TELECOM CUSTOMER CHURN PREDICTION ANALYSIS

Data Science Final Capstone Project

(Springboard Jul 2023 Cohort)

Prepared by : Thant Thiri Myo Kyi

March 2024

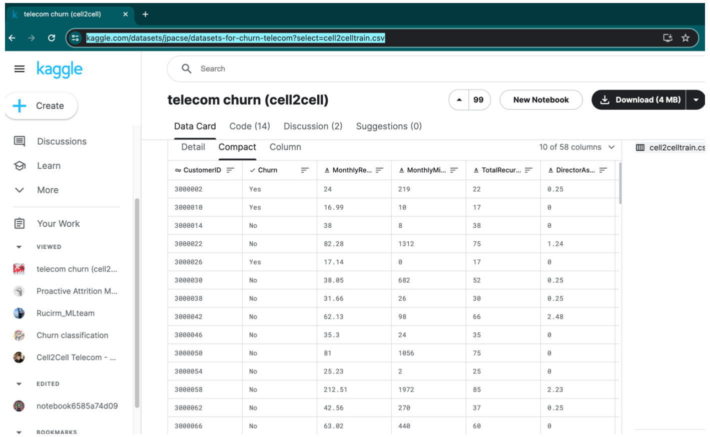
INTRODUCTION

- Problem Statement
  - To addresses the crucial challenge of developing a predictive model that can forecast the likelihood of customer churn for a telecom company.
  - By analyzing a multitude of factors—including customer demographics, service usage patterns, and prior churn rates—the model aims to pinpoint customers at a high risk of churning.
  - The ability to predict which customers are likely to leave enables the company to take strategic and timely actions to increase retention, thereby reducing churn rates and fortifying customer satisfaction and loyalty.

- Goal
  - To construct a dependable and precise predictive model that identifies from its insights which will lead the telecom company's ability to retain customers, leading customers who are potential churn risks.

DATA ACQUISITION & CLEANING

- Data Collection
  - Sourced from a Kaggle dataset, the data encompasses records of 51,047 telecom customers, providing a rich foundation for analysis. Variables cover a broad spectrum, including customer demographics, account details, usage metrics, and customer service interactions, encompassing 57 distinct fields ranging from monthly revenue to handset information.
  - Data Source: https://www.kaggle.com/datasets/jpacse/datasets-for-churn-telecom?select=cell2celltrain.csv



telecom churn (cell2cell)

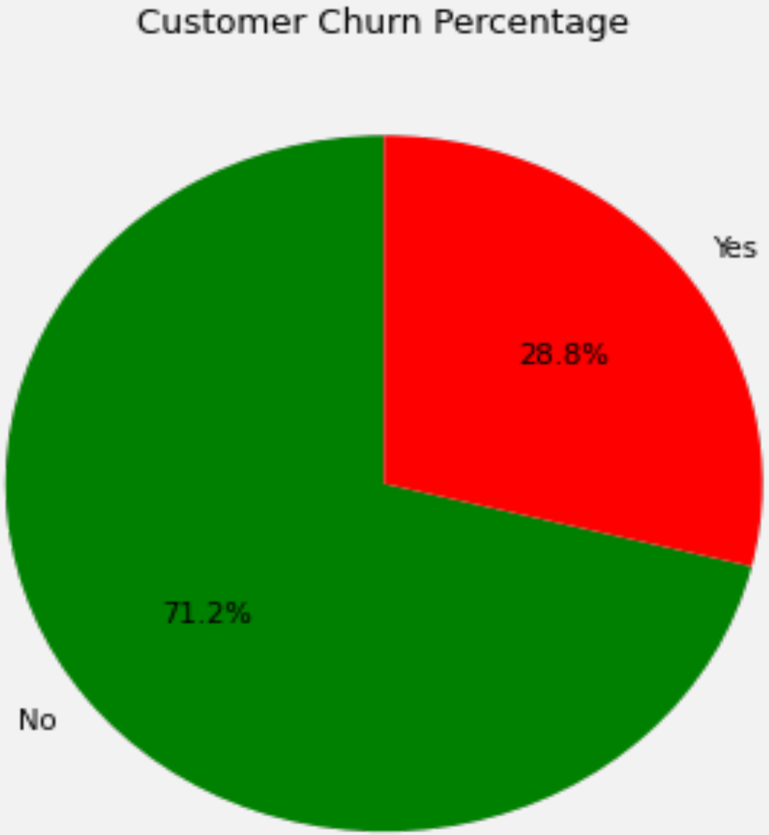
CustomerID	Churn	MonthlyRe...	MonthlyMi...	TotalRecur...	DirectorAs...
3000002	Yes	24	219	22	0.25
3000010	Yes	16.99	10	17	0
3000014	No	38	8	38	0
3000022	No	82.28	1312	75	1.24
3000026	Yes	17.14	0	17	0
3000030	No	38.05	682	52	0.25
3000038	No	31.66	26	30	0.25
3000042	No	62.13	98	66	2.48
3000046	No	35.3	24	35	0
3000050	No	81	1056	75	0
3000054	No	25.23	2	25	0
3000058	No	212.51	1972	85	2.23
3000062	No	42.56	270	37	0.25
3000066	No	63.02	440	60	0

DATA ACQUISITION & CLEANING

- Data Wrangling
  - The initial step involved a rigorous cleaning process to ensure data quality and readiness for subsequent analysis.
    - Identified and addressed missing data points, with key variables such as AgeHH1, AgeHH2, PercChangeMinutes, and PercChangeRevenues being affected—though these constituted less than 2% of the dataset.
    - Checked for duplicate records and took appropriate actions to remove redundancies, preserving the integrity of our dataset.
    - To enhance data granularity, we derived a new categorical feature, 'CityCode', from the Service Area using the first three characters, aligning with the IATA airport coding standard.
    - Analyzed the distribution of both categorical and numerical columns, identifying and deciding on the best approaches to handle outliers—through methods such as capping, imputation, or adjustments—acknowledging the inherent variability within telecom data due to diverse customer profiles.
  - The finalized, cleaned dataset consists of 51,047 records and 57 variables, providing a solid base for robust analytical exploration.
  - Stored the cleansed data in a file named '**cleaned_data.csv**', ready for deep statistical analysis and predictive modelling in the following phases of the project.

EXPLORATORY DATA ANALYSIS (EDA) OVERVIEW

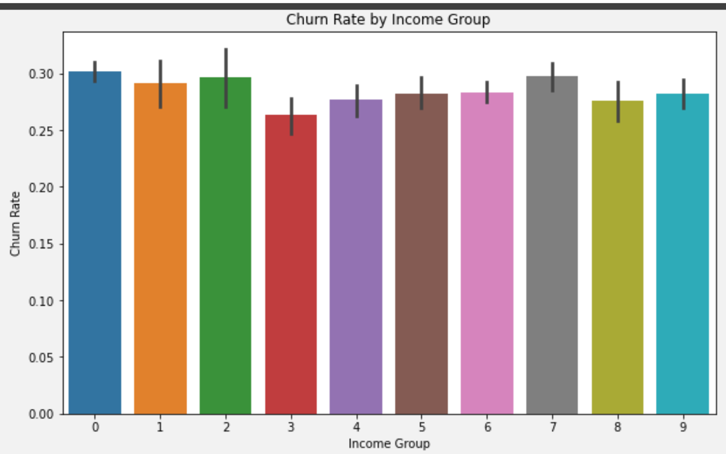
- Preliminary Data Examination
  - Initial step in EDA was to examine the churn status across our entire dataset, revealing that 28.8% of customers have churned.

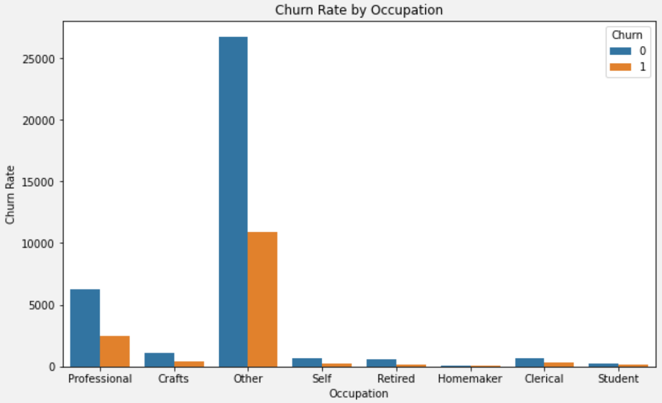


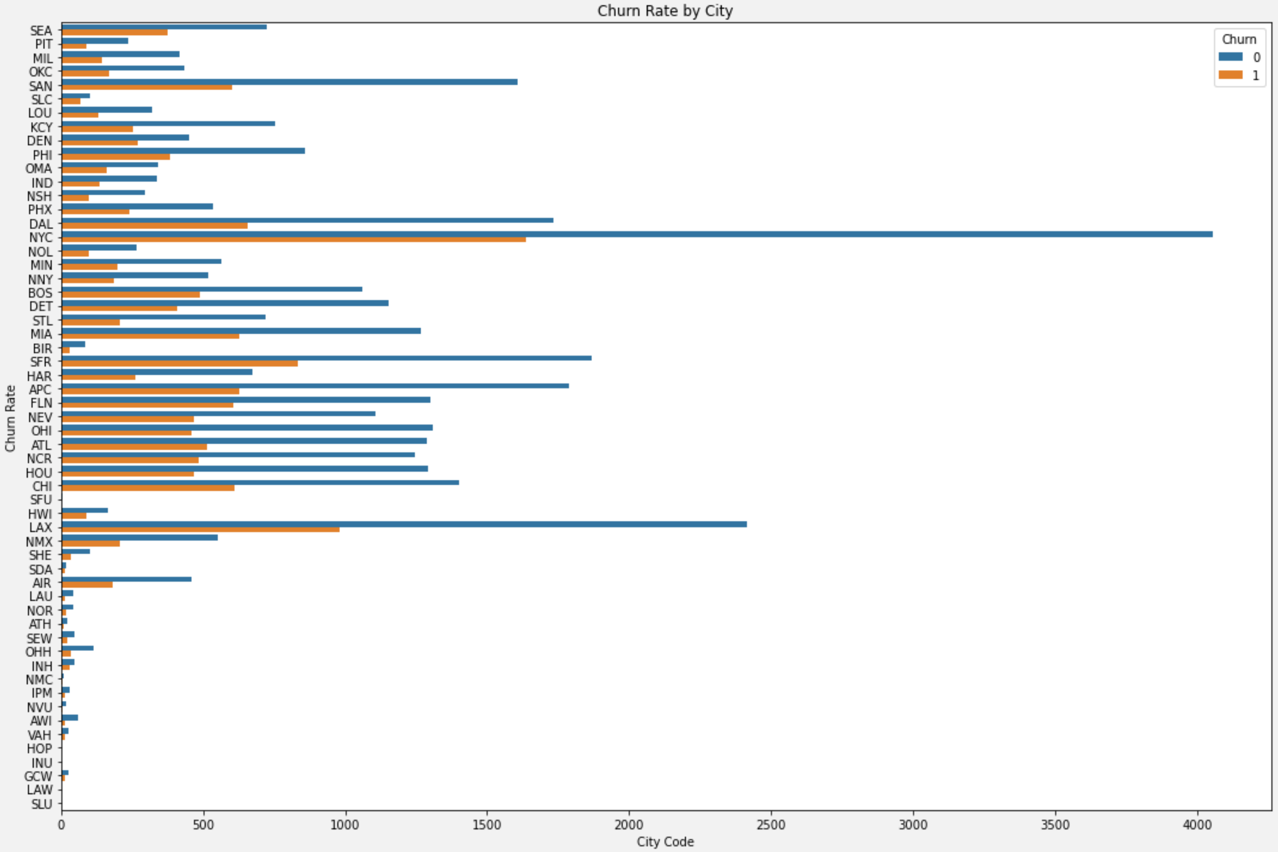
- Next step is driving into a variety of variables such as customer demographics, account specifics, and behavior patterns to lay the groundwork for an in-depth analysis.
  - Customer Analysis
    - The customer analysis was bifurcated into two key areas to gain a holistic understanding of what influences churn:
    - **Demographic Analysis:** Assessed how different customer segments—defined by age, income, marital status, and more—fare in terms of churn likelihood.
    - **Behavioural Analysis:** Investigated service usage and customer interaction metrics, aiming to pinpoint behavior patterns that signal a higher propensity for churn.

DEMOGRAPHIC ANALYSIS

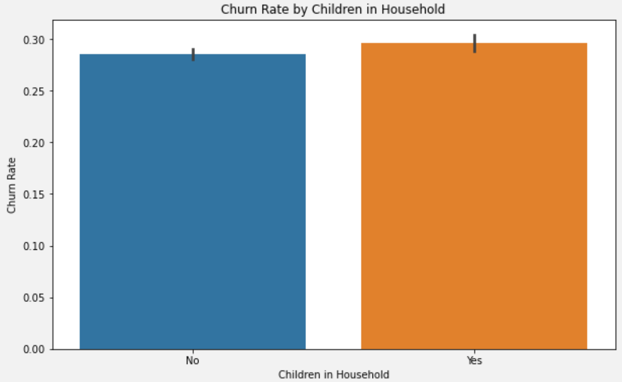
- Economic Influence
  - Analysis of churn across different economic brackets revealed a pattern where customers from specific income and occupation categories displayed higher churn rates.
  - A trend suggesting that customers with lower income levels were more prone to churn, potentially indicating a sensitivity to pricing and service costs.
  - Occupation types seem to correlate with churn; certain job categories, perhaps those with more transient lifestyles or erratic service requirements, showed a propensity to switch providers more frequently.

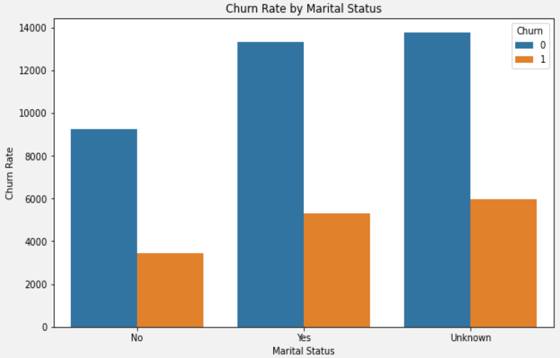






- Family Dynamics
  - Churn rates were notably different among households with and without children, suggesting that family needs play a significant role in retention.
  - Customers with children in the household had slightly higher churn rates, possibly due to shifting service needs or financial restraints.

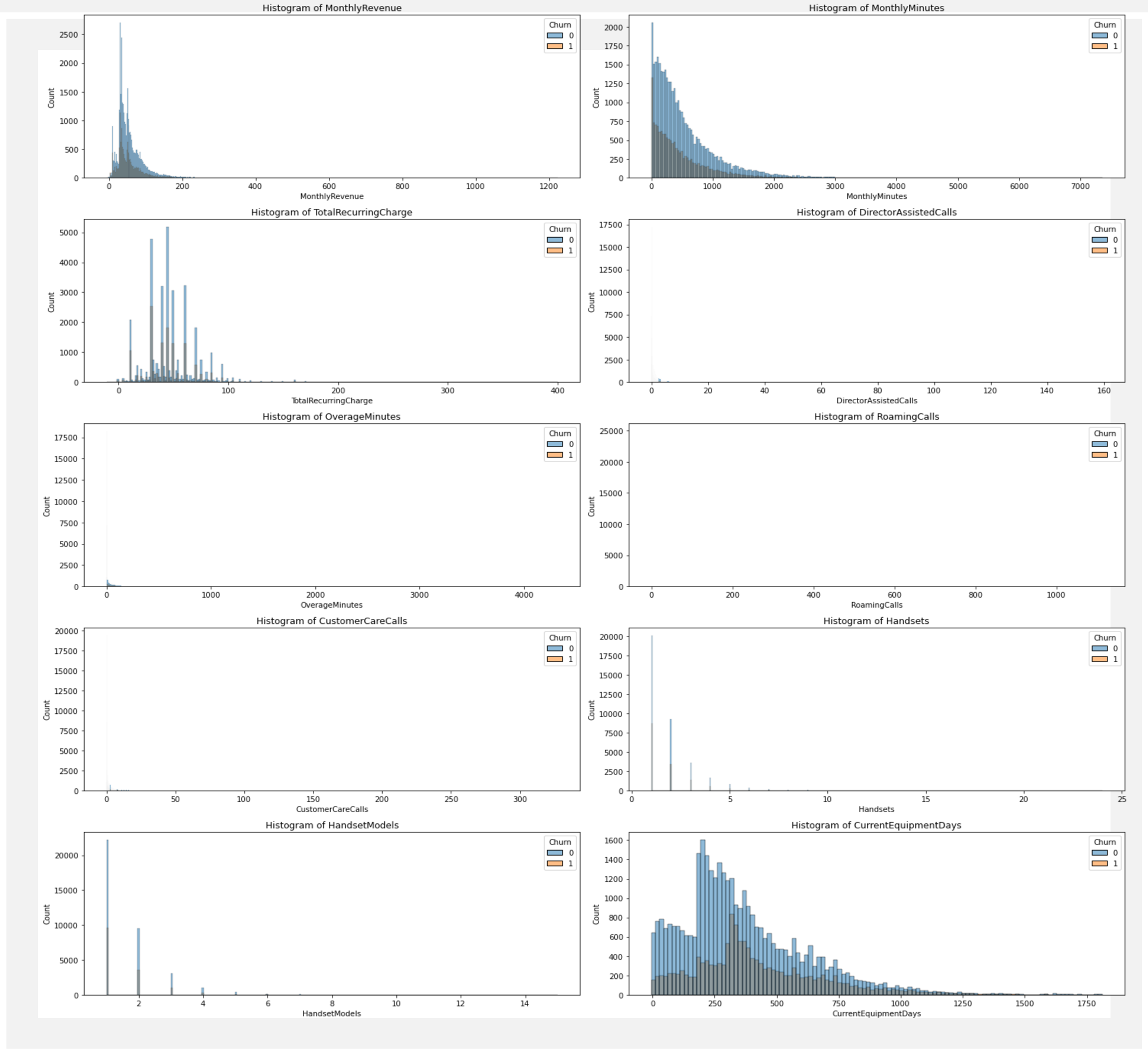




- Geographic Variance
  - A significant disparity in churn rates was observed when analyzing the data city-wise.
  - This variance could point towards external factors such as regional service quality, market competition, or economic conditions influencing customer loyalty.

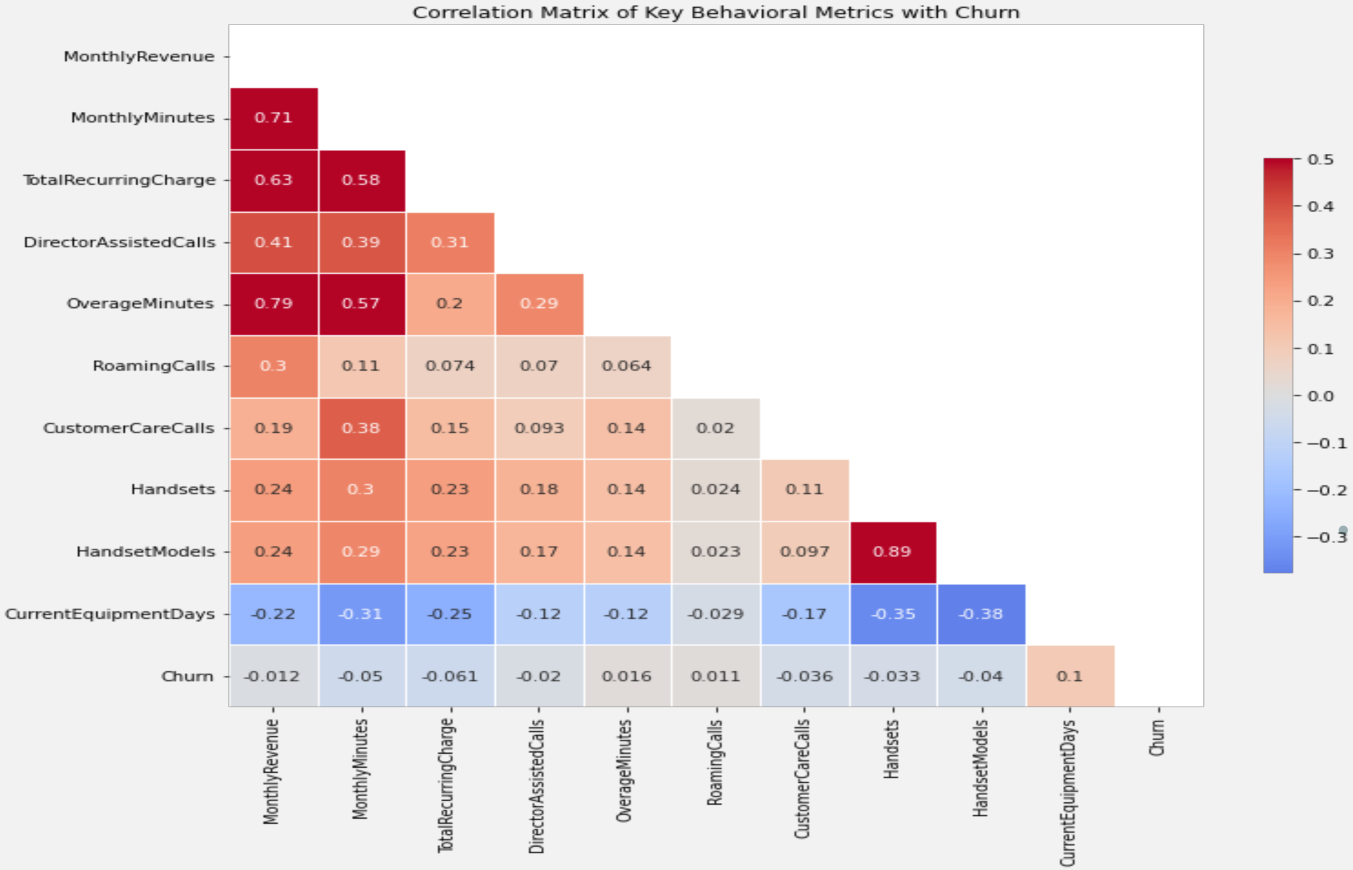
BEHAVIORAL ANALYSIS

- Service Usage
  - Customers who eventually churned exhibited lower service usage in terms of both monthly revenue and usage minutes. This suggests that lower engagement with the telecom services may be a precursor to customer departure, highlighting the need for plans that better match customer usage patterns.
  - A higher incidence of overage minutes among churned customers pointed to a possible mismatch between customer needs and their current plans.
- Customer Service Interactions
  - Churned customers were found to have a higher frequency of director-assisted calls and customer care interactions, implying dissatisfaction or unresolved issues.
  - The data suggests that addressing service-related concerns promptly and effectively could play a pivotal role in preventing churn.
  - A proactive customer service approach, focusing on resolving issues before they escalate, could enhance overall customer satisfaction and loyalty.
- Equipment and Engagement
  - The analysis also considered the age of customer handsets and the number of handset models used, uncovering a subtle yet notable correlation with churn rates.
  - Customers with older or outdated handsets showed a higher tendency to churn, possibly seeking newer technology available with competitors.
  - Engagement, as indicated by the number of handset models, also correlates with churn, suggesting that customers who are not taking advantage of the full range of services and updates may be at higher risk of churning.



CORRELATION & STATISTICAL INSIGHTS

- **Correlation Analysis**
  - Correlation matrix serves as a visual representation of the relationships between customer attributes and churn.
  - While no single factor is a dominant churn predictor, a combination of several weakly correlated factors may cumulatively have a strong influence.
  - Variables such as monthly minutes, customer service calls, and overage charges, though weakly correlated individually, together paint a compelling picture of potential churn risk.



Correlation Matrix of Key Behavioral Metrics with Churn

	MonthlyRevenue	MonthlyMinutes	TotalRecurringCharge	DirectorAssistedCalls	OverageMinutes	RoamingCalls	CustomerCareCalls	Handsets	HandsetModels	CurrentEquipmentDays	Churn
MonthlyRevenue											
MonthlyMinutes	0.71										
TotalRecurringCharge	0.63	0.58									
DirectorAssistedCalls	0.41	0.39	0.31								
OverageMinutes	0.79	0.57	0.2	0.29							
RoamingCalls	0.3	0.11	0.074	0.07	0.064						
CustomerCareCalls	0.19	0.38	0.15	0.093	0.14	0.02					
Handsets	0.24	0.3	0.23	0.18	0.14	0.024	0.11				
HandsetModels	0.24	0.29	0.23	0.17	0.14	0.023	0.097	0.89			
CurrentEquipmentDays	-0.22	-0.31	-0.25	-0.12	-0.12	-0.029	-0.17	-0.35	-0.38		
Churn	-0.012	-0.05	-0.061	-0.02	0.016	0.011	-0.036	-0.033	-0.04	0.1	

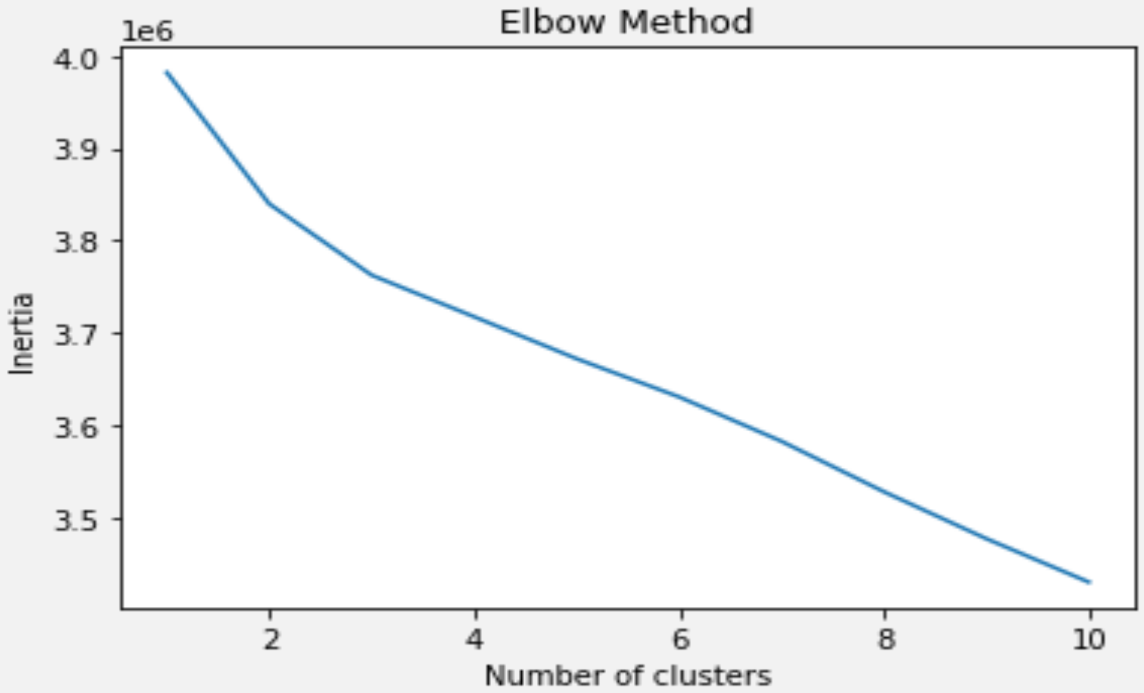
- **Statistical Significance**
  - The null hypothesis posited no significant difference between the means of churned and non-churned customers across various metrics.
  - Results indicated:
    - A significant difference in the mean monthly revenue and minutes between churned and non-churned customers ($p < 0.05$), rejecting the null hypothesis.
    - Customer service interactions, such as the number of director-assisted calls, were significantly higher in churned customers ($p < 0.05$), further rejecting the null.
    - Demographic factors like income group and marital status also showed significant differences in churn rates ($p < 0.05$), while occupation did not ($p > 0.05$).
  - These statistically significant differences validate the predictive power of certain behavioural and demographic factors in identifying potential churn.

PREPARING FOR PREDICTIVE MODELLING

- Refined Dataset
  - Based on our exploratory data analysis (EDA), we streamlined the dataset to concentrate on variables with significant relevance to churn. This refined dataset includes:
    - CustomerID: For tracking and analysis.
    - Churn: The target variable we aim to predict.
    - Service Usage Metrics: MonthlyRevenue, MonthlyMinutes, and TotalRecurringCharge.
    - Customer Interaction: DirectorAssistedCalls, CustomerCareCalls.
    - Usage Patterns: OverageMinutes, RoamingCalls.
    - Demographics: IncomeGroup, MaritalStatus, Occupation, CityCode, ChildrenInHH.
  - This selection is driven by our goal to create a predictive model that is both accurate and insightful, focusing on the variables most indicative of churn behavior.
- EDA Summary
  - Demographic Trends: Certain demographics, including income group, marital status, and the presence of children, have a pronounced impact on churn rates.
  - Service Usage and Interaction: There's a clear link between service usage patterns (such as monthly minutes and revenue) and churn, with higher incidences of churn among customers exhibiting lower usage or higher levels of service-related issues.
  - Geographic Variance: Churn rates vary significantly across different geographic regions, suggesting the influence of local competition and market saturation.
  - The insights derived from the EDA emphasize the multifaceted nature of customer churn, informing our approach to predictive modeling by highlighting the importance of considering a broad range of factors.

PRE-PROCESSING FOR MODEL TRAINING

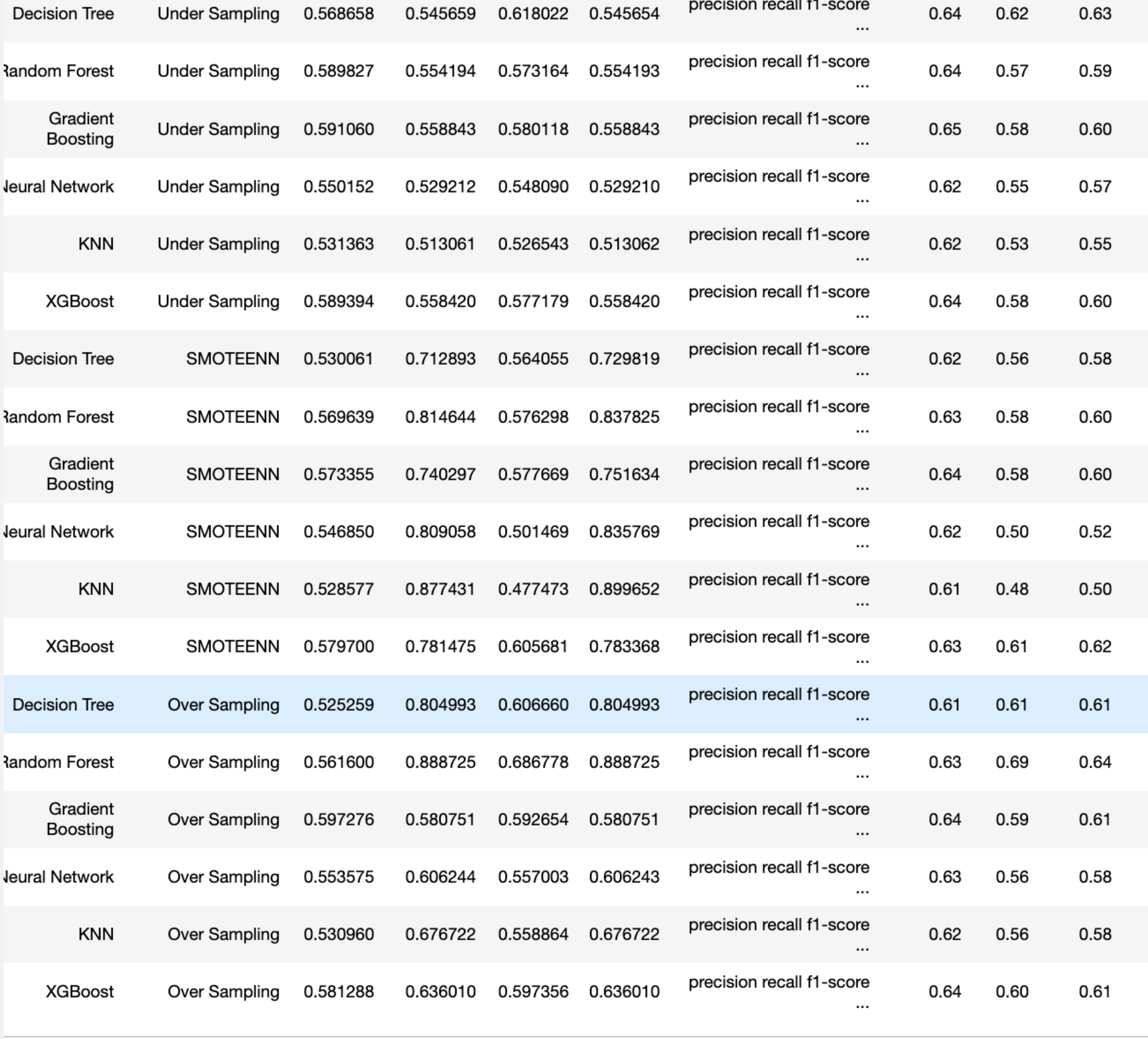
- Customer Segmentation
  - **K-Means Clustering:** To better understand our customer base, we employed K-Means clustering, a method that groups customers into clusters based on similarities in their data.
  - **Process & Outcome:** Relevant features such as service usage metrics and demographic information for the clustering process. Using the elbow method, the optimal number of clusters can be determined. This resulted in distinct customer segments, each with unique characteristics that inform tailored retention strategies.



- Feature Engineering
  - **Encoding Categorical Variables**: Applied one-hot encoding to transform categorical data (e.g., CityCode, MaritalStatus) into numerical formats for machine learning models.
  - **Refinement and Selection:** Removed redundant variables, focused on variables most predictive of churn. Created new features to better capture customer behavior and service usage.
  - **Preparing Dataset Split:** Training and Testing Sets: Data split into an 80% training set and a 20% testing set to accurately evaluate model performance.

MODELLING

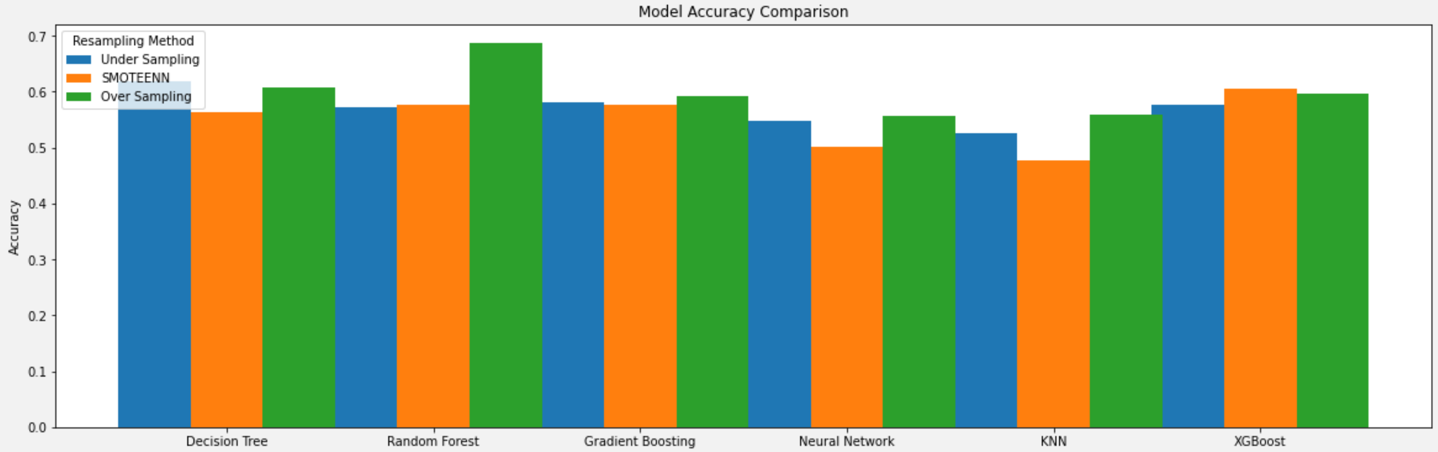
- Model Selection
  - Evaluated several machine learning models with Hyper-parameter Tuning to determine the most effective approach for predicting customer churn, including:
    - Decision Trees
    - Random Forest
    - Gradient Boosting
    - Neural Networks
    - K-Nearest Neighbors (KNN)
    - XGBoost
  - **Resampling Techniques:** Given the imbalance in our dataset between churned and non-churned customers, we applied resampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) and random under-sampling to balance the classes. This approach improves model sensitivity to churn by ensuring that the minority class is adequately represented.

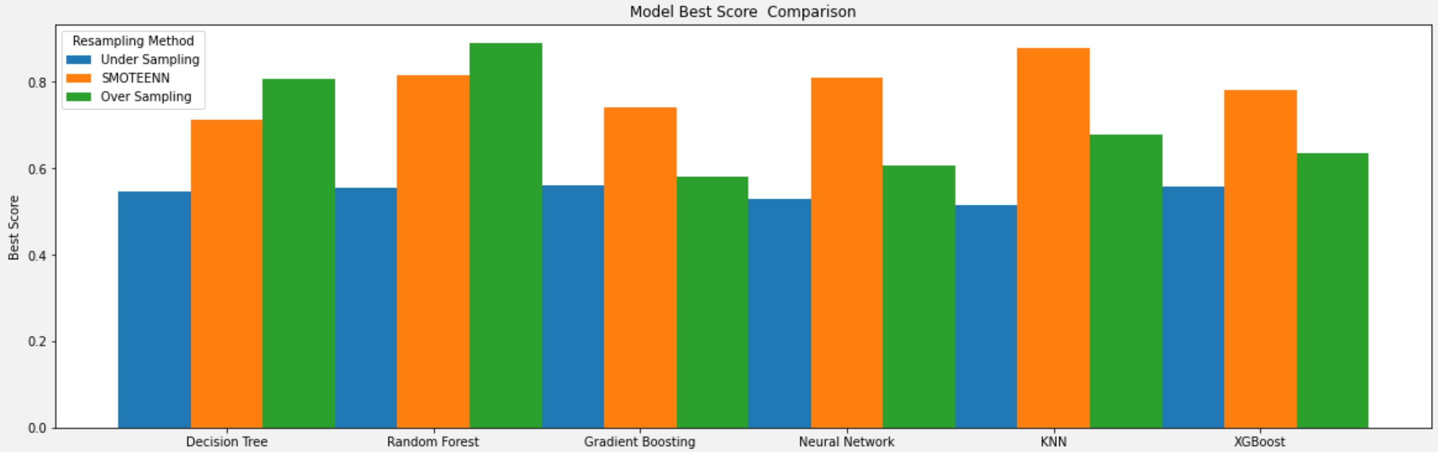


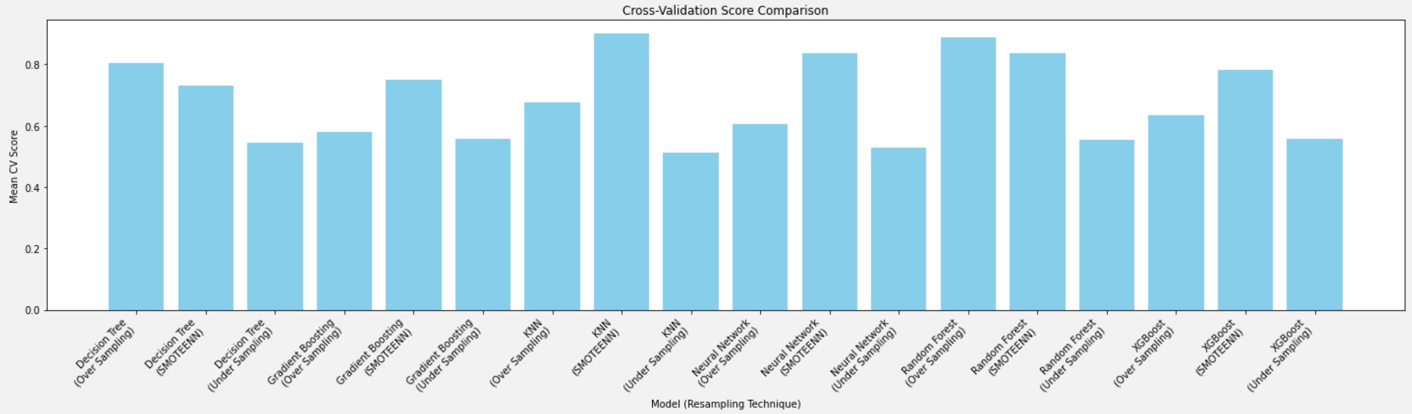
Decision Tree	Under Sampling	0.568658	0.545659	0.618022	0.545654	precision recall f1-score ...	0.64	0.62	0.63
Random Forest	Under Sampling	0.589827	0.554194	0.573164	0.554193	precision recall f1-score ...	0.64	0.57	0.59
Gradient Boosting	Under Sampling	0.591060	0.558843	0.580118	0.558843	precision recall 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

MODEL EVALUATION

- **Evaluation Metrics:** We assessed model performance using several key metrics, including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). These metrics provide a comprehensive view of model effectiveness, particularly in predicting churn accurately.
- **Final Model Selection:** Based on our evaluation criteria, the Random Forest Classifier, augmented with hyper-parameter tuning and oversampling techniques, emerged as the final model. This model demonstrated a strong balance between accuracy and generalizability, as evidenced by its performance metrics:
- High accuracy in distinguishing between churned and non-churned customers.
- A commendable AUC score, indicating excellent model sensitivity and specificity.
- The selection was grounded in the model's ability to deliver robust predictions while managing the complexities of our churn prediction task.


Model Accuracy Comparison

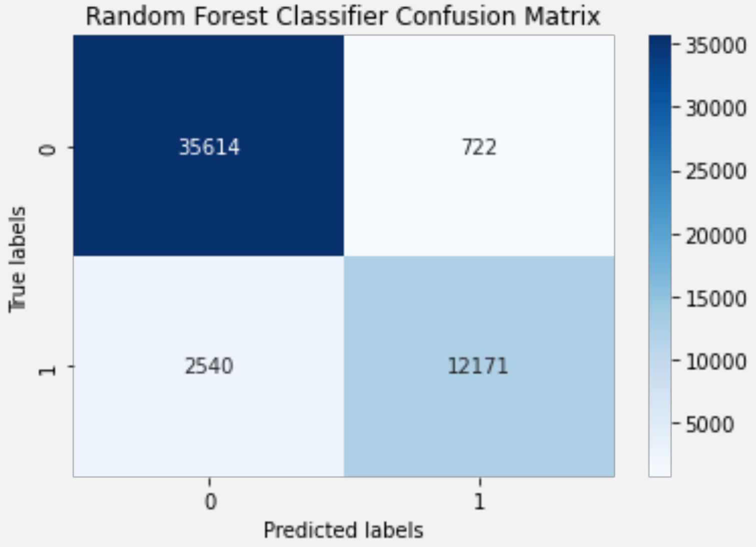

Model Best Score Comparison

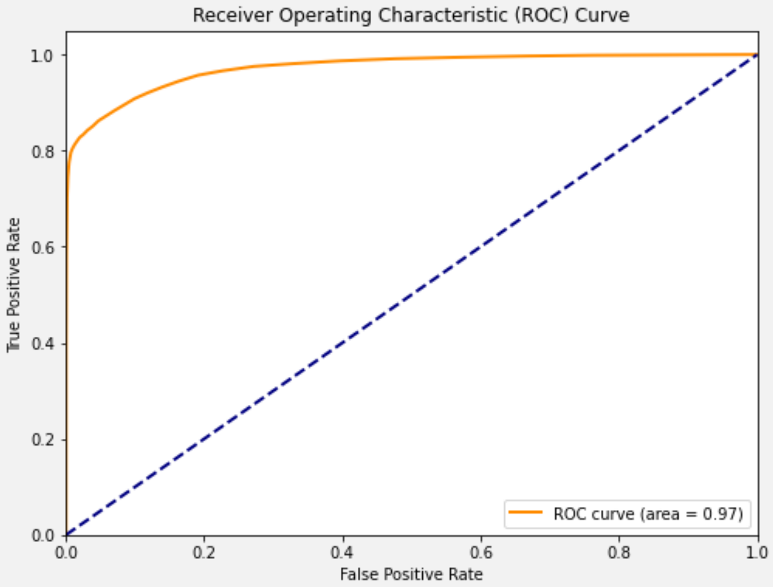

Cross-Validation Score Comparison

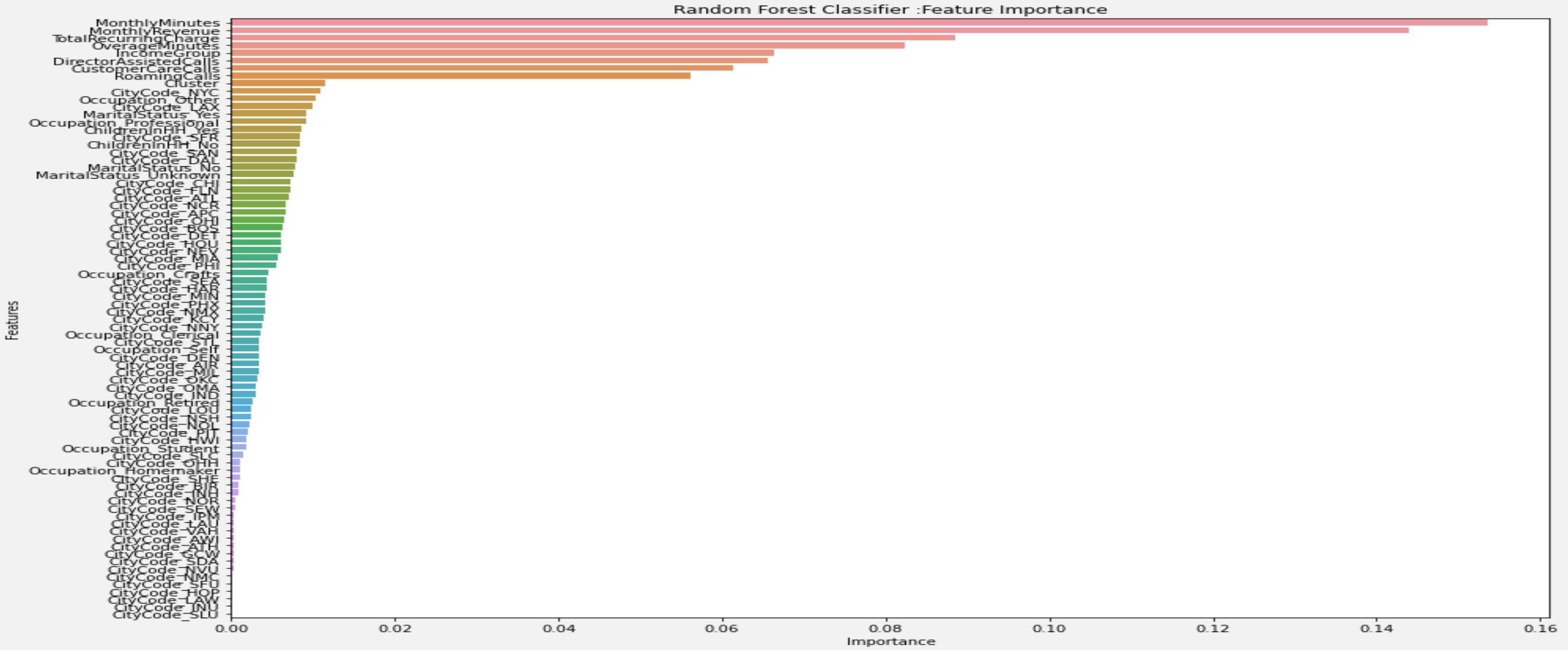
PREDICTION OUTCOME

The final model, the Random Forest Classifier, showcased impressive predictive outcomes, underlining its effectiveness in identifying potential churn customers.

- **Model Accuracy:** Demonstrated a high level of accuracy in classifying customers, ensuring reliable churn predictions.
- **ROC AUC Score:** Achieved an AUC score near 0.97, indicating exceptional model performance in distinguishing between churned and non-churned customers. A score closer to 1 signifies a high true positive rate and a low false positive rate.
- **Feature Importance:** Analysis revealed that service usage metrics (e.g., MonthlyMinutes, MonthlyRevenue) and customer interaction variables (e.g., CustomerCareCalls) were among the most significant predictors of churn. This insight directs focus towards critical areas for intervention to reduce churn.







RISK SEGMENTATION

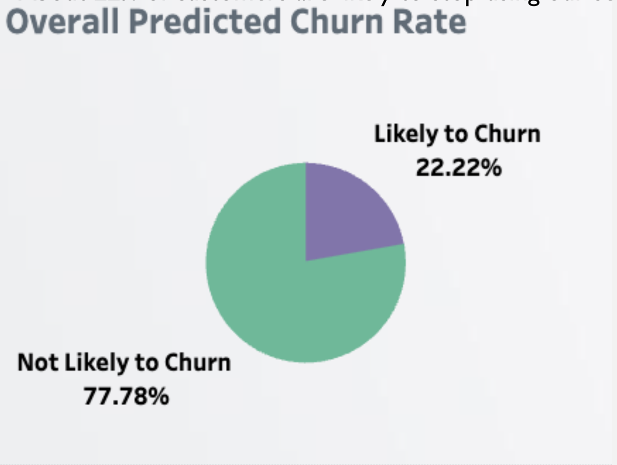
- Leveraging the predictive model, the customer base into distinct risk categories, enabling targeted retention strategies tailored to each group's likelihood of churn:
  - Low Risk (Probability of Churn < 30%): For customers with the lowest likelihood of churning, maintain satisfaction through regular service quality checks and personalized offers.
  - Medium Risk (Probability of Churn between 30% and 70%): Customers in this segment require more direct engagement strategies, such as personalized communication highlighting new features or loyalty programs that address their specific usage patterns.
  - High Risk (Probability of Churn > 70%): For those at highest risk, deploy intensive retention efforts, including special discounts, proactive customer service outreach, and customized plans designed to address their reasons for potential churn.
- By implementing these targeted strategies based on churn risk segmentation, we can more effectively allocate resources to retain customers and ultimately reduce overall churn rates.

DATA VISUALIZATION

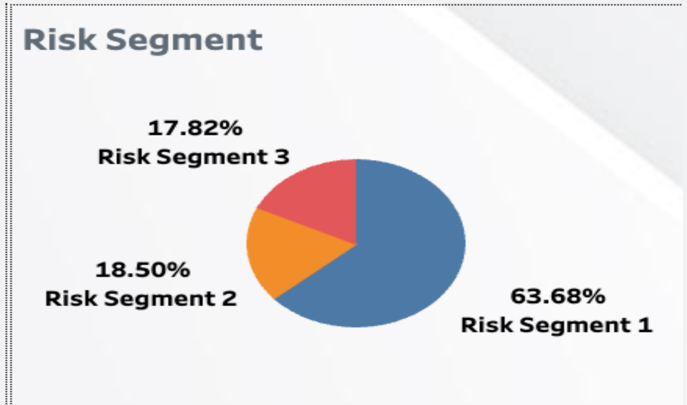
A series of compelling visualizations are designed using the prediction model data . Each graph illustrate key aspects of customer behavior, churn risk, and the impact of retention strategies. These visual insights are crucial for understanding the dynamics of customer churn and guiding strategic decisions.

https://public.tableau.com/app/profile/thant.thiri.kyi/viz/TelecomChurn_17094319695010/Story1

Churn Overview:
About 22% of customers are likely to stop using our services (churn).

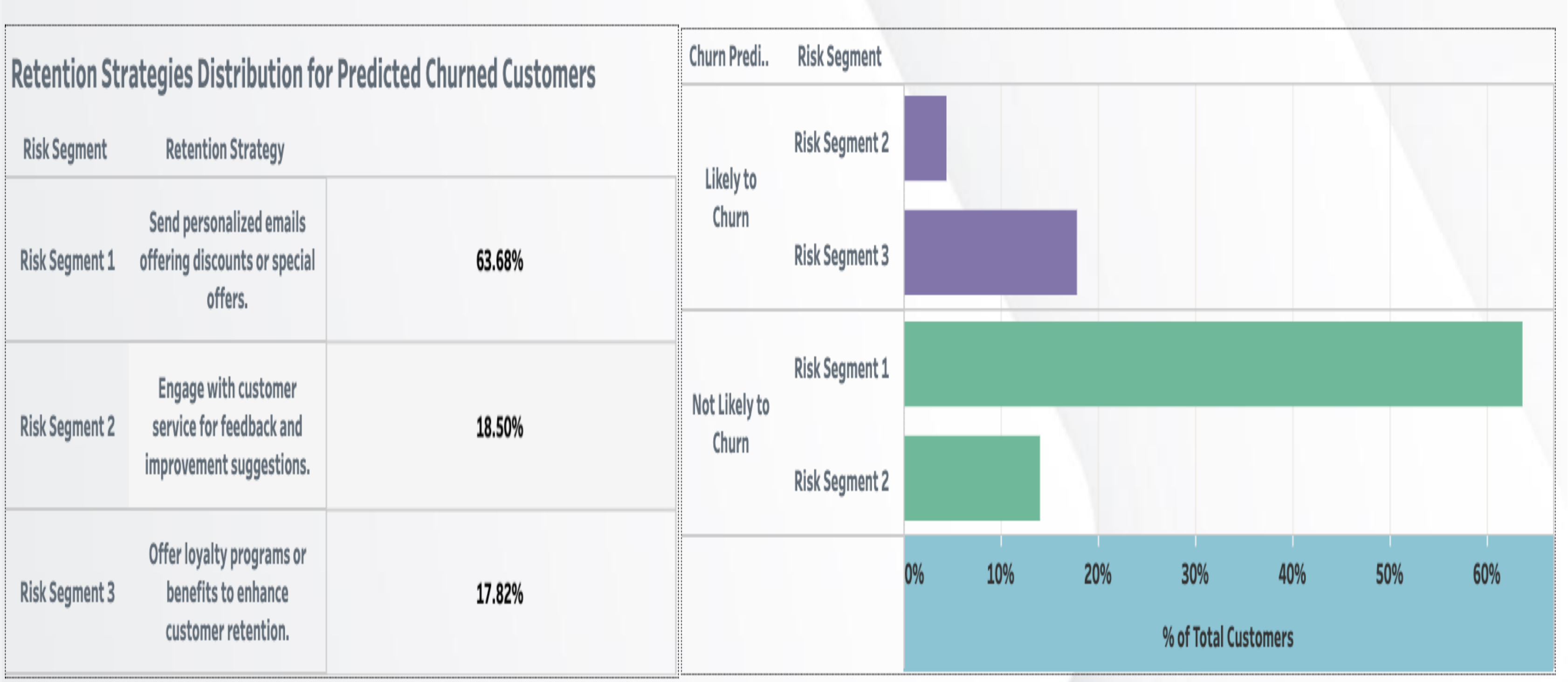

Overall Predicted Churn Rate

Churn Risk Segments:
Majority, nearly 64%, are in the low-risk group (Risk Segment 1).
About 19% are in the high-risk group (Risk Segment 3).


Risk Segment

Retention Strategies:
For low risk (Segment 1), sending personalized emails with deals.
For medium risk (Segment 2), asking for feedback and suggestions.
For high risk (Segment 3), offering loyalty programs.

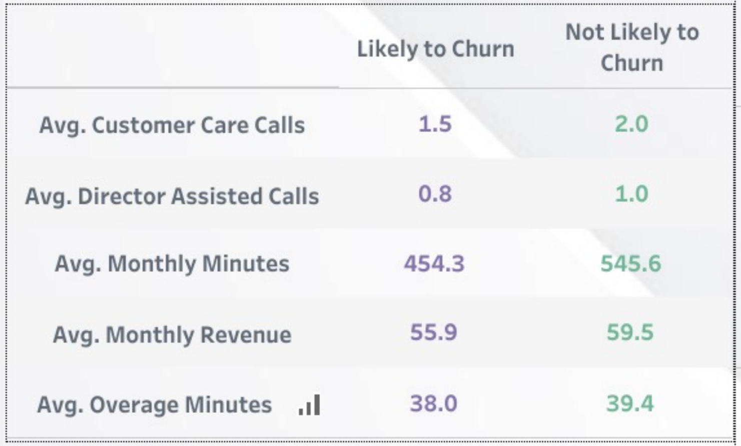
By comparing churn rates or customer satisfaction scores prior to and following the implementation of strategies, we can visually assess the impact of our actions.


Retention Strategies Distribution for Predicted Churned Customers

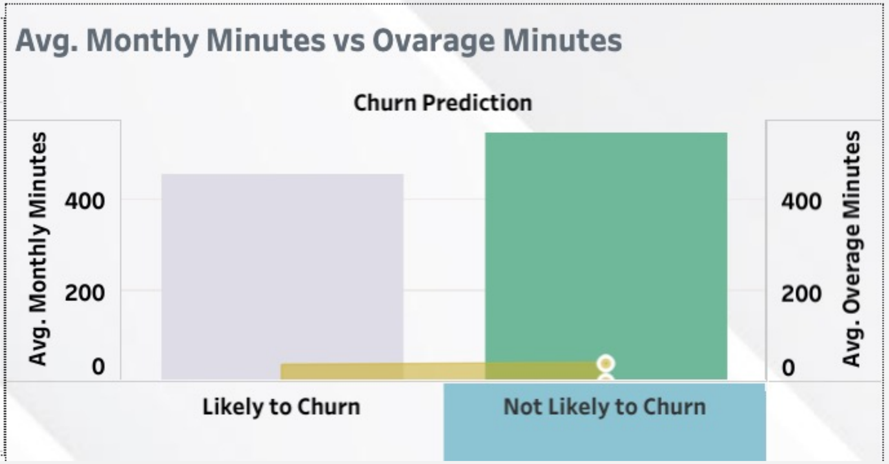
Risk Segment	Retention Strategy	
Risk Segment 1	Send personalized emails offering discounts or special offers.	63.68%
Risk Segment 2	Engage with customer service for feedback and improvement suggestions.	18.50%
Risk Segment 3	Offer loyalty programs or benefits to enhance customer retention.	17.82%

DATA VISUALIZATION

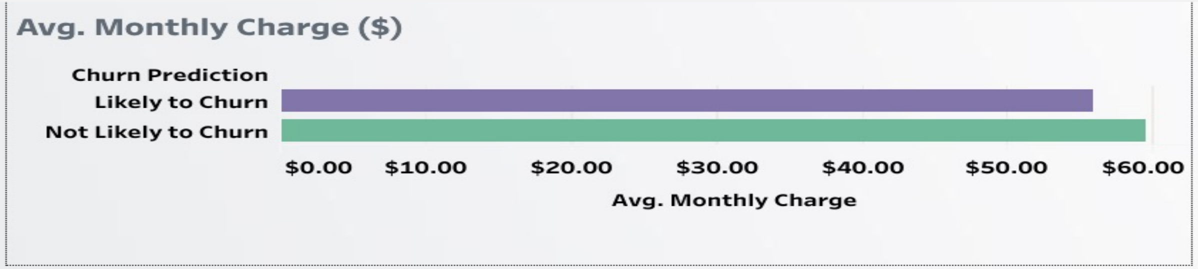
- Churn by Behavior:
  - Customers likely to churn make fewer customer care calls and have fewer director-assisted calls compared to others.
  - On average, they use more minutes and have higher monthly overages.
- Churn by Charges:
  - Higher total recurring charges link to a higher chance of churn.



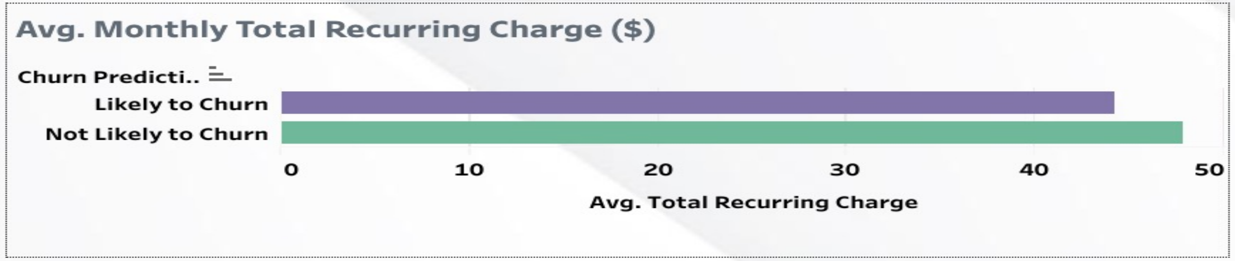
	Likely to Churn	Not Likely to Churn
Avg. Customer Care Calls	1.5	2.0
Avg. Director Assisted Calls	0.8	1.0
Avg. Monthly Minutes	454.3	545.6
Avg. Monthly Revenue	55.9	59.5
Avg. Overage Minutes	38.0	39.4



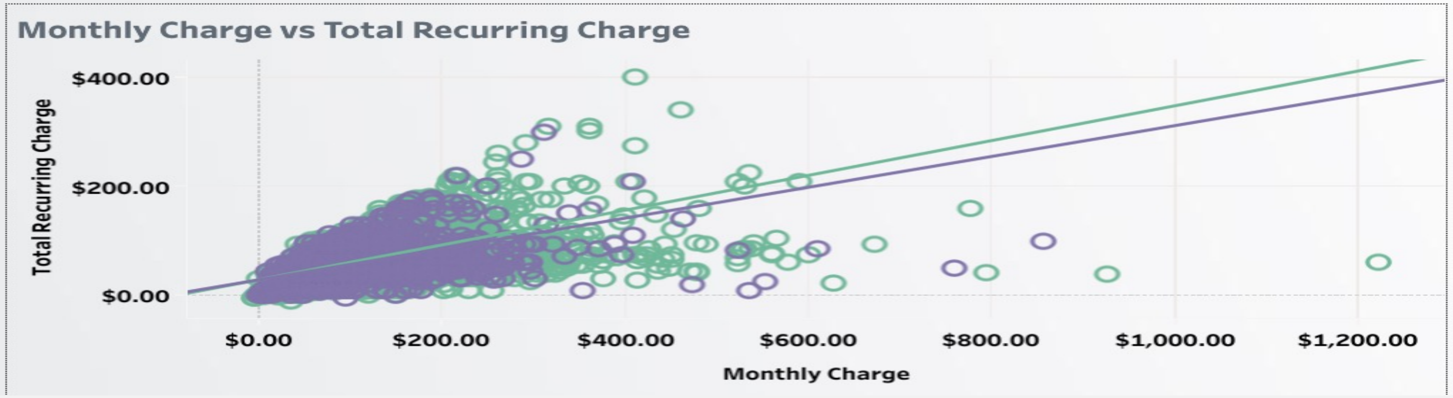
Avg. Monthy Minutes vs Ovarage Minutes



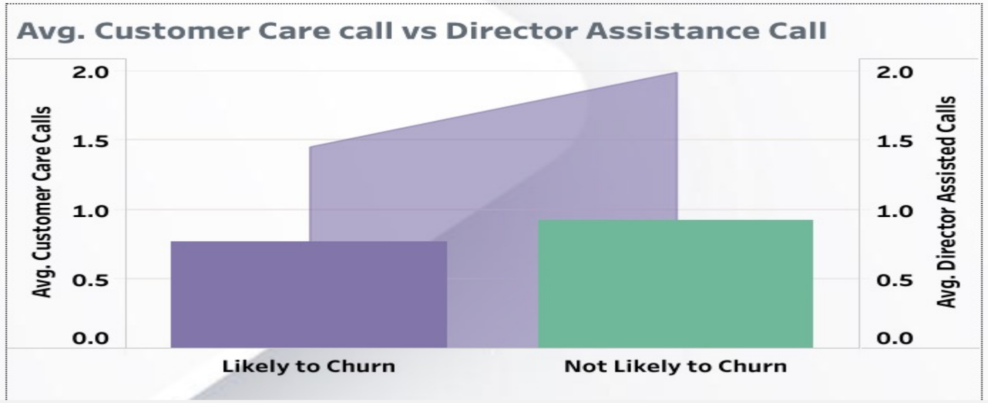
Avg. Monthly Charge ($)



Avg. Monthly Total Recurring Charge ($)



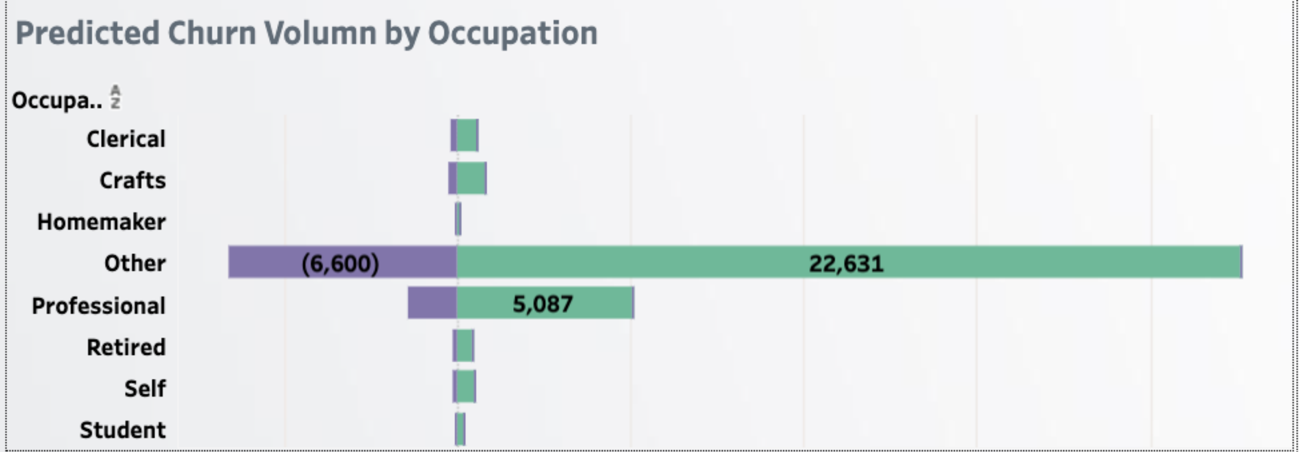
Monthly Charge vs Total Recurring Charge

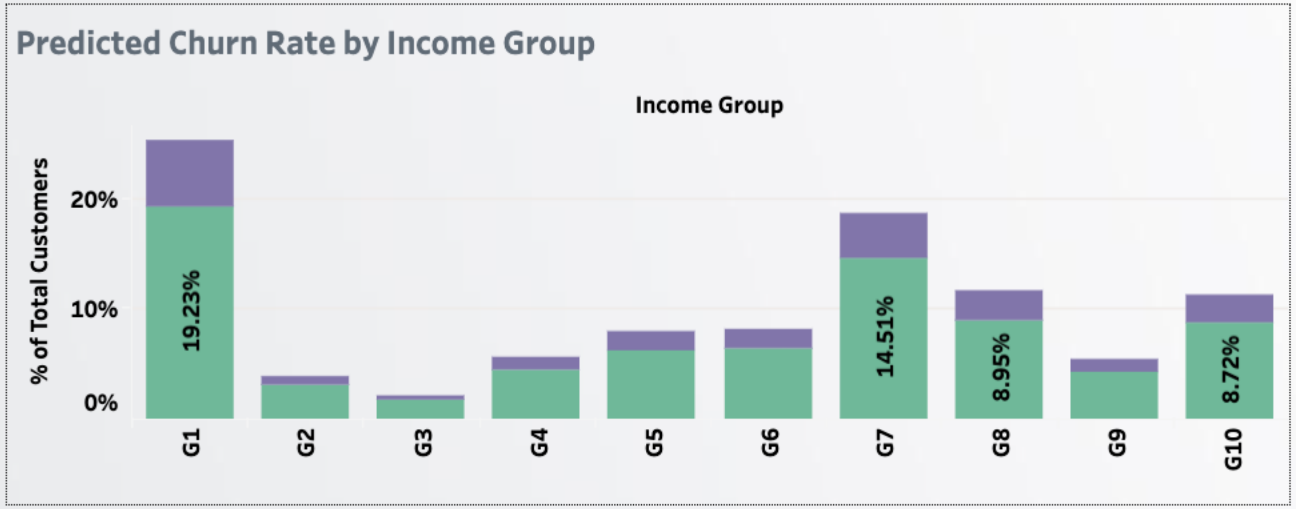


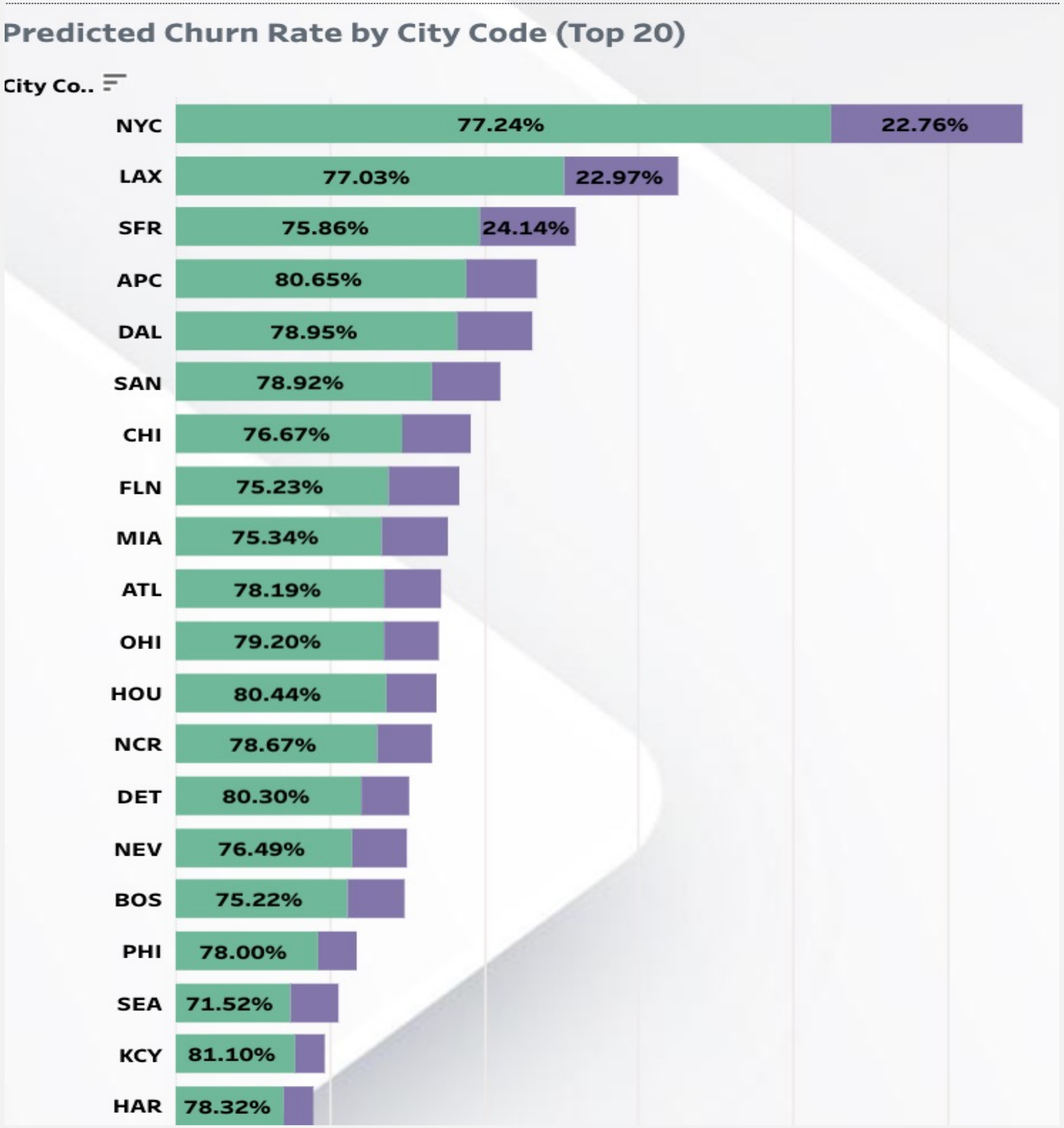
Avg. Customer Care call vs Director Assistance Call

DATA VISUALIZATION

- Churn by Demographics:
  - Churn does not vary significantly with marital status or children in the household.
  - The professional group has the highest number of customers likely to churn.
- Churn by Location:
  - City code 'KC' shows the highest likelihood of churn, over 81%.
- Income and Churn:
  - Lower-income groups show a higher rate of churn, especially the lowest income group at nearly


Predicted Churn Volumn by Occupation


Predicted Churn Rate by Income Group


Predicted Churn Rate by City Code (Top 20)

CONCLUSION

- Future Work - Continual Improvement Process
  - **Ongoing Model Evaluation and Updates:** Emphasize the necessity of regularly assessing the model's performance over time, incorporating new data, and adjusting to changes in customer behavior and market conditions.
  - **Feedback Loops:** Establish mechanisms for capturing the outcomes of retention strategies, allowing for data-driven refinements to both the predictive model and the strategies themselves.
  - **A/B Testing of Strategies:** Implement systematic testing of different retention interventions to quantitatively measure their effectiveness, informing more strategic deployment of retention efforts.
- Strategic Importance of Adaptability
  - **Adapting to Market Dynamics:** Acknowledge the rapidly changing telecom landscape and the importance of agility in both analytical approaches and customer engagement strategies.
  - **Leveraging Insights for Competitive Advantage:** Utilize the insights gained from predictive modeling and data analysis to not only reduce churn but also to enhance overall customer experience, driving competitive differentiation.
- Conclusion
  - The journey into predictive modeling and churn analysis marks the beginning, not the endpoint, of our efforts to understand and mitigate customer churn. The path forward involves continuous exploration, learning, and adaptation, ensuring our strategies remain effective and our customers stay engaged and loyal.

Q & A

- Thank you