IST 718: Big Data Analytics

Customer Segmentation and Churn Model for Telecom Industry

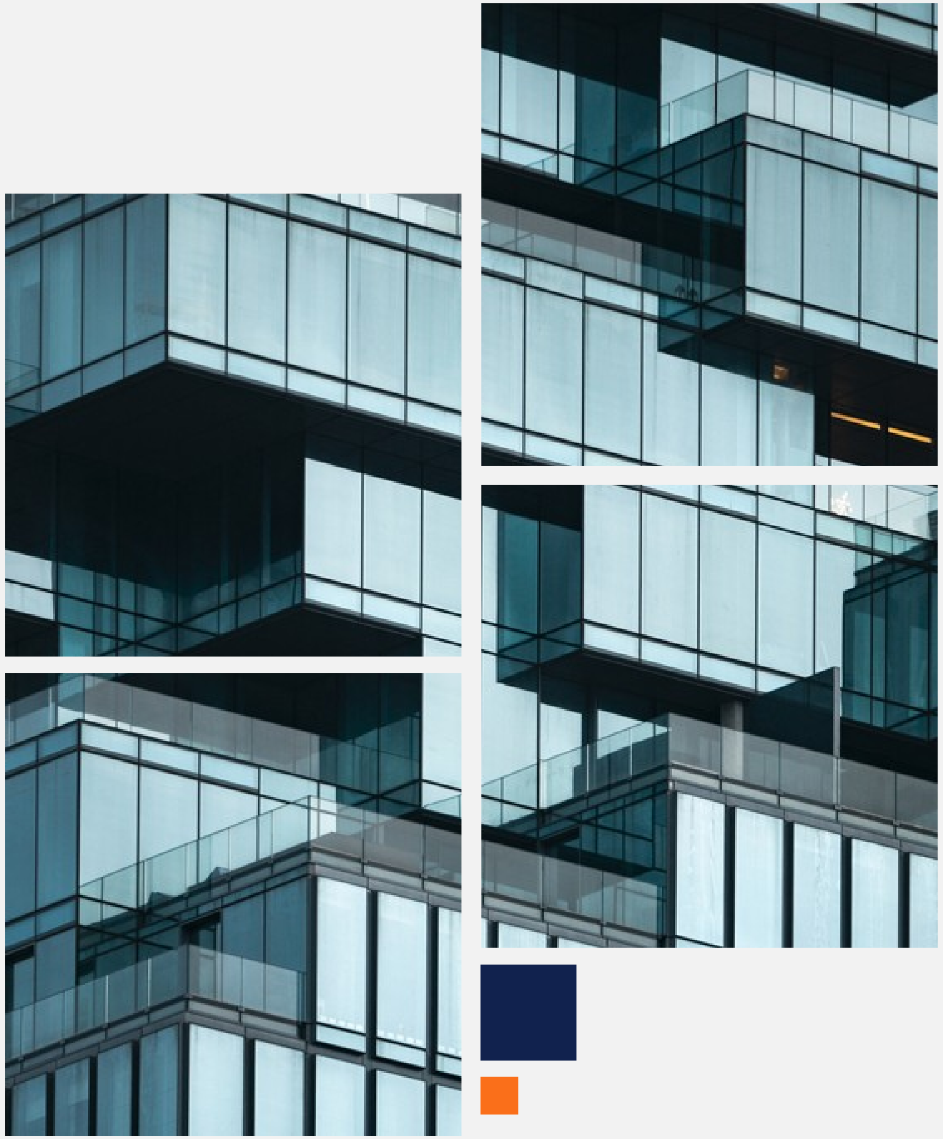
Group 3: Tejas Gawade, Harshit Joshi, Amisha Gangwar, Satwik Belaldavar

Agenda

- Introduction
- Data Exploration
- Prediction and Inference Goals
- Exploratory Data Analysis
- Predictions

Introduction

- The project focuses on customer segmentation and creating a churn model for the telecom industry.
- The main objective is to predict customer churn and manage it by identifying contracts likely to be canceled soon and targeting these customers with special incentives.
- Utilize advanced analysis and ML algorithms to uncover churn drivers, enabling the customization of retention incentives tailored to customer needs.



Data Exploration

Description

Dataset Description

- *Open-source data by Teradata center for customer relationship management at Duke University*
- *The Cell2Cell dataset includes customer demographics, billing information, usage statistics, and service interaction records*

Dimensions

Dataset Dimensions

- *The dataset comprises 71,047 instances (rows) and 58 attributes (columns)*

Features

Variables

- Predictors*: Total Recurring Charge, Customer Care Calls, Roaming, Handset Refurbished , Handset Web Capable*
- Target Variables*: Monthly Revenue, Churn, Overage Minutes*

Prediction and Inference Goals

Prediction Goals:

- Predict customer churn likelihood in the telecom industry by developing models that identify high-risk customers through analysis of usage patterns, billing history, and customer service interaction.
- Implement a model to forecast Monthly Revenue, aiding in recognizing revenue trends and identifying customers with potential changes in spending behavior.
- Aiming to forecast customers likely to exceed their plan minutes, enabling proactive adjustments to improve customer satisfaction and reduce churn

Inference Goals:

- Understanding Customer Behavior by analyzing data to understand the key factors that drive customer churn.
- Insight into Service Utilization through inferring patterns in service usage and customer satisfaction to identify areas for improvement in customer service and plan offerings.

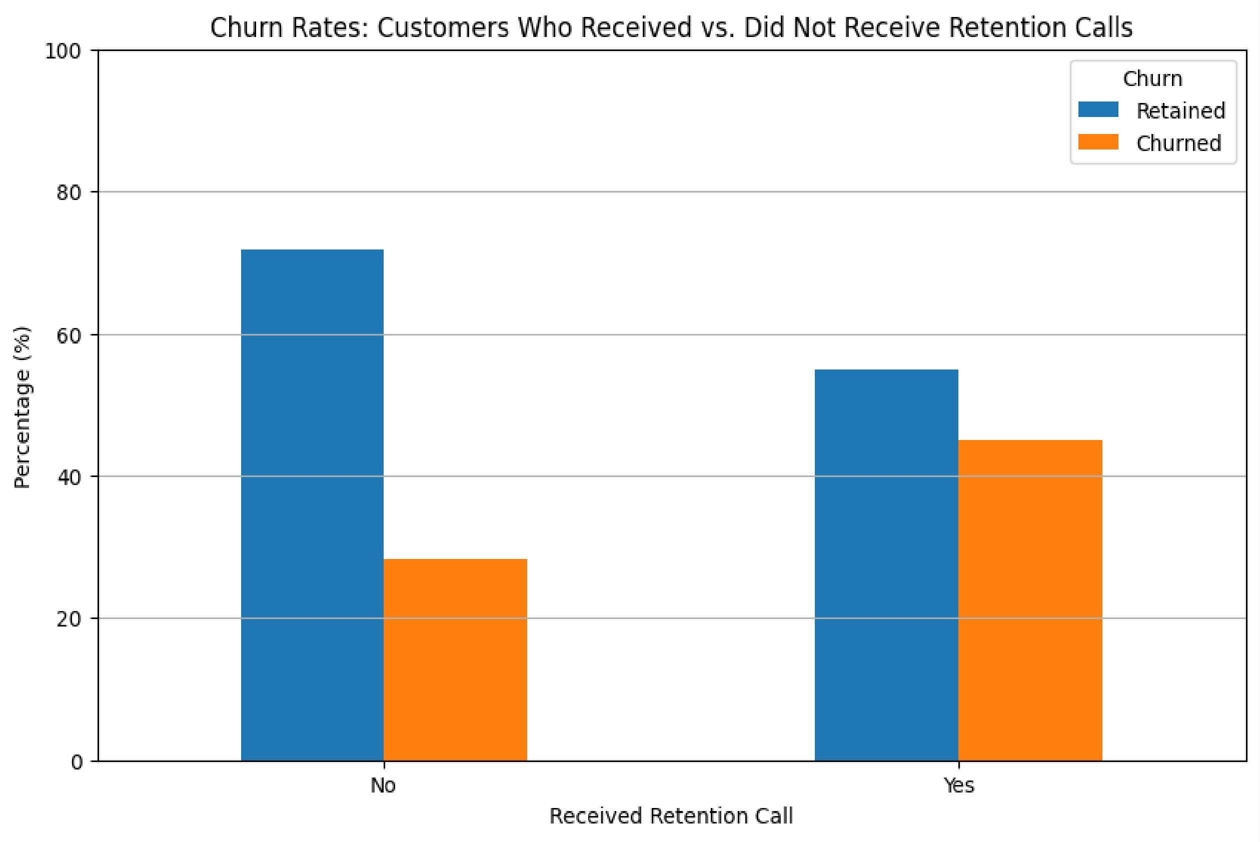
Exploratory Data Analysis

Customer Retention Efforts

- Retention calls lower churn, demonstrating their effectiveness in customer retention.

- Many customers stayed without retention calls, hinting at additional factors influencing loyalty.

- Retention calls reduce but don't eliminate churn, necessitating improved call quality and content analysis.

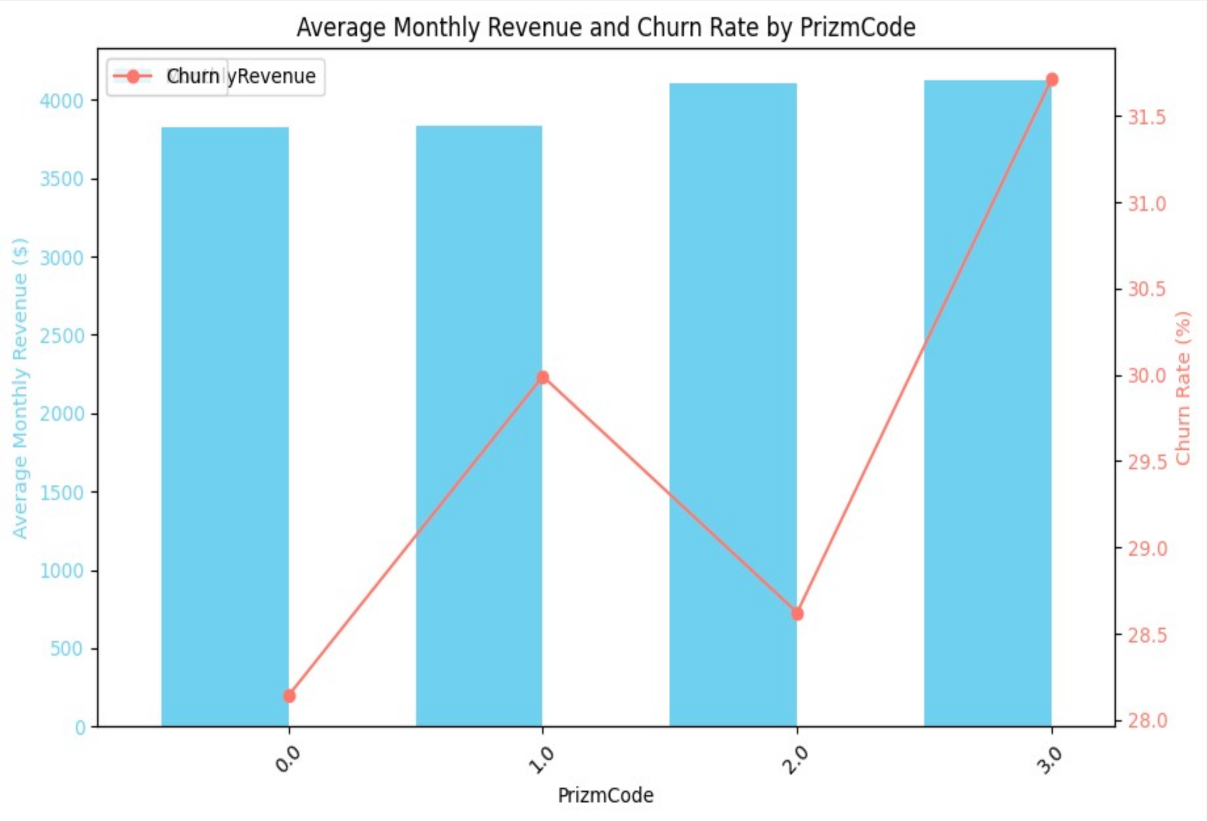


Demographic Insights

- The "Prizm Code" represents a classification of geographic areas into categories such as 'Suburban', 'Town', 'Other', and 'Rural' for demographic segmentation purposes.

- Analyze how different Prizm Code categories correlate with revenue generation and their respective churn tendencies.

- Utilize the preprocessed Cell2Cell dataset to derive meaningful insights on customer profitability and retention risks.



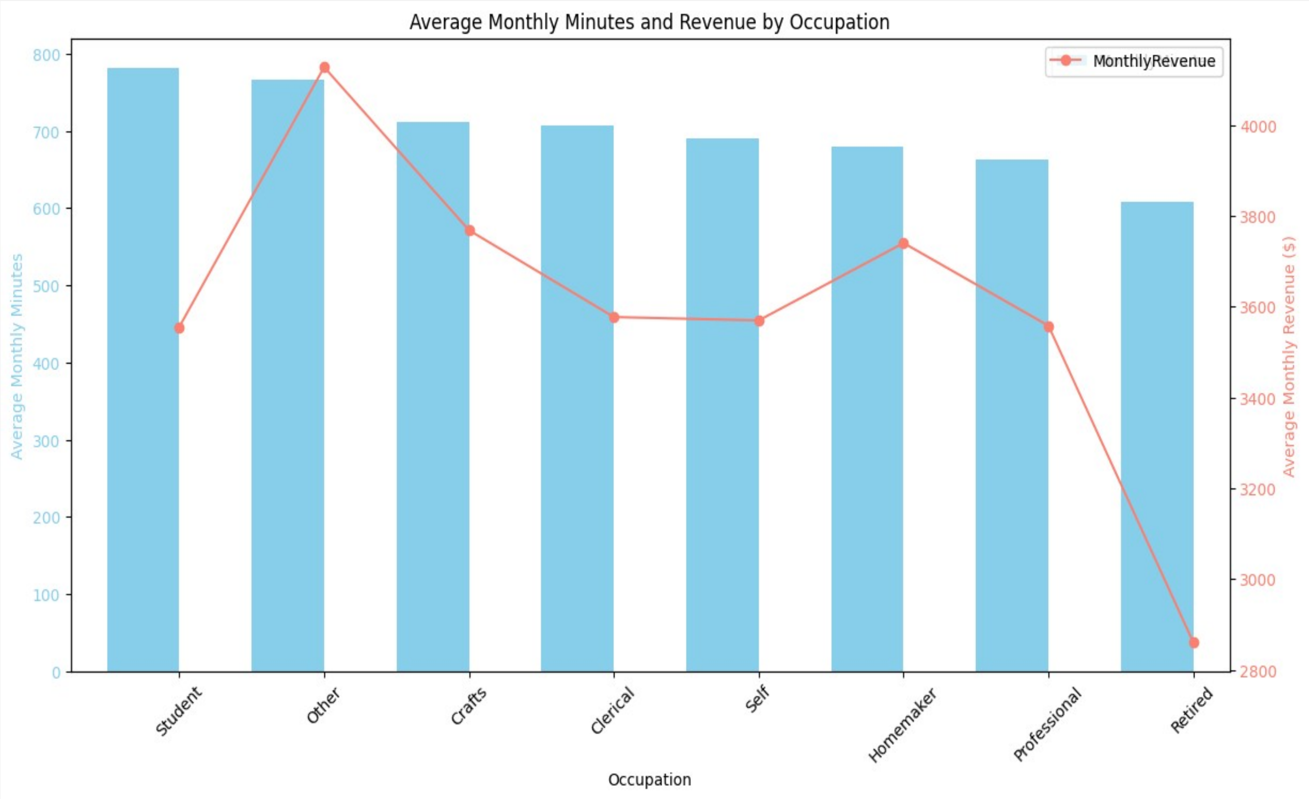
Occupation and Usage Patterns

- The attribute “Occupation” has unique values like student, crafts, clerical, self, homemaker, professional, retired and others

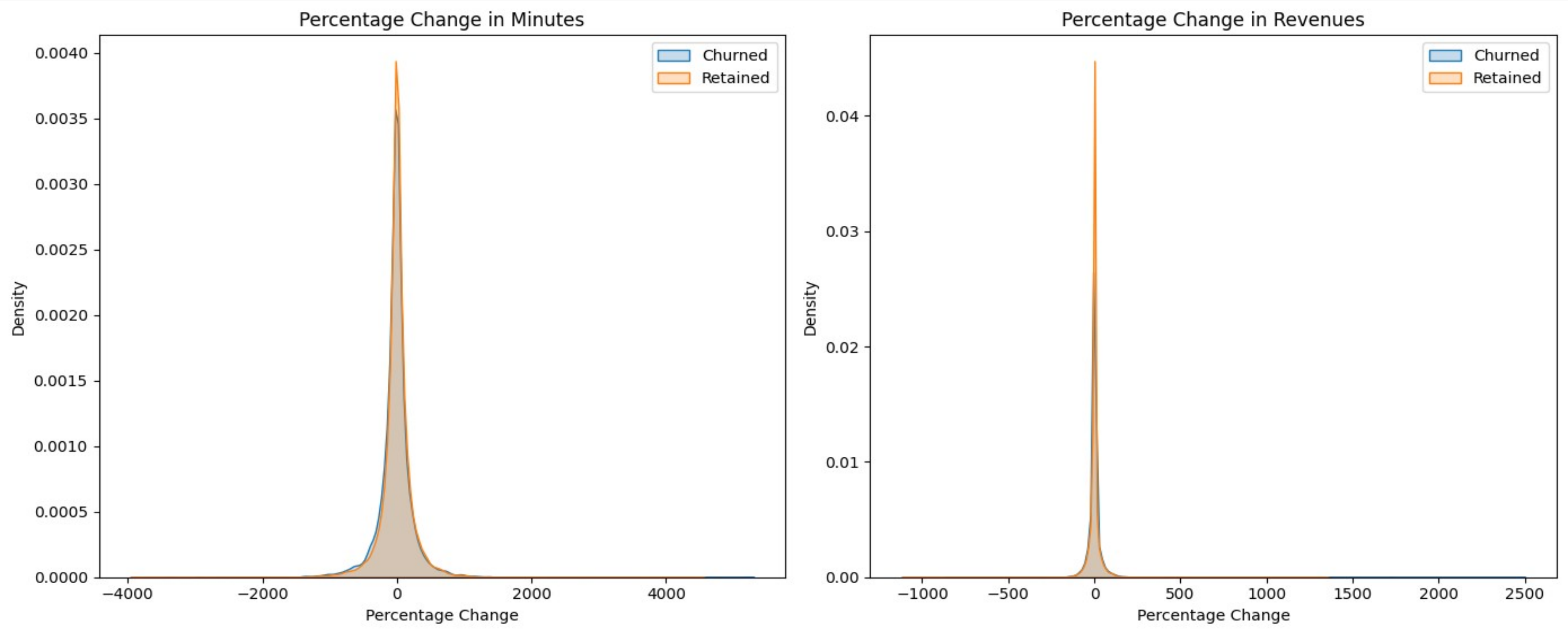
- The average monthly minutes used by customers varies across different occupations, with some occupations like 'Student' and 'Other' showing the highest usage.

- There is not a direct correlation between usage in minutes and revenue generated across occupations.

- There is a noticeable drop in both usage and revenue in the 'Professional' and 'Retired' categories, suggesting a potential area to investigate for retention strategies.



Trend Analysis

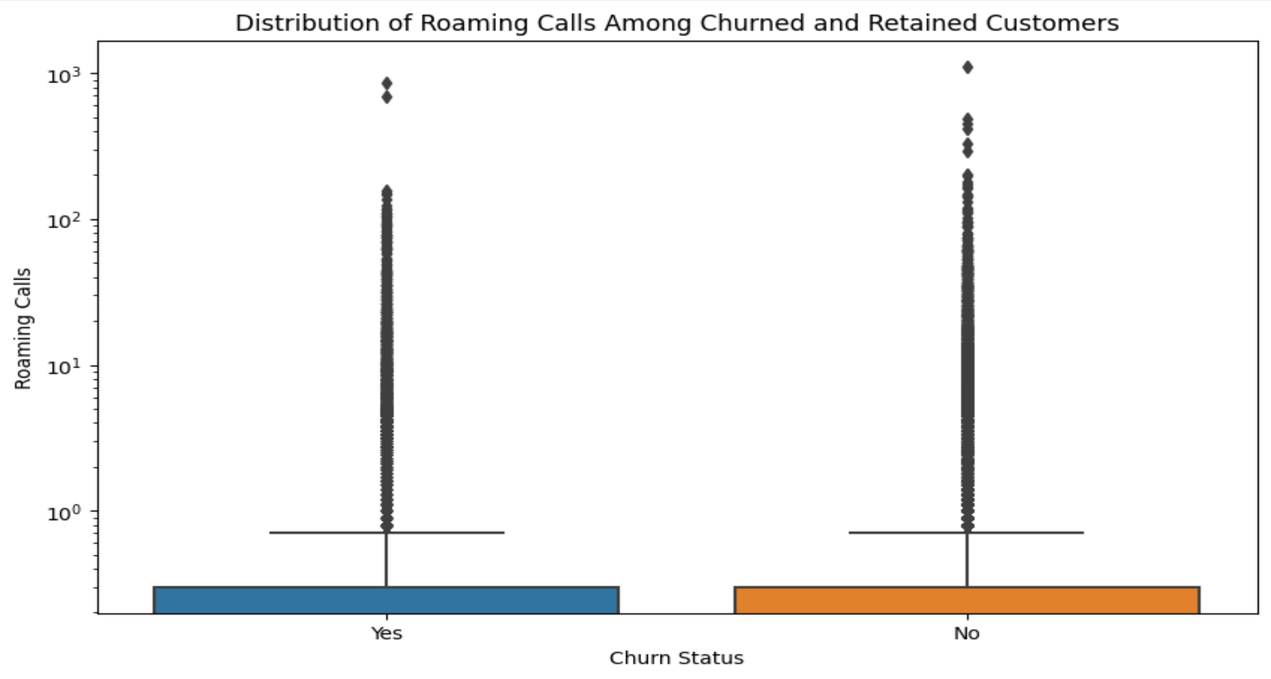


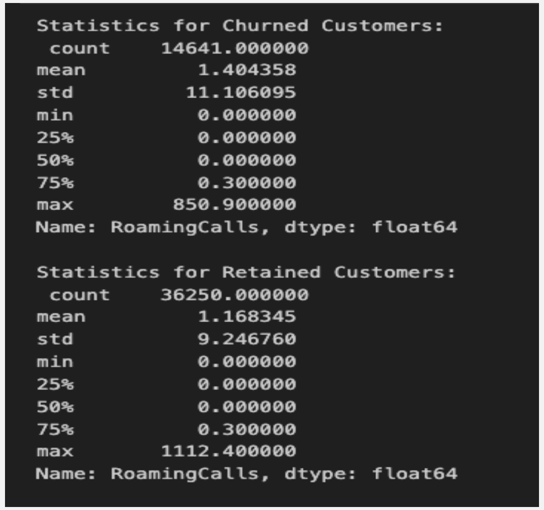
- The left-graph shows a spike around zero for both churned and retained customers, suggesting that most customers do not experience significant changes in service usage time. However, the churned customers' curve is slightly broader, indicating that while many churned customers maintain their usage.
- In the right-side graph there is a sharp peak at zero for both groups, which implies that revenue change is minimal for most customers. However, the distribution for churned customers is slightly wider than for retained customers, hinting at a higher variability in revenue change among churned customers.

Roaming Analysis

- There are significant outliers in both churned and retained customer groups, indicating that some customers have an exceptionally high number of roaming calls, which could be a segment requiring special attention or tailored roaming plans.

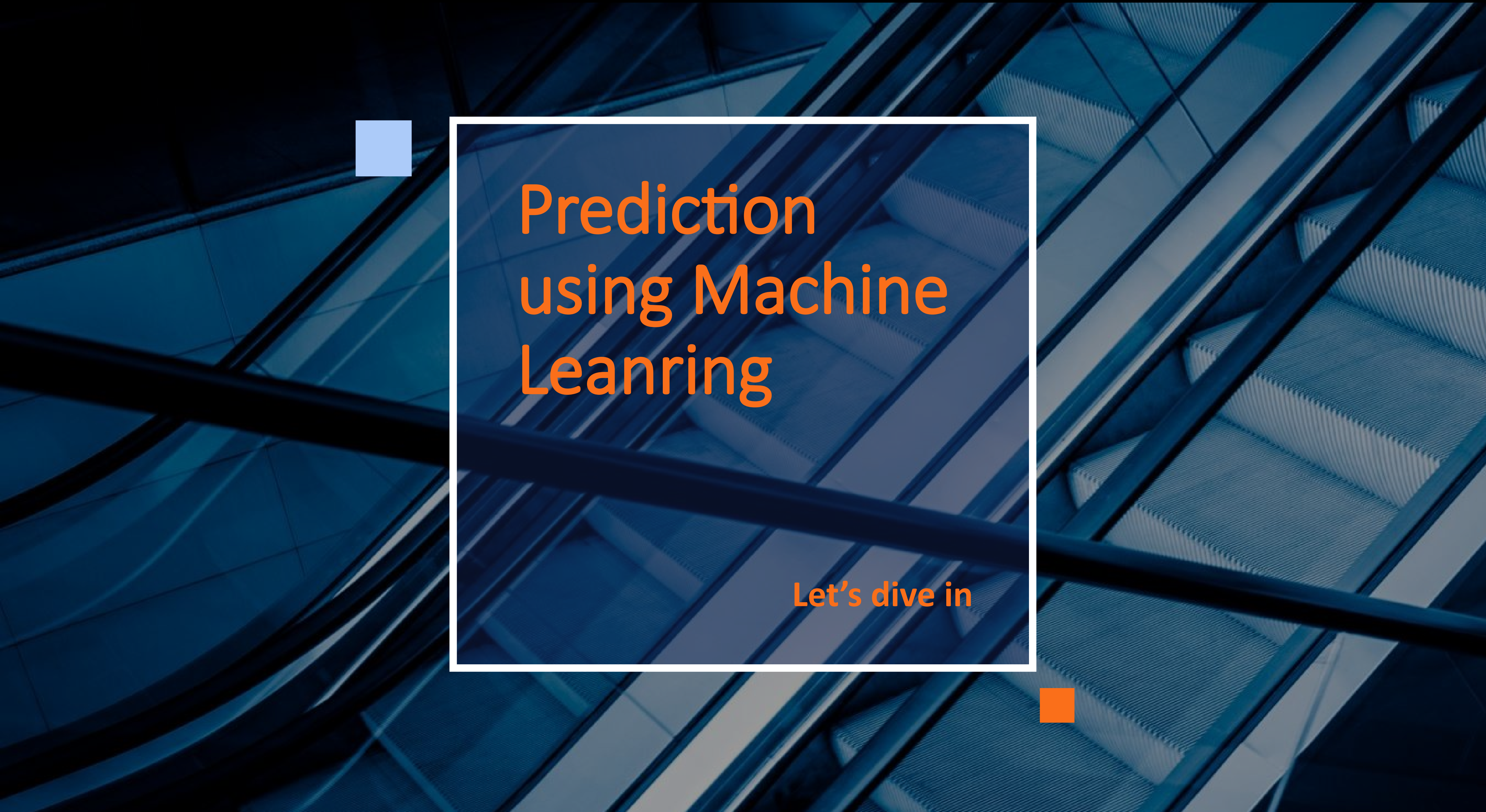
- Churned customers have a higher mean of roaming calls compared to retained customers, suggesting that roaming service experience or costs could be factors in customer churn.





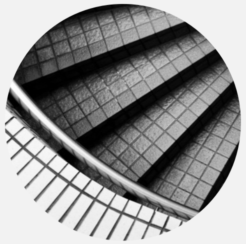
```
Statistics for Churned Customers:
 count     14641.000000
mean          1.404358
std          11.106095
min           0.000000
25%           0.000000
50%           0.000000
75%           0.300000
max         850.900000
Name: RoamingCalls, dtype: float64

Statistics for Retained Customers:
 count     36250.000000
mean          1.168345
std           9.246760
min           0.000000
25%           0.000000
50%           0.000000
75%           0.300000
max        1112.400000
Name: RoamingCalls, dtype: float64
```



Prediction using Machine Leanring

Let’s dive in



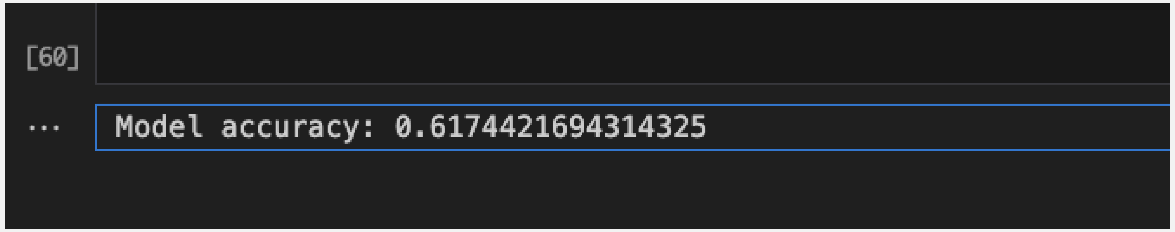
Customer Churn Prediction

Model Details

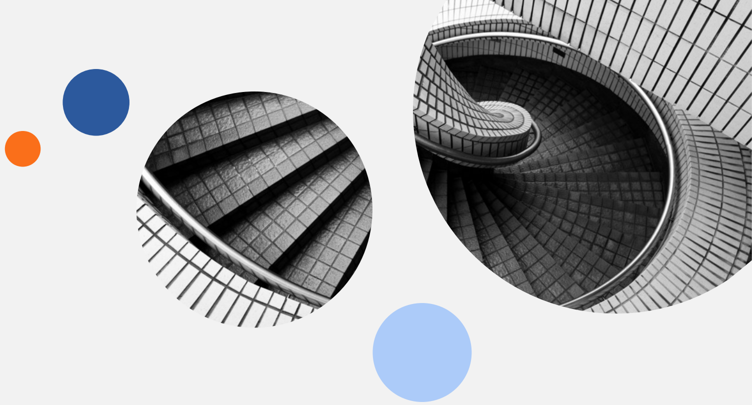
- Random Forest Classifier
- Data split into 80% training and 20% testing sets and model fitting on the training dataset.
- Data Preprocessing:
  - Binary encoding of the target variable 'Churn’.
  - Null value handling by dropping rows with nulls in key columns.
  - Feature engineering with both categorical and numerical data.

Model Evaluation

- Model accuracy obtained: 61.74%
- Evaluated using BinaryClassificationEvaluator



```
[60]
...   Model accuracy: 0.6174421694314325
```



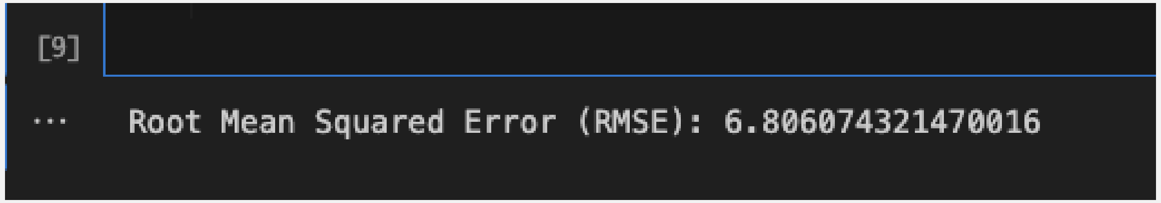
Monthly Revenue Forecasting

Model Details

- Random Forest Regressor
- Data split into 80% training and 20% testing sets and model fitting on the training dataset.
- Data Preprocessing:
  - Conversion of string to float where needed.
  - Handling null values with the 'keep' option in StringIndexer.
  - Combination of all feature columns into a single vector using VectorAssembler.

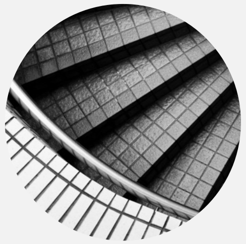
Model Evaluation

- Evaluation Metric: Root Mean Squared Error (RMSE)
- RMSE: 6.8067
- Interpretation: The RMSE value indicates the average error of the predictions made by the model.



```
[9]

...   Root Mean Squared Error (RMSE): 6.806074321470016
```



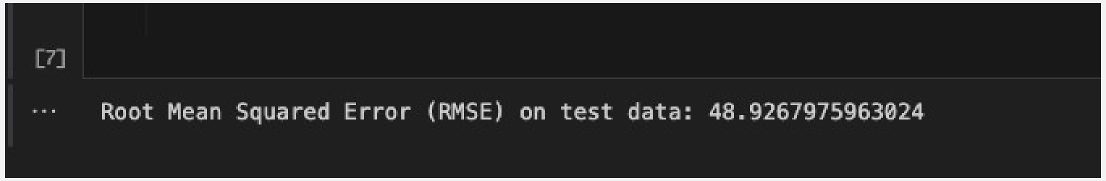
Overage Minutes Prediction

Model Details

- Random Forest Regressor
- Data split into 80% training and 20% testing sets and model fitting on the training dataset.
- Data Preprocessing:
  - Conversion of categorical variables using StringIndexer
  - Aggregation of features into a single vector using VectorAssembler

Model Evaluation

- Evaluation Metric: Root Mean Squared Error (RMSE)
- RMSE on test data: 48.93
- Interpretation: The RMSE value indicates the average error of the predictions made by the model.



```
[7]
...   Root Mean Squared Error (RMSE) on test data: 48.9267975963024
```

Conclusion

Prediction Outputs:

- Customer Churn Prediction:
  - Random Forest Classifier
  - Model accuracy obtained of 61.74%
- Monthly Revenue Forecasting:
  - Random Forest Regressor
  - RMSE 6.8067
- Overage Minutes Prediction:
  - Random Forest Regressor
  - RMSE on test data: 48.93

Challenges:

- Data Preprocessing Challenges: Handling of null values and data conversion (e.g., binary encoding, string to float conversions) might have been challenging, requiring careful consideration to maintain data integrity.
- The accuracy for the churn prediction model (61.74%) suggests there might be room for improvement, possibly due to model complexity, feature selection, or data quality issues.