CUSTOMER CHRN PREDICTION MODEL & ANALAYIS FOR A TELECOM COMPANY

PROJECT REPORT

Thant Thiri Myo Kyi
Springboard Data Science Bootcamp (Jul 2023 Cohort)
March 2024.

Table of Contents

1 Introduction

In the fiercely competitive telecommunications industry, companies are continuously striving to maintain and expand their customer base. However, one of the most significant challenges they face is customer churn, where customers cease their relationship with the company and move to a competitor. This phenomenon not only directly affects the company's revenue but also leads to increased expenses related to acquiring new customers, which is often more costly than retaining existing ones. In this context, the ability to predict customer churn becomes crucial for telecom companies to implement effective retention strategies.

1.1 Background

The telecom industry is characterized by rapid technological advancements and a high level of market saturation, leading to intense competition among service providers. In this environment, customer loyalty is volatile, and companies must go above and beyond to keep their customers engaged. Recognizing the patterns and behaviours that indicate a customer's likelihood to churn is essential for taking proactive steps towards retention. Utilizing data science and analytics, telecom companies can sift through vast amounts of customer data to identify these patterns, providing a foundation for targeted retention strategies.

1.2 Problem Statement

The primary challenge addressed in this project is the development of a predictive model capable of forecasting customer churn for a telecom company. By analysing various factors with customer demographics, service usage behavioural patterns, and previous churn rates, the model aims to identify customers who are at risk of leaving. This predictive capability is crucial for enabling the company to take timely and effective action to retain these customers, thus reducing churn rates and enhancing customer satisfaction and loyalty.

1.3 Goal

The goal of this project is to create a robust and reliable predictive model that can accurately identify potential churn customers, allowing the telecom company to focus its retention efforts more efficiently. Success will be measured by the accuracy and reliability of the model, as well as the effectiveness of the retention strategies that are implemented based on the model's insights. Ultimately, the project aims to contribute to a reduction in customer churn, leading to improved revenue stability and enhanced customer loyalty for the telecom company.

By leveraging data science techniques and analytics, this project seeks to provide actionable insights that directly impact the company's customer retention strategies. Through a detailed analysis of customer data and behaviour patterns, along with the development of a predictive churn model, the project endeavours to equip the telecom company with the tools necessary for making informed decisions to mitigate customer churn and foster sustainable business growth.

2 Data Acquisition & Cleaning

2.1 Data Collection

The dataset used was found on Kaggle. The data contains 51047 records of telecom customers. The following variables are found in the dataset:

1. CustomerID
2. MonthlyRevenue
3. MonthlyMinutes
4. TotalRecurringCharge
5. DirectorAssistedCalls
6. OverageMinutes
7. RoamingCalls
8. PercChangeMinutes
9. PercChangeRevenues
10. DroppedCalls
11. BlockedCalls
12. UnansweredCalls
13. CustomerCareCalls
14. ThreewayCalls
15. ReceivedCalls
16. OutboundCalls
17. InboundCalls
18. PeakCallsInOut
19. OffPeakCallsInOut
20. DroppedBlockedCalls
21. CallForwardingCalls
22. CallWaitingCalls

23. MonthsInService
24. UniqueSubs
25. ActiveSubs
26. ServiceArea
27. Handsets
28. HandsetModels
29. CurrentEquipmentDays
30. AgeHH1
31. AgeHH2
32. ChildrenInHH
33. HandsetRefurbished
34. HandsetWebCapable
35. TruckOwner
36. RVOwner
37. Homeownership
38. BuysViaMailOrder
39. RespondsToMailOffers
40. OptOutMailings
41. NonUSTravel
42. OwnsComputer
43. HasCreditCard
44. RetentionCalls
45. RetentionOffersAccepted
46. NewCellphoneUser
47. NotNewCellphoneUser
48. ReferralsMadeBySubscriber
49. IncomeGroup
50. OwnsMotorcycle
51. AdjustmentsToCreditRating
52. HandsetPrice
53. MadeCallToRetentionTeam
54. CreditRating
55. PrizmCode
56. Occupation
57. MaritalStatus

2.2 Data Wrangling

The data wrangling process involved thorough cleaning, transformation, and premilitary data exploration, setting a solid foundation for deeper statistical exploration and data-driven insights.
The following steps performed at this stage:

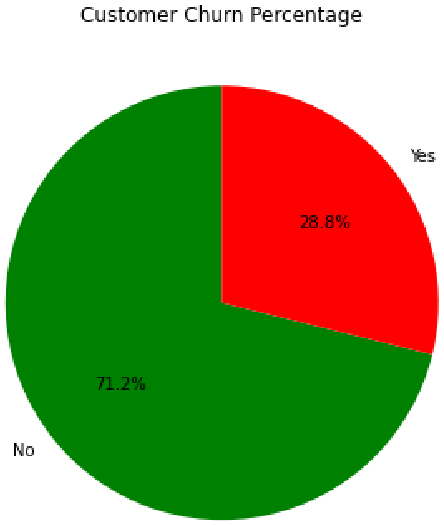
1. Libraries for data manipulation were imported.
2. The dataset was loaded for initial exploration.
3. Missing data were identified, with key missing entries including AgeHH1, AgeHH2, PercChangeMinutes, and PercChangeRevenues, accounting for less than 2% of the data.
4. Duplicate records were checked and handled accordingly.
5. A new categorical column, CityCode, was created from the first three characters of the Service Area, encoded by the IATA standard.
6. Both categorical and numerical columns and their distributions were analysed, identifying outliers and deciding on methods to handle them, such as removal, capping, imputation, or adjustment.Data suggests The presence of outliers can be seen in all selected variables, especially in "MonthlyMinutes" and "OverageMinutes," which show a considerable number of points lying far above the upper quartile, indicating unusually high values compared to the rest of the data. Given the variability in telecom data resulting from diverse customer lifestyles, it is essential to recognize that outliers may vary from one customer to another. Consequently, the decision to remove or impute outliers should be approached with caution and may necessitate careful consideration. With that, no further change required to data.
7. The cleaned data contains 51047 records and 58 columns including index.
8. Finally, the cleaned data is saved as 'cleaned_data.csv' for further statistical analysis in the next step.

3 Exploratory Data Analysis (EDA)

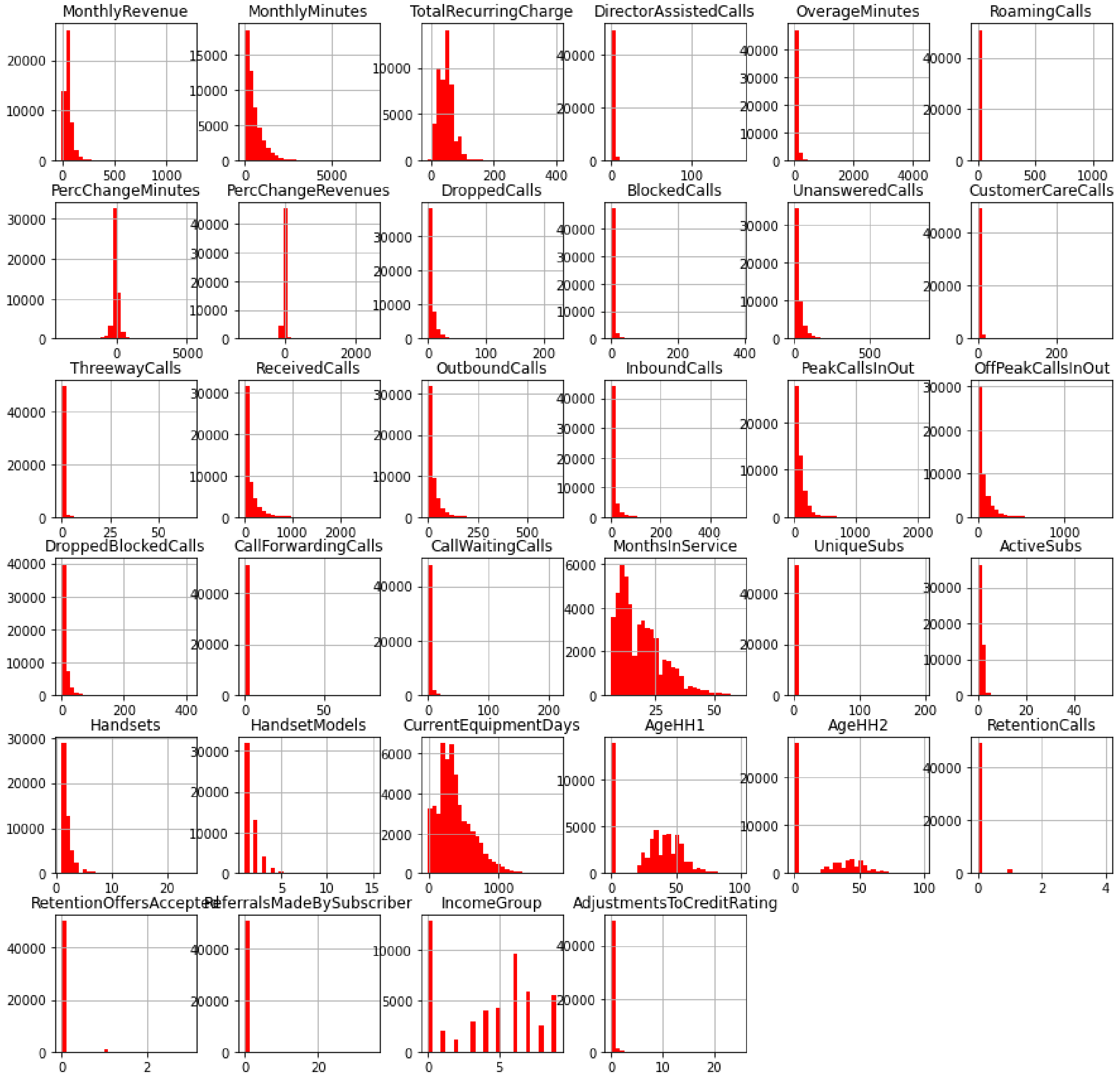
For next step , EDA is performed on the cleaned dataset file , 'cleaned_data.csv', containing customer churn data which comprises 51,047 entries across 58 columns, providing a rich matrix of demographic and behavioural customer data.

3.1 Preliminary Data Examination

Data Exploration started with Churn or Non-Churn based on the whole dataset. And the historical data navigate currently 28.8% of total customer left.



Initial data exploration reveals a diverse range of variables, including customer demographics (e.g., age, income, marital status), account details (e.g., monthly revenue, handset information), and behavioural patterns (e.g., usage metrics, service interactions). A detailed examination of these variables lays the foundation for our in-depth analysis.



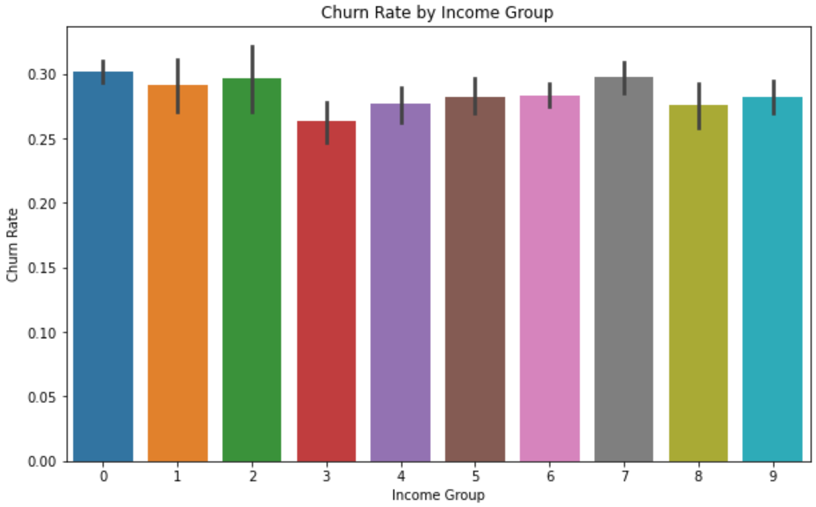
3.2 Customer Analysis

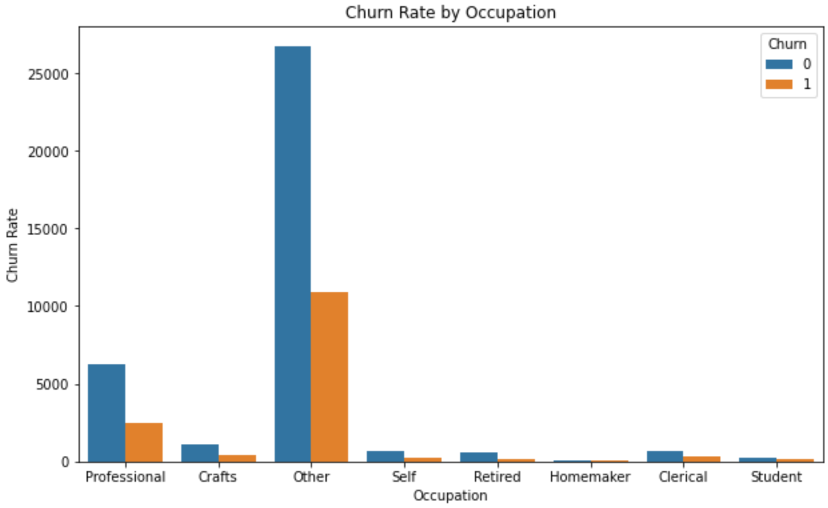
Customer Analysis is structured around two critical lenses: demographic analysis and behavioural analysis, providing a comprehensive understanding of factors driving customer churn.

3.2.1 Demographic Analysis

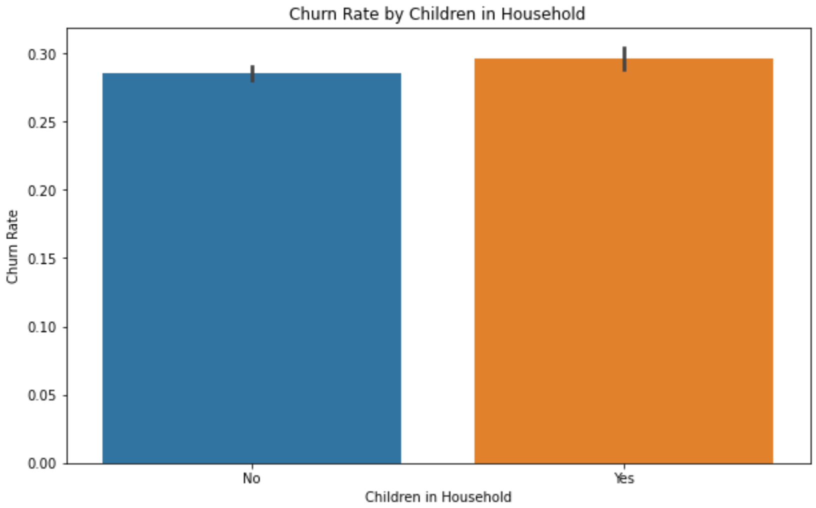
By comparing churn rates across various demographic groups, we unearth several key insights:

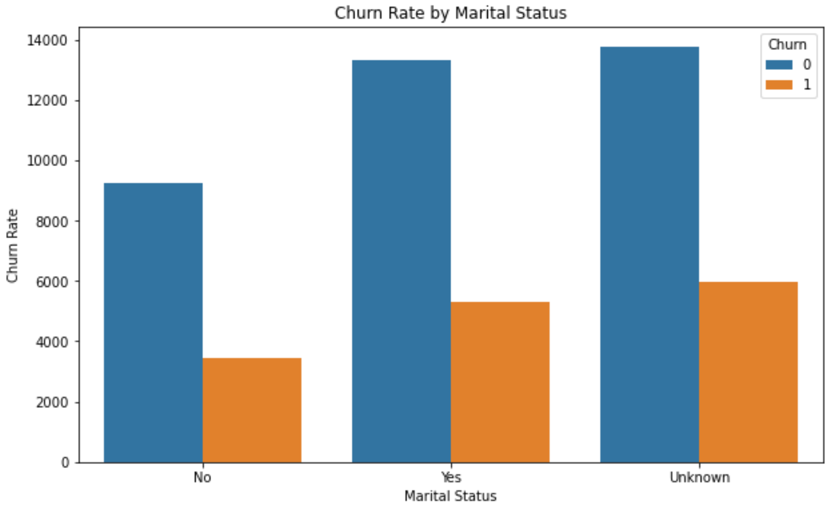
1. Economy and Churn: A nuanced relationship exists, suggesting that certain income and occupation brackets exhibit higher churn rates, highlighting economic factors' role in churn.



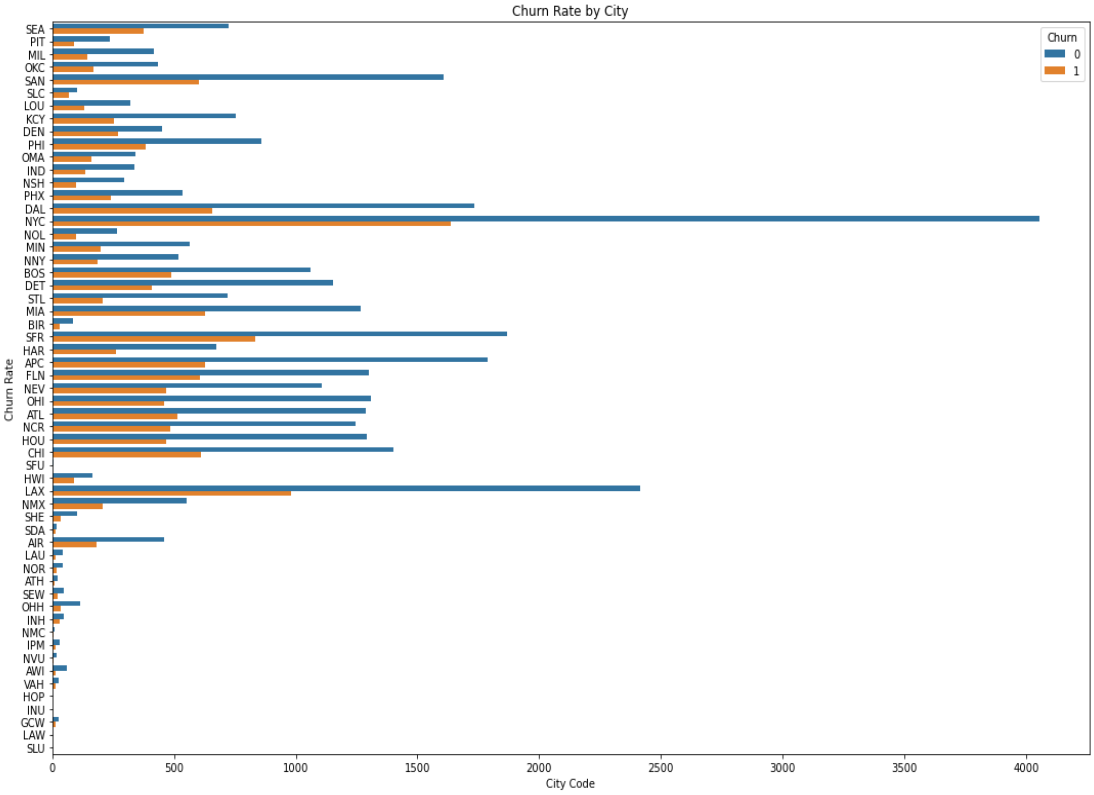


2. Family Dynamics: The presence of children in a household emerges as a notable churn predictor, possibly due to changing service needs or financial priorities.





3. Geographical Variance: Churn rates significantly differ across cities, suggesting local factors like service quality and competition play a crucial role.



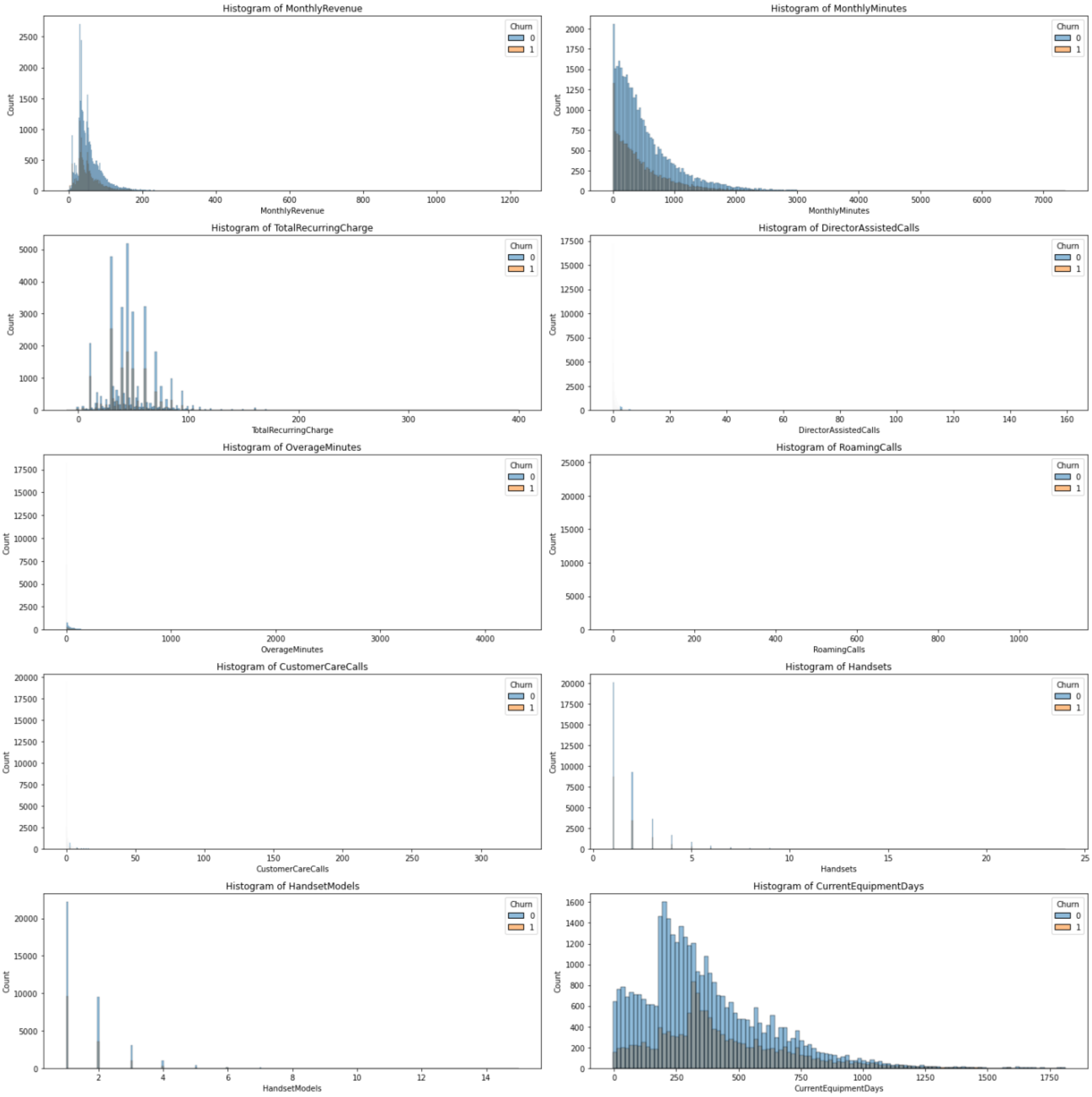
In a nutshell, Income, marital status, children presence, and occupation influence churn. Besides ,Churn rates vary significantly by city due to local factors. These findings underscore the importance of demographic considerations in understanding churn, pointing to targeted strategies that address specific customer segments' needs and preferences.

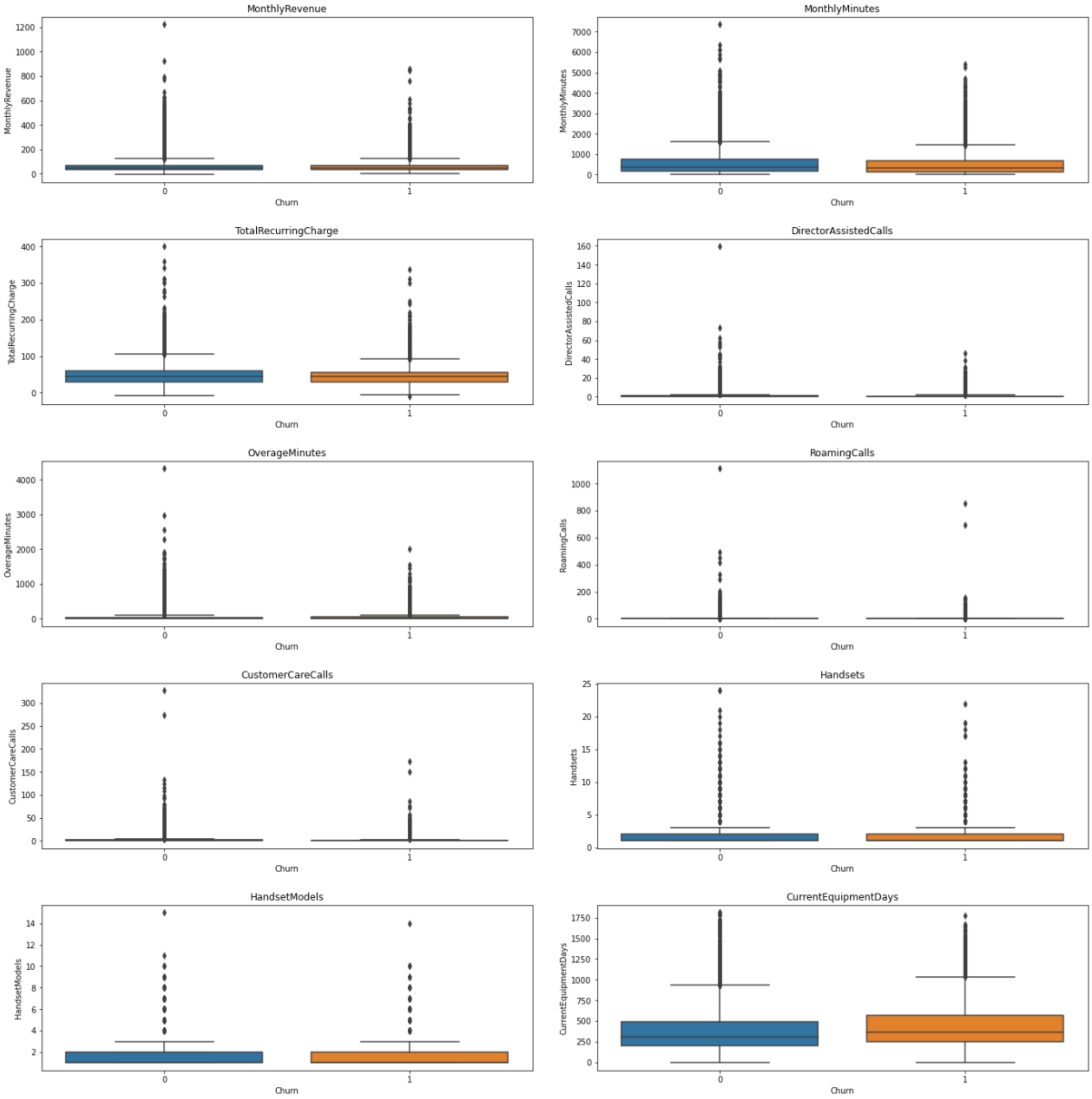
3.2.2 Behavioural analysis

Exploring customer behaviour, particularly around service usage and interactions, reveals several impactful trends:

1. Service Usage: Lower monthly revenue and usage metrics are characteristic of churned customers, indicating potential dissatisfaction or unmet needs.
2. Customer Service Interactions: Higher frequencies of director-assisted and customer care calls among churned customers suggest that service-related issues contribute to churn.

3. Equipment and Engagement: The age of equipment and number of handset models are subtly linked with churn, hinting at the importance of keeping customers engaged and technologically updated.

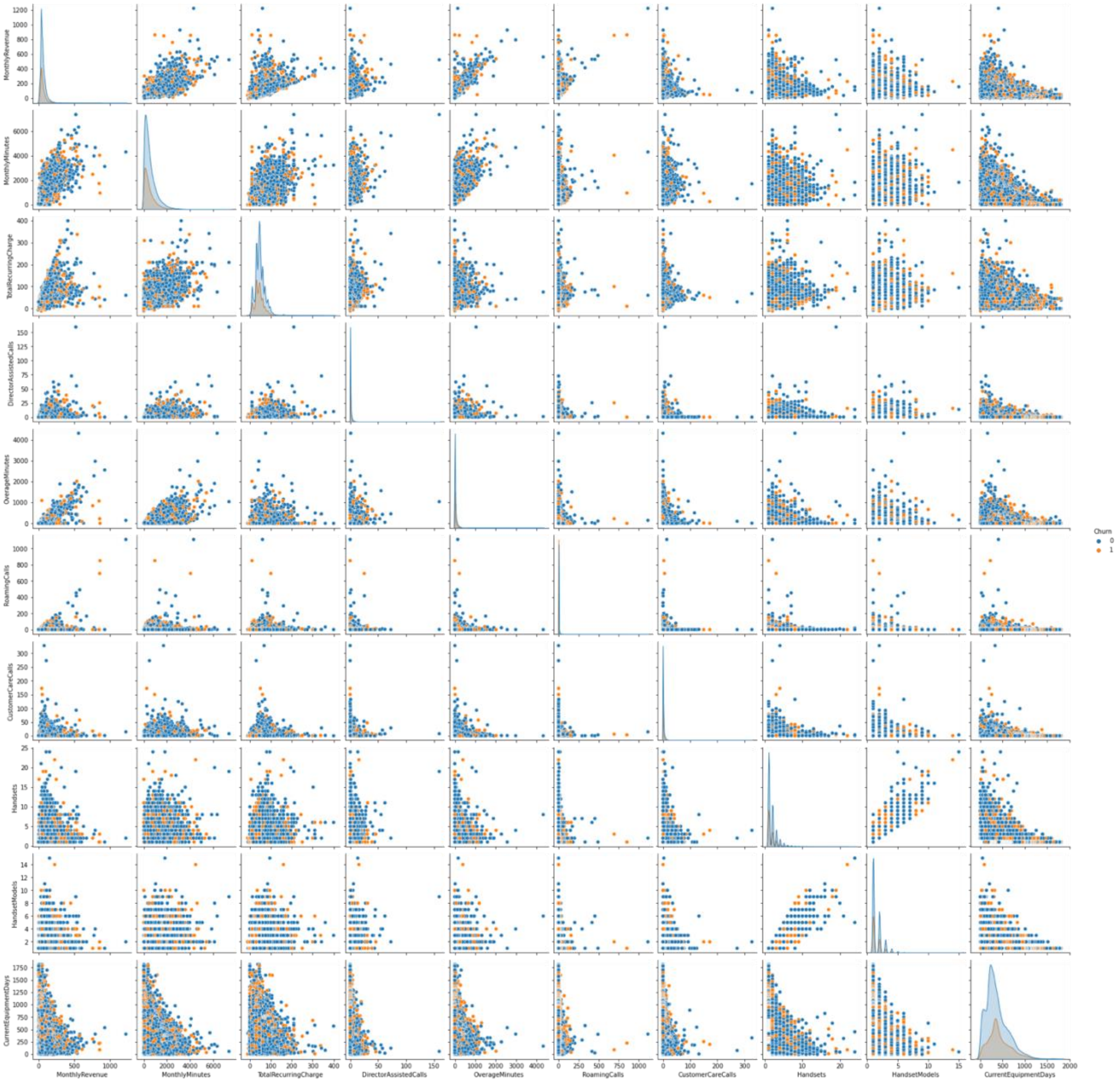




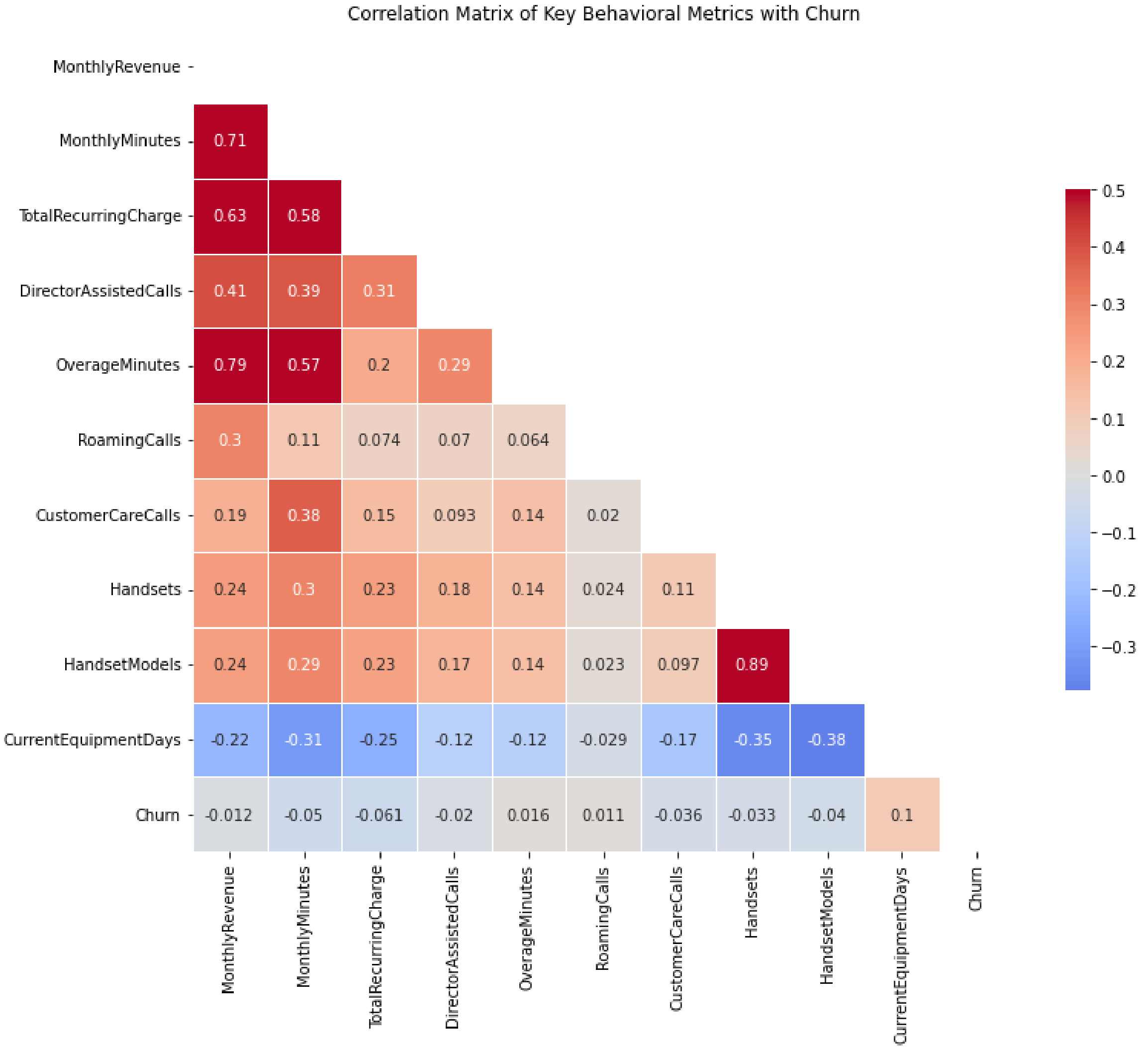
Significant indicators of churn include monthly revenue, usage metrics (minutes), service-related issues (director-assisted and customer care calls), and satisfaction indicators (overage minutes).Churned customers are characterized by lower usage/revenue and potentially more service dissatisfaction. Equipment age and engagement levels (e.g., handset models) also correlate with churn, albeit less strongly. Income and age do not appear to be primary factors in churn.

3.3 Correlation Analysis

The investigation into the interrelationships between variables through correlation analysis further illuminates the complex web of factors influencing churn. While no single metric emerges as a dominant predictor, the collective influence of several weakly correlated factors highlights the multifaceted nature of customer churn.



The below correlation matrix provides insights into how different numerical variables relate to each other and to churn.



Correlation Matrix of Key Behavioral Metrics with Churn

The insights suggest that demographic factors do have some influence on churn, but the effects are nuanced. The presence of children in the household seems to be a more distinguishable factor in predicting churn, whereas age and income group show less clear-cut patterns. These factors, combined with the behavioural analysis on service usage and customer service interactions, provide a more comprehensive understanding of what might influence a customer's decision to churn.

3.4 Statistical Insights

To further investigate the statistical significance of the findings, Hypothesis testing is conducted. A common hypothesis related to churn are:

- Null Hypothesis : There is no significant difference in the mean value between churned and non-churned customers.
- Alternative Hypothesis: There is a significant difference in the mean value between churned and non-churned customers.

3.4.1 Behavioural Influence on Churn

- MonthlyRevenue: Mean difference is significant (t-statistic = -2.761, p-value = 0.00577),
- MonthlyMinutes: Mean difference is significant (t-statistic = -11.724, p-value = 0.00000),
- TotalRecurringCharge: Mean difference is significant (t-statistic = -14.158, p-value = 0.00000)',
- DirectorAssistedCalls: Mean difference is significant (t-statistic = -4.816, p-value = 0.00000).
- OverageMinutes: Mean difference is significant (t-statistic = 3.695, p-value = 0.00022),
- RoamingCalls: Mean difference is significant (t-statistic = 2.244, p-value = 0.02481),
- CustomerCareCalls: Mean difference is significant (t-statistic = -8.695, p-value = 0.00000)

3.4.2 Demographical Influence on Churn

- IncomeGroup:

  Chi-square Statistic: 31.88523845163498, p-value: 0.00020843103577175245

  IncomeGroup Mean difference is significant
- MaritalStatus:

  Chi-square Statistic: 36.925487881712506, p-value: 9.588091848858717e-09

  MaritalStatus Mean difference is significant
- Occupation:

  Chi-square Statistic: 10.316148268152288, p-value: 0.17135426532792455

  Occupation Mean difference is not significant
- CityCode:

  Chi-square Statistic: 187.82898412451533, p-value: 3.845522641131109e-16

  CityCode Mean difference is significant
- ChildrenInHH:

  Chi-square Statistic: 4.61817626131507, p-value: 0.03163485621782633

  ChildrenInHH Mean difference is significant

Through hypothesis testing, the significance of observed patterns are validated. T-tests and chi-square tests confirm that both demographic factors (e.g., income group, marital status) and behavioural indicators (e.g., monthly revenue, customer care calls) significantly differ between churned and non-churned customers. These statistical tests strengthen our confidence in the identified churn drivers.

3.5 Preparing for Predictive Modelling

As a final step, the dataset to focus on variables significantly related to churn is defined, setting the stage for predictive modelling. The variables include:

1. CustomerID
2. Churn
3. MonthlyRevenue
4. MonthlyMinutes
5. TotalRecurringCharge
6. DirectorAssistedCalls
7. OverageMinutes
8. RoamingCalls
9. CustomerCareCalls
10. IncomeGroup
11. MaritalStatus
12. Occupation
13. CityCode
14. ChildrenInHH

This streamlined dataset with 14 columns above, saved as ‘refined_data.csv’ for further analysis, represents a foundation for developing models aimed at predicting and mitigating customer churn.

3.6 EDA Summary

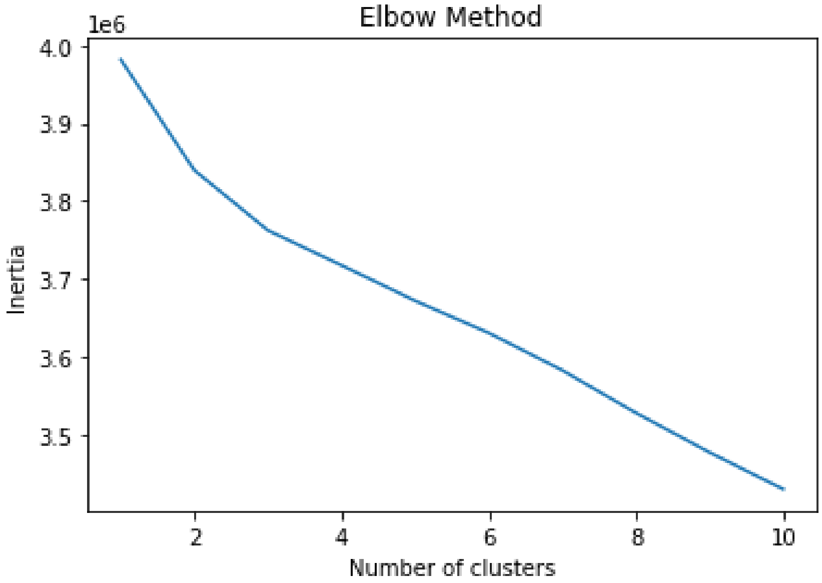
Exploratory analysis paints a detailed picture of the dynamics underlying customer churn. It's clear that churn is a multifactorial issue, with both demographic and behavioural dimensions playing critical roles. The insights gleaned from our analysis suggest targeted intervention strategies, such as personalized customer engagement plans, service adjustments based on family dynamics, and proactive customer support enhancements. The insights outlined herein provide a solid foundation for further analysis, predictive modelling, and strategic intervention planning.

4 Pre-processing

The purpose of pre-processing step is to prepare the data for modelling by encoding categorical variables, handling missing values, and normalizing or standardizing numerical values. It begins with loading refined dataset from EDA step.

4.1 Customer Segmentation

To understand the diverse customer base better, a segmentation exercise using K-Means clustering was embarked. This technique aimed to identify distinct customer groups based on their service usage patterns, demographic characteristics, and behavioural traits.



- A selection of demographic and behavioural features was made for customer segmentation.
- Categorical variables within the dataset were encoded using one-hot encoding.
- Data standardization was performed to normalize the feature scales.
- The K-Means clustering algorithm was applied, with the optimal number of clusters determined by the elbow method.
- Cluster labels were assigned back to the dataset to segment customers into distinct groups.

- The average churn rate and customer counts for each cluster were calculated, revealing differences in churn behaviour among segments.
- The K-Means clustering has successfully segmented the customers into three distinct clusters, with the following sizes:
    - Cluster 0: 17509 customers
    - Cluster 1: 28320 customers
    - Cluster 2: 5218 customers
- The 'Cluster' feature was added to the dataset, incorporating segmentation insights into the feature set.

4.2 Feature Engineering

4.2.1 Encoding Categorical Variables

- With the segmentation complete, the next step was to transform categorical variables using one-hot encoding.

4.2.2 Refinement and Feature Selection

- Categorical columns, now redundant due to one-hot encoding, were removed from the dataset.
- The 'Cluster' feature was added to the dataset, incorporating segmentation insights into the feature set.
- Features relevant for classification, excluding the churn label and customer ID, were selected.

4.3 Preparing for Model Training

4.3.1 Dataset Split

The final prelude to modelling involved splitting dataset into training 20% and testing 80% sets to ensure that future models are evaluated on an unbiased platform, simulating their performance in real-world scenarios.

4.3.2 Saving Pre-processed Data

Recognizing the importance of reproducibility and ease of access, pre-processed data is saved as “train_test_split.pkl”. Both the cluster-labelled dataset and the split datasets were preserved for future modelling and predictive analysis.

5 Modelling

This is final step to build and evaluate models to predict customer churn, comparing different algorithms to identify the best performer. A series of essential steps were undertaken, involving the importation of libraries, data loading, processing, model selection, training, evaluation, and the final model selection.

5.1 Preparation

- Libraries necessary for data manipulation, visualization, statistical analysis, and machine learning were imported.
- The data, previously split into training and test sets, was loaded from a pickle file.
- NaN and Infinity Value were handled by imputing them with the mean as necessary.
- Feature scaling was applied to both the training and test datasets using the **StandardScaler** from scikit-learn.

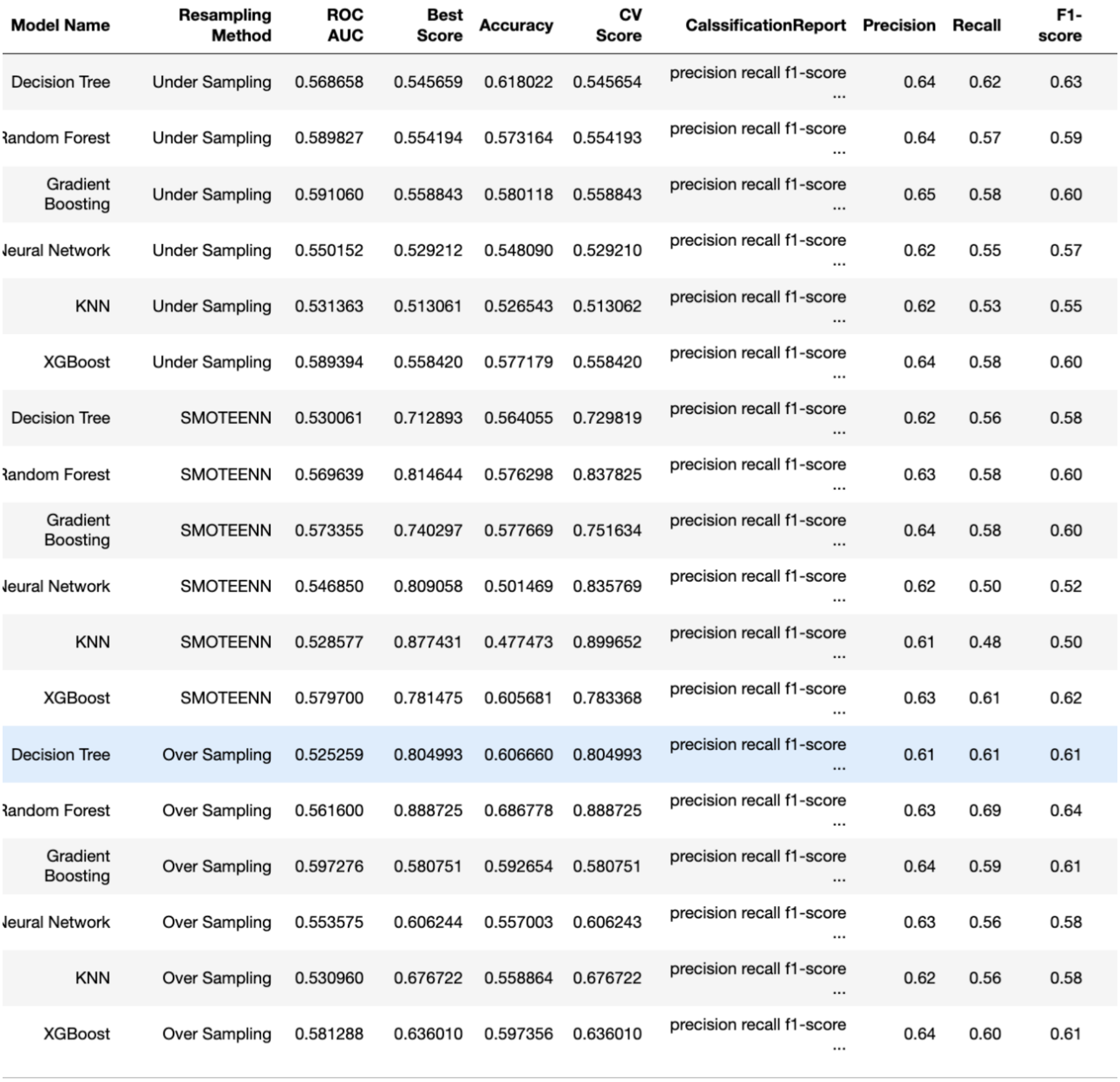
5.2 Model Selection and Training

- A diverse array of models was considered.
    - Decision Trees,
    - Random Forest,
    - Gradient Boosting,
    - Neural Networks,
    - KNN, and
    - XGBoost.

- Each model was subjected to hyper-parameter tuning to optimize its performance.
- Given the dataset's imbalanced nature, resampling techniques such as Random Under Sampling, SMOTEENN, and Random Over Sampling were employed to balance the classes, thereby enhancing the model's ability to learn from both classes effectively.

5.3 Model Evaluation

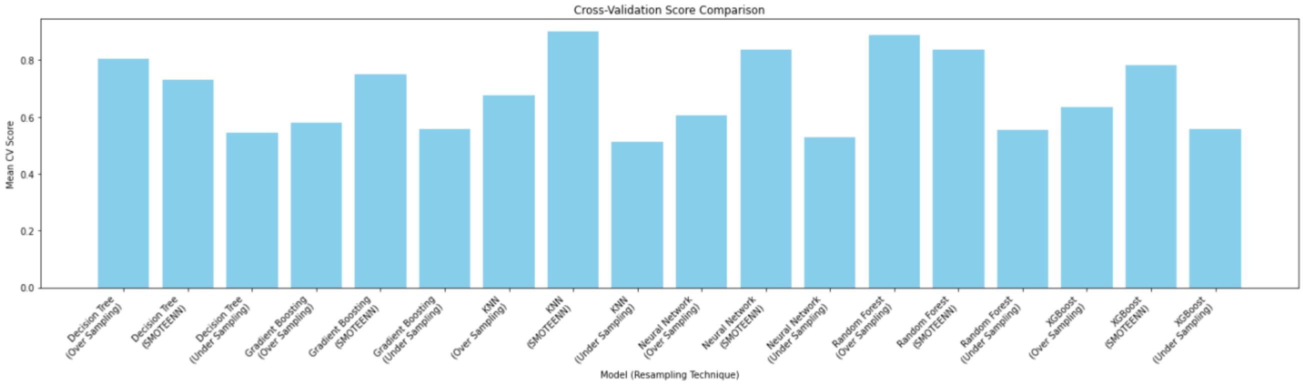
Models were evaluated based on their ROC AUC scores, accuracy, precision, recall, and F1-score. Cross-validation scores were also calculated to assess the models' generalizability. The result as below.

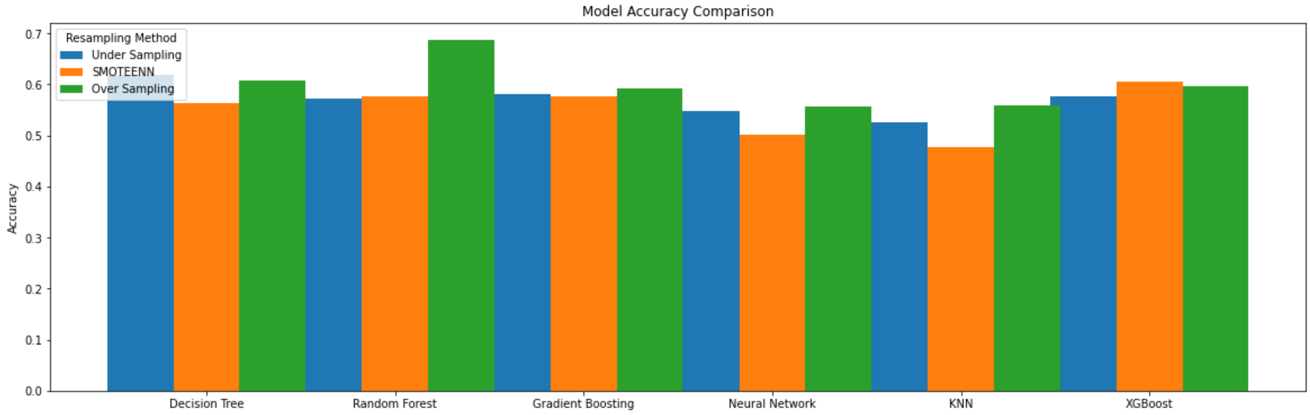


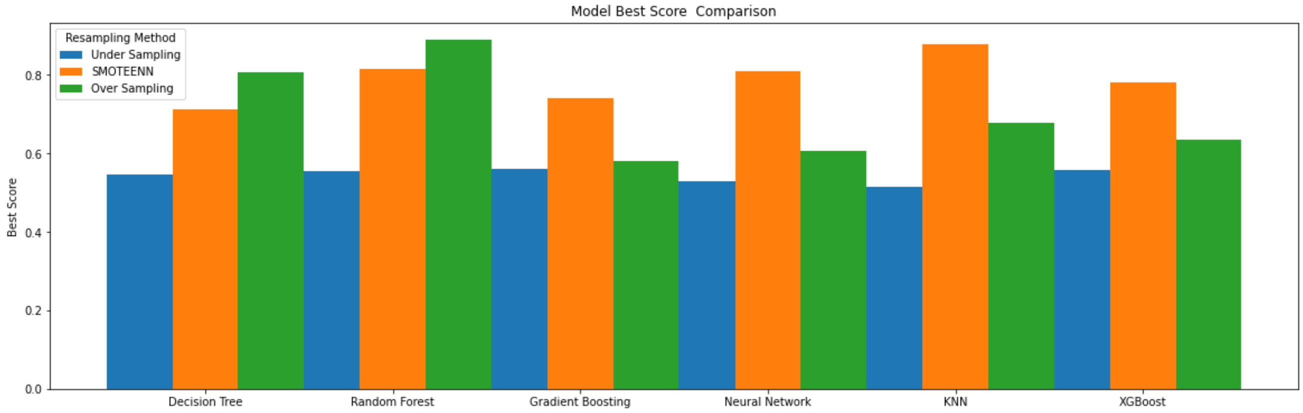
Model Name	Resampling Method	ROC AUC	Best Score	Accuracy	CV Score	CalssificationReport	Precision	Recall	F1-score
Decision Tree	Under Sampling	0.568658	0.545659	0.618022	0.545654	precision recall f1-score ...	0.64	0.62	0.63
Random Forest	Under Sampling	0.589827	0.554194	0.573164	0.554193	precision recall f1-score ...	0.64	0.57	0.59
Gradient Boosting	Under Sampling	0.591060	0.558843	0.580118	0.558843	precision recall f1-score ...	0.65	0.58	0.60
Neural Network	Under Sampling	0.550152	0.529212	0.548090	0.529210	precision recall f1-score ...	0.62	0.55	0.57
KNN	Under Sampling	0.531363	0.513061	0.526543	0.513062	precision recall f1-score ...	0.62	0.53	0.55
XGBoost	Under Sampling	0.589394	0.558420	0.577179	0.558420	precision recall f1-score ...	0.64	0.58	0.60
Decision Tree	SMOTEENN	0.530061	0.712893	0.564055	0.729819	precision recall f1-score ...	0.62	0.56	0.58
Random Forest	SMOTEENN	0.569639	0.814644	0.576298	0.837825	precision recall f1-score ...	0.63	0.58	0.60
Gradient Boosting	SMOTEENN	0.573355	0.740297	0.577669	0.751634	precision recall f1-score ...	0.64	0.58	0.60
Neural Network	SMOTEENN	0.546850	0.809058	0.501469	0.835769	precision recall f1-score ...	0.62	0.50	0.52
KNN	SMOTEENN	0.528577	0.877431	0.477473	0.899652	precision recall f1-score ...	0.61	0.48	0.50
XGBoost	SMOTEENN	0.579700	0.781475	0.605681	0.783368	precision recall f1-score ...	0.63	0.61	0.62
Decision Tree	Over Sampling	0.525259	0.804993	0.606660	0.804993	precision recall f1-score ...	0.61	0.61	0.61
Random Forest	Over Sampling	0.561600	0.888725	0.686778	0.888725	precision recall f1-score ...	0.63	0.69	0.64
Gradient Boosting	Over Sampling	0.597276	0.580751	0.592654	0.580751	precision recall f1-score ...	0.64	0.59	0.61
Neural Network	Over Sampling	0.553575	0.606244	0.557003	0.606243	precision recall f1-score ...	0.63	0.56	0.58
KNN	Over Sampling	0.530960	0.676722	0.558864	0.676722	precision recall f1-score ...	0.62	0.56	0.58
XGBoost	Over Sampling	0.581288	0.636010	0.597356	0.636010	precision recall f1-score ...	0.64	0.60	0.61

5.4 Final Model Selection

The **RandomForestClassifier**, with specific hyperparameters and Oversampling, was identified as the final model after comprehensive testing and evaluation. This model exhibited a promising balance between accuracy and generalizability, as evidenced by its performance metrics such as CV scores, ROC AUC, and classification report outcomes.



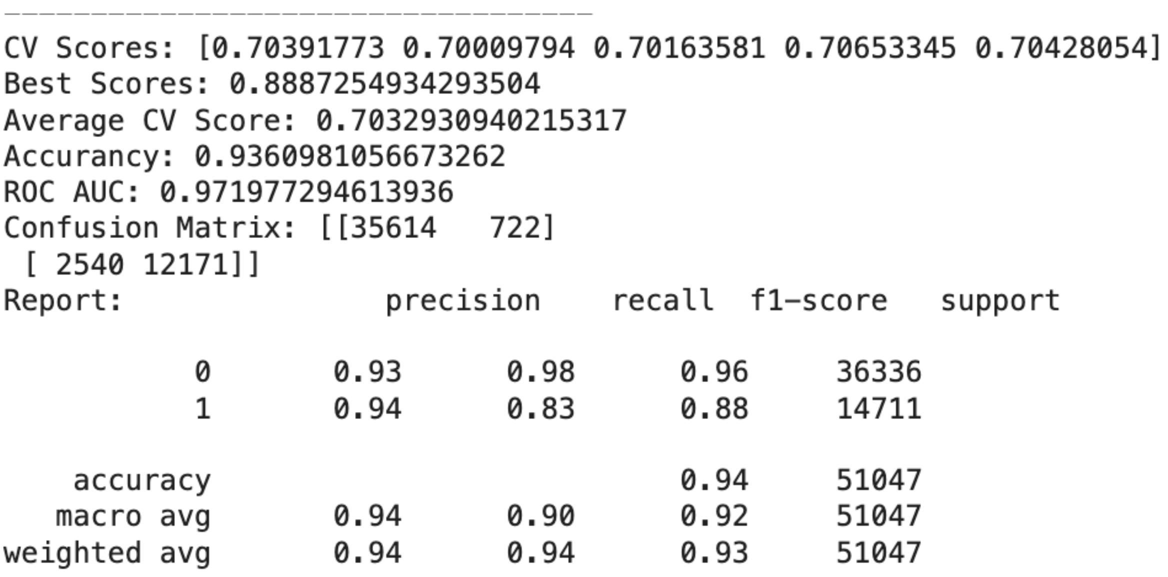




5.5 Outcomes

5.5.1 Performance Metrics

The RandomForestClassifier demonstrated robust performance across various metrics, achieving high accuracy and ROC AUC scores. The model's classification report and confusion matrix provided insights into its precision, recall, and F1-score, reflecting its efficacy in correctly classifying the instances.



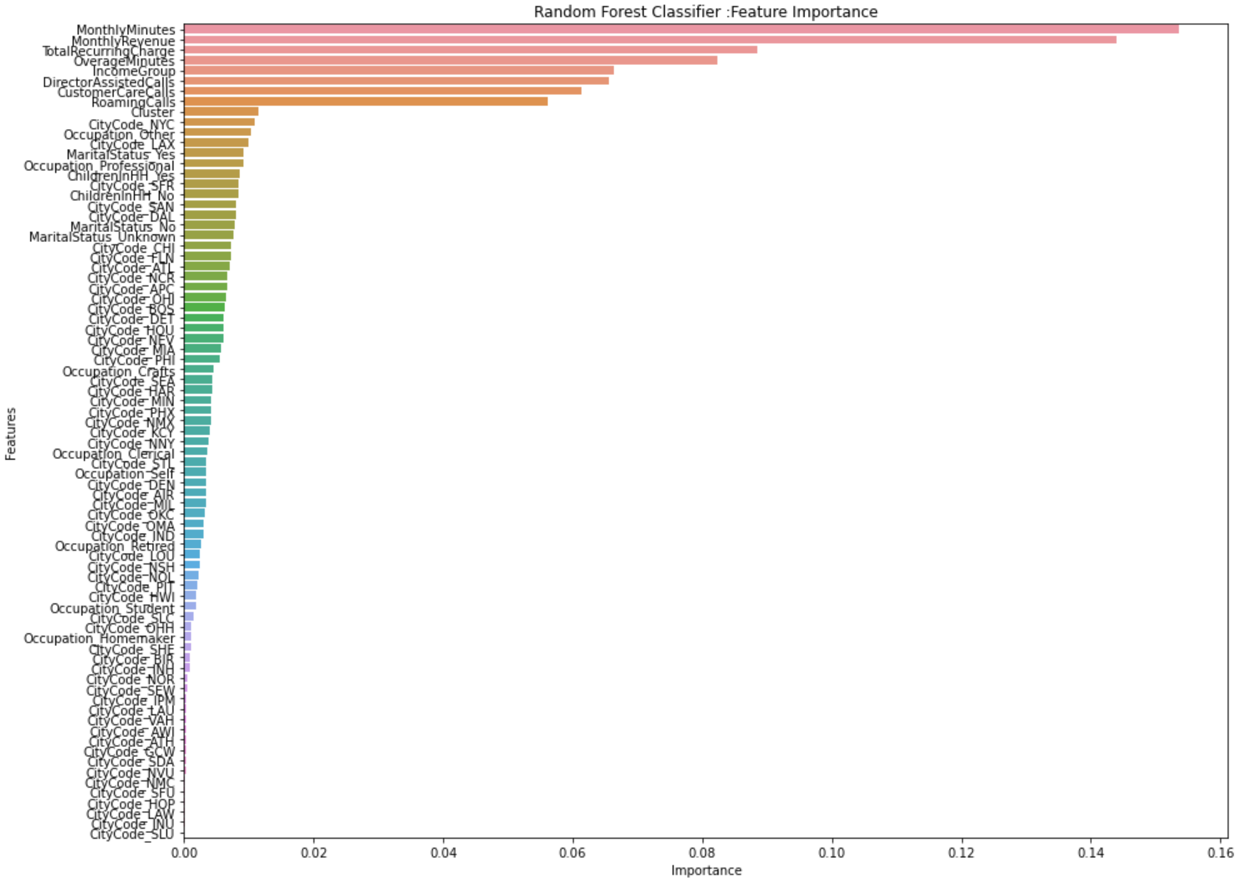
```
-----------------------------------------------------------------
CV Scores: [0.70391773 0.70009794 0.70163581 0.70653345 0.70428054]
Best Scores: 0.8887254934293504
Average CV Score: 0.7032930940215317
Accurancy: 0.9360981056673262
ROC AUC: 0.971977294613936
Confusion Matrix: [[35614   722]
 [ 2540 12171]]
Report:               precision    recall  f1-score   support

           0       0.93      0.98      0.96     36336
           1       0.94      0.83      0.88     14711

    accuracy                           0.94     51047
   macro avg       0.94      0.90      0.92     51047
weighted avg       0.94      0.94      0.93     51047
```

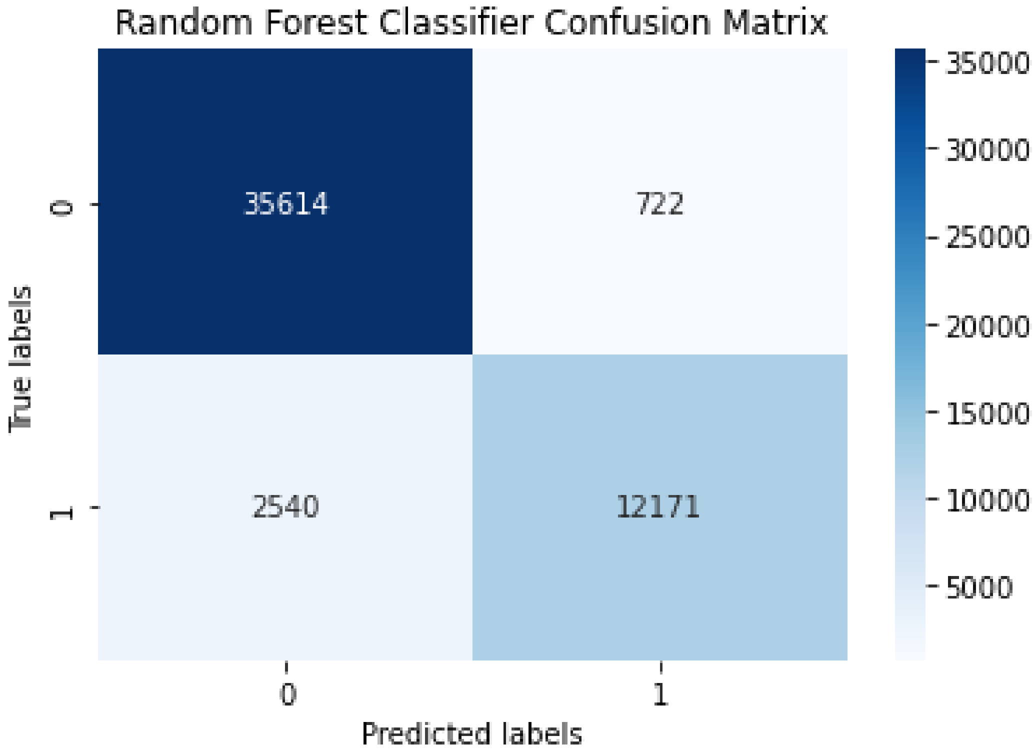
5.5.2 Feature Importance

An analysis of feature importance was conducted. It outlined that Customer Usage patterns such as Monthly minute, Monthly revenue and Total Recurring Charge are the most important factors to impact customer churn.

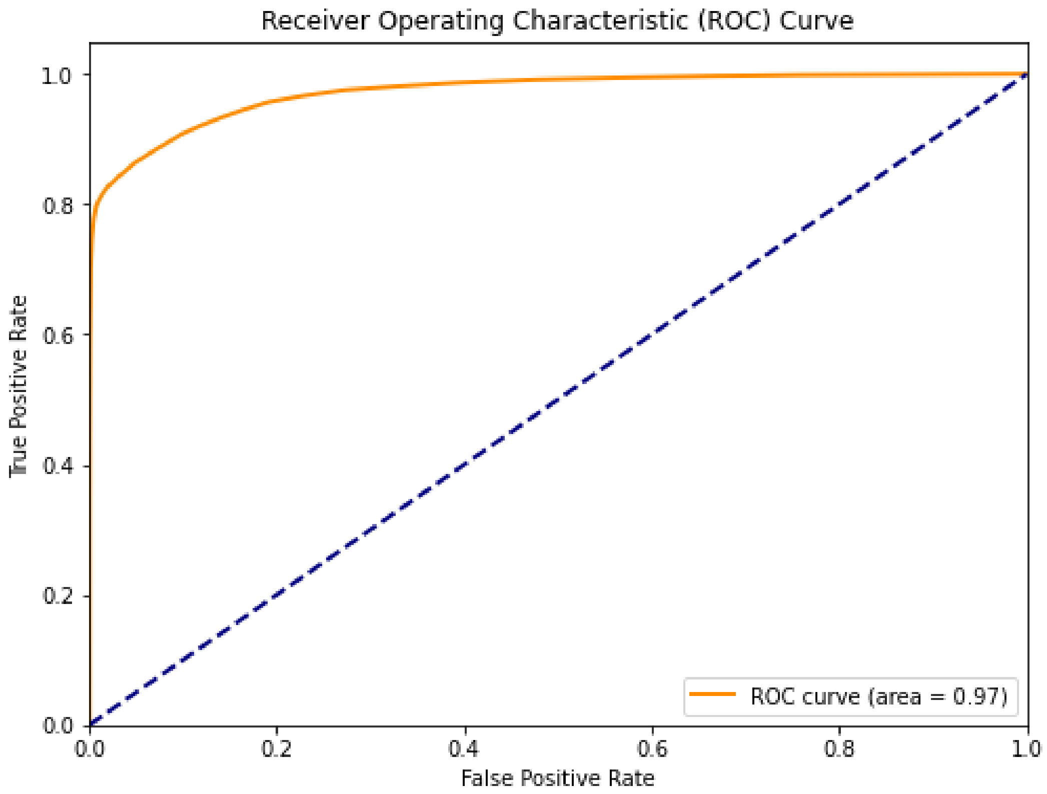


5.5.3 Confusion Matrix and ROC Curve

The confusion matrix and ROC curve further validated the model's performance.Confusion Matrix suggests that that the classifier has a high degree of accuracy and precision with a reasonably high recall. The F1 score being closer to 1 is indicative of a good balance between precision and recall, which is particularly important in cases where there is an imbalance between the positive and negative classes.



The AUC value ranges from 0 to 1.An AUC of 0.5 suggests no discriminative ability, equivalent to random guessing. An AUC of 1.0 represents perfect discrimination, where the classifier can perfectly differentiate between the two classes. An AUC less than 0.5 suggests worse than random predictions, but this is typically observed only when there's a problem with the way the classifier is being used. The graph suggested an **AUC of 0.97** indicates a very good predictive model with high sensitivity and specificity.



5.6 Predictions and Risk Segmentation

The final model was applied to make predictions on the dataset, which were then used to segment the customers based on their risk of churn.
Using churn probabilities from the classification model to further segment customers based on their risk of churning. This segmentation can help in strategizing targeted interventions for different risk groups. Low Risk ($p < 0.3$) Medium Risk ($0.3 \leq p < 0.7$) High Risk ($p \geq 0.7$). Recommendations were made to implement targeted strategies based on risk segmentation derived from model predictions below.

- Segment 0: 31238 customers
  - Strategy: Send personalized emails offering discounts or special offers.
- Segment 1: 9074 customers
  - Strategy: Engage with customer service for feedback and improvement suggestions.

- Segment 2: 10735 customers
    - Strategy: Offer loyalty programs or benefits to enhance customer retention.

5.7 Conclusion

This phase culminated in the preparation for targeted strategies aimed at mitigating churn and enhancing customer retention, underpinned by a deep understanding of the model's predictive capabilities and the underlying data patterns. The predicted result was saved as 'pred_result.csv' for visualisation.

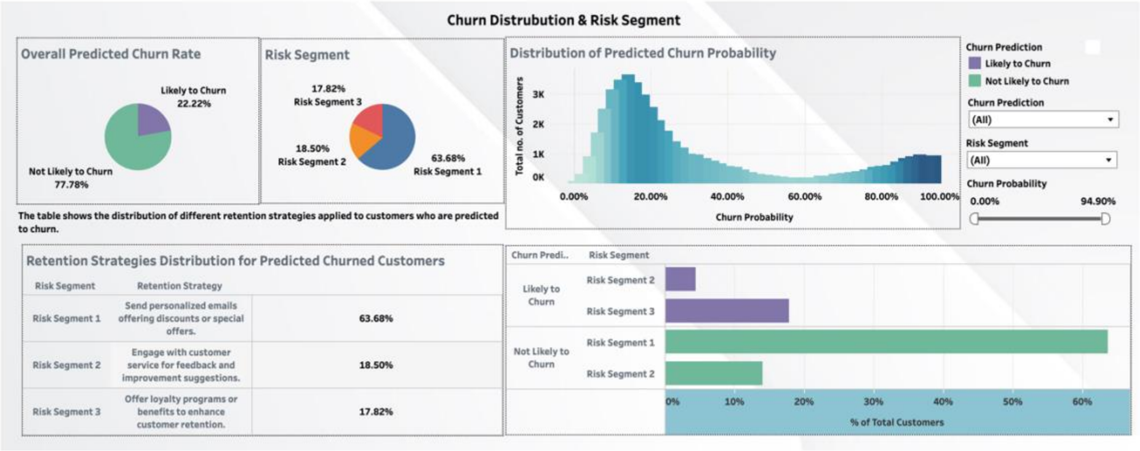
6 Data Visualisation

As final step of the project, the data visualization of the Telecom Customer Churn Prediction Model Analysis was built based on the predication data. The data intricate details about risk segmentation, customer behaviour and usage pattern which are elucidated through a variety of graphical representations.

6.1 Churn Probability and Risk Segmentation

- Overall Predicted Churn Rate: A pie chart reveals that 22.22% of customers are likely to churn, while a significant majority, 77.78%, are not likely to churn.
- Risk Segment Distribution: Another pie chart divides the risk segments, with the largest segment being Risk Segment 1 at 63.68%, followed by Risk Segment 2 at 18.50% and Risk Segment 3 at 17.82%. This segmentation allows for the identification of customers at different risk levels of churning.
- Churn Distribution & Risk Segment: A histogram displays the distribution of predicted churn probabilities, providing a visual representation of how churn risk is spread across the customer base.

- Retention Strategies Distribution for Predicted Churned Customers: Bar charts indicate the retention strategies recommended for each risk segment. Risk Segment 1, the largest group, is targeted with personalized emails offering discounts or special offers. For Risk Segment 2, engagement with customer service for feedback and improvement suggestions is advised. Finally, for Risk Segment 3, offering loyalty programs or benefits to enhance customer retention is suggested.



Churn Distrubution & Risk Segment

Churn Prediction: (All)

Risk Segment: (All)

Churn Probability: 0.00% – 94.90%

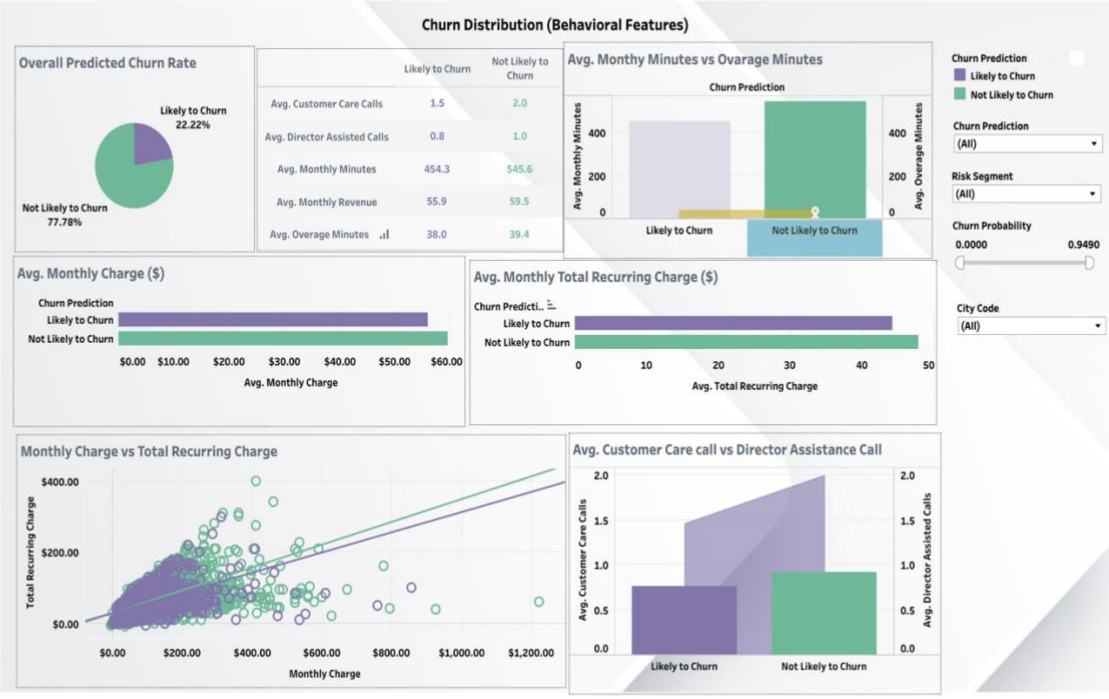
The table shows the distribution of different retention strategies applied to customers who are predicted to churn.

Retention Strategies Distribution for Predicted Churned Customers

Risk Segment	Retention Strategy	
Risk Segment 1	Send personalized emails offering discounts or special offers.	63.68%
Risk Segment 2	Engage with customer service for feedback and improvement suggestions.	18.50%
Risk Segment 3	Offer loyalty programs or benefits to enhance customer retention.	17.82%

6.2 Behavioural Features

- Average Behavioral Metrics: A side-by-side comparison of the average behavioral metrics for customers likely to churn and those not likely is presented. The metrics include average customer care calls, average director-assisted calls, average monthly minutes, and average monthly revenue.
- Monthly Charge Analysis: The graphs elucidate the relationship between monthly charges and total recurring charges, displaying a scatter plot with a trend line indicating a positive correlation between the two metrics.
- Comparison of Customer Care Calls and Director-Assisted Calls: A set of bar graphs contrasts the average customer care calls and director assistance calls between customers likely to churn and those not likely to churn.

- 

Churn Distribution (Behavioral Features)

	Likely to Churn	Not Likely to Churn
Avg. Customer Care Calls	1.5	2.0
Avg. Director Assisted Calls	0.8	1.0
Avg. Monthly Minutes	454.3	545.6
Avg. Monthly Revenue	55.9	59.5
Avg. Overage Minutes	38.0	39.4

Churn Prediction: (All)

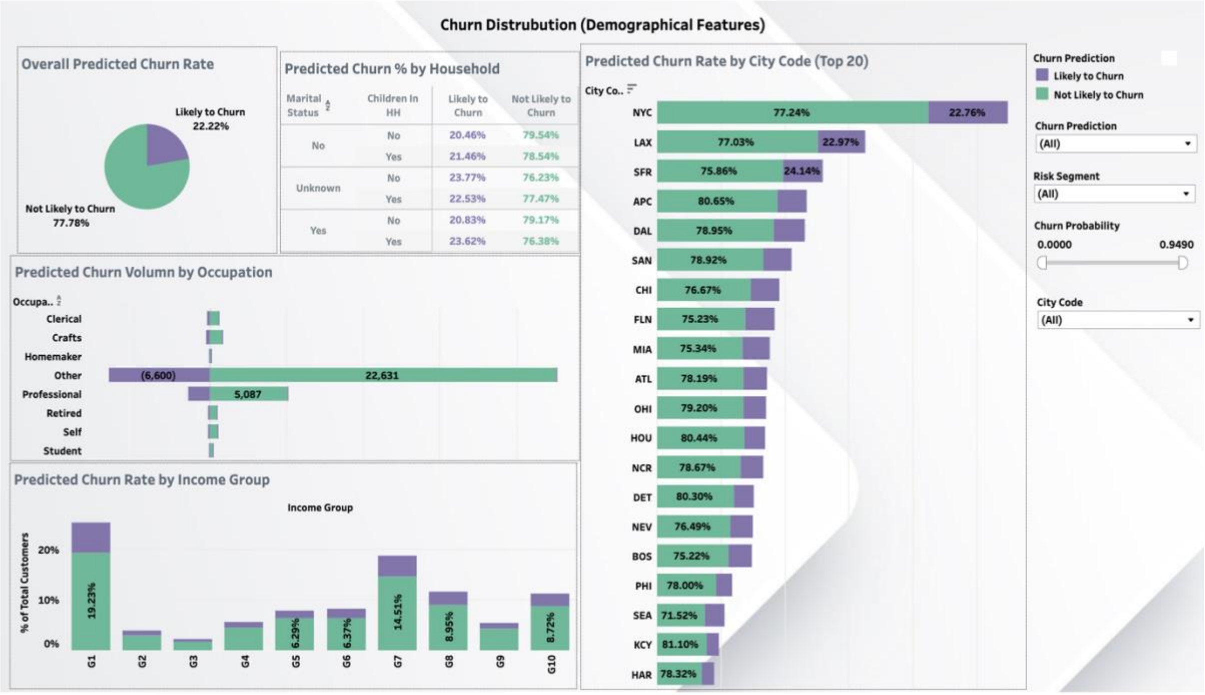
Risk Segment: (All)

Churn Probability: 0.0000 – 0.9490

City Code: (All)

6.3 Demographical Features

- **Predicted Churn by Household Composition:** The data is dissected to reflect churn probability in relation to marital status and the presence of children in the household. A nuanced pattern emerges, where households with children tend to show slightly higher churn rates across marital statuses.
- **Predicted Churn Volume by Occupation:** A bar graph quantifies the churn volume across different occupations. Professions categorized as 'Other' demonstrate the highest churn volume, followed by 'Professional' and 'Clerical' jobs, indicating that churn tendencies may vary significantly across occupational lines.
- **Predicted Churn Rate by Income Group:** Income groups are analyzed, revealing that lower income groups exhibit higher churn rates. The bar chart distinctly portrays a decline in churn rates as income groups ascend, suggesting that higher economic stability might be linked to lower churn rates.
- **Predicted Churn Rate by City Code (Top 20):** A bar chart details the churn rate by city code, with cities such as KCY showing the highest churn rate and SEA the lowest among the top 20. This geographical analysis allows for region-specific strategic planning to mitigate churn.



Predicted Churn % by Household

Marital Status	Children In HH	Likely to Churn	Not Likely to Churn
No	No	20.46%	79.54%
	Yes	21.46%	78.54%
Unknown	No	23.77%	76.23%
	Yes	22.53%	77.47%
Yes	No	20.83%	79.17%
	Yes	23.62%	76.38%

Churn Prediction
(All)

Risk Segment
(All)

Churn Probability
0.0000 0.9490

City Code
(All)

By leveraging predictive insights, telecom providers can refine their customer engagement and retention strategies, tailoring them to meet the specific needs and characteristics of different customer segments.

7 Future Work

For future iterations of the project, it would be beneficial to explore additional data sources, feature engineering techniques, and advanced modelling approaches, such as ensemble methods or deep learning, depending on the project's complexity and requirements. Further analysis could also focus on refining the models based on feedback and new data, continuously improving their accuracy and reliability.

This report provides derails of the project's approach and findings. For detailed analysis, insights, and technical discussions, refer to the individual notebooks covering each project phase.