IST 718: Big Data Analytics

Customer Segmentation and Churn Model for Telecom Industry

Final Project Report

Group 3:
Tejas Gawade
Amisha Gangwar
Satwik Belaldavar
Harshit Joshi

Project Overview

In the competitive landscape of the telecom industry, where constant connectivity is key, managing customer churn is essential. The churn prediction model developed using advanced analytics offers telecom operators a strategic edge. By pinpointing potential churn risks through customer behavior and demographics, operators can craft targeted retention strategies, improving loyalty and securing their market position in an industry where every customer's choice has a significant impact on revenue and long-term success.

This project introduces a churn prediction model to enhance customer retention strategies in the telecom sector. Using customer segmentation and machine learning within a Spark and Python framework, we analyzed patterns and predictors of churn, including service usage and demographic factors. Our data-driven approach led to the development of a Random Forest Classifier, known for its accuracy in predicting customer behavior and Random Forest Regressor.

Ultimately, our findings support targeted retention initiatives to mitigate churn and improve customer loyalty. This project not only contributes to academic knowledge but also provides practical solutions for the telecom industry, aiming to reduce turnover and strengthen market presence.

Data Exploration Overview

Our Data is 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

The dataset for this telecom project comprises 51,047 customer records, each with 58 attributes, providing a comprehensive view of customer profiles and interactions. Key features include roaming calls, monthly revenue, service usage in minutes, charges, call details, and percentages of change in usage. The dataset also encapsulates demographic information such as the primary and secondary account holder's age, household characteristics, and tenure of service.

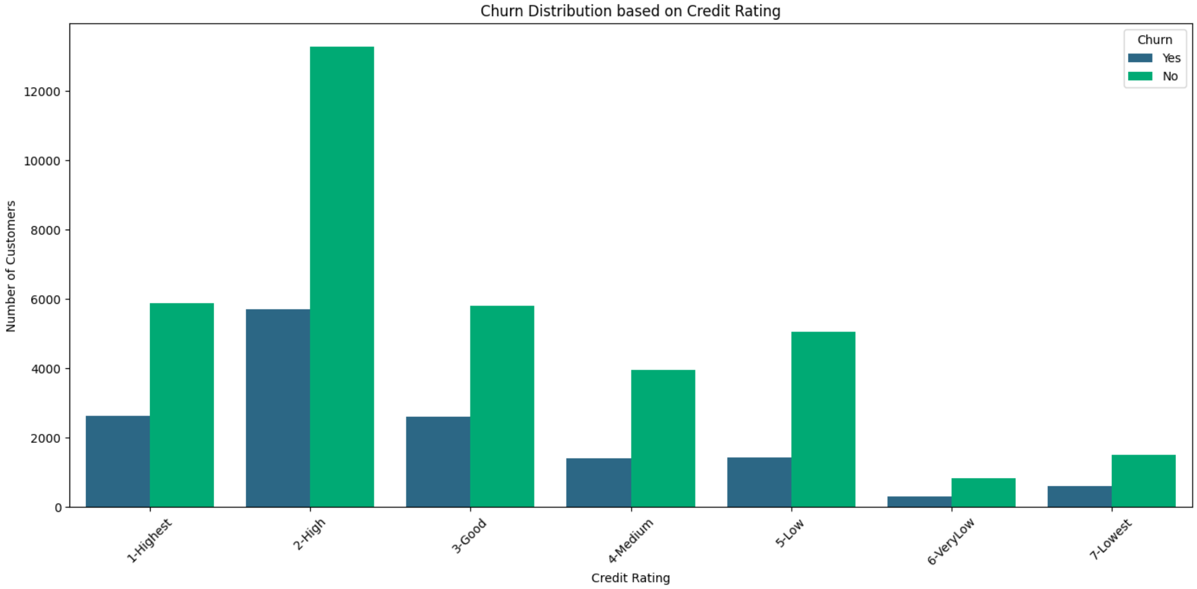
For the report, we would note that the exploratory data analysis aimed to identify key factors influencing customer churn. Visualization techniques such as bar charts for categorical data and histograms for continuous variables were employed to discern patterns and anomalies. The exploration set the stage for deeper analytics, enabling us to refine features for the churn prediction model.

Potential Variables for Machine Learning Prediction:

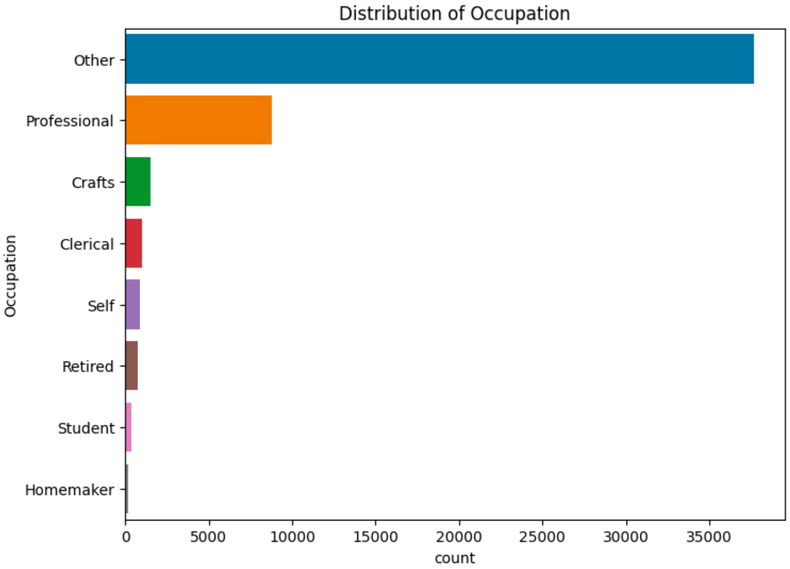
- Predictors*: Total Recurring Charge, Customer Care Calls, Roaming, Handset Refurbished, Handset Web Capable*
- Target Variables*: Monthly Revenue, Churn, Overage Minutes*

Inference Goals

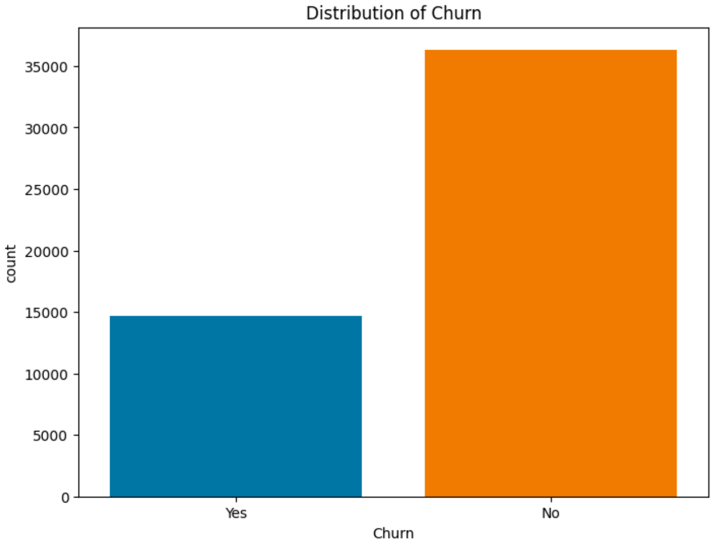
1) Identification of Key Churn Drivers:
The visualizations suggest that customers' credit ratings significantly impact churn, with majority of retained customers being in top 3 rating categories. This insight directs us to consider credit rating as a key driver of customer churn, possibly due to misalignment between service offerings and the financial expectations or capabilities of various customer segments.



2) Insights for Service Improvement:
The occupation distribution chart hints at a potential relationship between customers' occupations and their service usage patterns. This can inform targeted improvements in customer service and product offerings, especially to the professional and other high-frequency segments shown in the data.



3) Strategic Customer Retention Planning:
The stark contrast in churn numbers between the 'Yes' and 'No' categories underscores the need for strategic planning in customer retention initiatives. Understanding the specific factors that contribute to customer satisfaction, as indicated by the low churn in certain segments, can guide the development of more effective retention strategies.



Prediction Goals:

- Customer Churn Prediction

The overarching goal of this model will be to determine the likelihood of customer churn within the telecom industry. By leveraging a Random Forest Classifier, we aim to scrutinize customer data to identify those at high risk of ending their service. The dependent variable, churn, is a binary outcome signifying whether a customer will leave or stay, derived from the 'Churn' column in the dataset.

- Monthly Revenue Forecasting

The objective of this model will be to forecast monthly revenue in the telecom sector, which is essential for discerning revenue trends and pinpointing customers who might exhibit significant changes in spending patterns. The dependent variable for our predictions is 'MonthlyRevenue', a continuous measure that reflects the monetary value accrued from each customer per month.

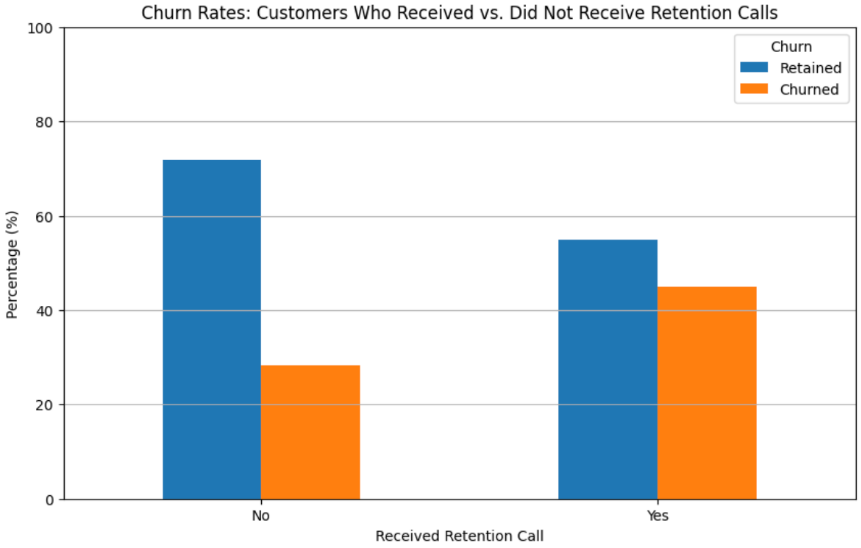
- Overage Minutes Prediction

This model will predict instances when customers are likely to exceed their allocated plan minutes. This forecast serves as a strategic tool for telecom companies to identify customers who may need plan adjustments or targeted communication to enhance satisfaction and minimize churn. The dependent variable in this scenario is 'OverageMinutes', quantifying the extent to which customers surpass their plan limits.

Exploratory Data Analysis:

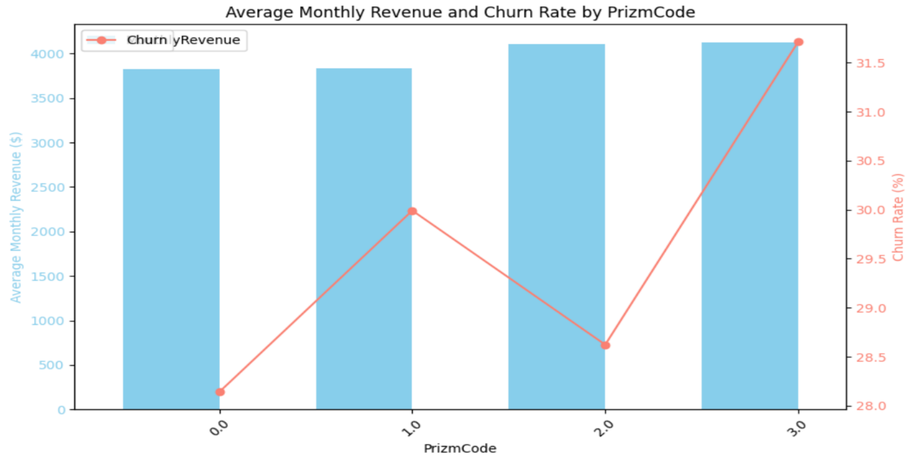
- Customer Retention Efforts

In our first data exploration, we find that retention calls play a vital role in mitigating customer churn. However, the persistence of churn despite these calls and the retention of many customers without such intervention points to other factors influencing customer loyalty. This suggests that while retention calls are useful, there's room to enhance their effectiveness and investigate other loyalty drivers.

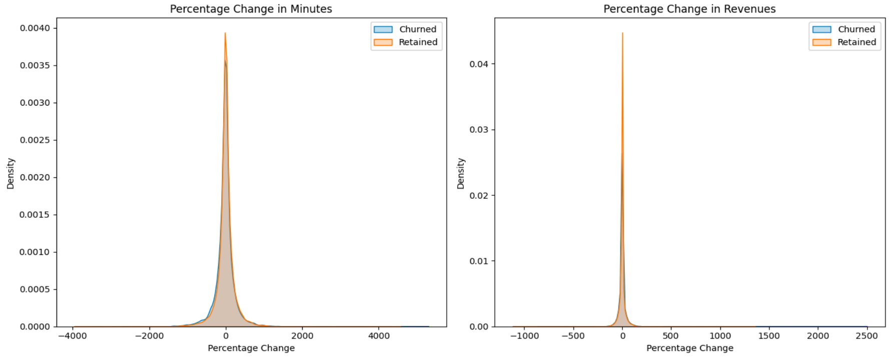


- Occupation and Usage Patterns

The chart presents an analysis of the Cell2Cell dataset, showing average monthly revenue and churn rate by "Prizm Code," which categorizes geographic areas for demographic segmentation. It indicates the relationship between demographic segments and revenue generation against customer churn tendencies. This analysis is crucial for identifying regions with high revenue but also high churn, informing strategies for targeted customer retention and profitability enhancement, which in this case is Prizm code 3.0. This data exploration can be vital for tailoring communication plans and service options to various occupational demographics to optimize usage and revenue.

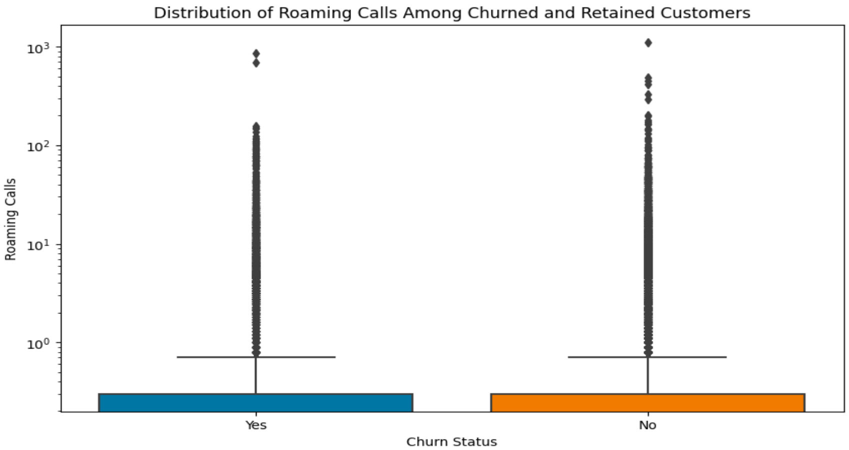


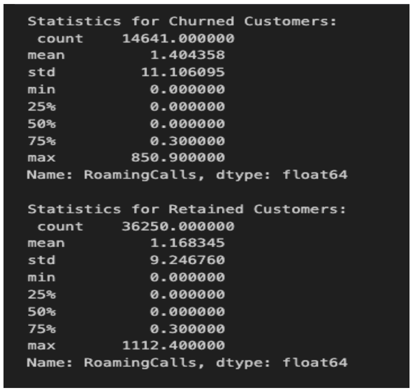
- Trend Analysis:



For analyzing the trend between Churned and retained customers, 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:
The 'Roaming Analysis' section of the report elucidates the distribution of roaming calls between churned and retained customers, as evidenced by the provided visualizations and statistics. Notably, there are outliers in both customer groups, with some individuals making an extremely high number of roaming calls; this suggests a niche market that may benefit from customized roaming plans. The statistical summary reveals that churned customers have a higher average number of roaming calls than those retained, implying that roaming service quality or cost could be contributing to customer turnover. These insights underscore the importance of analyzing roaming usage patterns to address potential dissatisfaction and to develop retention strategies tailored to high roaming users' needs.





```
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
```

Predictions:

- Customer Churn Prediction:

The 'Customer Churn Prediction' report segment details a predictive model using a Random Forest Classifier to forecast customer attrition, with an achieved accuracy of 62.48%. The model, trained on 80% of the dataset with preprocessing like binary encoding and null value handling, helps identify customers likely to discontinue service. This prediction enables targeted actions to enhance retention, which is crucial for maintaining revenue and customer base. The moderate accuracy indicates potential for model optimization to improve prediction reliability.

- Monthly Revenue Forecasting:

The "Monthly Revenue Forecasting" report segment describes a predictive model constructed using a Random Forest Regressor to estimate future revenue, achieving a Root Mean Squared Error (RMSE) of 8.538. The model was trained on 80% of the data, which underwent preprocessing such as conversion of string data to float and handling null values to ensure quality inputs. This process of forecasting monthly revenue is vital for the business to make strategic decisions, allocate resources effectively, and understand market trends. The RMSE indicates the average difference between the revenues predicted by the model and the actual figures, providing a quantitative measure of the model's accuracy. The obtained RMSE suggests a degree of precision that could be improved through further model refinement, aiming to enhance the reliability of the revenue forecasts.

- Overage Minutes Prediction:

The "Overage Minutes Prediction" report segment outlines a predictive model using a Random Forest Regressor to estimate the number of overage minutes — time exceeding the plan's allotted minutes — with a Root Mean Squared Error (RMSE) of 52.58% on the test data. Trained on 80% of the dataset and preprocessed to handle categorical variables and null values, the model is designed to forecast the additional minutes customers are likely to use. Understanding overage patterns is beneficial for optimizing plan structures and potentially increasing revenue. The given RMSE offers insight into the model's prediction accuracy and highlights opportunities for further tuning to enhance forecast precision.

Challenges:

- Data Preprocessing Difficulties: Faced significant challenges in handling null values and converting data types, such as binary encoding and transforming strings to floats. These steps were critical to ensure the integrity and reliability of the data being fed into the churn prediction model.
- Model Optimization Opportunities: The churn prediction model achieved an accuracy of 62.48%, indicating potential areas for improvement. This may involve increasing model complexity, refining feature selection, or enhancing data quality to boost predictive performance.

Conclusion:

Our project achieved its prediction and inference goals with moderate success. The churn prediction model, using a Random Forest Classifier, attained a 62.48% accuracy, suggesting a foundation for targeted retention actions but also highlighting room for optimization. Revenue forecasting with a Random Forest Regressor yielded an RMSE of 8.538, indicating a reasonable prediction accuracy, while overage minutes prediction achieved an RMSE of 52.58%, pointing to further opportunities for model tuning. These results show a good starting point for practical applications and future enhancements in churn prediction and revenue forecasting within the telecom industry.

References:

- Lemmens, Aurélie, and Christophe Croux. "Bagging and Boosting Classification Trees to Predict Churn." Journal of Marketing Research, vol. 43, no. 2, 2006, pp. 276-286.

- Probst, Philipp, Anne-Laure Boulesteix, and Bernd Bischl. "Random Forest Versus Logistic Regression: A Large-Scale Benchmark Experiment." BMC Bioinformatics, vol. 20, no. 1, 2019, p. 242.

- Probst, Philipp, Anne-Laure Boulesteix, and Bernd Bischl. "Tune Random Forest." arXiv, arXiv:1508.04409, 2015.

- Vafeiadis, T., et al. "A Comparison of Machine Learning Techniques for Customer Churn Prediction." Simulation Modelling Practice and Theory, vol. 55, 2015, pp. 1-9.

- Zhunqiang. "Customer Segmentation and Churn Model for Telecom." Kaggle, Kaggle Inc., www.kaggle.com/code/zhunqiang/customer-segmentation-and-churn-model-for-telecom. Accessed 11/28/2023.