Transforming Data into Actional Insights: Unveiling Customer Churn in the Telecom Sector

Sahib Bhatti
190319889
Nafi Ahmad
MSc Big Data Science

***Abstract*— This data science project offers an in-depth analysis of customer churn within the California-based telecom industry, emphasizing its impact on revenue and brand reputation. Customer churn presents significant risks to market share and public image. The project details a comprehensive churn analysis pipeline, including data exploration, preparation, and visualization, while highlighting key customer attributes and financial indicators. Statistical tests (T-test) and regression analysis explore two hypotheses: longer subscriptions reduce churn rates, and higher monthly bills increase churn. The findings link high churn rates to higher fees, competitive offers, and service dissatisfaction. Behavioral patterns, such as preferences for long-term contracts and additional services, are identified. Four machine learning models—Logistic Regression, Random Forest, Gradient Boosting, and XGBoost—are trained to predict churn and key predictors. XGBoost emerges as the most effective model for predicting and mitigating churn, ultimately enhancing customer satisfaction and loyalty. Feature importance further confirms churn determinants. Strategic recommendations include targeting top-tier clients, securing market advantages, incentivizing longer contracts, improving internet service, expanding internet and streaming services, revising ineffective deals, and reducing customer turnover to improve business performance.**

Keywords—customer churn, telecom, revenue, XGBoost, retention

I. Introduction

The telecommunications industry is facing growing challenges in customer retention, where losing clients to competitors—known as customer churn—poses a significant risk to revenue and brand reputation. Retaining existing customers has become as crucial as acquiring new ones, as it is more cost-effective in today's competitive market [1] [2]. Reducing churn is essential to ensure long-term profitability and market stability.

This study focuses on a telecom provider in California that is experiencing increasing customer churn rates. This issue not only impacts revenue but also erodes market share and damages the company's public image. Understanding the underlying causes of churn, such as customer dissatisfaction or competitive offers, is vital for developing strategies to enhance customer loyalty and retention [7]. High churn rates, particularly among long-term, high-value customers, indicate deeper issues within the company's services or management, leading to significant revenue loss and increased vulnerability to competitors [9].

Recent breakthroughs in customer churn prediction emphasize the integration of advanced techniques like hybrid models and deep learning, along with real-time data processing, to enhance accuracy. While complex models are gaining traction, studies show that machine learning models can achieve remarkable accuracy for a moderate level of features, highlighting their relevance and effectiveness in practical applications for churn prediction. These advancements underscore the growing importance of machine learning in customer retention strategies, enabling timely actions that help maintain market share and enhance customer satisfaction [18] [15].

The driving force behind this research is the urgent need to address rising customer churn in the telecom industry, which threatens long-term business viability. By carrying out data analysis, hypothesis tests, applying advanced machine learning techniques, and determining feature importance, this study seeks to create effective predictive models that can pinpoint the factors leading to churn and provide actionable strategies to improve customer retention [10] [11].

This project advances churn prediction by with a dual focus on technical rigor and practical application, making its contribution particularly valuable. It enhances telecom industry strategies with robust methodologies, actionable recommendations, and key churn indicators. By evaluating models, integrating hypothesis testing, and using advanced techniques like XGBoost, it sets a benchmark for future research and offers data-driven solutions to reduce churn and improve retention, benefiting researchers and practitioners.

Paper Organization: Section II reviews customer churn literature; Section III outlines methodology, including data exploration, hypothesis testing, and model development. Section IV presents results and implications; Section V offers strategic recommendations based on analysis. Section VI discusses insights, and Section VII concludes with a summary and future work suggestions.

II. Related Work

Customer churn is a significant challenge for the telecom industry, affecting both profitability and growth. Dahiya and Bhatia explored this issue by developing a predictive framework using data mining techniques, specifically decision trees and logistic regression. Their structured approach involved data acquisition, aggregation, and preprocessing, ultimately finding that decision trees outperformed logistic regression in accuracy, making them more effective for predicting customer churn [1].

Similarly, Ahn, Han, and Lee conducted a thorough analysis of customer churn in the South Korean mobile telecommunications industry using a dataset of customer transactions and billing records. They applied multinomial

and binary logistic regression models to test hypotheses on churn determinants, focusing on factors like call quality, loyalty points, and service usage. Their findings revealed that higher call drop rates increased churn risk, while loyalty points reduced it, although membership card holders unexpectedly had higher churn rates. The study also identified changes in customer status, such as moving to non-use, as early indicators of churn, highlighting the need to enhance retention strategies by improving loyalty programs and monitoring status changes as key churn signals [2].

In their paper, Çelik and Osmanoglu compare several machine learning algorithms, including logistic regression, decision trees, and neural networks, to determine their effectiveness in predicting customer churn. They emphasize that deep learning succeeds amongst complex data but no single technique consistently outperforms others across all datasets. Instead, the effectiveness of each method can vary depending on the specific characteristics of the data and the industry context [3].

These studies collectively highlight the effectiveness of boosting in predicting customer churn across various industries. Lu, Lin, Lu, and Zhang developed a tailored churn prediction model for the telecom industry using Gentle AdaBoost, which improved accuracy by clustering customers based on the algorithm's weight and applying separate logistic regression models for each cluster, particularly benefiting high-risk customers [4]. Qing Liu, QiuYing Chen, and Sang-Joon Lee demonstrated that Gradient Boosting outperformed other models on the ELEME delivery platform, achieving an F1-Score of 0.942 and an AUC of 0.955, showcasing its superiority in non-contractual settings [5]. Similarly, Chenggang He, Chris H.Q. Ding, Sibao Chen, and Bin Luo created an intelligent machine learning system where Gradient Boosting delivered the highest accuracy of 95.32% and an F1-score of 97.3%, further proving its efficacy in churn prediction [6]. Abhikumar Patel and Amit G Kumar focused on telecom customer retention, with XGBoost standing out by achieving the highest accuracy of 94% and effectively handling large datasets [7]. Finally, Abhishek Gaur and Ratnesh Dubey confirmed the dominance of Gradient Boosting Trees in their study, where it achieved the highest AUC value of 84.57%, solidifying its role in accurate churn prediction and strategic retention efforts [8]. Baburao Markapudi, Kunchaparthi Jyothsna Latha and Kavitha Chaduvula introduced the Boosted Leaf Model (BLM), combining Decision Trees and Gradient Boosting, which improved churn prediction and achieved superior AUC scores by effectively analysing key attributes like usage patterns and billing information [9]. Collectively, these studies underscore the consistent superiority of boosting algorithms in enhancing churn prediction accuracy and informing effective customer retention strategies.

These studies focus on neural networks and advanced machine learning approaches. Fujo, Subramanian, and Khder address customer churn prediction in the telecom industry using deep learning models, specifically implementing a Deep-BP-ANN model with techniques like variance thresholding, Lasso regression, early stopping, and Random Oversampling to handle imbalanced data. Their model outperforms other techniques in predicting customer churn [10]. Amin et al. introduce a novel approach using classifier certainty, grouping data into zones with varying certainty levels to improve prediction accuracy [11]. Larasati, Ramadhanti, Chen, and Muid enhance a Deep Learning ANN model for churn prediction by optimizing parameters, achieving 76.35% accuracy and 89.99% precision through fine-tuning of epochs, hidden layers, and activation functions [12].

These studies highlight the effectiveness of techniques such as SMOTE to increase accuracy in the data preparation stage and models such as Logistic Regression and Random Forest Model in correctly predicting churn. Jesmi Latheef and Vineetha S explored the use of Long Short-Term Memory (LSTM) models in the banking sector, showing that combining SMOTE with LSTM improved accuracy to 88%, outperforming other models. They assessed attributes like tenure, account balances, demographics, and credit scores, with LSTM effectively capturing temporal dependencies crucial for accurate churn prediction [13]. Harish A S and Malathy C compared a k-means clustering-based model with a conventional RFM-based model, finding Random Forest most accurate in predicting churn (0.9875) due to its robust ensemble learning and feature importance analysis [14]. Liwen Ou's study found that Random Forest and Extra Tree classifiers provided high accuracy and valuable insights into telecom churn [15]. Another telecom study by Preetha further highlighted Random Forest's effectiveness, achieving top accuracy and AUC by splitting the dataset into four U.S. population subgroups, which enhanced prediction accuracy and customer behavior insights [16]. In the insurance industry, Jajam Nagaraju and colleagues applied logistic regression, random forest, and a hybrid LGBM and XGBoost model, with logistic regression achieving the highest AUC, making it the preferred model [17].

The literature showcases how machine learning is just as effective as complex neural or hybrid models, with the advantage of saving computational resources since this problem domain does not carry an abundance of features or parameters. The studies show the benefits of hypothesis testing and machine learning as separate approaches, but it has influenced me to combine them. By integrating both methods, I can validate assumptions and identify key churn factors, while also leveraging machine learning for precise predictions. This combined strategy, enhanced with SMOTE and a focus on feature importance, offers an effective approach for accurately predicting churn and developing targeted retention strategies. Most literature lacks data-driven strategies for addressing churn, and focus more heavily on predicting it, thus I plan to fill this gap.

III. Methodology

This section details our approach to solving the problem, and the rationale behind it. The dataset named Maven Churn, sourced from Kaggle, shows customer churn at a California based telecom company in the second quarter of the 2022.

A. Dataset

The dataset includes two tables: 'Customer Churn' with data from 7,043 customers covering over 35 metrics (tenure, location, demographics, subscription types, customer status) and 'Zip Code Population' providing demographic details for the zip codes mentioned. Merging these tables helps us understand regional characteristics and develop effective retention and marketing strategies for different demographic and geographical segments, improving customer loyalty.

B. *Data Exploration*

The first analysis phase determines if the data supports the issue we aim to address. This involves thoroughly examining the data to confirm the presence and impact of customer attrition on the company's revenue, taking a detailed look at the initial rows, column names, and any missing values.

The descriptive statistics confirm that we have quite a broad age range of customers where customers have diverse family sizes. Geographic diversity is shown through a wide range of zip codes and location data. Subscription durations average at 32.4 months, indicating a mix customer dynamic that are new as well as long terms subscribers. Monthly fees range widely, from -10 USD (likely an error) to 118.75 USD, averaging 63.60 USD. The mean total charge is 2,280.38 USD, with overall revenue ranging from 21.36 to 11,979.34 USD, highlighting financial variation among customers.

C. *Data Pre-processing*

During data exploration, we discovered numerous missing values and potential outliers. Addressing these is crucial to prevent skewed analysis, overfitting, and inaccurate machine learning predictions before proceeding with detailed analysis.

1) Missing Value Exploration

During the data exploratory analysis, we found many missing values in the dataset, especially in the 'Churn Category' and 'Churn Reason' columns. Columns related to internet services also had numerous null values, which shows that some customers did not use these services. To address this, we used two main strategies to ensure data completeness. For average monthly charges and data usage, we filled missing values with the average. For service options and subscription types, we marked missing data as 'No'. For customer retention categories, we labeled missing data as 'Other'.

2) Outliers Exploration

Outliers are data points that can skew analysis, but in our dataset, they weren't extreme, remaining within boxplot ranges.

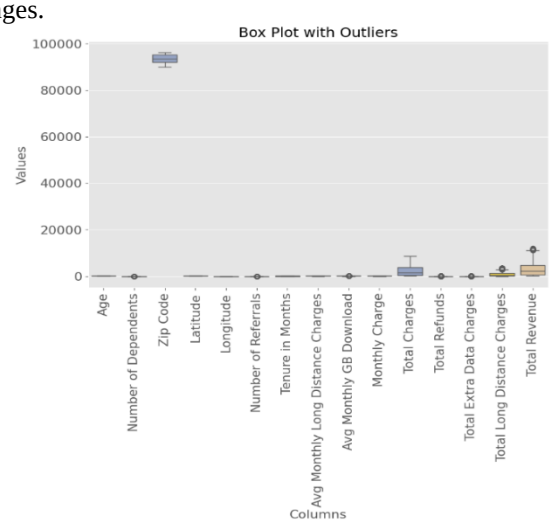

Figure 01. Box Plots Showcasing Outliers

Initially, we removed them through methods such as Interquartile Range (IQR) Outlier Detection, Winsorization, and Capping, but this led to distorted analysis and inaccurate metrics for total revenue and total customers, which misrepresented the company's data and true reality, it further impaired algorithm effectiveness since features were not truly weighted. After reconsideration, we retained the outliers, which proved beneficial. They contributed to accurate metrics, enhanced machine learning model accuracy, and provided valuable insights into high-value customers, making them crucial for informed realistic business decisions.

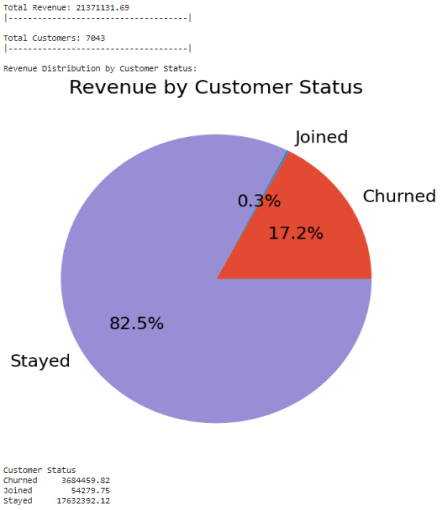

Figure 02. Revenue of the company by customers

D. *Visualisations and Analysis*

Figure 2 shows that the company has generated a substantial total revenue of $21,371,131.69 from 7,043 customers, indicating strong financial health. Of this, $3,684,459.82 comes from customers who have since churned, highlighting the revenue impact of customer attrition. The below pie chart verifies that a significant part of the company's income is decreased because of customer churn. We will now continue with more in-depth analysis.

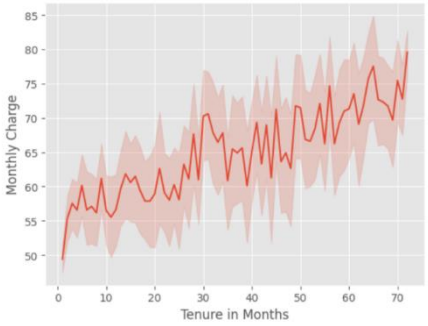

Figure 03. Monthly charge with tenure in months

E. *Customer Churn Reasons Analysis*

We investigated the main factors contributing to customer churn. Figure 3 shows that customers who churned were paying higher average monthly fees, around $73.35, while new customers paid about $42.78, and loyal customers averaged $61.74. This suggests that pricing affects retention. Gender does not significantly influence pricing differences, as both male and female customers had similar average charges. Figure 4 shows churned customers stayed for about 18 months on average, whereas loyal customers had an average tenure of 41 months, indicating that longer tenures may lead to higher monthly charges.

```
df.groupby('Customer Status')['Monthly Charge'].describe()
```

Customer Status	count	mean	std	min	25%	50%	75%	max
Churned	1869.0	73.347592	26.373764	-10.0	55.3000	79.500	94.20	118.35
Joined	454.0	42.775991	24.612329	-8.0	20.1625	43.975	64.35	107.95
Stayed	4720.0	61.737415	32.198758	-10.0	25.1000	65.600	89.55	118.75

```
df.groupby(['Customer Status','Gender'])['Monthly Charge'].describe()
```

Customer Status	Gender	count	mean	std	min	25%	50%	75%	max
Churned	Female	939.0	73.606656	26.093400	-10.0	59.6250	79.500	93.625	117.45
	Male	930.0	73.086022	26.665339	-9.0	53.8750	79.500	94.500	118.35
Joined	Female	211.0	43.374645	23.839314	-8.0	20.2250	44.000	69.225	95.85
	Male	243.0	42.256173	25.301925	-8.0	20.1000	43.850	60.800	107.95
Stayed	Female	2338.0	62.343926	32.085856	-10.0	25.2500	65.625	89.750	118.75
	Male	2382.0	61.142107	32.304853	-10.0	25.0125	65.550	89.375	118.20

```
df.groupby(['Customer Status'])['Tenure in Months'].describe()
```

Customer Status	count	mean	std	min	25%	50%	75%	max
Churned	1869.0	17.979133	19.531123	1.0	2.0	10.0	29.0	72.0
Joined	454.0	1.720264	0.818486	1.0	1.0	1.0	2.0	3.0
Stayed	4720.0	41.041525	22.363674	4.0	21.0	42.0	63.0	72.0

Figure 04. Statistics for monthly charges by customer status, tenure in months and gender

Further analysis presented by figure 5 shows that the main reasons for customer churn are superior devices from competitors (313 customers) and better offers elsewhere (311 customers). This highlights the need to focus on device quality, competitive pricing and promotional strategies. Also, poor interactions with support staff (220 cases) and service providers (94 cases) have led to churn, which requires a need to work on improving customer service. A small number of customers left due to personal reasons like passing away or not being comfortable with technology. Therefore, to reduce churn, we must focus on enhancing device quality, offering attractive pricing, and improving customer service.

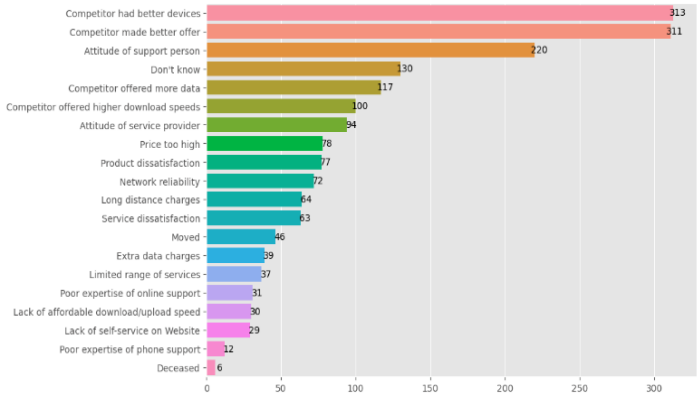

Figure 05. Reasons why customers churned.

Examining the bar graphs in figure 6 provide more detailed key insights. Many customers are not responding to the company's marketing efforts, especially Offer E, which has a high churn rate, reflecting the need for more effective conversion strategies. Increased churn rates in internet services point to issues like slow connectivity, so this furthers the need for service quality improvements. Internet services resulted in significant churn indicated poor internet performance and internet type such as Fiber Optic internet, despite its high speed has surprisingly high churn rates, suggesting poor network reliability, which again spoils customer experience. In terms of contract arrangements, long-term contracts retain customers better than month-to-month ones. Finally, additional internet services like online security enhance customer safety improving loyalty, highlighting their role in reducing churn and increasing satisfaction.

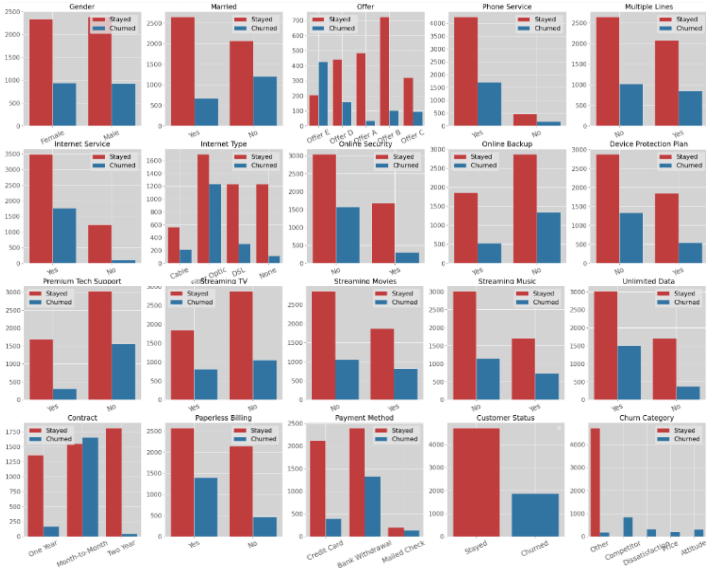

Figure 06. Bar charts for categorical attributes

Based on visualization of stayed customers in Figure 7, several notable trends emerge. The majority (56.12%) of high-value customers are married, with only few selecting "Offer E." A significant 67.22% are committed to long-term contracts—38.41% have opted for two-year contracts, while 28.81% have chosen one-year contracts. In stark contrast, figure 8 shows that only 35.79% of churned customers are married, suggesting married couples are more likely to stay due to stability reasons. "Offer E" appears to be popular amongst the churned. 88.55% prefer monthly payments and 66.13% of churned customers used fiber optic internet. Thus we can generate the insight that there seems to be an issue with this internet service type. It can be interpreted now that customers churn due to monthly contracts since these average higher fees than longer term payment plans, meaning customers who prefer paying monthly are likely to churn to competitors with more competitive monthly prices. Thus, these insights can guide strategies to reduce churn and improve customer satisfaction.

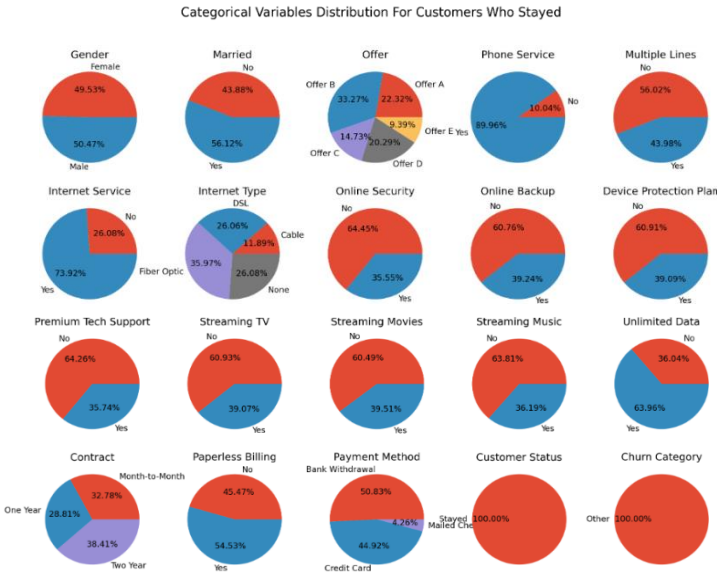

Figure 07. Pie Chart Distribution of Variables Stayed Customers

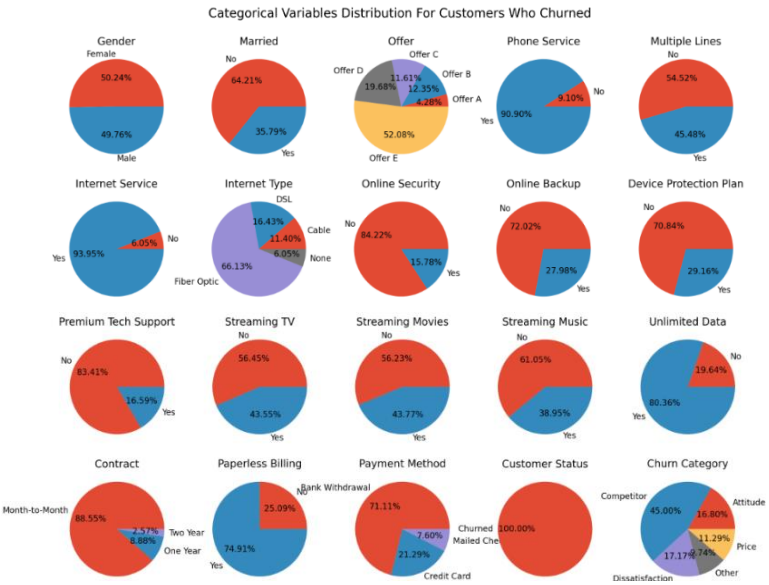

Figure 08. Pie Chart Distribution of Variables for Churned Customers

F. Hypothesis Analysis

Hypothesis testing reveals significant relationships between variables. By selecting key factors from prior analysis, we will test two hypotheses to gain data-driven insights into customer churn. This rigorous validation helps identify and confirm specific churn determinants, enabling targeted strategies to improve retention and reduce churn rates. [2].

1) Hypothesis 1:

For the first hypothesis, guided by the contract chart in figure 6, we test if longer subscriptions reduce churn. We visualize subscription durations by customer status with a count plot. Then we will carry out a t-test, which is a statistical test used to determine whether there is a significant difference between the means of two groups, churned and active customers. This will be conducted across contract types where a p value of less than 0.05 (commonly used threshold) will indicate that the difference is statistically significant, and that the hypothesis does not occur due to random chance.

2) Hypothesis 2:

Our second hypothesis, influenced by findings of figure 3, is to test if higher monthly bills lead to higher churn. We test this with a logistic regression model. First, we will convert customer status to binary values (0 for 'Stayed', 1 for 'Churned') and split the data into training and testing sets. We will judge by the accuracy of the model if the correlation between the variables matches the hypothesis. Then statsmodel api will be used as it provides detailed statistical analysis, including coefficients, p-values, and confidence intervals, which help carefully validate the relationship between monthly billing amounts and customer churn.

G. Predictive Modelling:

The literature recommends various approaches for training predictive models, including machine learning algorithms, neural networks, boosting techniques, deep learning models, and hybrid methods for enhanced accuracy and robustness. We will use Logistic Regression, Random Forest, Gradient Boosting, and XGBoost to predict customer churn and identify key factors, as these models have demonstrated proven accuracy and effectiveness across various industries.

Data preparation involved converting categorical data to numerical values (e.g., 'Yes' to 1, 'No' to 0), categorizing columns like contract types, and removing irrelevant information. The target variable 'Customer Status' was label-encoded into a numerical format, (0 as churned, 1 for stayed). We scale numerical features using Min-Max Scaling and split the data into training and testing sets. To address class imbalance, we apply the SMOTE technique to the training set, creating synthetic samples for the minority class to enhance model accuracy and generalizability [13]. This made the data ready for model training.

We split the dataset into an 80:20 train test ratio. K-fold cross-validation was used to help tune hyperparameters and validate the Machine Learning Models. It was chosen over base validation since it evaluates a model's performance by dividing the data into multiple subsets, enhancing generalization and reducing overfitting. We implement 10-fold cross-validation for Logistic Regression, Random Forest, Gradient Boosting and XGBoost, averaging metrics like accuracy, precision, recall, and F1 scores across folds. This method was effective since it would train on 9 folds and test on the 10th, iterating 10 times on different folds. The best performing model was then chosen and executed on the unseen test data to see how well the model generalizes [18].

G. 1.1 Logistic Regression Model:

Logistic Regression is an efficient statistical method for binary classification, modeling the relationship between dependent and independent variables by estimating probabilities with a logistic function. This approach offers clear interpretability, especially in predicting customer churn. We employ a pipeline that begins with StandardScaler preprocessing step to standardize features, ensuring each has a mean of 0 and a standard deviation of 1—crucial for models like Logistic Regression that are sensitive to feature scaling, enhancing learning and performance. The model is refined with L2 regularization to prevent overfitting by avoiding overemphasis on any single feature and a tuned regularization strength (C=10) to balance bias and variance, ensuring accuracy and generalization to new data [17].

G. 1.2 Random Forest Classifier Model:

Random Forest is a powerful model for classification and regression that boosts accuracy by creating multiple decision trees from random data samples and features, then combining their predictions. We fine-tuned the hyperparameters: n_estimators=26 (number of trees) to balance capturing key data patterns with training efficiency, max_depth=14 (depth of trees) to prevent overfitting while capturing essential features, and random_state=42 to ensure consistent results. This method effectively handles complex, large-scale data, reduces tree correlation, and prevents overfitting.

G. 1.3 Gradient Boosting Model:

Gradient Boosting is a powerful ensemble method for classification and regression, iteratively refining models by correcting previous errors by minimizing a loss function through gradient descent, resulting in high accuracy and effective handling of complex, non-linear data. We fine-tuned the hyperparameters as follows: n_estimators=70 to steadily enhance accuracy over more boosting stages, learning_rate=0.05 to slow down learning to ensure better generalization, and max_depth=5 to limit tree complexity and prevent overfitting. We used min_samples_split=5 and min_samples_leaf=4 to maintain meaningful splits and leaf nodes, avoiding overfitting. The subsample=0.7 introduces 70% data randomness per tree for added robustness, while random_state=42 ensures consistent, reproducible results.

G. 1.4 XGBoost Model:

XGBoost is a powerful, efficient version of Gradient Boosting, ideal for classification tasks like customer churn prediction. It excels by reducing overfitting through regularization and speeding up computations with parallel processing. Its ability to handle large datasets and missing data makes it a go-to for precise churn predictions and effective retention strategies. We fine-tuned XGBoost's hyperparameters: binary:logistic for binary classification, logloss as the evaluation metric to avoid classification errors, and max_depth=8 to avoid overfitting. We used a learning_rate of 0.01 for gradual learning, subsample=0.8 to generalize using a random 80% of the data per tree, and colsample_bytree=0.8 to reduce overfitting by using random 80% of the features. With n_estimators=1000, the model undergoes numerous boosting rounds, and seed=42 ensures consistent results. A mild 0.1 value for Reg_alpha (L1 regularization) simplifies the model by reducing less important feature weights towards zero, while reg_lambda=1 (L2 regularization) was used to control weight size to prevent them from becoming too large. Both help reduce overfitting and improve the model's generalization. [7].

IV. RESULTS

A. Hypothesis Testing Results

1) Hypothesis 1:

Figure 9 results show that month-to-month customers have shorter tenures than those with longer contracts, supporting the hypothesis that longer subscriptions reduce churn. A large absolute value of the t-statistic indicates a large difference between the group means. A p-value less than 0.05 (commonly used threshold) indicates that the difference is statistically significant, where across all three contract types (0.023, 7.02e-75, 0.00), the p value was less than 0.05, thus rejecting the assumption of it being random chance that longer subscriptions have less churn. The monthly t-statistic value of 18.81 compared to the negative t-statistic values of the yearly contacts can confirm that the t-test supports the hypothesis that longer subscription durations are associated with lower churn rates, due to true differences in customer behavior, and not random chance.

Figure 9. T-test results for subscription duration vs churn

2) Hypothesis 2:

The regression model achieved an accuracy of 72.4%, (figure 10) demonstrating a correlation between higher bills and churn. Using the Statsmodels API, we confirmed that higher monthly charges significantly predict churn, as evidenced by a positive coefficient for 'MonthlyCharge' and a low p-value. Specifically, the coefficient for 'MonthlyCharge' is 0.0127, indicating that higher monthly billing amounts are associated with increased log-odds of churn. The p-value of 0.000 strongly suggests that this finding is statistically significant and not due to random chance, indicating that customers with higher monthly billing amounts are indeed more likely to churn, as per hypothesis.

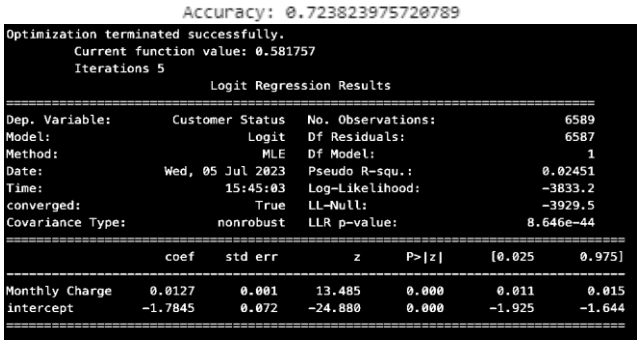

Figure 10. Regression Results

B. 1. Predictive Modelling Results

1) Model Performance using K-fold Cross-Validatoin

Performance	Machine Learning Model			
	Logistic Regression	***Random Forest***	***Gradient Boosting***	***XGBoost***
Accuracy	0.8393	0.8970	0.8876	0.9108
Precision	0.8724	0.9033	0.8982	0.9099
Recall	0.7952	0.8893	0.8745	0.9118
F1 Score	0.8318	0.8962	0.8861	0.9108

Figure 11. Performance metrics for models k-fold cross-validation

Figure 11 demonstrates that XGBoost outperforms Random Forest, Gradient Boosting and Logistic Regression in predicting customer churn. All models were evaluated using 10-fold cross-validation, which averaged accuracy, precision, recall, and F1 scores across folds. This method internally managed model fitting by training on 9 folds and testing on the 10th, iterating 10 times. XGBoost achieved the highest accuracy (0.9108), along with strong precision

(0.9099), recall (0.9118), and F1 score (0.9108). Random Forest and Gradient Boosting performed nearly as well, with accuracy only slightly lower. Surprisingly, compared to its other performance metrics, Logistic Regression has a high precision (0.8724), despite its lower recall, meaning it was reliable in the case it did identify as churn. Overall, XGBoost's superior accuracy and balance makes it the optimal choice for minimizing false positives while accurately identifying churned customers.

2) Chosen XGBoost Model Performance on the Test Data

Performance	**Machine Learning Model**
	XGBoost
Accuracy	0.8801
Precision	0.9123
Recall	0.9219
F1 Score	0.9171

Figure 12. Performance metrics for XGBoost on the Test Data

The XGBoost model demonstrated consistent performance across the k-fold cross-validation phase and the test data. It achieved a slightly lower accuracy on the test data of 88.01% in comparison to 0.9108 from the training set. The accuracy is still very close and this consistency indicates that XGBoost generalizes very well to unseen data, making it a reliable choice for predicting customer churn. These minor variations indicate the model's stability and reliability, confirming that it was not overfitted during validation and can accurately predict customer churn in new datasets.

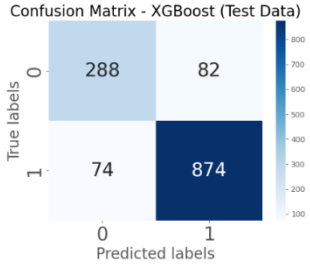

Figure 13. Comparing Confusion Matrix for XGBoost (Test Data)

The confusion matrix offers clear insights into the model's performance by detailing correct and incorrect predictions, which helps evaluate precision, recall, and manage class imbalances. In this case, 0 stands for the number of customers churned, and 1 for the number of customers that stayed. The confusion matrix for the XGBoost model on the test data correctly identified 874 stayed customers (true positives) and 288 churned customers (true negatives). However, it also made 82 false positive predictions (incorrectly predicting stayed for churned customers) and 74 false negative predictions. Minimizing false negatives is most crucial because failing to identify churned customers is the most costly. These results demonstrate the model's strong accuracy and reliability in predicting customer churn, with only minor discrepancies when applied to unseen data.

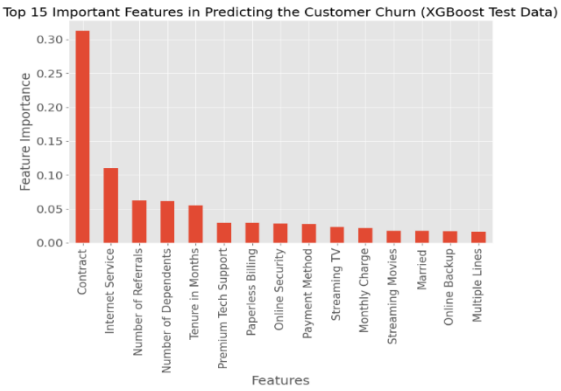

Figure 14. Feature Importance Bar Chart (XGBoost)

A feature importance chart highlights which features have the most influence on predicting whether a customer churns or not. From the bar chart, we can see that contract remains as the top feature which is common amongst the feature importance charts of the other models (refer to appendix), yet the second and third features are the Internet Service and Number of Referrals. Internet service from our analysis earlier had a high churn rate, possibly due to poor connection and disconnection rates, suggesting that this feature significantly affects churn as customers seek out better connection speeds from competitors. Number of Referrals can be a key reason why customers stay, suggesting that customers that refer others are engaged customers, meaning they are satisfied with the service provided boosting loyalty. In a close 4th place is the Number of Dependents feature. This could suggest that customers with more dependents might be either more or less likely to churn, potentially influenced by financial responsibilities or the necessity for reliable services to support their dependents. For the feature importance conducted on Random Forest and Gradient Boosting, Number of Referrals and Tenure in Months was a feature consistently appearing second and third, whilst internet service was surprisingly not a significant feature for these models. Contract always being top and Tenure in Months being significant, further support both the hypothesis about longer subscriptions reducing churn, and higher monthly bills increasing churn, indicated how significant these factors are to effecting churn.

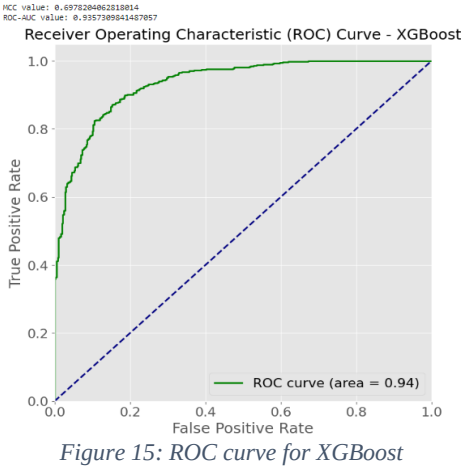

Figure 15: ROC curve for XGBoost

The ROC curve compares models based on their true positive rates (proportion of correctly identified positives)

versus false positive rates (proportion of actual negatives incorrectly identified as positives), revealing their effectiveness in predicting customer churn. With an ROC-AUC score of 0.94, the model demonstrates strong performance in distinguishing between customers likely to churn and those who will stay, effectively balancing sensitivity (the true positive rate) and specificity (the false positive rate). Additionally, the MCC of 0.7 indicates a solid agreement between predicted and actual outcomes, confirming the XGBoost model's reliability and accuracy in forecasting customer churn.

V. Recommendations

To improve customer engagement and reduce customer churn, we recommend the following retention strategies on our most valuable customers based on the above analysis.

- By introducing a special loyalty scheme, we can offer exclusive benefits to the top third of our customer base who are not only loyal but also contribute significantly to our revenue. This will make them feel valued and help maintain their connection with our brand.
- It's also important to consistently meet the high expectations of customers with superior product and service offerings evidenced in figure 5. By doing so, we can secure a competitive edge that minimizes the risk of customers switching to competitors.
- Encouraging customers to switch from month-to-month to long-term plans, as supported by analysis and hypothesis tests, can reduce churn by emphasizing long-term value and cost savings.
- Given the importance of internet services, enhancing their quality and reliability should be a priority, especially for fiber optic services, which despite high speeds, show the highest churn rate (as seen in figures 6 and 8). Improving these services can boost customer satisfaction and retention, particularly for those who rely heavily on them.
- Users with online security have lower churn rates, implying a revenue opportunity by upselling this service. Targeted offers could boost adoption, reduce churn, and increase satisfaction (figure 6, 7, 8)
- Customers with multiple dependents may have different service needs or financial considerations. The company should consider developing targeted strategies, such as family plans or discounts, to better cater to these customers and reduce their likelihood of churning (figure 14).
- Customers who refer others tend to be more engaged and loyal. The company should implement or enhance referral programs that reward customers for bringing in new clients, thereby creating a positive feedback loop that supports customer retention (figure 14).
- Our analysis shows that Offer E is not performing well, with a high churn rate. We recommend a thorough review of this offer, comparing it with more successful ones to pinpoint areas for improvement (figure 6, 7, 8).
- Maintaining customer satisfaction is crucial. Regular surveys can identify early dissatisfaction, enabling proactive responses. Involving customers in feedback and adapting services fosters community and loyalty.

VI. Discussion

The analysis identifies key factors driving customer churn that, if addressed, can greatly improve retention. The company should focus on retaining its most valuable customers by enhancing service quality, offering fair pricing, and incentivizing longer contracts through loyalty programs. The XGBoost model proves most effective in predicting churn, balancing accuracy and customer retention. All models place the feature of contracts as the most important feature whilst the XGBoost model also highlights the importance of features like the internet service, improving this can retain customers with good network reliability, number of dependents and referrals, suggesting that customers with more dependents value stability, while referrals indicate strong satisfaction and loyalty. This underscores the need for family-friendly plans and enhanced referral programs. Regularly updating strategies based on feedback and ongoing analysis is crucial for maintaining customer relationships and reducing churn. Furthermore, relating to some research from related works section, Ahn, Han, and Lee's study used traditional regression models, focusing mainly on specific churn factors, while Çelik and Osmanoglu compared basic machine learning methods without exploring advanced techniques [2] [3]. My project surpassed these by incorporating XGBoost, which captures complex data relationships more effectively. Additionally compared to the studies utilizing advanced models, I applied k-fold cross-validation to ensure model robustness and conducted feature importance analysis to identify key churn drivers. By integrating hypothesis testing (absent in majority of studies) I pinpointed precise, actionable insights for immediate impact. This combination of advanced modeling, validation, and analysis made my contribution and findings more impactful in a real-world context.

VII. Conclusion

In conclusion, the project successfully addresses the company’s concern about high customer churn rates. The data analysis confirmed the issue of high customer churn rates and revealed key causes. Hypothesis testing showed that longer subscriptions reduce churn, while higher monthly bills increase it. Machine learning models, particularly XGBoost, can effectively predict customer churn, allowing for proactive retention strategies and improved customer satisfaction. Thus, this study offers a new, all-round and comprehensive approach to handling customer churn. Future work can explore more advanced machine learning techniques, such as deep learning models, to improve churn prediction accuracy [19]. Additionally, incorporating real-time data and feedback loops could enhance the responsiveness of predictive models [20]. Investigating the impact of emerging technologies, like 5G and IoT, on customer behavior and churn rates could provide further insights [21]. These technologies are quickly transforming how consumers experience services, making it crucial to understand their impact on customer satisfaction and churn.

REFERENCES

[1] K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2015, pp. 1-6, doi: 10.1109/ICRITO.2015.7359318.

[2] Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications policy, 30(10-11), 552-568.

[3] Çelik, O., & Osmanoglu, U. O. (2019). Comparing to techniques used in customer churn analysis. Journal of Multidisciplinary Developments, 4(1), 30-38.

[4] N. Lu, H. Lin, J. Lu and G. Zhang, "A Customer Churn Prediction Model in Telecom Industry Using Boosting," in IEEE Transactions on Industrial Informatics, vol. 10, no. 2, pp. 1659-1665, May 2014, doi: 10.1109/TII.2012.2224355.

[5] Q. Liu, Q. Chen and S. -J. Lee, "A Machine Learning Approach to Predict Customer Churn of a Delivery Platform," 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia, 2023, pp. 733-735, doi: 10.1109/ICAIIC57133.2023.10067108.

[6] C. He, C. H. Q. Ding, S. Chen and B. Luo, "Intelligent Machine Learning System for Predicting Customer Churn," 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Washington, DC, USA, 2021, pp. 522-527, doi: 10.1109/ICTAI52525.2021.00085.

[7]A. Patel and A. G. Kumar, "Predicting Customer Churn In Telecom Industry: A Machine Learning Approach For Improving Customer Retention," 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), Rajkot, India, 2023, pp. 558-561, doi: 10.1109/R10-HTC57504.2023.10461822.

[8] A. Gaur and R. Dubey, "Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-5, doi: 10.1109/ICACAT.2018.8933783.

[9] B. Markapudi, K. J. Latha and K. Chaduvula, "A New hybrid classification algorithm for predicting customer churn," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2021, pp. 1-4, doi: 10.1109/ICSES52305.2021.9633795.

[10] Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer churn prediction in telecommunication industry using deep learning. Information Sciences Letters, 11(1), 24.

[11] Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research, 94, 290-301.

[12] A. Larasati, D. Ramadhanti, Y. W. Chen and A. Muid, "Optimizing Deep Learning ANN Model to Predict Customer Churn," 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Malang, Indonesia, 2021, pp. 1-5, doi: 10.1109/ICEEIE52663.2021.9616714.

[13] J. Latheef and S. Vineetha, "LSTM Model to Predict Customer Churn in Banking Sector with SMOTE Data Preprocessing," 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Ernakulam, India, 2021, pp. 86-90, doi: 10.1109/ACCESS51619.2021.9563347.

[14] H. A. S and M. C, "Evaluative study of cluster based customer churn prediction against conventional RFM based churn model," 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 2023, pp. 1-6, doi: 10.1109/ICEEICT56924.2023.10156962.

[15] L. Ou, "Customer Churn Prediction Based on Interpretable Machine Learning Algorithms in Telecom Industry," 2023 International Conference on Computer Simulation and Modeling, Information Security (CSMIS), Buenos Aires, Argentina, 2023, pp. 644-647, doi: 10.1109/CSMIS60634.2023.00120.

[16]S. Preetha and R. Rayapeddi, "Predicting Customer Churn in the Telecom Industry Using Data Analytics," 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, 2018, pp. 38-43, doi: 10.1109/ICGCIoT.2018.8753096.

[17] Jajam Nagaraju , Abbaraju Sai Sathwik, Bheri Saiteja, Nagendra Panini Challa, Beebi Naseeba, "Predicting Customer Churn in Insurance Industry Using Big Data and Machine Learning," 2021.

[18]Lalwani, P., Mishra, M.K., Chadha, J.S. *et al.* Customer churn prediction system: a machine learning approach. *Computing* **104**, 271–294 (2022). https://doi.org/10.1007/s00607-021-00908-y

[19] Feng Xie, Han Yuan, Yilin Ning, Marcus Eng Hock Ong, Mengling Feng, Wynne Hsu, Bibhas Chakraborty, Nan Liu, Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies, Journal of Biomedical Informatics, Volume 126, 2022, 103980, ISSN, 1532-0464, https://doi.org/10.1016/j.jbi.2021.103980

[20] E. Mehmood and T. Anees, "Challenges and Solutions for Processing Real-Time Big Data Stream: A Systematic Literature Review," in IEEE Access, vol. 8, pp. 119123-119143, 2020, doi: 10.1109/ACCESS.2020.3005268.

[21] M. R. Palattella et al., "Internet of Things in the 5G Era: Enablers, Architecture, and Business Models," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 3, pp. 510-527, March 2016, doi: 10.1109/JSAC.2016.2525418.

APPENDIX (FIGURES NOT INCLUDED IN REPORT BODY)

	Age	Number of Dependents	Zip Code	Latitude	Longitude	Number of Referrals	Tenure in Months	Avg Monthly Long Distance Charges	Avg Monthly GB Download	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges	Total Long Distance Charges	Total Revenue
count	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	6361.000000	5517.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000
mean	46.509726	0.468692	93486.070567	36.197455	-119.756684	1.951867	32.386767	25.420517	26.189958	63.596131	2280.381264	1.962182	6.860713	749.099262	3034.379056
std	16.750352	0.962802	1856.767505	2.468929	2.154425	3.001199	24.542061	14.200374	19.586585	31.204743	2266.220462	7.902614	25.104978	846.660055	2865.204542
min	19.000000	0.000000	90001.000000	32.555828	-124.301372	0.000000	1.000000	1.010000	2.000000	-10.000000	18.800000	0.000000	0.000000	0.000000	21.360000
25%	32.000000	0.000000	92101.000000	33.990646	-121.788090	0.000000	9.000000	13.050000	13.000000	30.400000	400.150000	0.000000	0.000000	70.545000	605.610000
50%	46.000000	0.000000	93518.000000	36.205465	-119.595293	0.000000	29.000000	25.690000	21.000000	70.050000	1394.550000	0.000000	0.000000	401.440000	2108.640000
75%	60.000000	0.000000	95329.000000	38.161321	-117.969795	3.000000	55.000000	37.680000	30.000000	89.750000	3786.600000	0.000000	0.000000	1191.100000	4801.145000
max	80.000000	9.000000	96150.000000	41.962127	-114.192901	11.000000	72.000000	49.990000	85.000000	118.750000	8684.800000	49.790000	150.000000	3564.720000	11979.340000

Figure 1: Descriptive statistics of dataset

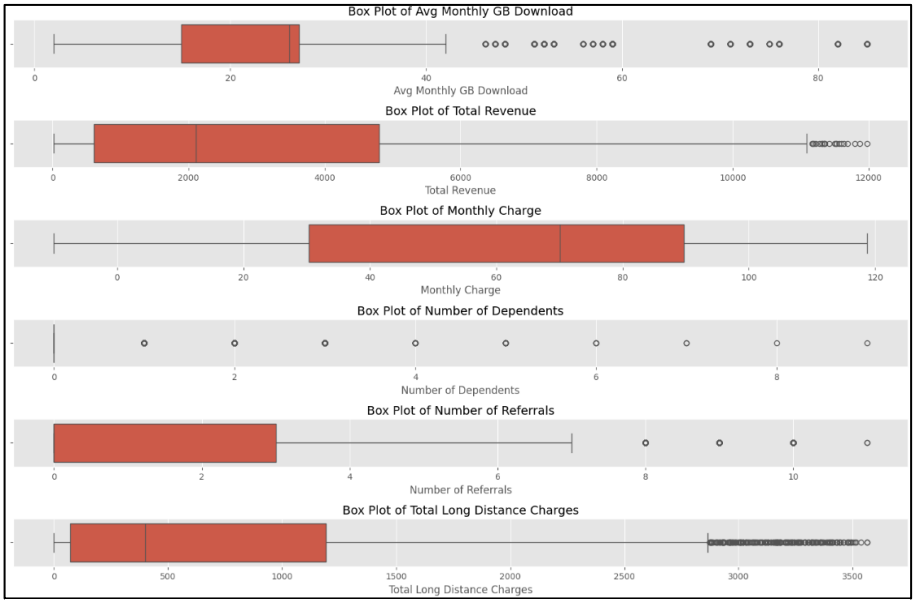

Figure 2. Box Plot of features with outliers

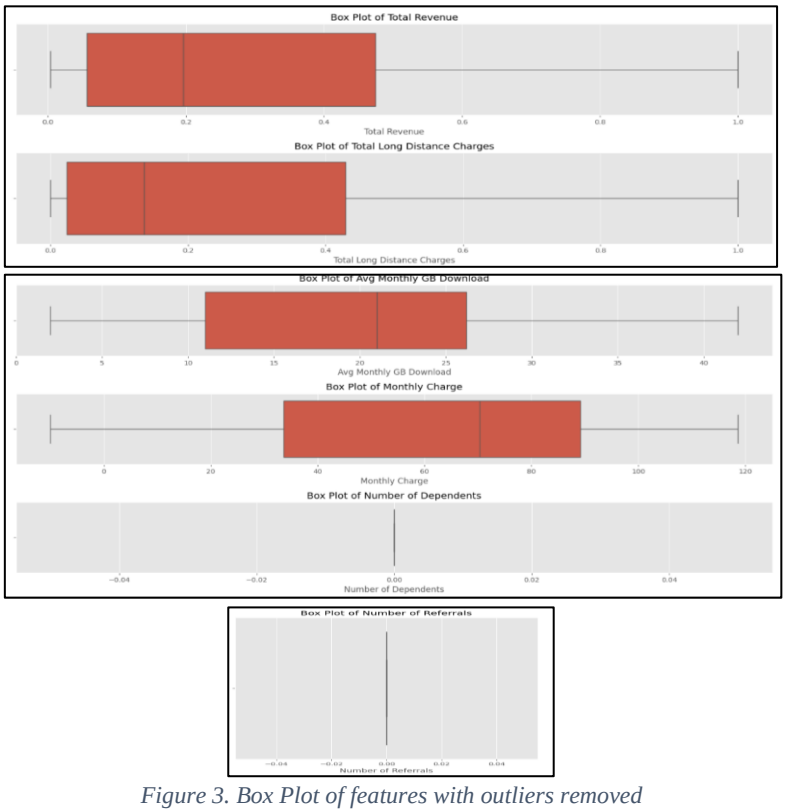

Figure 3. Box Plot of features with outliers removed

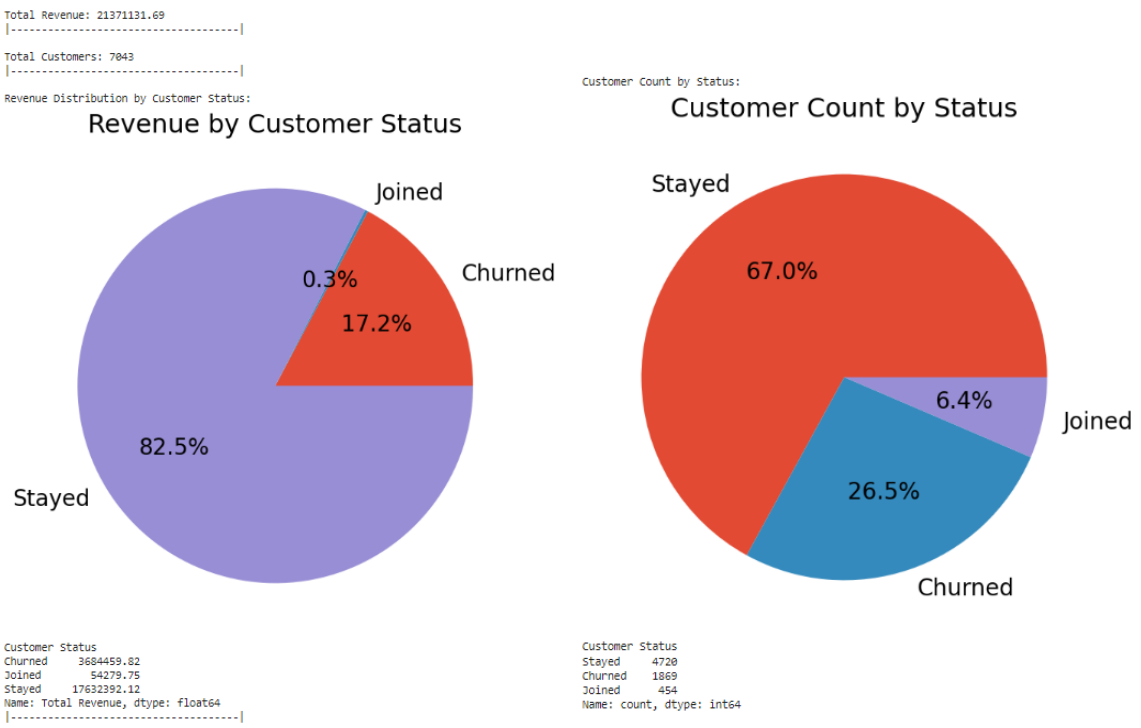

Figure 4. Pie chart distribution by customer status

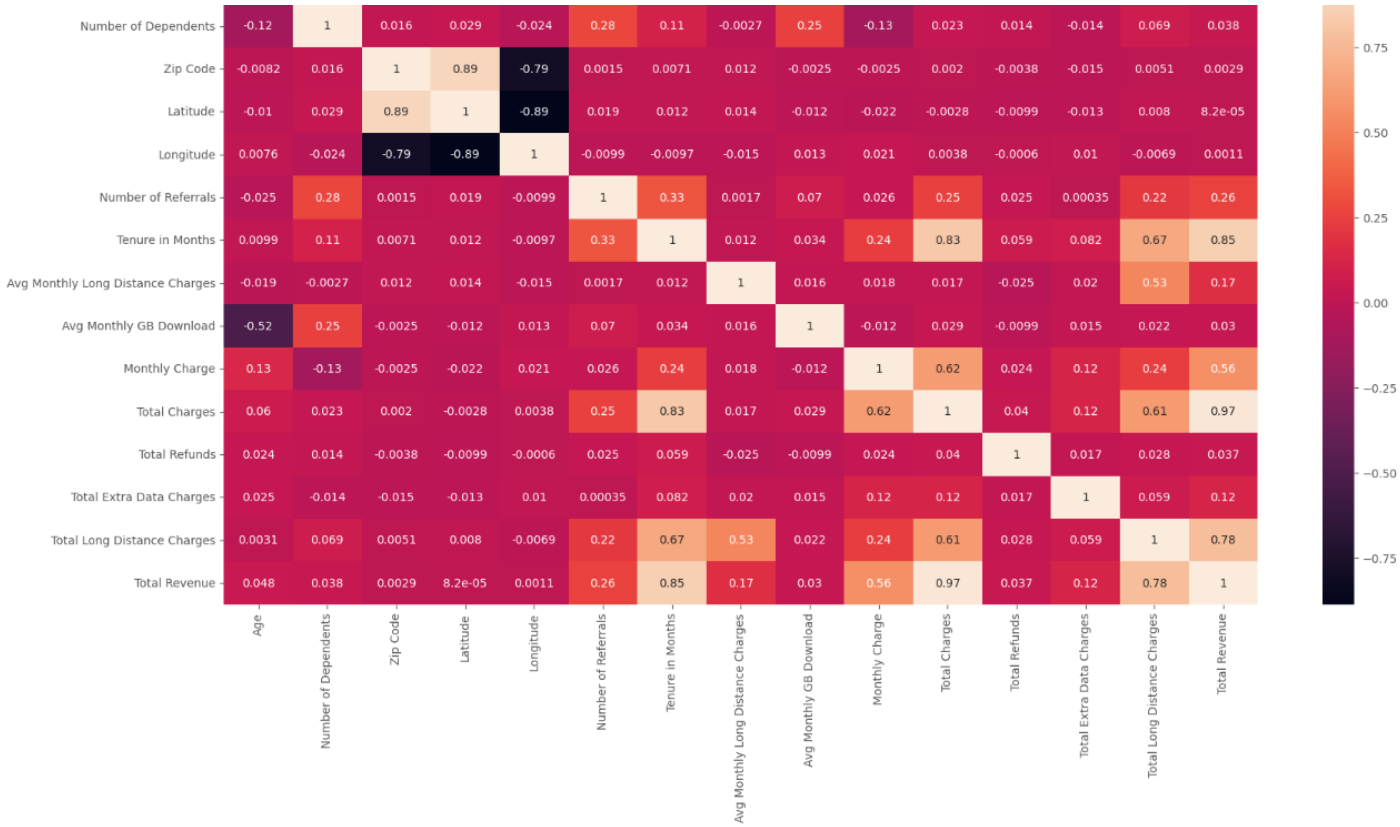

Figure 5: Correlation Matrix heatmap

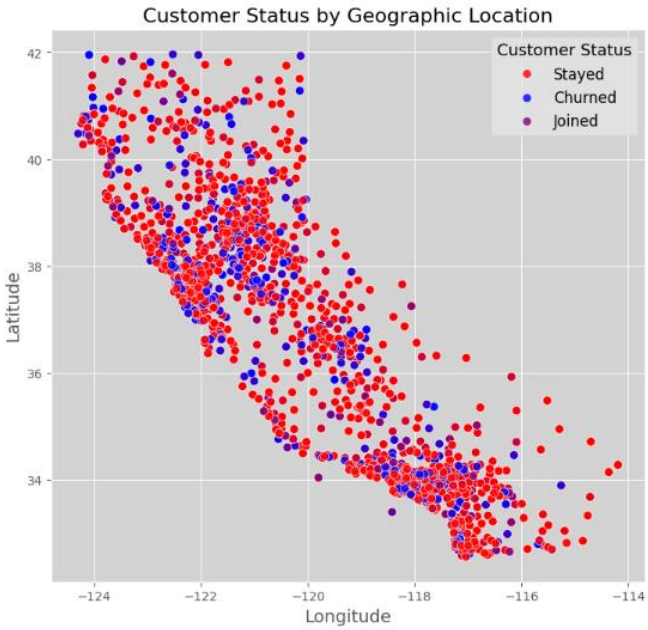

Figure 6: Scatterplot – customer status by geographic location

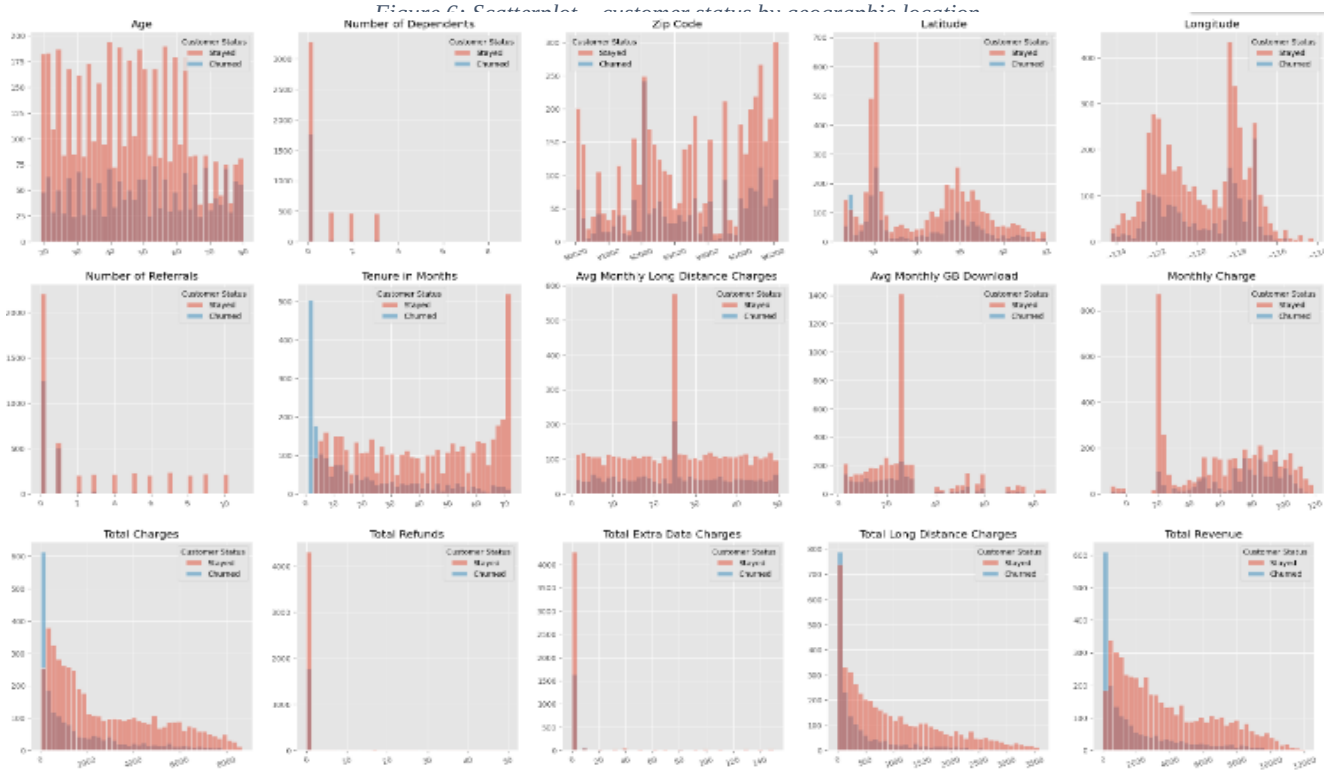

Figure 7: Numerical columns by customer status

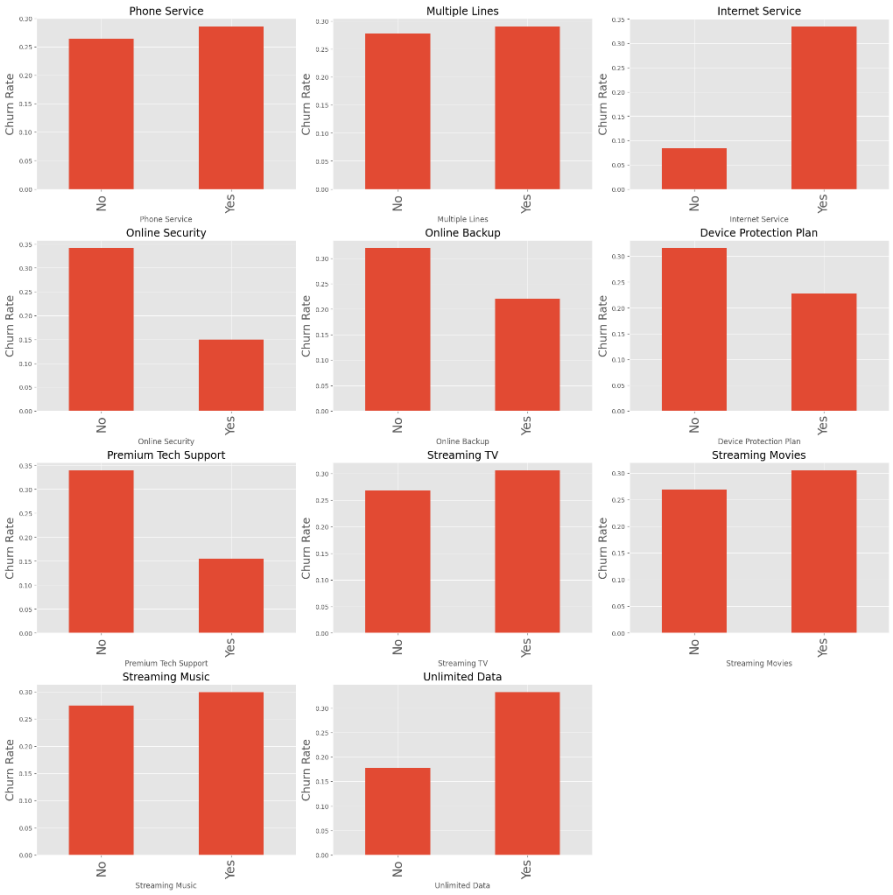

Figure 8: Services by customer status (churned or not)

Hypothesis 1

- Customers with longer subscription durations are less likely to churn.

```
sns.countplot(data=data, x='Contract', hue='Customer Status')
plt.xlabel('Contracts')
plt.ylabel('Count')
plt.title('Distribution of Subscription Durations with Customer Status')
plt.show()
```

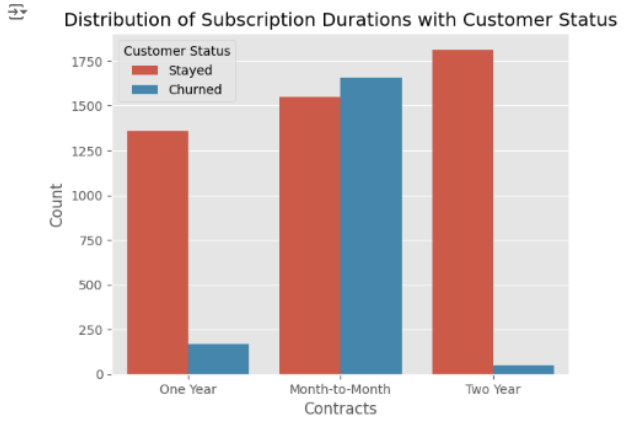

Figure 9: Hypothesis 1 count plot

Hypothesis 2:

- Customers with higher monthly billing amounts are more likely to churn

```
sns.boxplot(x='Monthly Charge', y='Customer Status', data=data)
plt.xlabel('Monthly Charge')
plt.ylabel('Customer Status')
plt.title('Boxplot for Monthly Charge with Respect to Customer Status')
plt.show()
```

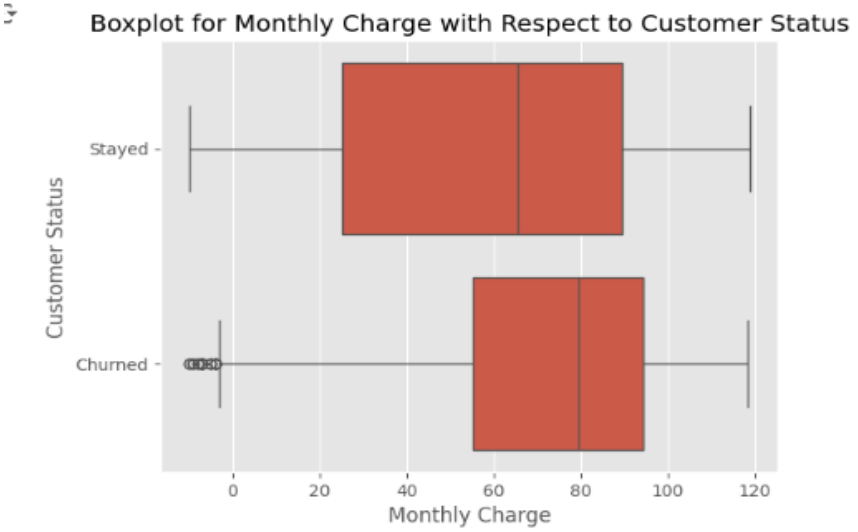

Figure 10: Hypothesis 2 box plot

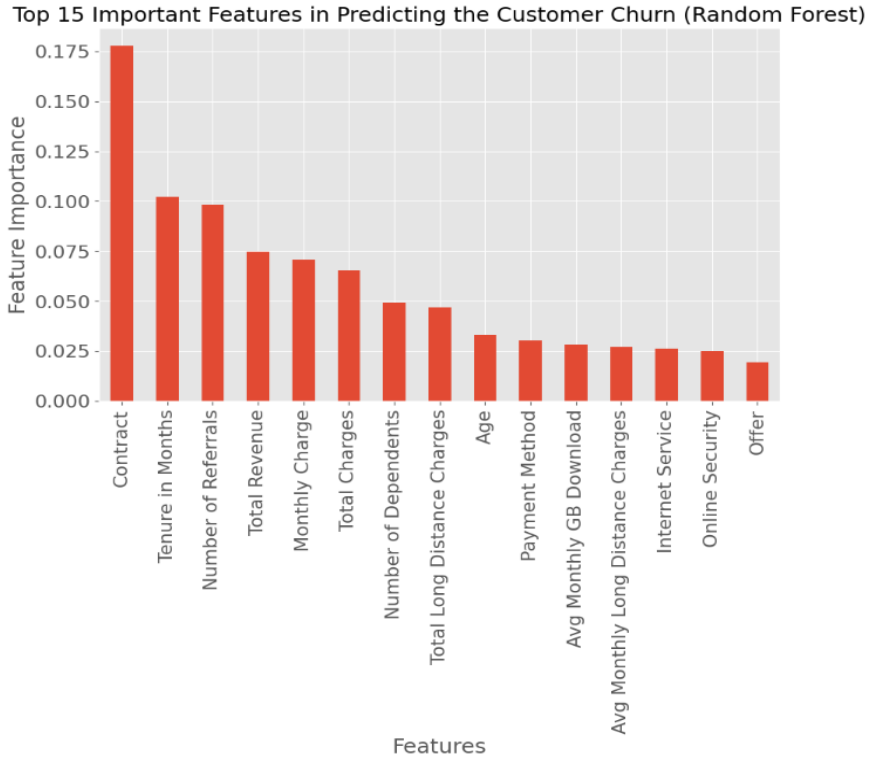

Figure 11: Feature importance for Random Forest

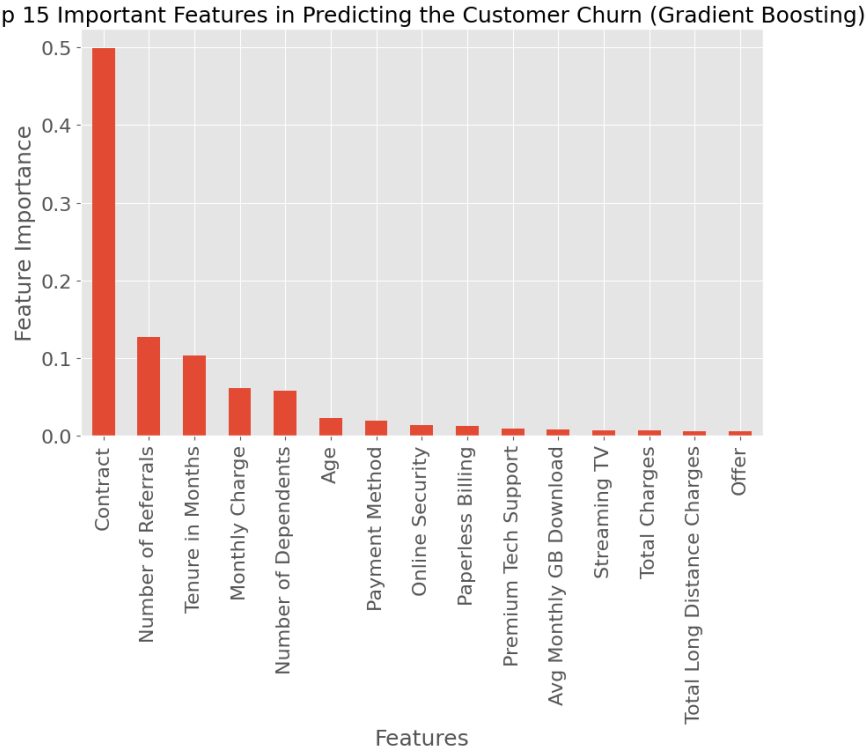

Figure 12: Feature importance for Gradient Boosting

MSc Project - Reflective Essay

Project Title:	Transforming Data into Actional Insights: Unveiling Customer Churn in the Telecom Sector
Student Name:	Sahib Bhatti
Student Number:	190319889
Supervisor Name:	Nafi Ahmad
Programme of Study:	MSc Big Data Science

Introduction

This project has been a journey of discovery and personal development, not just in data science but also in understanding its profound impact on real-world business challenges. The goal was to analyse customer churn in the telecom sector and establish retention strategies through conducting insights from hypothesis testing and predictive modelling. I became deeply engaged by both the technical challenges and the stories behind the data. This essay reflects on the strengths, weaknesses, and ethical considerations encountered, sharing how this experience has enhanced my technical skills and deepened my appreciation for how data can drive meaningful change.

Working on this project was fulfilling, blending my passion for data with understanding human behaviour. Choosing customer churn wasn't just technical; it was personal. With my background in financial services, I've always been intrigued by what drives customer loyalty. The telecom industry provided a perfect setting to explore these dynamics, aligning with my long-term goals. This project was more than an academic exercise; it was an opportunity to contribute to a field I'm passionate about, with the potential to make a tangible difference in business strategies. I aspire to apply the skills learnt in this project in the future, simulating these types of methods in real world applications such as a professional workplace one day.

Strengths

A key strength of my project was that it achieved a remarkable accuracy in predicting customer churn with an 88% success rate, and further exceptional performance metrics. This was achievable due to the approach taken; after understanding the data through various visualisations and analysis, I was able to preprocess the data effectively for the algorithms and utilised feature selection to employ the most relevant variables from the dataset that contributed significantly to the predictive accuracy. Then through tuning the model hyperparameters, I improved model performance and reduced overfitting, and then compared model performances where XGBoost had the greatest success, and thus assessed its feature importance. This combined with my hypothesis tests and visualisation analysis enabled me to effectively create targeted data-driven strategies the telecom company should take to mitigate customer churn. These strategies were mentioned in accordance with the specific determinants and features that caused significant loss of customers.

I was driven not only by a desire to apply data science techniques but also by a passion to develop something that could genuinely help businesses gain deeper insights into their customers and enhance the service they provide. Thus, a major strength of my project was the visualisations I used to manipulate and understand the data. Most modern research in this topic, would conduct some form of data analysis, and then

proceed to shift more effort on the predictive modelling. In my case however, I treated both with equal focus, conducting thorough visualisations and applying hypotheses, then performing predictive modelling. Visualisations enabled me to pinpoint exact areas of focus where targeted interventions can be developed for an instant impact, for example such as revising offer E which had a high churn rate, or improving internet service. The literature emphasized the value of these methods individually, but by integrating them in my project, I was able to validate assumptions, pinpoint key churn factors, and leverage machine learning for highly accurate predictions.

Initially, I incorporated a standard validation set in my code, using a 60% training, 20% validation, and 20% testing ratio. After a progress meeting with my supervisor, I was advised to explore k-fold cross-validation. This shift proved advantageous, as k-fold cross-validation allowed for more effective hyperparameter tuning and model validation. I implemented 10-fold cross-validation for Logistic Regression, Random Forest, Gradient Boosting, and XGBoost, averaging metrics like accuracy, precision, recall, and F1 scores across folds. This method, which repeatedly trains on 9 folds and tests on the 10th, enhanced model generalization and reduced overfitting, ultimately leading to a higher test set accuracy than the standard validation approach. This change significantly strengthened my project by improving overall model performance.

A key strength of this project, which some may consider subjective or even a weakness, lies in my decision to retain outliers. While initially removing them resulted in distorted results, retaining these outliers proved critical. This approach preserved the integrity of key metrics, such as total revenue, directly influencing the model's accuracy and offering more realistic and actionable insights into customer behaviour. These outliers were not extreme; they significantly contributed to key statistics and ensured that the analysis remained true to the company's reality. By keeping them, I enhanced our machine learning model’s accuracy, particularly in predicting customer churn, and identified key features since the outliers improved their weight, total revenue appeared as a top 4 feature in the Random Forest feature importance chart, and as lower yet still relevant feature in the Gradient Boosting feature importance chart. This decision allowed the analysis to reflect the real-world scenario of the telecom industry, preventing the oversight of important patterns in the data and leading to more informed business decisions.

To justify my reasons for keeping outliers, I demonstrate effects of removing them here:

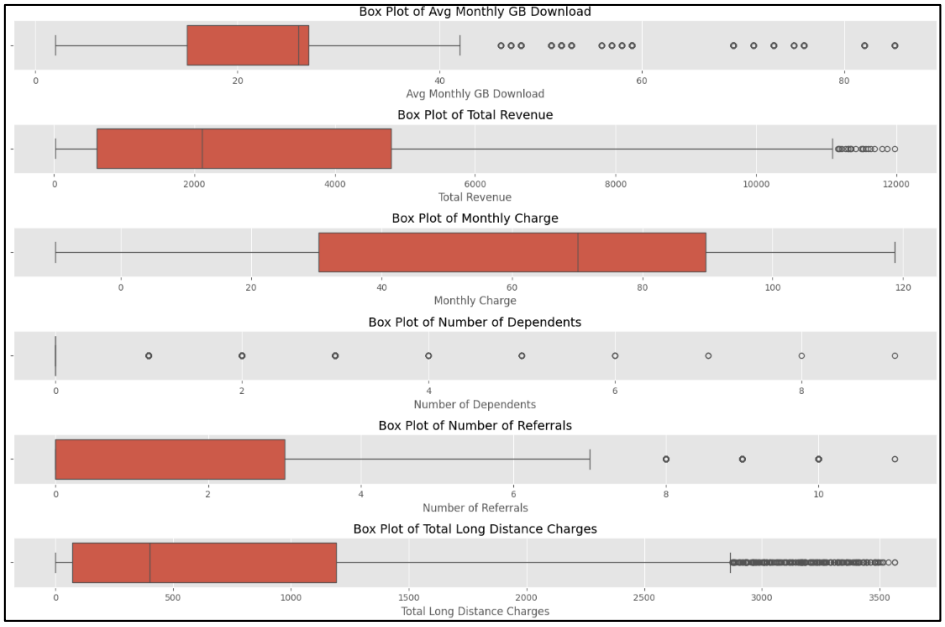

Figure 1. Box Plot of features with outliers

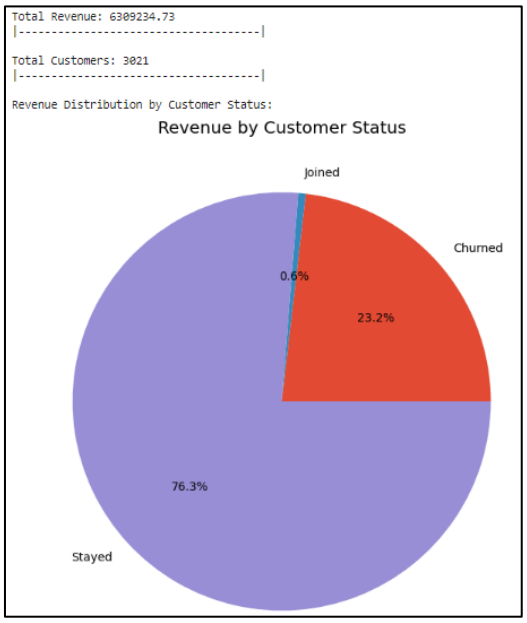

Figure 3. Distorted key Metrics due to outlier removal

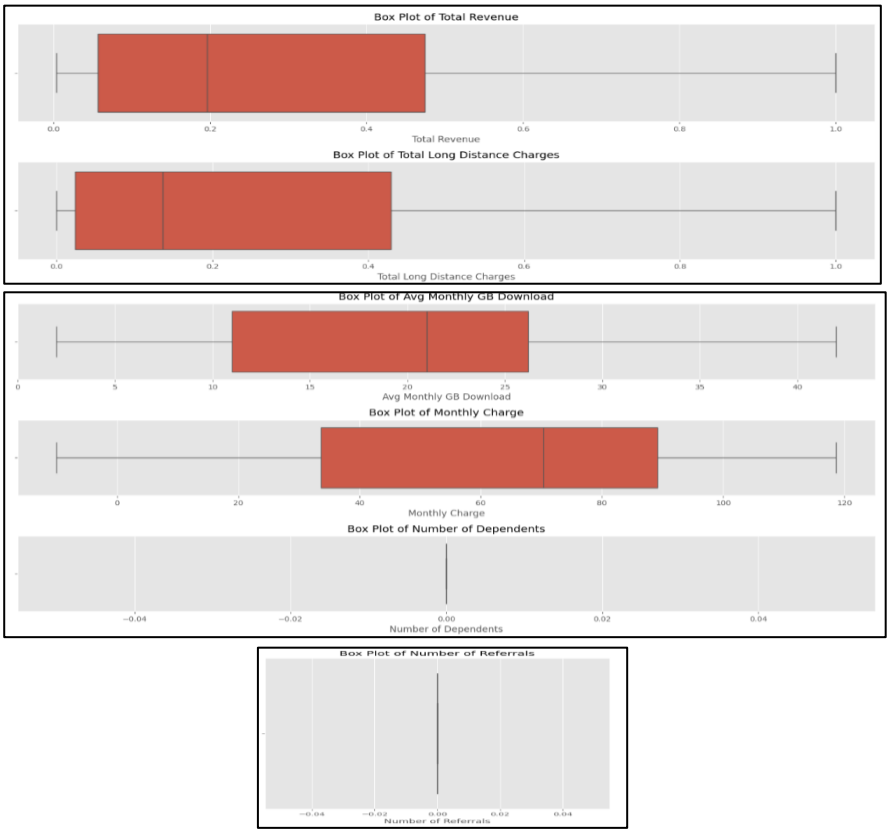

Figure 2. Box Plot of features with outliers removed

```
XGBoost - Accuracy: 0.8257
XGBoost - Precision: 0.8537
XGBoost - Recall: 0.8614
XGBoost - F1 Score: 0.8576
```

Figure 4. Reduced final model accuracy on test set due to outlier removal

As we can see, removing them would have led to an inaccurate representation of the data, analysis and inaccurate (more than 5% accuracy lost) algorithms in this case.

Weaknesses

One aspect of the project that could have been enhanced is the scope of feature engineering. While the existing features were effectively utilized, the project could have further explored additional feature creation or transformation. Feature engineering plays a crucial role in uncovering complex patterns within the data. By primarily relying on the original features, the model's ability to capture nuanced customer behaviours may have been limited. Expanding this area might have led to more robust insights and improved predictive performance. While the initial features provided a strong base, deeper exploration remained untapped. Developing or transforming features could have revealed intricate customer behaviour patterns. For instance, creating interaction terms, polynomial features, or segmenting customers based on derived metrics could have added layers of insight. By not fully expanding on feature engineering, the project may have missed out on identifying some complex relationships within the dataset, which could have further strengthened the overall analysis and outcomes. Nonetheless, the existing features were sufficient to achieve significant results, making this a subtle area for improvement rather than a glaring weakness.

If interpreted as a general customer churn study, and not a study just focused on the specific telecom company residing in California then it can be said that my project missed the opportunity to incorporate external data sources that could have provided a richer understanding of customer churn. Factors like economic trends, industry shifts, or competitor activities were not considered, which could have offered additional context and deeper insights into the customer behaviours and churn patterns that emerged from the internal data alone. Including such external data could have revealed correlations or trends not apparent when only internal data is analysed.

Furthermore, the project adhered to traditional data analysis and machine learning methods. While effective in achieving high accuracy and meeting the project's goals, this may have limited the exploration of potentially more innovative or cutting-edge approaches. Techniques like deep learning, neural networks, or more advanced feature engineering could have been explored to push the analysis boundaries. These methods might have uncovered new patterns or insights, offering a different perspective on customer churn. However, the decision to focus on well-established methods was deliberate, ensuring that the project aligned with its primary objective of developing actionable retention strategies. This approach was successful, but exploring newer methodologies could have potentially added more layers of understanding and value to the findings.

Further Work

Expanding the scope of feature engineering is a promising direction for further work. By delving into additional feature creation and transformation, such as interaction terms, polynomial features, or customer segmentation based on derived metrics, the analysis could uncover previously hidden relationships within the data [1]. This could lead to more accurate predictions and deeper insights into customer behaviour patterns. Feature engineering is crucial in capturing the underlying factors that influence customer decisions, and further exploration in this area could significantly enhance the model's performance and the overall quality of insights derived from the data.

Another valuable avenue for future work involves incorporating external data sources. By integrating data such as social media sentiment analysis, economic indicators, or competitor actions, the analysis could provide a more comprehensive understanding of the factors influencing customer churn [2]. External factors often play a critical role in

shaping customer decisions, and understanding these influences could offer a broader perspective on the challenges faced by the telecom industry. Enriching the dataset with external data would allow for a more holistic analysis, potentially revealing correlations or trends that are not immediately apparent when relying solely on internal data.

Lastly, integrating real-time data streaming into predictive models could transform churn prediction. Real-time data processing would enable the company to identify at-risk customers immediately and take swift, targeted actions to retain them [3]. This dynamic and responsive approach could significantly enhance the effectiveness of customer retention strategies, providing a competitive advantage. Additionally, exploring more advanced machine learning techniques, such as deep learning models, could further improve the accuracy and robustness of churn predictions. These methods could uncover new patterns or insights, offering a more sophisticated understanding of customer behaviour in real-time, ultimately improving decision-making and customer satisfaction.

Work That Could Have Been Conducted with More Time

Given more time, I would have expanded the scope of my project by incorporating additional external datasets from various telecom companies across California. This would have enriched the dataset, allowing for a more comprehensive analysis and improving the generalization of my machine learning algorithms. By integrating data from different geographical locations within California, I could have uncovered regional trends and patterns in customer behaviour that my project did not fully explore. This broader perspective would have provided deeper insights into the factors influencing customer churn across different regions.

I would have also delved deeper into advanced machine learning techniques, particularly deep learning models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. These models are adept at capturing temporal dependencies and long-term patterns in customer behaviour, which could have led to more precise churn predictions [4]. Exploring such techniques would have been an opportunity to push the boundaries of the project's predictive capabilities.

Furthermore, a detailed exploration of the impact of emerging technologies, such as 5G and IoT, on customer behaviour could provide critical insights into future retention strategies [5]. These technologies are rapidly shaping consumer experiences, making it essential to understand their influence on customer satisfaction and churn rates.

Critical Analysis of the Relationship Between Theory and Practical Work

In this project, the blend between theory and practice was essential. While theoretical knowledge provided a strong basis for choosing machine learning algorithms and statistical methods, practical challenges often dictated how these were actually applied. For example, even though theory suggested removing outliers, I opted to keep them to maintain the accuracy and relevance of key metrics, ensuring that the insights reflected the real-world environment of the company. This choice highlighted the need for adaptability when translating theoretical concepts into real-world applications. Throughout the project, I realized that data doesn't always fit neatly into theoretical frameworks. Addressing data quality issues, managing outliers, and emphasizing feature importance required more than just textbook knowledge; it involved making thoughtful, context-specific decisions to achieve meaningful results.

Another point to consider in the critical analysis of the relationship between theory and practical work is the challenge of model interpretability versus performance. In theory, complex models like deep learning can offer superior predictive power. However, in practice, these models often lack transparency, making it difficult to explain predictions to stakeholders. This project had to balance the trade-off between using interpretable models, like logistic regression, and more complex models, like XGBoost, which offered better accuracy but at the cost of reduced interpretability. This balance is crucial in real-world applications where stakeholders need to understand and trust the decisions made by machine learning models.

Awareness of Legal, Social, Ethical Issues, and Sustainability

Throughout the project, I remained acutely aware of the legal, social, and ethical implications of handling customer data. Ensuring compliance with data protection regulations, such as GDPR, was a priority. All data processing steps were designed to protect customer anonymity and prevent the misuse of sensitive information, which is crucial in maintaining customer trust and adhering to legal standards.

From a social and ethical perspective, the project emphasized the importance of balancing targeted marketing efforts with ethical considerations. Predicting customer churn can lead to more effective retention strategies, but it is essential to avoid manipulative tactics that could harm vulnerable customer groups. The project also underscored the need for transparency in how customer data is used, ensuring that customers are aware of and consent to the analysis of their data.

Sustainability was another key consideration. By focusing on long-term customer retention strategies, the project aimed to minimize the resources spent on acquiring new customers, leading to more sustainable business practices. Understanding and addressing the reasons behind customer churn can lead to improved customer satisfaction and loyalty, contributing to a more sustainable customer base.

In conclusion, the project provided valuable insights into the factors driving customer churn in the telecom industry. The decision to retain outliers, combined with the rigorous application of machine learning techniques, ensured that the analysis remained true to the company’s reality and delivered actionable recommendations for improving customer retention. Moving forward, further exploration of advanced techniques and external data sources could provide even deeper insights, while maintaining a strong focus on legal, social, and ethical considerations will be crucial for the ongoing success of customer churn analysis.

References:

[1] S. Boeschoten, C. Catal, B. Tekinerdogan, A. Lommen, and M. Blokland, "The automation of the development of classification models and improvement of model quality using feature engineering techniques," Expert Systems With Applications, vol. 213, p. 118912, 2023.

[2] S. Lamrhari, H. El Ghazi, M. Oubrich, and A. El Faker, "A social CRM analytic framework for improving customer retention, acquisition, and conversion," Technological Forecasting and Social Change, vol. 174, 2022, Art. no. 121275. doi: 10.1016/j.techfore.2021.121275.

[3] E. Mehmood and T. Anees, "Challenges and Solutions for Processing Real-Time Big Data Stream: A Systematic Literature Review," in IEEE Access, vol. 8, pp. 119123-119143, 2020, doi: 10.1109/ACCESS.2020.3005268.

[4] Feng Xie, Han Yuan, Yilin Ning, Marcus Eng Hock Ong, Mengling Feng, Wynne Hsu, Bibhas Chakraborty, Nan Liu, Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies, Journal of Biomedical Informatics, Volume 126, 2022, 103980, ISSN, 1532-0464, https://doi.org/10.1016/j.jbi.2021.103980

[5] M. R. Palattella et al., "Internet of Things in the 5G Era: Enablers, Architecture, and Business Models," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 3, pp. 510-527, March 2016, doi: 10.1109/JSAC.2016.2525418.