|  |  |
| --- | --- |
| Magnifying glass showing decling performance  Customer Churn Prediction Report  Date: 23/07/2023 | Abstract  This report presents an analysis of customer churn prediction using machine learning techniques for a US telecommunication service. The objective of the report is to identify factors that influence customer churn, build predictive models to accurately forecast churn behaviour as well as establish target customer profile and provide some recommendations to their proactive churn-management program.  Thao Van Phung – z5353483  COMM3501 |

# Table of Content

## Introduction

## Data Collection and Limitations

## Exploratory Data Analysis

3.1 Descriptive Statistics

3.2 Data Visualisation

3.4 Important Variables

## Model Comparison

4.1 Methodology

4.2 K-Nearest Neighbours (KNN)

4.3 Random Forest

4.4 Logistic Regression

4.5 Best Models

## Resampling Methods and Model Results

6.1 Performance Comparison of Resampling Methods

6.2 Interpretation of Results

## Recommendations and Conclusions

## References

## Introduction

The client is a major U.S. wireless telecommunications company which provides cellular telephone service, facing fierce competition in satisfying its customers. Our task is to develop a statistical model to predict churn, establish a target customer profile for a proactive churn-management program, providing the solution to the client’s customer-facing call centres.

The desired model should accurately predict churn and provide insights into customer behaviours, helping develop better customer retention strategies and improve overall satisfaction improvement.

## Data Collection and Preprocessing

To ensure the accuracy of the analysis, rows containing any missing data points were removed. This process resulted in a reduction of the total number of rows from 30,000 to 29,377, where each remaining row is now complete with all required information.

While removing NA values allowed us to proceed with a more complete dataset, it is important to acknowledge the potential effects of this data reduction. There is a possibility that some valuable information could have been excluded from the analysis, potentially impacting the model performance or generalisability of the results if the missing values are huge (Kumar 2020). However, for this dataset, the number of NA rows removed were small, which is likely to have limited impacts on the results.

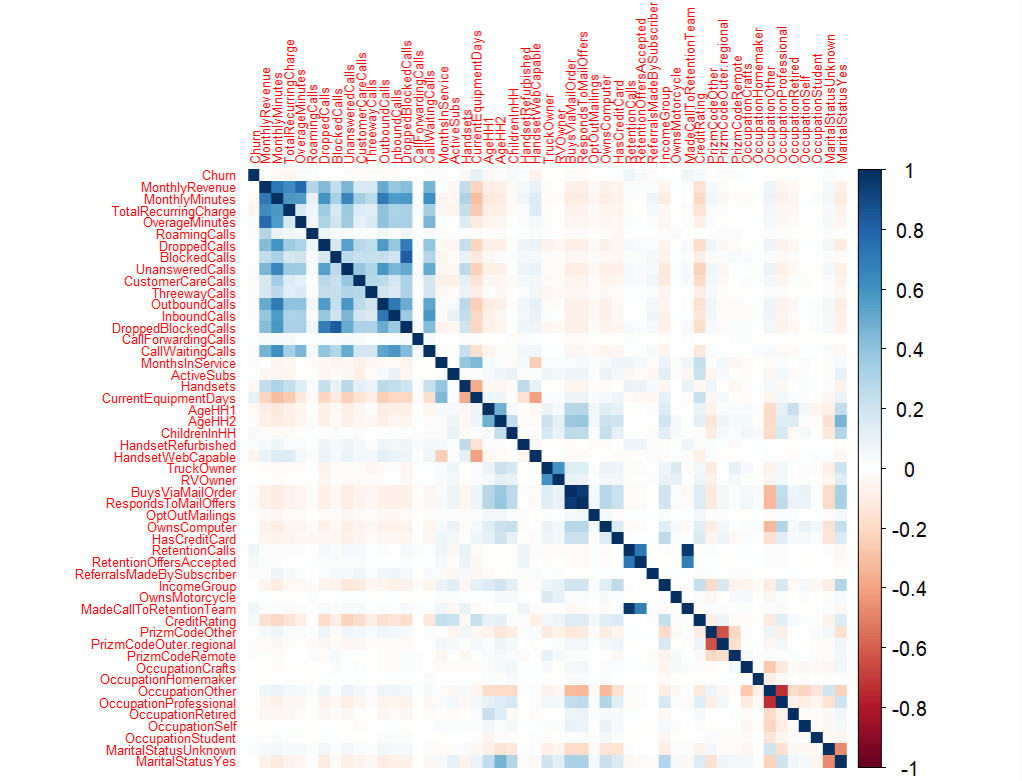
Thus, categorical variables were converted to dummy variables for better analysis. For categorical variables that only have 2 classes (e.g Yes/No), they were converted to binary representation 0, 1 and for those with more than 2 classes, a process called one-hot encoding was employed. One-hot encoding involved creating new binary columns for each unique category within the original categorical variable (Brownlee 2020). Each binary column was then labelled with 0 or 1, indicating the presence or absence of the corresponding category for each observation.

## Exploratory Data Analysis

3.1 Descriptive Analysis

The goal is to identify the most influential predictors for the churn prediction models. Therefore, both correlation tests and chi-square tests to examine the relationships between the target outcome, "Churn", and various predictor variables in the dataset.

Additionally, a correlation test was performed to quantify the strength and direction of the linear relationship between each predictor variable and the target outcome (Soetewey 2020). The correlation test underscores the potential linear associations, suggesting the predictors that have the highest correlation with the target outcome to be considered for inclusion in the final models. Nonetheless, a heat map was utilised to visually represent these relationships as highlighted in graph 1, with darker colours indicating stronger correlations. Thus, this visualisation provided a comprehensive overview of the relationships between variables, highlighting the potential impact of the predictors on the churn prediction.



Graph 1. Heatmap of correlation test among variables

The predictor variables that have the strongest influence on “Churn” variables are:

Blocked Calls, Income Group, Roaming Calls, Retention Calls, Active Subs, Months In Service, AgeHH1, AgeHH2, Overage Minutes, Credit Rating, Customer Care Call.

However, the correlation test only captures linear relationships and is not optimal to capture the categorical variable's influence on churn prediction as “Churn” is also categorical. Hence, a chi-square test was conducted, which specifically assesses the association between categorical variables (Kent State University 2023). Hence, predictors that have significant association with Churn are: Made Call To Retention Team, Responds To Mail Offers, Handset Refurbished, Buys Via Mail Order, Handset Web Capable, Owns Computer, Children In Household and Marital Status.

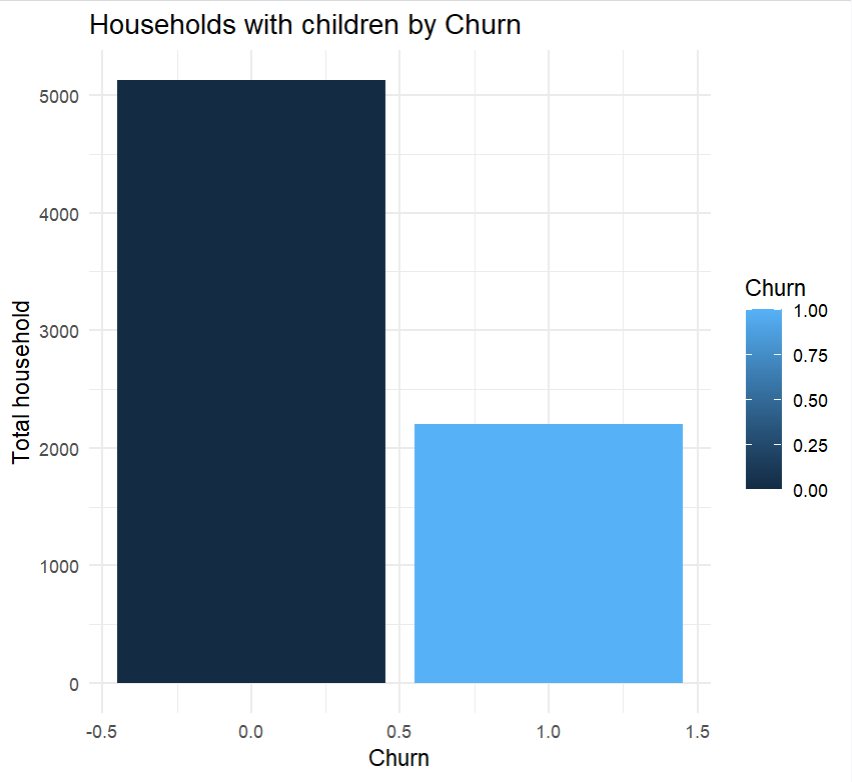
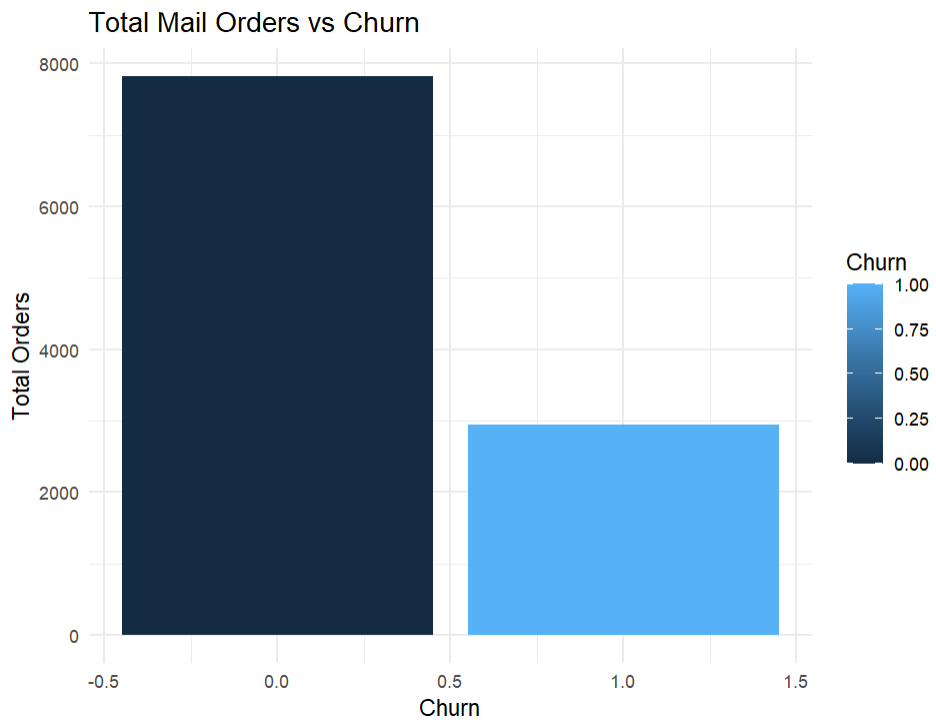
3.2 Data Visualisation

A graph of a number of blue squares

Description automatically generated with medium confidence

Graph 2. Imbalanced class of Churn

Graph 2 visualises the number of customers who churned versus non-churn customers in the training dataset. The non-churn number of customers are double the number of churn customers. Although this is a good sign for the client’s current performance, this imbalanced outcome can lead to bias issues when building predictive models as they perform poorly on the minority class – churn customer (Tang, 2023). Hence, it is necessary to employ some resampling methods to improve the models’ performance.



Graph 3. Difference between churn vs non-churn customer

Nonetheless, several visualisations were employed to highlight possible different characteristics between churn and non-churn customers. For example, in graph 2, it suggests that churn customers are those who received less customer care calls and buy less via mail order than those who stayed, which is consistent with the chi-square test results about Buys Via Mail Order and Children in Household variable’s impact.

3.4 Important Variables

Furthermore, the predictors should not have high correlation to avoid multicollinearity problems when building prediction models. Hence, our final predictors selected are Blocked Calls, Income Group, Roaming Calls, Retention Calls, Active Subs, Handset Web Capable, Buys Via Mail Order, Owns Computer, Months In Service, AgeHH1, AgeHH2, Overage Minutes, Credit Rating, Children In Household, Marital Status Unknown, Marital Status Yes, Customer Care Calls. Another heat map was utilised to visually represent the relationships and patterns among predictors, ensuring that they are not highly correlated with each other highlighted in Appendix 1.

## Model Comparison

4.1 Methodology

For the prediction modelling phase, cross-validation techniques are employed to ensure a robust and reliable performance of 4 types of models: K-Nearest Neighbour, Random Forest, Classification Tree, and Logistic Regressions. The dataset is divided into two subsets: an 80% training set and a 20% validation set. This hold-out method aims to train the models on the larger training set and evaluate their performance based on the independent validation set, evaluating their ability to generalise to unseen data (Lakshana 2021).

As Churn prediction is a classification task, the models were validated using the following metrics: Accuracy, Sensitivity, Specificity and Area Under the ROC Curve (AUC) (Goyal 2021). Accuracy measures the proportion of correctly predicted outcomes out of the total instances. Additionally, sensitivity evaluates the model's ability to correctly identify positive outcomes (the true positive rate), which is Churn equals 1 (Yes) in this scenario. In contrast, specificity assesses the model's capability to correctly identify negative outcomes. Lastly, the AUC summarises the model's classification ability, which should be higher than 50% at least to be considered better than random chance (Dembla 2020). Nonetheless, the desired model should correctly predict customers who will churn, accuracy, sensitivity and AUC will be prioritise when comparing models.

Finally, as the target outcome had imbalanced class, multiple resampling methods were applied to improve the overall performance of the selected best models.

4.2 K-Nearest Neighbours (KNN)

The K-Nearest Neighbours (KNN) model works by measuring the similarity between a new customer (a datapoint) and the existing customers in the dataset. It does this by calculating the distance between the chosen features/predictors of the new customer and those of the existing customers (Srivastava 2023). Hence, all predictors were scaled before feeding it into the model. This KNN model then selected k nearest neighbours based on the smallest distance. Once the nearest neighbours are identified, it looks at the churn status of these neighbours. If most of the neighbours have churned, then the new customer is assigned a churn label, indicating that they are likely to churn as well.

Out of the 3 KNN models conducted with different predictors (Appendix), this model appears to be the most optimal as it has the highest performance metrics as highlighted in table 1. Its optimal k is 10 and it only includes Blocked Calls, Income Group, Roaming Calls, Retention Calls, Active Subs, Handset Web Capable, Buys Via Mail Order, Months In Service, AgeHH1, AgeHH2, Overage Minutes, Credit Rating, Customer Care Call as predictor.

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Sensitivity | Specificity | AUC |
| 69% | 12% | 92% | 55% |

Table 1. Performance metrics of the most optimal KNN model

The model's accuracy of 69% highlights that it correctly predicts the outcome for 69% of the instances in the dataset. Thus, the sensitivity of only 12% suggests that the model performs poorly in customer who will churn. Additionally, the specificity of 92% suggest that the model is only good at predicting non-churn customers. Finally, the AUC of 55% highlights that the model's overall performance is only slightly better than random guessing.

Although KNN model is easy to understand and has the highest sensitivity level, it provides no information on predictor importance and can be computationally expensive, especially with large datasets. Hence, it is not useful in capturing customers’ characteristics and factors that influence churn tendency.

4.3 Random Forest

The decision tree starts at the top with the most important question based on the dataset. Depending on the answer, it moves down to the next question, and so on, making up the decision nodes in the tree and acting as a means to split the data until it reaches a final category (IBM n.d).

Random Forest is a smart algorithm that combines the opinions of many decision trees to make accurate predictions, each trained on different data samples and predictors. By taking a majority prediction of all the trees, it arrives at the final prediction (IBM n.d). This approach prevents overfitting, handles large datasets, and provides robust results, making it a popular choice for predicting outcomes in various fields.

Out of the 3 models of the same family performed, this random forest model is the most optimal as it has the highest performance metrics as highlighted in table 2. The model excludes Owns Computer predictor to achieve higher performance.

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Sensitivity | Specificity | AUC |
| 71% | 5% | 97% | 58% |

Table 2. Performance metrics of the most optimal Random Forest model

The model's accuracy of 71%, which is slightly higher than the overall accuracy of the KNN model. However, the sensitivity of only 5% suggests that the model still performs poorly in identifying churn customers. Thus, the specificity of 97% implies that the model is good at predicting non-churn customers. Lastly, the AUC of 58% indicates that the model's classification ability is slightly better than KNN.

Random Forest also provides a mean decrease accuracy, which highlights the importance of each predictor by evaluating the reduction in accuracy of the model when a specific predictor is randomly shuffled as highlighted in table 3 (Ignacio 2020).

A graph with text overlay

Description automatically generated

Table 3. Predictors Importance of Random Forest Model

Based on the table above, it highlights from the most important predictors: Age of the 2 members in the household, Credit Rating, Retention Calls to the least important predictors: Roaming Calls, which is helpful in selecting the most optimal predictors to improve the overall performance. However, the Random Forest model does not showcase the partial effects of these predictors on Churn. Thus, it does not provide a tree diagram or some visualisation, resulting in a lack of interpretability (Singh 2020). Hence, the random forest should only be selected for prediction purposes and cannot be used alone to investigate the characteristics of churn customers.

4.4 Logistic Regression

Logistic regression is a statistical method used for predicting binary outcomes and is designed to classify data into two categories (IBM n.d). Its outcome is a probability of an event occurring, bounding between 0 and 1.

Three logistic regressions with different predictors were conducted to find the best model. Additionally, the best model had the highest performance metrics when using the 9 most significant predictors: Retention Calls, Handset Web Capable, Buys Via Mail Order, Months in Service, AgeHH1, Overage Minutes, Credit Rating, Children in Household, Customer Care Calls.

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Sensitivity | Specificity | AUC |
| 70% | 1% | 99% | 58% |

Table 4. Performance metrics of the most optimal Logistic Regression model

Based on table 4, the model accuracy is quite similar to the Random Forest model. However, its ability to predict churn customer is extremely low and it is much better at predicting non-churn customers. Thus, the AUC is equals to Random Forest’s.

Thus, Logistic regression model is computational efficient and informative as it provides information about important predictors and probabilistic interpretation (Lawton n.d). Hence, even though the sensitivity is not as high as the other two models, this Logistic Regression is useful for investigating customers behaviour and characteristics.

4.5 Best Models

Given the objective is to accurately identify customers that will churn and understand their characteristics and behaviours, both the Random Forest model and Logistic Regression model should be utilised together for the most optimal performance. The Random Forest model can avoid overfitting, making it robust and reliable for predictive tasks. On the other hand, the Logistic Regression model provides probabilistic interpretability, enabling us to understand the partial effects of individual predictors on the likelihood of churn.

## Resampling Methods and Model Results

6.1 Performance Comparison of Resampling Methods

In the process of comparing different resampling methods, various techniques were performed on the selected Logistic Regression and Random Forest models, including random under sampling, over sampling, cutoffs at prevalence and SMOTE with both scaled and unscaled data (Appendix 5). Among these methods, random under sampling with unscaled data yielded the highest performance metrics for both models.

The chosen method randomly reduced the number of non-churn customer to match the number of churn customer, creating a balanced dataset that allows the predictive models to give equal importance to both classes (Pykes 2020). Hence, this method handled the imbalanced data effectively, resulting in an improved sensitivity up to 52% and 59% for the 2 models and predicting the churn customer in the validation dataset and the whole dataset better as highlighted in table 5.

|  |  |  |  |
| --- | --- | --- | --- |
| Models | Accuracy | Sensitivity | Specificity |
| Performance on **Validation Dataset** | | | |
| Logistic Regression | 51% | 52% | 50% |
| Random Forest | 55% | 59% | 53% |
| Performance on the **whole Dataset** | | | |
| Logistic Regression | 57% | 52% | 59% |
| Random Forest | 75% | 91% | 68% |

Table 5. Logistic Regression and Random Forest performance after resampled.

Although the accuracy and specificity of the models were decreased when performing with validation data, the models can predict churn customer much better than before, which suits the objective in this case. Thus, the Random Forest model’s accuracy and sensitivity when performing with the whole dataset was significantly improved. Hence, it is used to predict 3000 customers who are likely to churn.

6.2 Interpretation of Results

|  |  |
| --- | --- |
|  | Coefficients |
| Intercepts | 0.22 |
| RetentionCalls | 0.64 |
| HandsetWebCapable | -0.44 |
| BuysViaMailOrder | -0.12 |
| MonthsInService | -0.003 |
| AgeHH1 | -0.005 |
| OverageMinutes | 0.0004 |
| CreditRating | 0.07 |
| ChildrenInHH | 0.21 |
| CustomerCareCalls | -0.01 |

Table 6. Logistic Regression results after resampled.

Based on table 6, predictors that have inverse relationship with churn are desirable as it can help establishing strategies to prevent churn. These predictors are highlighted in green on the table above. Customers with web-capable handsets may be more engaged with their mobile services, utilising various features and staying connected, resulting in lower churn rates. Thus, although it is subtle, customers who has been using the service longer, receiving care calls and often make purchases via mail might have a more stable and long-term relationship with the company, leading to lower chances of churn.

In contrast, factors like Retention Calls, Credit Rating, and Households with children have positive relationships churn customer. Potential reasons can be: households with more children might have specific communication needs that the company has not met; customers who make more retention calls may be dissatisfied with the customer service or constantly face issues with their service; customers with high overage minutes might feel they are not on the most suitable plan for their needs; customers with high credit rating might have more options, flexible budget and higher expectations from their mobile service provider. All this dissatisfaction could lead them to consider switching to a different provider, resulting in potential churning.

Recommendations and Conclusions

To prevent churn and retain customers, the telecommunications company can implement a proactive churn-management program that focuses on personalised offers, responsive customer support, and customer feedback surveys. Here is the link to the churn prediction web application using the random forest model <https://vanphug.shinyapps.io/3501project/>. Thus, a screenshot of the app is attached at the end of the report (Appendix 6).

The company can create personalised offers and incentives for at-risk customers to encourage them to stay with the company. This can include discounts, or suggestions of new mobile plans based on the overall usage of that customer. However, repetitive discounts might condition the customers to believe that the regular prices are a rip-off in comparison to discounted prices, reducing trusts from both loyal customers and customers who are more likely to churn (Simpkins 2022).

In addiontion, the company can conduct regular customer feedback surveys to gauge customers’ satisfaction and identify areas for improvement (Stec 2023), which can be done online or in-store. By act on the feedback received, the company can improve the overall customer experience and meet customers’ specific needs by recommending them with the right mobile plan (e.g Household with children). A limitation with this approach can be survey fatigue, which might lead to unreliable results (Proprofssurvey 2023). Hence, the company can utilise incentives like gift card to encourage customers fill out surveys and give feedbacks.

Nonetheless, the company’s app should employ an AI chatbot so that customers’ concerns and questions can be addressed promptly, improving customer experience. Although it is a costly approach, a chatbot may reduce customers’ needs to call the retention team, making it more convenient for both customers and the company as it saves time as well as operation costs (Sumrak 2022).

By tailoring incentives, addressing concerns promptly, and acting on customer feedback, the company may foster customer loyalty and reduce churn rates.

## References

1. Brownlee, J. (2020). *Ordinal and One-Hot Encodings for Categorical Data*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/.
2. Cornell, J. (2021). *Top Advantages & Disadvantages of Customer Feedback Surveys*. [online] ProProfs Survey Blog. Available at: https://www.proprofssurvey.com/blog/advantages-disadvantages-of-customer-feedback-surveys/.
3. Dembla, G. (2021). *Intuition behind ROC-AUC score*. [online] Medium. Available at: https://towardsdatascience.com/intuition-behind-roc-auc-score-1456439d1f30.
4. Goyal, S. (2021). *Evaluation Metrics for Classification Models*. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-models-e2f0d8009d69.
5. IBM (n.d.). *What is Random Forest? | IBM*. [online] www.ibm.com. Available at: <https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly>.
6. IBM (n.d.). *What is Logistic Regression? | IBM*. [online] Available at: https://www.ibm.com/topics/logistic-regression.
7. Kumar, S. (2020). *7 Ways to Handle Missing Values in Machine Learning*. [online] Medium. Available at: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e.
8. Lakshana, G.V. (2021). *Cross-Validation Techniques in Machine Learning for Better Model*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-with-python-code/.
9. Lawton, G. (2022). *What is Logistic Regression? - Definition from SearchBusinessAnalytics*. [online] SearchBusinessAnalytics. Available at: <https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression>.
10. Martinez-Taboada, F. and Renondo, J.I. (2020). Variable importance plot. doi:https://doi.org/10.1371/journal.pone.0230799.g002.
11. Pykes, K. (2020). *Oversampling and Undersampling*. [online] Medium. Available at: https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf.
12. Simpkins, T. (2021). *The Advantages and Disadvantages of Discounts Omnisend Blog*. [online] Omnisend Blog. Available at: https://www.omnisend.com/blog/the-advantages-and-disadvantages-of-discounts/.
13. Singh, J. (2020). *Random Forest: Pros and Cons*. [online] Medium. Available at: https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04.
14. Soetewey, A. (2020). *Correlation coefficient and correlation test in R*. [online] Stats and R. Available at: https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/.
15. Srivastava, T. (2019). *Introduction to KNN, K-Nearest Neighbors : Simplified*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.
16. Stec, C. (2021). *How to Reduce Customer Churn: 7 Helpful Tips to Try*. [online] Hubspot. Available at: https://blog.hubspot.com/service/how-to-reduce-customer-churn.
17. Sumrak, J. (2022). *13 Undeniable Benefits of Chatbots (Plus Challenges) | Twilio*. [online] Twilio Blog. Available at: https://www.twilio.com/blog/benefits-of-chatbots.
18. Tang, T. (2023). *Class Imbalance Strategies — A Visual Guide with Code*. [online] Medium. Available at: https://towardsdatascience.com/class-imbalance-strategies-a-visual-guide-with-code-8bc8fae71e1a [Accessed 6 Aug. 2023].
19. Yeager, K. (2021). *LibGuides: SPSS Tutorials: Home*. [online] libguides.library.kent.edu. Available at: https://libguides.library.kent.edu/SPSS.

## Appendix

Exploratory data analysis

1. Correlation test of selected features

A graph with numbers and text

Description automatically generated

Model Comparison

2. KNN

Model 1 includes all predictors, using 10-fold cross validation,

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Sensitivity | Specitivity | AUC |
| 68% | 10% | 91% | 53% |

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

Models 2 excludes dummy

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Model 3 EXCLUDES: ChildrenInHH, OwnsComputer, MaritalStatusUnknown, MaritalStatusYes

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

3. Random Forest

Model 1 with all predictors

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Sensitivity | Specificity | AUC |
| 71% | 5% | 97% | 58 |

A close-up of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with text overlay

Description automatically generated

A graph with a line

Description automatically generated

Model 2 exclude the last variable:

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Sensitivity | Specificity | AUC |
| 71% | 5% | 97% | 58% |

A black text on a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with text overlay

Description automatically generated

A graph with a line

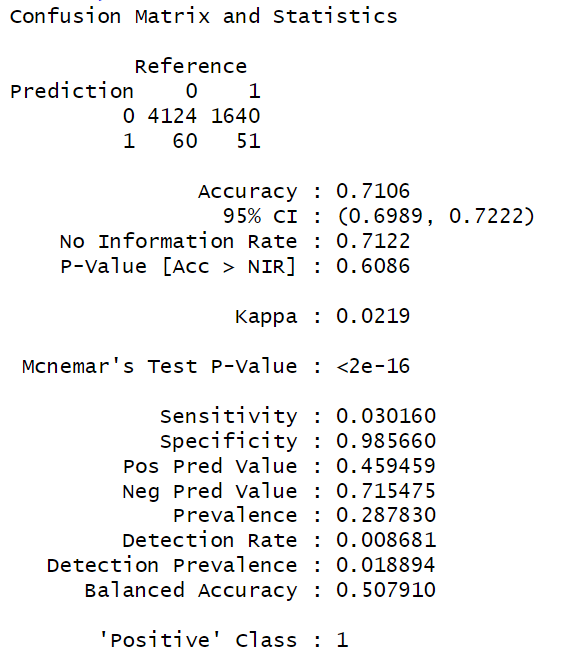
Description automatically generated

Model 3 excludes the last variable: roaming calls

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Sensitivity | Specificity | AUC |
| 71% | 3% | 98% | 57 |

A graph with numbers and letters

Description automatically generated with medium confidence

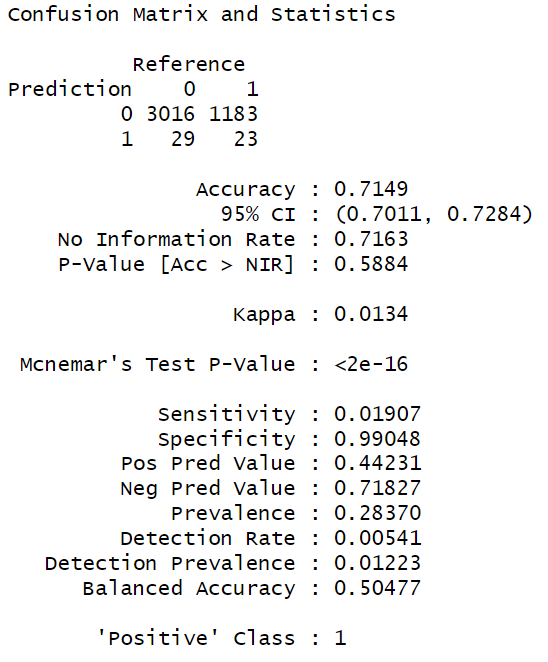


A computer screen shot of a code

Description automatically generated

4. Logistic Regression

Model 1 with all predictors



Model 2

A graph of a curve

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Model 2 excludes:

A screenshot of a computer

Description automatically generated

A graph of a curve

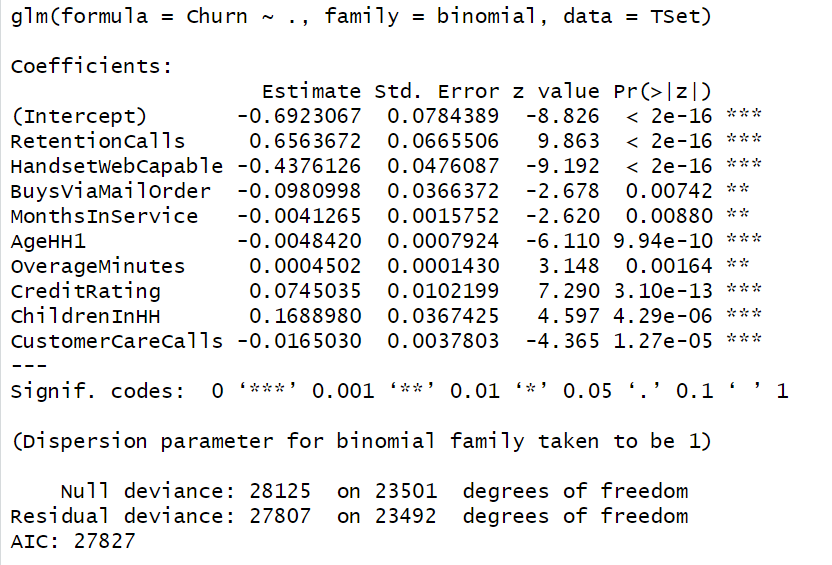
Description automatically generated

A screenshot of a computer

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Sensitivity | Specificity | AUC |
| 70% | 1% | 99% | 58 |

Model 3: 9 most significant predictors



A graph of a curve

Description automatically generated

A screenshot of a computer

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Sensitivity | Specificity | AUC |
| 70 | 1% | 99% | 58 |

5. Performance Comparison of Resampling Methods

Random Forest (without data scaling):

A screenshot of a computer

Description automatically generated

A group of graphs showing different types of data

Description automatically generated

A screenshot of a computer

Description automatically generated

Random Forest (with data scaling):

A screenshot of a graph

Description automatically generated

A group of graphs showing different types of trees

Description automatically generated

Logistic with scaling

A screenshot of a computer

Description automatically generated

Logistic without scaling

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

6. Screenshot of the working webapp

A screenshot of a computer

Description automatically generated