AAMD – Term Project

Telecom Customer Churn Prediction

Group 7:

Hafsa Ghuyoor Syed 19262

Javeria Malik 19292

Maryam Fatima 18817

Sheza Malik 18715

Submitted to:

Hassan Khalid

Spring 2023

Contents

Executive Summary 1
Introduction 1
Data Description 1
Assumptions 2
Problem Statement 2
Methodology 2
Data Import 3
Data Cleaning 3
Modeling 3
Logistic Regression 3
Customer Segmentation 4
Model Performance 4
Profitability 4
Results 6
Ideal Thresholds 6
Customer Segmentation 6
ROC Curve 7
PR Curve 8
Policy Recommendations 9
Limitations 10
Future Scope 10
References 11
Appendix 12
Exhibit A 12
Exhibit B 12
Exhibit C 13
Exhibit D 13
Exhibit E 14
Exhibit F 14
Exhibit G 15
Exhibit H 15
Exhibit I 16
Exhibit J 16
Exhibit K 17
Contribution Sheet 18

Executive Summary

This project focused on addressing customer churn in a telecom company by developing a logistic regression model and implementing customer segmentation using a decision tree. The logistic regression model successfully predicted customer churn, allowing the company to take proactive measures to retain customers and reduce churn rates. By analyzing customer data and identifying key characteristics associated with churn, customer segmentation provided valuable insights for targeted marketing strategies and personalized customer experiences.

Furthermore, the project evaluated the performance of the logistic regression model and decision tree predictions through various metrics, including the calculation of expected profit for churners. This analysis provided a comprehensive understanding of the effectiveness of the models and their potential impact on the company's profitability.

The findings of this project emphasize the importance of proactive customer retention strategies based on the logistic regression model's predictions. By identifying customers at high risk of churn, the company can implement personalized offers, improved customer support, and tailored marketing campaigns to enhance customer loyalty and reduce churn rates.

Additionally, the customer segmentation derived from the decision tree analysis enables the company to understand different customer groups and their unique characteristics. This information can guide the development of targeted marketing initiatives, customized service offerings, and tailored communication strategies to meet the specific needs and preferences of each customer segment.

Overall, this project provides actionable insights and recommendations for the telecom company to mitigate customer churn, improve customer satisfaction, and drive business growth. By leveraging the predictive power of the logistic regression model and customer segmentation, the company can optimize its resources and strategies to retain valuable customers, enhance profitability, and foster long-term customer relationships.

Introduction

In this project, we have a telecom firm that has gathered comprehensive data on its customers, consisting of 7042 data points and 21 predictor variables. The main objective is to predict whether a particular customer will switch to another telecom provider or not, using logistic regression. Additionally, we aim to conduct research on the potential pricing and cost factors associated with customer decisions, which will enable us to calculate the profitability of different strategies. The ultimate goal is to optimize profitability by identifying an optimal threshold that maximizes profit. By undertaking this project, we aim to provide valuable insights and actionable recommendations to the telecom firm, facilitating effective decision-making and customer retention strategies.

Data Description

The dataset contains 7,042 rows and 21 columns. The target variable is "Churn," which is a binary variable indicating whether a customer has churned or not. The remaining 20 variables are independent. The data was available in 3 parts, these were Customer Data, Internet Data, Churn Data. This dataset was downloaded from Kaggle.com, and contains the following columns:

Column Name	Description
customerID	Unique customer ID
tenure	Number of periods customers have existed
PhoneService	Whether or not a customer had phone service
Contract	Type of contract, 3 levels
PaperlessBilling	Whether or not a customer has paperless billing
PaymentMethod	Type of payment method, 4 levels
MonthlyCharges	Monthly charges for the customer
TotalCharges	Total charges for the customer
Churn	Whether or not a customer churned
gender	Gender of the customer, female or male
SeniorCitizen	Whether or not a customer classified as a senior citizen
Partners	Whether or not a customer had partners
Dependents	Whether or not a customer has dependents
MultipleLines	Whether or not a customer has multiple lines
InternetService	Type of internet service, 3 levels (includes no internet service)
OnlineSecurity	Whether or not a customer had online security
OnlineBackup	Whether or not a customer had online backup
DeviceProtection	Whether or not a customer had device protection
TechSupport	Whether or not a customer had tech support
StreamingTV	Whether or not a customer had TV streaming
StreamingMovies	Whether or not a customer had movie streaming

Assumptions

In order to measure the model's profitability an assumption is made that the company will try to retain those likely to churn by giving them free internet for a month. Internet deals start with approximately $55/month (refer to Exhibit A)[1]. An estimate of what it might cost a US based service provider is 50% of this ($27.5) and to account for any other charges, we kept the cost at $30 (CPOC).

The number of people expected to stay or not churn after giving the offer is taken to be 20%.

The average customer lifetime value (ALTV) was calculated by taking an average of the LTV of customers who have actually churned (churn = 1).

Problem Statement

The objective of this project is to build a logistic regression model that predicts whether a customer of a telecom firm will churn or not. By analyzing the past information of customers, the model will predict whether a particular customer will switch to another telecom provider or not. The model's profitability will also be assessed. Furthermore, customers will be divided into segments to identify customer cohorts that are more likely to churn.

Methodology

The following explains the steps taken in order to clean the data, get it ready for logistic regression, running the logistic regression, finding customer segments, assessing the performance of the logistic regression and decision tree models using ROC and PR curves and finally calculating the profitability

[1] https://www.allconnect.com/providers/att

of the logistic regression model used to decide which customers to try to retain with a free internet offer.

Data Import

The data was imported as three separate datasets, Customer Data, Internet data and Churn data. First these datasets had to be merged, this was done using the merge command and the customerId column which was present in all datasets.

Data Cleaning

The data was checked for NA values, and it was found that there were 11 NA values, all present in the TotalCharges column. These missing values were replaced by the median of the TotalCharges column (without the NAs). After making a boxplot (refer to Exhibit B) we saw that this column has outliers, hence using the mean would not give a fairly accurate picture of the data and median is used instead[2].

Modeling

The variables that appeared as character variables were converted to factor variables in order to transform qualitative or categorical data into a suitable format for statistical modeling. Then these variables were recoded to be **binary variables**, where “Yes” = 1 and “No” = 0. Some variables that were dependent on the InternetService column had 3 levels, here “No internet service” = 0.

Next variables representing different types of the same thing (e.g., types of contracts, Month-to-month, One year, Two year) were broken up with each level (except one, to avoid multicollinearity) being converted into a binary variable in a process called **one-hot encoding**. By leaving out one level, we establish a baseline or reference category against which the other levels are compared. The coefficients associated with the other levels represent the difference in relation to the reference level. These newly created variables were converted into factor variables and then the redundant old columns containing the initial categories were deleted from the dataset.

The data was then assigned random integers and then sorted by the integers assigned. This data was then saved as csv (leaving out the random integer column), for it to be used in Excel to find the optimal threshold after getting predictions from the logistic regression.

Logistic Regression

To run the logistic regression on the model, we first split the data into testing and training data. The training data was taken to be 75% of the data whereas the testing data was 25% of the data. Initially, we trained the model through k-fold cross-validation(k=10) to avoid overfitting in the data.

The logistic regression model was then applied on the training data. The y-variable ‘Churn’ was taken into account and all the other variables were considered to be the x variables. After that, the model underwent the stepwise regression model. Next, we calculated the training and testing predictions using the stepwise model that was trained. The testing and training predictions were then combined into one data frame and the outcome was saved as a csv file, for it to be used in Excel to find the optimal threshold and profitability of the model.

[2] https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/

Customer Segmentation

We ran the Decision Tree Model on the training data to come up with some generalizable customer segments for the telecom company. The Decision Tree Model for customer segmentation was built and interpreted as follows:

1. **Data:**
   The train dataset, containing relevant variables and the Churn variable, is used as the input data for model construction. The Churn variable serves as the dependent/target variable, representing the churn status of customers.
2. **Model Construction:**
   The rpart() function, a decision tree algorithm, is employed to build the decision tree model. The formula Churn~. is specified, indicating that the Churn variable is to be predicted based on all other variables present in the dataset. The control parameter is set using rpart.control(), allowing for the customization of specific control parameters for the decision tree construction. The min bucket parameter is set to 10, defining the minimum number of observations required in a terminal (leaf) node of the tree. The complexity parameter(cp) is set to 0.004, controlling the complexity of the decision tree by adjusting the pruning threshold. Which means, that the model is looking for even small improvements in quality to perform additional splits. A smaller cp value allows for more splits in the tree, potentially leading to a more complex and detailed model.
3. **Decision Tree Analysis:**
   The resulting decision tree structure is obtained, displaying the hierarchy of nodes and splits. Each node represents a decision point based on a specific variable and cutoff value. The numbers within the nodes indicate the number of observations falling into each node. The values in parentheses represent the predicted class probabilities for each node.
4. **Interpretation:**
   In the decision tree, each path represents a set of conditions that classify the observations into different segments or classes. The class probabilities in the leaf nodes indicate the likelihood of belonging to a specific churn category (0 or 1).

Model Performance

To understand the performance of our model, we plot ROC and PR curves for both the training and the testing data. The area under the graphs is considered to determine the accuracy and precision-recall balance of the model.

First, ROC and PR curves were plotted and the area under their curves were considered for the training and testing data of the logistic regression model to assess the model's performance.

Additionally, we trained the decision tree model on the training data and predicted training and testing values through the decision tree model. For the decision tree model, ROC and PR curves were plotted and the area under the curve was considered to compare the decision tree model's performance as opposed to the logistic regression model.

Profitability

Predicted values by the model were brought into Excel so we could compare them to the original churn values. Since the predicted values are in decimals between 0 and 1, a threshold was used to define above which the predicted values will be considered "1" or the customer will be predicted to churn. This was used as an input at 7%.

Then the table was made on the pattern of a confusion matric which had the threshold, the true positive, false positive, true negative, false negative, recall, precision, F-1 Score and lastly profitability.

If the actual churn is 1, and our model predicted also predicted 1 (based on the threshold), then that was taken as a True Positive. Similarly, if Actual = 0, Prediction =1, it was taken as False Positive; if Actual =1, prediction = 0, it was taken as False Negative and lastly, if actual = 0, predication =0, it was taken as True Negative.

Recall, Precision, and f-1 Score was calculated for each level of threshold using the following formulas:

- Recall = TP / (TP + FN)
- Precision = TP / (TP + FP)
- F-1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Furthermore, we calculated the profitability using the Average Lifetime Value (ALTV) and estimated cost of giving a free service to a customer predicted to churn, to retain them. Following treatments were used:

- True Negative: (customers our model correctly predicted wouldn't churn) Since we continue to possess their LTV, the profit from them was taken as product of ALTV and number of TN customers.
- False Positive: (customers our model incorrectly predicted would churn, but in actual they didn't) we continue to possess their LTV as well, but because of our incorrect prediction, we spent the CPOC ($30) to retain them. The profitability from them would be the ALTV less the CPOC.
- False Negative: (customer our model incorrectly predicted would not churn, but in actual they did). Since our model predicted they wouldn't churn, we didn't spend any money to retain them. However, when they churned, we lost their LTV. So, the profitability was the loss of their ALTV.
- True Positive: (customers our model correctly predicted would churn). Since we expected them to churn, we assumed that the company would take an action to prevent these customers from churning, such as free internet. In this example we assumed the company would give 5 GBs of internet for free. According to our estimates stated in the assumptions, this cost is assumed to be $30. We also assumed that because of this offer, 20% of the potential churners decided to continue the service and didn't churn. The profit would therefore be:
    - For the 20% of retained customers: (#TP x 20%) x (ALTV – CPOC)
    - For the 80% that still churned, a loss equal to CPOC (#TP x 80%) X CPOC (for ease of calculation, the nod of TP * retained customers was multiplied by the ALTV, and then the product of #TP and CPOC was deducted from it)

Maximizing Profitability

- Conditional formatting was applied to the profit's column, to find which threshold gave us the maximum profit.
- This came up to $ 6274026.752, at a threshold of 0.07 or 7%.

Results

Ideal Thresholds

Training Data

Maximizing Profit at a Churn Probability Threshold

When the probability of churning surpasses 7%, a customer is considered to churn. At this level, the profit reaches its peak, amounting to $6,273,997. We use the threshold of 7% as our priority as business is maximizing profit.

Optimizing F-1 Score at a Churn Probability Threshold

When the probability of churning exceeds 33%, a customer is categorized as a churner. This threshold ensures that both Precision and Recall are optimized, resulting in a maximized F-1 score.

Different Threshold Selections: F-1 Score vs. Profit Maximization

The reason for selecting different thresholds for maximizing the F-1 score and profit is due to the varying costs associated with false negatives (failure to correctly identify a churner) and false positives. The cost of a false negative ($1,531.80) is higher than that of a false positive ($30), hence the distinct threshold choices.

Testing Data

Maximizing Profit at a Churn Probability Threshold

When the probability of churning surpasses 10%, a customer is considered to churn. At this level, the profit reaches its peak, amounting to $ 2,113,223. We use the threshold of 7% as our priority as business is maximizing profit.

Optimizing F-1 Score at a Churn Probability Threshold

When the probability of churning exceeds 27%, a customer is categorized as a churner. This threshold ensures that both Precision and Recall are optimized, resulting in a maximized F-1 score.

Model Performance Variation on Unseen Data

This discrepancy suggests that the model's performance may vary when applied to unseen data. Therefore, the model's ability to generalize and accurately predict churners may differ in real-world scenarios.

Customer Segmentation

Based on the Decision Tree (Exhibit C), we segmented the customers of the telecom company into two groups namely, churners and non-churners. The following are the characteristics of both segments.

Churners:

1. **Price-Sensitive Churners:** Customers with a tenure of less than 17.5 periods and a monthly charge of greater than or equal to $68.62. This group of churners is more sensitive to pricing and may be inclined to switch to competitors offering more affordable plans or better deals.
2. **Digitally Disconnected Churners:** Customers without DSL internet service and lacking technical support. These churners heavily rely on technology and are more likely to switch

providers if they encounter issues with their internet service or lack proper technical support.

3. **Unsettled High-Value Churners:** Customers with shorter tenure and higher monthly charges when DSL internet service is unavailable. This category comprises customers who are considered valuable due to their higher monthly charges, but their shorter tenure and dissatisfaction with the lack of DSL internet service make them more prone to churn.

Non-Churners:

1. **Loyal Customers:** Customers with a longer tenure (17.5 periods or more) are more likely to be non-churners. This category represents customers who have established stability and loyalty to the company, indicating their inclination to stay and continue their relationship.
2. **Tech-Supported DSL Subscribers**: Among non-churners with DSL internet service, those with technical support are less likely to churn compared to those without technical support. This category comprises customers who not only have DSL internet service but also benefit from reliable technical support, which enhances their overall satisfaction and reduces the likelihood of switching providers.

By using these characteristics, the telecom company can identify and target specific customer segments. For example, they can focus retention efforts on customers with shorter tenures, higher monthly charges, and no DSL internet service, as they are more likely to churn. They can also prioritize providing technical support to non-churners with DSL internet service to further reduce the churn rate.

The tree structure shows that all leaf nodes have more than 10 observations, indicating that the min bucket value of 10 did not affect the splitting process. This suggests that no nodes reached the minimum size threshold, and all nodes were able to split further.

The specified cp value of 0.004 influenced the decision tree's splitting decisions, as evidenced by the branching and splitting of nodes based on the provided conditions.

ROC Curve

Logistic Regression

Training Data

The ROC Curve as shown in Exhibit D, evaluates the performance of a customer churn prediction model. It shows the trade-off between correctly identifying churners (sensitivity) and incorrectly classifying non-churners as churners (false positives) at different classification thresholds. The area under the ROC curve (AUC-ROC) summarizes the model's performance, with a higher score indicating better discrimination between churners and non-churners.

Our AUC-ROC is 0.85, hence, the model demonstrates a high ability to discriminate between churners and non-churners. This means that when the model is applied to a set of customers, it can correctly rank the majority of churners higher than non-churners.

Testing Data

As can be seen in Exhibit E, the AUC-ROC for testing data is slightly less than the AUC-ROC for Training Data (0.83). The slightly lower AUC-ROC score indicates that the model's performance might have slightly decreased when applied to new, unseen data. This difference could be attributed to the presence of variations or differences between the training and testing datasets. Nevertheless, an AUC-ROC score of 0.83 on the testing data still indicates a reasonably good performance in predicting customer churn. It implies that the model retains a significant ability to discriminate between churners and non-churners in real-world scenarios, although it might not be as strong as observed during training.

Decision Tree

Training Data

As shown in Exhibit F, the training AUC-ROC score of 0.712 for a decision tree model indicates its performance in distinguishing between churners and non-churners on the training dataset. It suggests that the decision tree can discriminate between the two classes of churners and non-churners better than randomly guessing (which would mean AUC-ROC of 0.5)

Testing Data

As shown in Exhibit G, the AUC-ROC on testing data is lower than training (0.679). It indicates the model's ability to generalize to new instances beyond the training data. Since the testing AUC-ROC is lower than the training AUC-ROC, it suggests that the decision tree model may not generalize as well to unseen data. However, it is still better than random guessing, since it is above 0.5.

PR Curve

Logistic Regression

Training data

The PR curve, as shown in Exhibit H, evaluates the performance of a customer churn prediction model in terms of precision and recall. It showcases the trade-off between accurately identifying churners (precision) and correctly capturing all actual churners (recall) at different classification thresholds. In this case, the PR curve has an area under the curve (AUC-PR) of 0.664. This value summarizes the model's overall performance in terms of precision and recall. The AUC-PR of 0.664 suggests that the model can correctly prioritize a significant portion of churners over non-churners when applied to a set of customers.

Testing data

The AUC-PR value of 0.6312744, as shown in Exhibit I, for the logistic regression model's PR curve indicates a moderate level of performance in terms of precision and recall. The curve demonstrates that the logistic regression model is able to reasonably prioritize churners over non-churners and capture a significant proportion of the actual churners across different classification thresholds.

While the AUC-PR value is not exceptionally high, it suggests that the logistic regression model has some capability in accurately identifying churners and capturing a meaningful portion of the actual churners.

Decision Tree

Training Data

The AUC-PR value, as shown in Exhibit J, is 0.5197845 for the decision tree PR curve suggests that the model's performance in terms of precision and recall is relatively low. The curve indicates that the model struggles to strike a balance between correctly identifying churners (precision) and capturing all actual churners (recall) across different classification thresholds.

A low AUC-PR value indicates that the decision tree model may not be effective in accurately prioritizing churners over non-churners or capturing a significant proportion of the actual churners. This could potentially lead to misclassification of churners as non-churners or missing out on a large number of actual churners.

Testing Data

The AUC-PR value of 0.6312744 for the decision tree, as shown in Exhibit K, PR curve indicates a moderate level of performance in terms of precision and recall. The curve suggests that the decision tree model can reasonably prioritize churners over non-churners and capture a considerable proportion of the actual churners across different classification thresholds.

While the AUC-PR value is not exceptionally high, it suggests that the decision tree model shows some capability in accurately identifying churners and capturing a significant portion of the actual churners.

Policy Recommendations

Based on the provided decision tree model, we can derive some policy recommendations for the telecom company to address customer churn:

- Implement targeted pricing strategies and promotional offers for customers who are price sensitive. This includes customers with a tenure of less than 17.5 periods and higher monthly charges. By offering competitive pricing and attractive promotions, the company can encourage these customers to stay and reduce churn rates.
- Improve DSL internet service availability and reliability for customers who currently lack DSL internet service and technical support. By addressing their connectivity issues and providing efficient technical assistance, the company can enhance the overall customer experience and reduce the likelihood of churn among this segment.
- Focus on personalized retention efforts for high-value customers with shorter tenure and higher monthly charges, particularly when DSL internet service is unavailable. Understanding their specific needs and concerns will allow the company to tailor retention strategies, such as offering alternative solutions, providing additional benefits, or resolving any dissatisfaction they may have.
- Continuously monitor customer satisfaction and feedback to identify pain points and areas for improvement. By actively listening to customer feedback and addressing their concerns promptly, the company can enhance overall customer satisfaction and loyalty.
- Invest in targeted marketing campaigns to educate customers about the benefits and value-added services offered by the company. By highlighting unique features, superior customer service, and exclusive offers, the company can reinforce its value proposition and strengthen customer loyalty.

Limitations

The limitations of this project include:

- The model assumes that the independent variables are not highly correlated with each other. High multicollinearity can lead to unstable estimates.
- We are not aware of the pricing of the telecom products offered to the customers, which has an impact on the accuracy of the profitability of the model.
- The handling of missing values in the TotalCharges by replacing them with the median might introduce some level of imputation bias.
- There was some level of ambiguity in the model ie the variable 'tenure' was not well defined. We were unsure if these periods were given in years or months or some other measure. If we were well aware of what the tenure signified, we might have been able to interpret the model and the findings in a more accurate way.
- The model does not account for external factors or other events that can have a significant influence on y-variable (Churn) but are not included in the given dataset. These factors could include industry trends, competitor actions, economic conditions, or other circumstances that impact customer behavior and decisions in the Telecom industry.

Future Scope

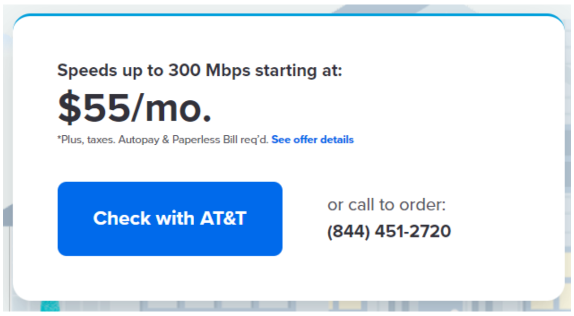
Moving forward, there are several areas for future development and enhancement in this project. Advanced predictive analytics techniques, such as random forests or neural networks, can be explored to improve churn prediction accuracy. Real-time churn monitoring, by taking the above segmentation to decide if a particular customer has the likeliness of churn or not based on the services they opt for, can be implemented to proactively identify at-risk customers, and take immediate action. Refining customer segmentation by incorporating additional variables or advanced clustering techniques can enable more targeted retention strategies. Conducting a more comprehensive analysis of customer lifetime value can provide insights into profitability and help prioritize retention efforts. Integration of social media sentiment analysis can aid in addressing customer concerns and improving satisfaction. Enhancing the customer experience across touchpoints and incorporating collaborative filtering and recommender systems can drive engagement and cross-selling opportunities. Establishing a feedback loop and integrating churn prediction with CRM systems ensures ongoing improvement and targeted marketing campaigns. These enhancements will enable proactive churn prevention, personalized engagement, and revenue maximization.

References

- Kaggle. (n.d.). LogisticRegression - Telecom Customer Churn Prediction. Retrieved from https://www.kaggle.com/datasets/dileep070/logisticregression-telecomcustomer-churmpredictio
- Allconnect. (n.d.). AT&T. Retrieved from https://www.allconnect.com/providers/att
- Vitalflux. (n.d.). Pandas - Impute Missing Values (Mean, Median, Mode). Retrieved from https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/

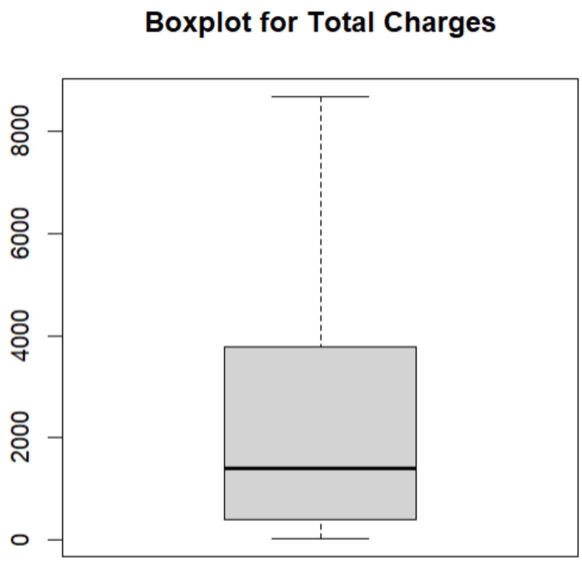
Appendix

Exhibit A



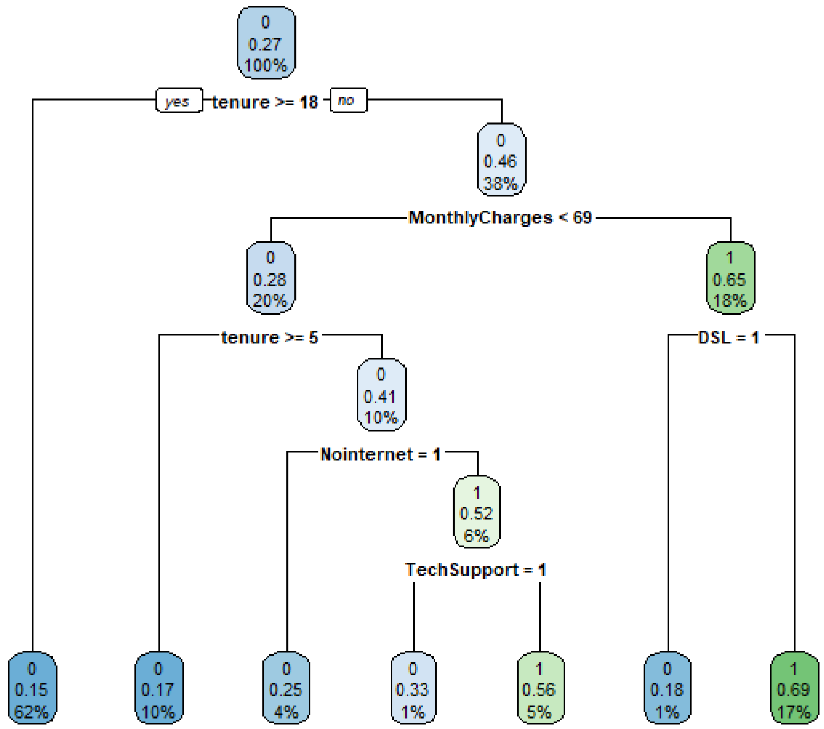
(Click here to go back)

Exhibit B



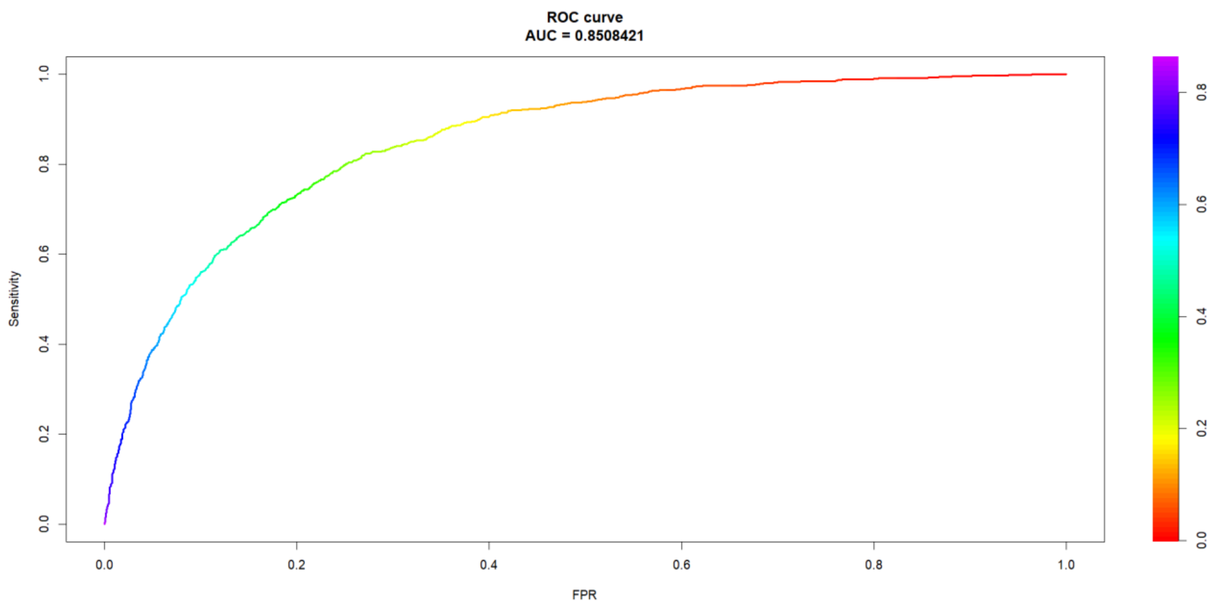
(Click here to go back)

Exhibit C



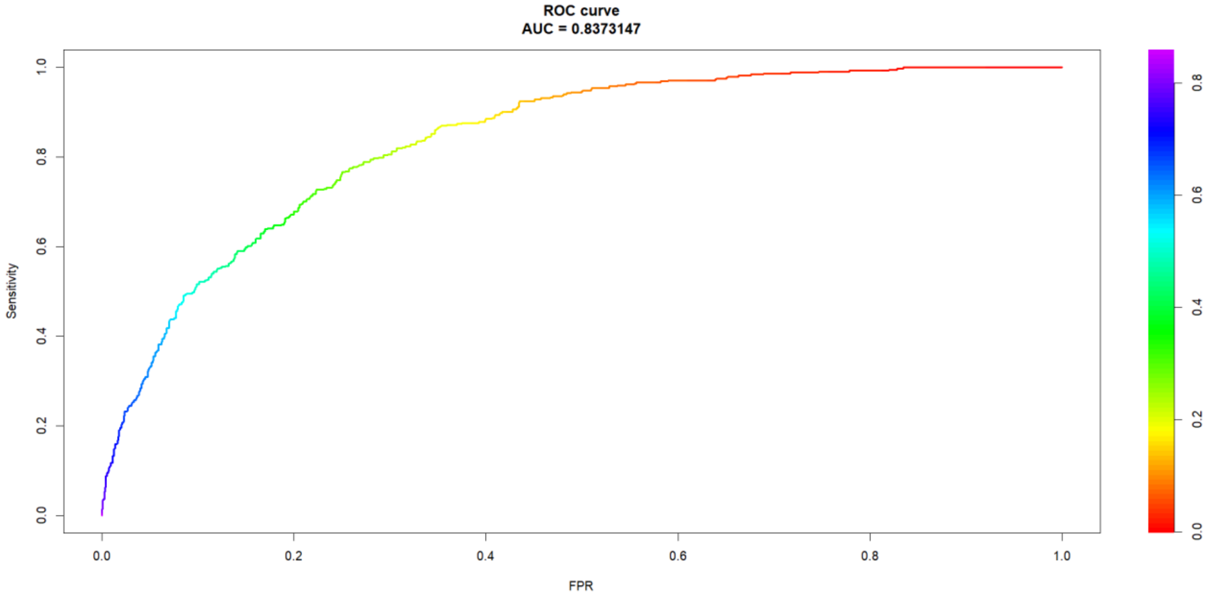
(Click here to go back)

Exhibit D



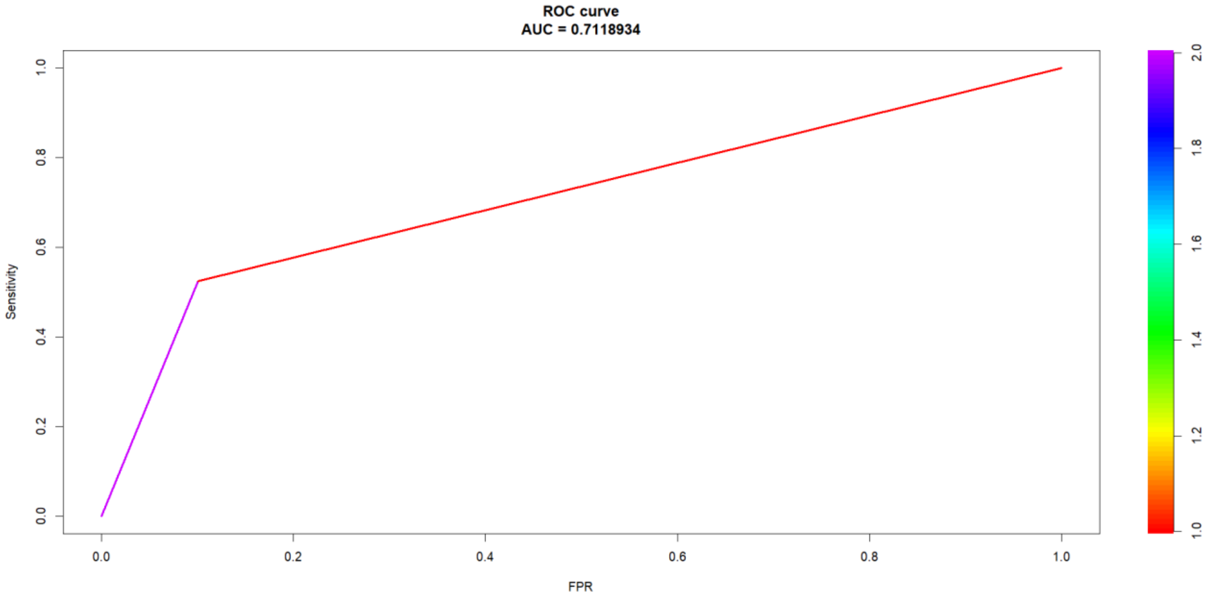
(Click here to go back)

Exhibit E



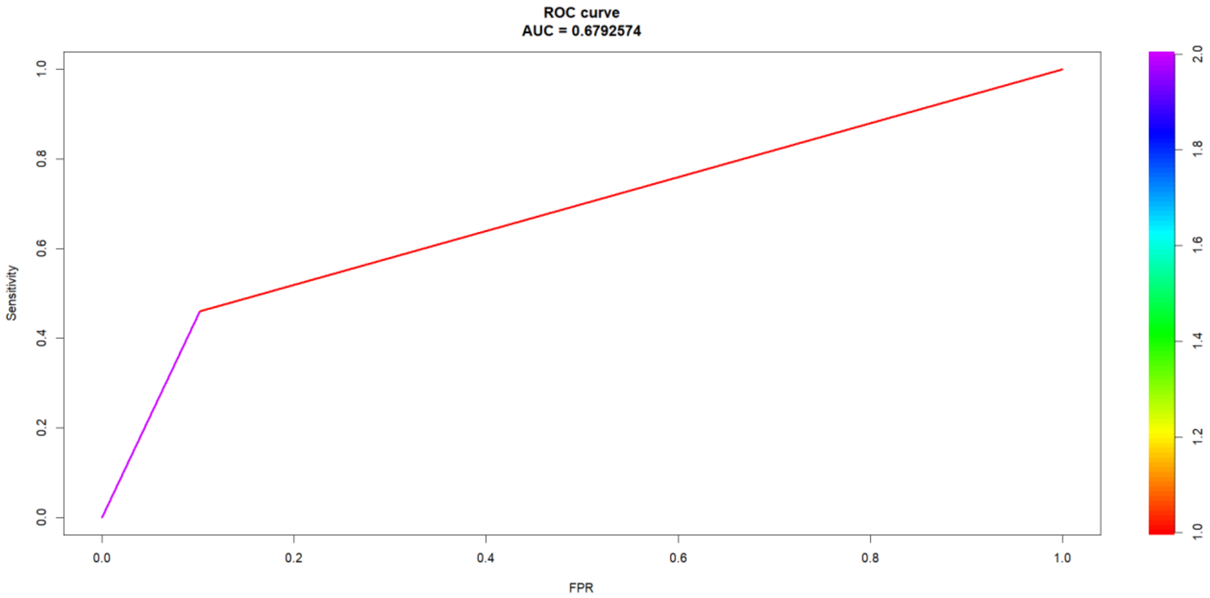
(Click here to go back)

Exhibit F



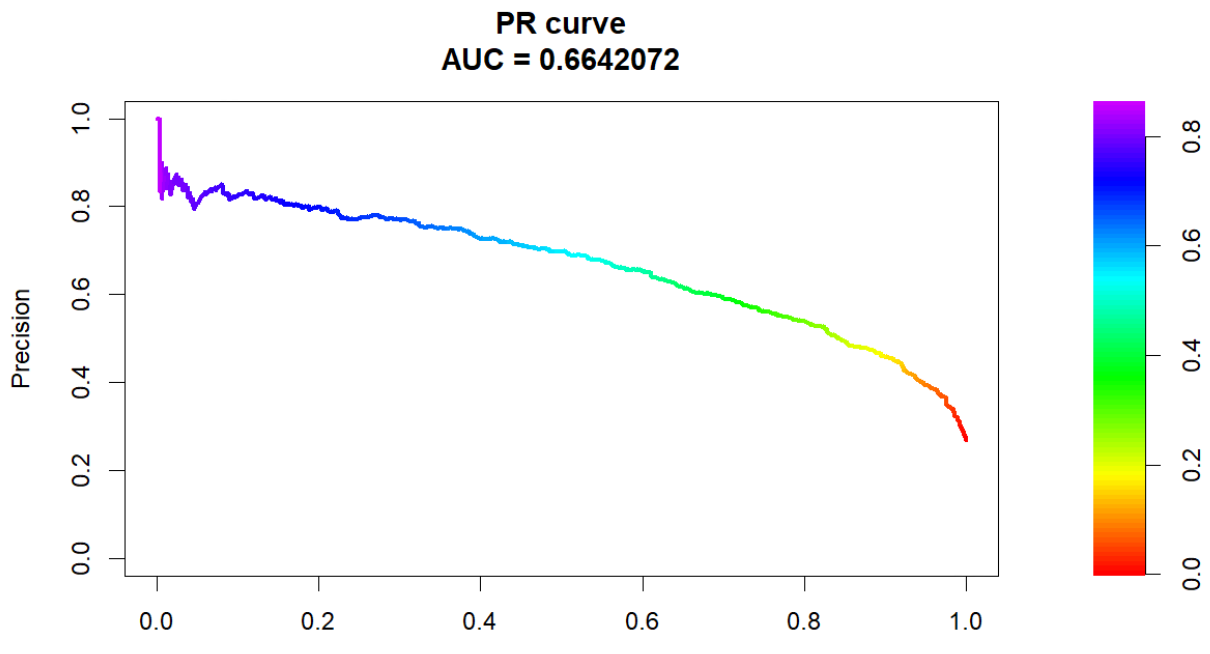
(Click here to go back)

Exhibit G



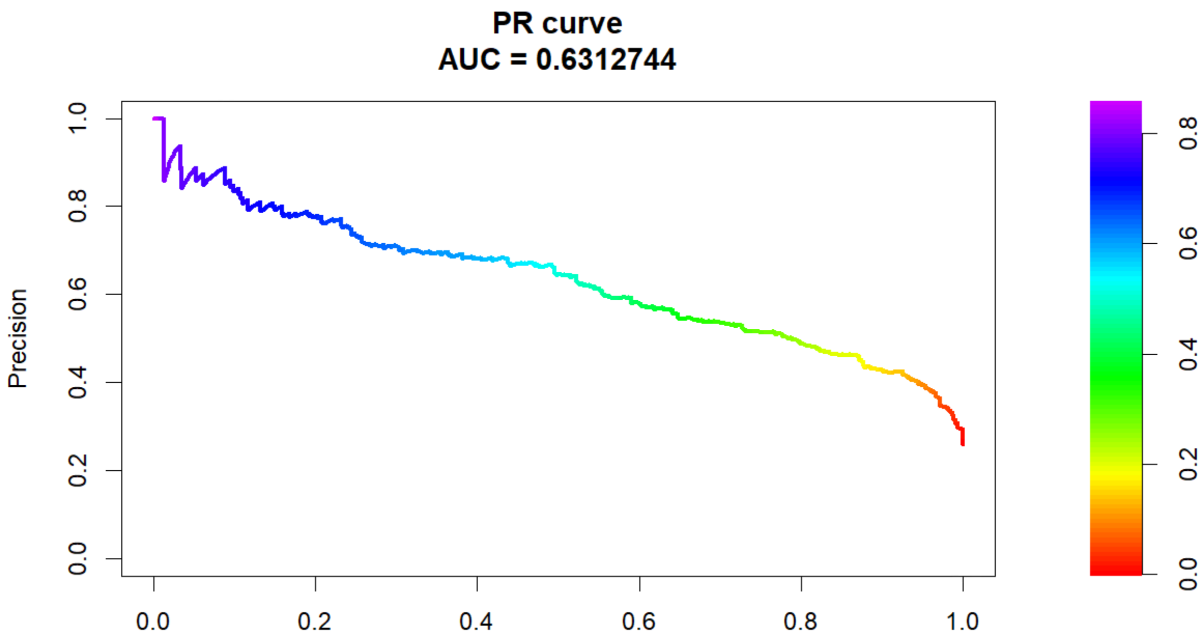
(Click here to go back)

Exhibit H



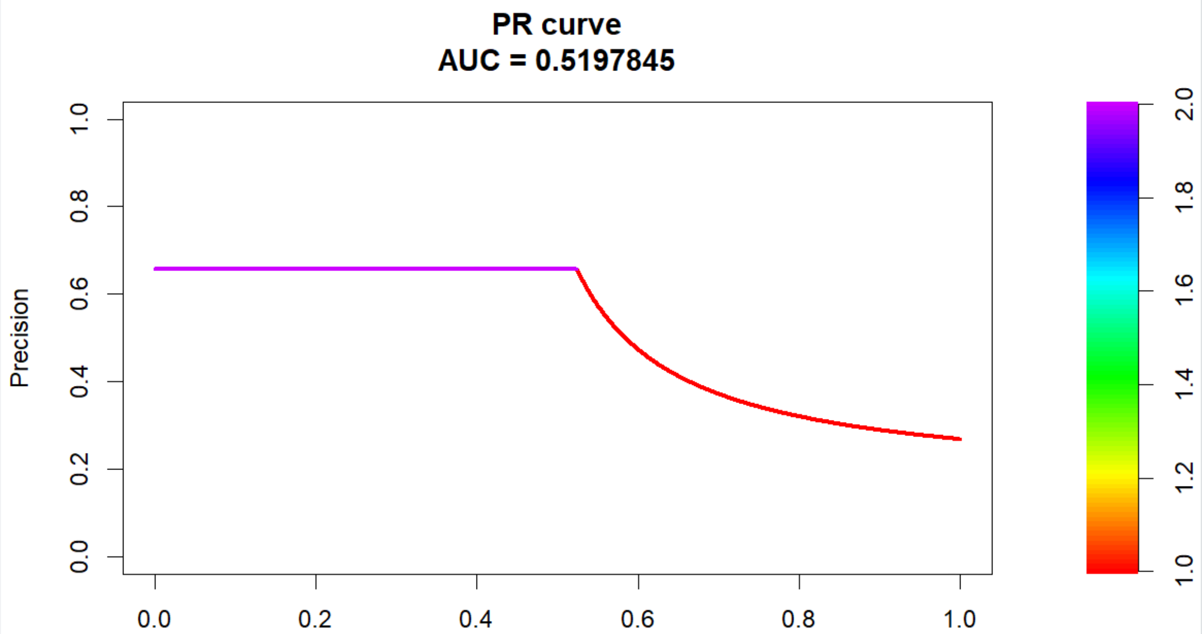
(Click here to go back)

Exhibit I



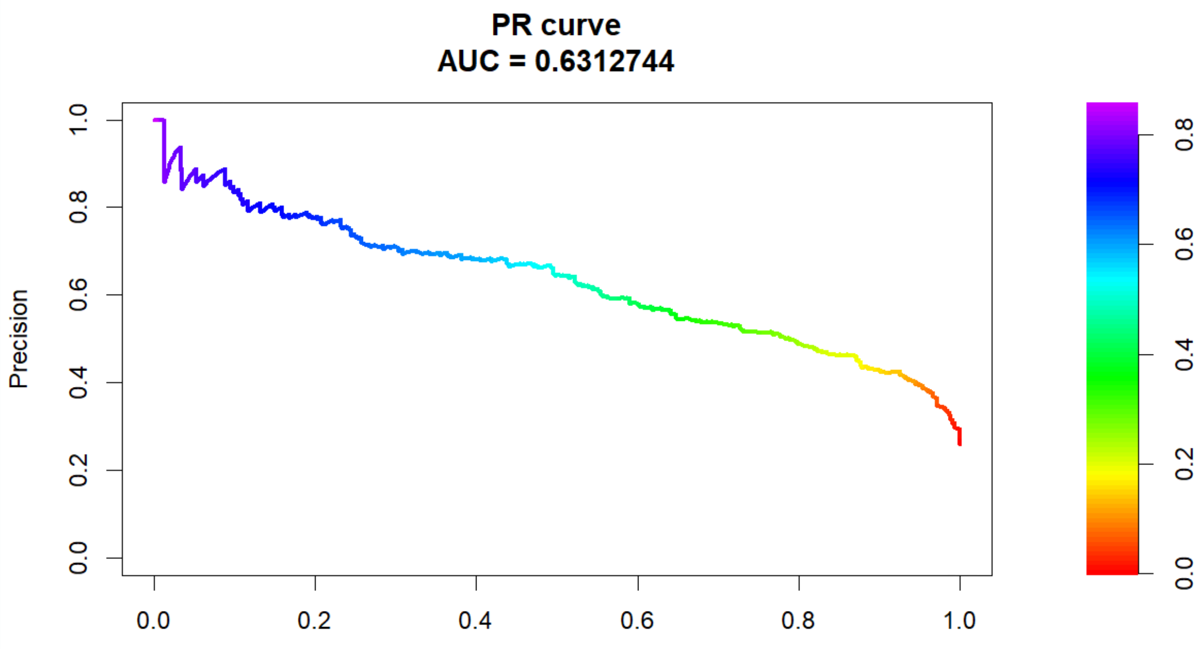
(Click here to go back)

Exhibit J



(Click here to go back)

Exhibit K



(Click here to go back)

Contribution Sheet

Member	Contribution
Hafsa Ghuyoor Syed	Data importation and merging, Data cleaning, Data Imputation, Binary and one-hot encoding, getting predictions from the model, Confusion Matrix, Recall, Precision, F-1 Score, and Profitability, Customer Segmentation using Decision Tree, Introduction, Data Description, Assumptions, Problem statement, Interpretation of Threshold, Explanation of Data Cleaning and Modeling
Javeria Malik	Randomizing and splitting training and testing data, Construction of Logistic Regression model, Step wise regression, K-fold CV, getting predictions from the model, Confusion Matrix, Recall, Precision, F-1 Score, and Profitability, Customer Segmentation using Decision Tree, Making ROC and PR curves for Logistic Regression Model & Decision Tree, explanation of Logistic Regression Limitations
Maryam Fatima	Binary and one-hot encoding, getting predictions from the model, Confusion Matrix, Recall, Precision, F-1 Score, and Profitability, Customer Segmentation using Decision Tree, Exec. Summary, Explanation of Segmentation, Explanation of PR curves, Policy Recommendations, Future Scope,
Sheza Malik	Data Cleaning, getting predictions from the model, Confusion Matrix N, Recall, Precision, F-1 Score, and Profitability, Customer Segmentation using Decision Tree, Assumptions, Explanation of ROC Curves f Logistic regression and Decision Tree, Explanation of Calculation of Profitability