found that flexible contracts negatively affect churn, it did not include contract length. This new finding from Boosting suggests that, in addition to having a fixed contract, the duration of the fixed contract a consumer holds with the supplier may provide a more accurate prediction of churn. The supplier should also utilize contract type and contract length for customer segmentation to identify potential loyal customers versus those who are more likely to switch brands, allowing for the application of targeted and effective retention strategies.

Some potential variables to consider include relationship length and income. In addition to contract length, the duration of the customer relationship with the supplier reflects how familiar the customer is with the supplier, which could positively influence churn predictions. Income, as demonstrated in the baseline model, has a negative impact on churn, a trend supported by most of the models. This suggests that the supplier should implement targeted retention strategies for low-income customers, especially if they form a significant part of the customer base.

Despite its high accuracy and consistent performance, the supplier must also consider the limitations of Boosting. First, it only demonstrates correlations with the dependent variable and indicates the relative importance of factors, but it does not provide insight into the causal direction of these relationships. As a result, it does not clarify how the factors are related to churn, making interpretation challenging. Second, it is difficult to explain why certain variables are more significant than others. Therefore, depending on the specific managerial objectives, it would be beneficial for the company to use the variables identified by Boosting and conduct causal analysis (such as the Granger causality test) to gain a deeper understanding of how these factors contribute to churn, including the direction and strength of their impact.

# Reference


Amador, F. J., González, R. M., & Ramos-Real, F. J. (2013). Supplier choice and WTP for electricity attributes in an emerging market: The role of perceived past experience, environmental concern and energy saving behavior. Energy Economics, 40, 953–966. https://doi.org/10.1016/j.eneco.2013.06.007

Ahn, J.-H., Han, S.-P., & Lee, Y.-S. (2006). Customer churn analysis: Churn determinants and mediation effects of service quality and satisfaction. Telecommunications Policy, 30(1), 42–55. https://doi.org/10.1016/j.telpol.2005.06.005 (s11301-023-00335-7).

Coussement, K., & Van Den Poel, D. (2006). 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. https://doi.org/10.1016/j.eswa.2006.09.038

Ek, Kristina and Patrik So¨derholm. (2008). "Households' switching behavior between electricity suppliers in Sweden." Utilities Policy 16(4): 254–261. https://doi.org/10.1016/j.jup.2008.04.005.

Haridasan, A.C.; Fernando, A.G.; Balakrishnan, S. Investigation of consumers' cross-channel switching intentions: A push-pullmooring approach. J. Consum. Behav. 2021, 20, 1092–1112.

Gamble, Amelie E. Asgair Juliusson and Tommy Ga¨rling. (2009). "Consumer attitudes towards switching supplier in three deregulated markets." The Journal of Socio-Economics 38(5): 814–819. https://doi.org/10.1016/j.socec.2009.05.002.

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

Marmolejo-Duarte, C., & Bravi, M. (2017). Does the energy label (EL) matter in the residential market? a stated preference analysis in Barcelona. Buildings, 7(4), 53. https://doi.org/10.3390/buildings7020053

Oetama, R. S. (2023). Unveiling churn prediction at bank ivory. Jurnal Informatika Dan Teknik Elektro Terapan, 11(3s1). https://doi.org/10.23960/jitet.v11i3s1.3394

Deller, D., Giulietti, M., Loomes, G., Waddams, C. P., Moniche, A., & Jeon, J. Y. (2021). Switching energy suppliers: It's not all about the money. The Energy Journal, 42(3). https://doi.org/10.5547/01956574.42.3.ddel

de Haan, E. (2024). Data science methods for MADS: Session 2—Tree models, bagging, and boosting. Faculty of Economics and Business, Department of Marketing. University of Groningen.



Haridasan, A.C.; Fernando, A.G.; Balakrishnan, S. Investigation of consumers' cross-channel switching intentions: A push-pullmooring approach. J. Consum. Behav. 2021, 20, 1092–1112.

He, X., & Reiner, D. (2017). Why consumers switch energy suppliers: The role of individual attitudes. Energy Journal, 38, 25–54. https://doi.org/10.5547/01956574.38.6.hxia

Hübner, F., Herberger, T. A., & Charifzadeh, M. (2023). Determinants of customer recovery in retail banking—lessons from a german banking case study. Journal of Financial Services Marketing. https://doi.org/10.1057/s41264-023-00224-w

Lipowska, I., Lipowski, M., Dudek, D., & Mącik, R. (2024). Switching behavior in the Polish energy market—The importance of resistance to change. Energies, 17(306). https://doi.org/10.3390/en17020306

Ribeiro, H., Barbosa, B., Moreira, A. C., & Rodrigues, R. G. (2023). Determinants of churn in telecommunication services: A systematic literature review. Management Review Quarterly, 74(1327–1364). https://doi.org/10.1007/s11301-023-00335-7 (s11301-023-00335-7)

Seo, D., Ranganathan, C., & Babad, Y. M. (2008). Two-level model of customer retention in the US mobile telecommunications service market. Telecommunications Policy, 32(3–4), 182–196. https://doi.org/10.1016/j.telpol.2008.01.001 (s11301-023-00335-7)

Svendsen, G. B. and Prebensen, N. K. (2013). The effect of brand on churn in the telecommunications sector. European Journal of Marketing, 47(8), 1177-1189. https://doi.org/10.1108/03090561311324273

Giulietti, M., Waddams Price, C., & Waterson, M. (2013). Consumer choice and competition policy: A study of UK energy markets. The Economic Journal, 113(491), 488–517. https://doi.org/10.1111/1468-0297.t01-1-00129

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2010). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems With Applications, 38(3), 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023

Keaveney, S. M. (1995). Customer switching behavior in service industries: An exploratory study. Journal of Marketing, 59(2), 71–82. https://doi.org/10.2307/1252074 (Switching_Behavior_in_t…)

Rabih, R., Sun, W., Ayoubi, M., & Jamal, W. (2024). Highly accurate customer churn prediction in the telecommunications industry using MLP. International Journal of Integrated Science and Technology, 2(10), 907–918. https://doi.org/10.59890/ijist.v2i10.2564

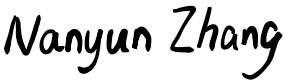
Umayaparvathi, V., & Iyakutti, K. (2012). Applications of data mining techniques in telecom Churn prediction. International Journal of Computer Applications, 42(20), 5-9.

## AI statement

We used AI for checking spelling, grammar, and rewriting. We also used AI to improve our R-code. For all of this, we used ChatGPT 4o.

## Signature

Ny Ha Dao - S5677858: 

Nanyun Zhang - S5658292: 

Yichong Tao - S5065623: 

Tim van Noord - S3734978: