PROJECT 3

Project Overview

Background Information

Domain

In the ever-evolving landscape of the telecommunications industry, SyriaTel, a key player, confronts the critical challenge of customer churn—a phenomenon where subscribers discontinue their association with the provider. Against the backdrop of rapid technological advancements, shifting consumer preferences, and intense market competition, the need to understand and predict customer behavior is more pressing than ever. This project emerges as a response to this imperative, aspiring to construct a sophisticated binary classification model that not only anticipates customer churn but provides actionable insights for SyriaTel's strategic approach to customer retention.

The telecommunications sector is characterized by its dynamic nature, where intricate interactions between service quality, pricing structures, and customer expectations influence the longevity of customer relationships. Customer churn, beyond being a financial concern, is a strategic puzzle that demands nuanced solutions. Our project is conceived to delve into the multifaceted aspects influencing churn, acknowledging that customer behavior is shaped by geographical variations, plan utilization patterns, and subtle indicators embedded in customer service interactions.

Amidst this complexity, challenges such as data heterogeneity, feature identification, and nuanced indicator interpretation underscore the need for a comprehensive solution. Our proposed approach transcends the conventional by integrating geographical insights, granular usage metrics, and the subtle nuances of customer service interactions into the predictive model. The objective is not solely to predict churn but to empower SyriaTel with a strategic compass—enabling the anticipation of customer attrition and guiding tailored, preemptive measures for customer retention. As we embark on this journey, the goal is to contribute to SyriaTel's resilience, fostering a satisfied and loyal customer base in the dynamic landscape of the telecommunications market.

Challenges

- Data Variability The SyriaTel customer churn dataset reflects diverse customer profiles, usage patterns, and service interactions. This variability poses a challenge in accurately predicting churn and assessing the impact of different factors on customer retention.
- Non-Normality of Residuals:Violation of the assumption of normally distributed residuals in the SyriaTel dataset can impact the validity of statistical tests and confidence intervals in the customer churn prediction model.
- Model Evaluation: Selecting appropriate evaluation metrics for the SyriaTel customer churn model and defining acceptable performance levels are complex tasks. The dynamic nature of the telecom industry requires a robust modeling approach that considers changing customer behaviors over time.
- Outliers Challenge: Unusual customer profiles or unique interactions, such as high-value corporate accounts or specialized service demands, may introduce outliers affecting the performance of the customer churn prediction model.

- Data Imbalance: Underrepresentation of certain customer segments, such as corporate clients or specific usage categories, in the SyriaTel dataset can lead to biased predictions, impacting the reliability of the customer churn model.
- Categorical Data Exploration: Handling and exploring categorical variables in the SyriaTel dataset, such as service plans or customer segments, is challenging, especially when there are numerous categories. Establishing meaningful relationships with the target variable (churn) requires careful consideration.
- Missing Data Handling: Incomplete or missing data in the SyriaTel dataset, such as gaps in customer service records, can impact the quality of exploratory analysis and subsequent modeling. Deciding how to impute or handle missing data is a significant challenge.
- Multicollinearity: High correlation among predictor variables in the SyriaTel dataset, such as call duration and data usage, complicates the assessment of each variable's unique contribution and can lead to unstable coefficient estimates in the customer churn prediction model.

Solution

To effectively address the challenges posed by the SyriaTel customer churn dataset, a structured approach involving data preparation, exploratory data analysis (EDA), and multiple linear regression analysis techniques will be employed:

- Data Preparation: Commence with data preparation, focusing on handling missing values and encoding categorical variables. This ensures that the dataset is in a standardized and analyzable format, laying a robust foundation for subsequent analysis.
- EDA (Exploratory Data Analysis): Undertake a comprehensive exploratory data analysis to delve into the nuances of the dataset. This process involves understanding data variability, identifying outliers, assessing heteroscedasticity, and detecting multicollinearity. EDA provides crucial insights, guiding the selection of relevant features for the subsequent regression analysis.
- Regression Analysis: Proceed to build a multiple linear regression model, utilizing the features identified during EDA. This step involves addressing multicollinearity challenges and assessing the model's performance. The goal is to ensure that the regression model aligns with the assumptions of linearity, normality of residuals, and homoscedasticity. Rigorous model evaluation will be conducted to guarantee its reliability.

By systematically implementing these steps, we aim to not only mitigate the challenges posed by data variability, non-normality of residuals, model evaluation complexities, outliers, data imbalance, categorical data intricacies, missing data, and multicollinearity but also to create a robust predictive model tailored to the nuances of the SyriaTel customer churn dataset. This comprehensive approach ensures that the model not only accurately predicts customer churn but also provides actionable insights for SyriaTel's strategic customer retention initiatives in the dynamic telecom industry.

Conclusion

In the culmination of our project, our focus lies in addressing the distinct challenges presented by the SyriaTel customer churn dataset. The primary goal is to derive actionable insights that empower the telecom industry, particularly SyriaTel, to proactively manage and reduce customer churn. By

aligning our project objectives with the specific challenges inherent in the dataset, we have devised a structured and comprehensive approach encompassing data preparation, exploratory data analysis (EDA), and multiple linear regression analysis.

Problem Statement

In the realm of telecommunications, customer churn remains a critical challenge, and SyriaTel, a prominent player in the industry, is not immune to its impact. The dataset at hand encapsulates a diverse array of customer attributes, usage patterns, and service interactions, presenting a complex scenario for predicting customer churn accurately. The multifaceted challenges inherent in the dataset include data variability, non-normality of residuals, model evaluation complexities, outliers, data imbalance, categorical data intricacies, missing data, and multicollinearity.

Objectives

1. To construct a predictive model for customer churn that takes into account geographical variations, providing insights into regional variations in churn likelihood.
2. To explore usage metrics within the SyriaTel dataset and integrate them into the churn prediction model, enhancing its predictive accuracy by considering customer usage patterns.
3. To analyze customer service call data and extract relevant features, incorporating them into the churn prediction model. This ensures that customer service interactions are integral to predicting churn.
4. To develop a predictive model that anticipates customer churn for SyriaTel, proactively identifying potential churners before they discontinue their association with the telecom company.
5. To select suitable machine learning algorithms for the churn prediction model and train them on the preprocessed dataset. Consideration will be given to hyper parameter tuning to optimize algorithm performance.
6. To rigorously evaluate the performance of the developed predictive model, emphasizing interpretability. Insights gained will be translated into actionable recommendations for SyriaTel to enhance customer retention strategies in the telecom business.

Data Understanding

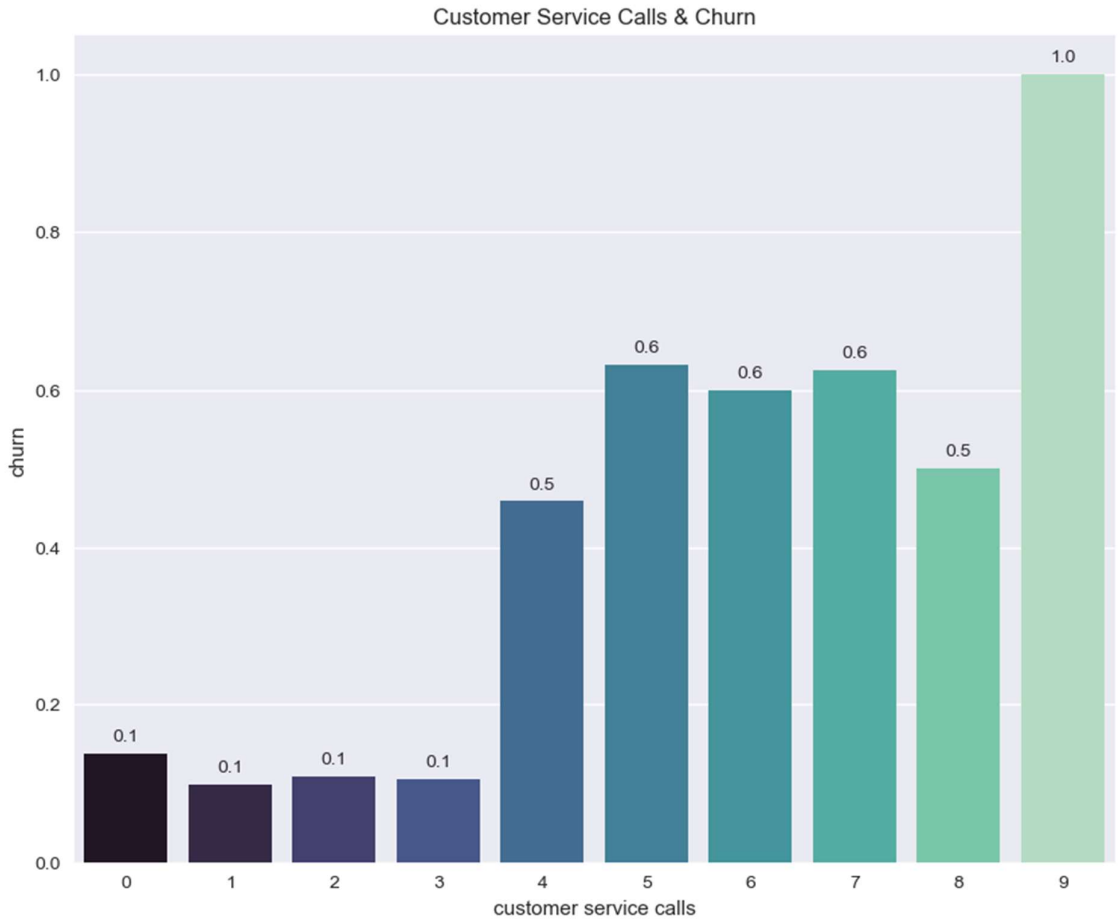
This project utilizes the SyriaTel Customer Churn dataset, accessible in the file 'bigml_59c28831336c6604c800002a.csv' . The dataset comprises a comprehensive collection of information pertaining to customer interactions, usage patterns, and demographic details within the telecommunications domain. The dataset serves as the foundation for predicting customer churn and formulating actionable insights for SyriaTel in the dynamic telecom industry

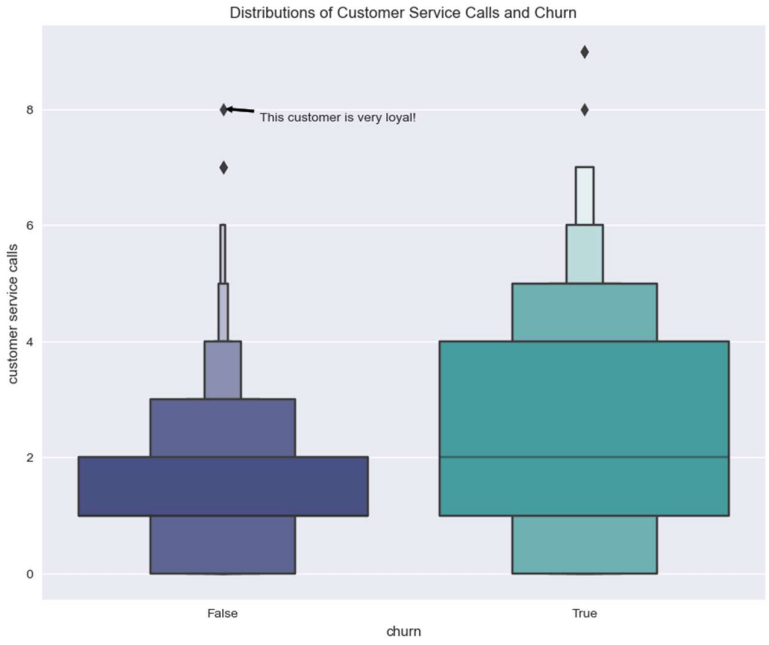
Data Preparation

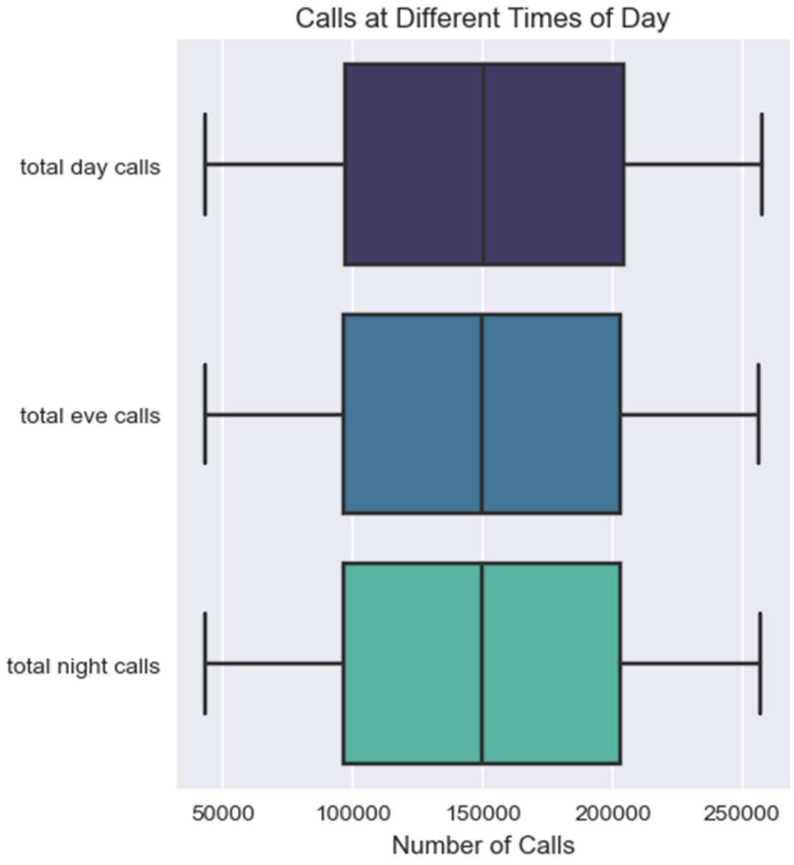
The data utilized in our analysis was sourced from a CSV file, syriatel_churn_data.csv, and Pandas library was employed to import and structure the data into a DataFrame. The DataFrame encapsulates various features such as customer ID, date, pricing details, bedrooms, bathrooms, floor information, among others. In our pursuit of data accuracy and consistency, we conducted crucial data cleaning and preprocessing tasks. This involved addressing missing values, eliminating unnecessary rows, and ensuring uniform data types throughout the dataset. These measures were

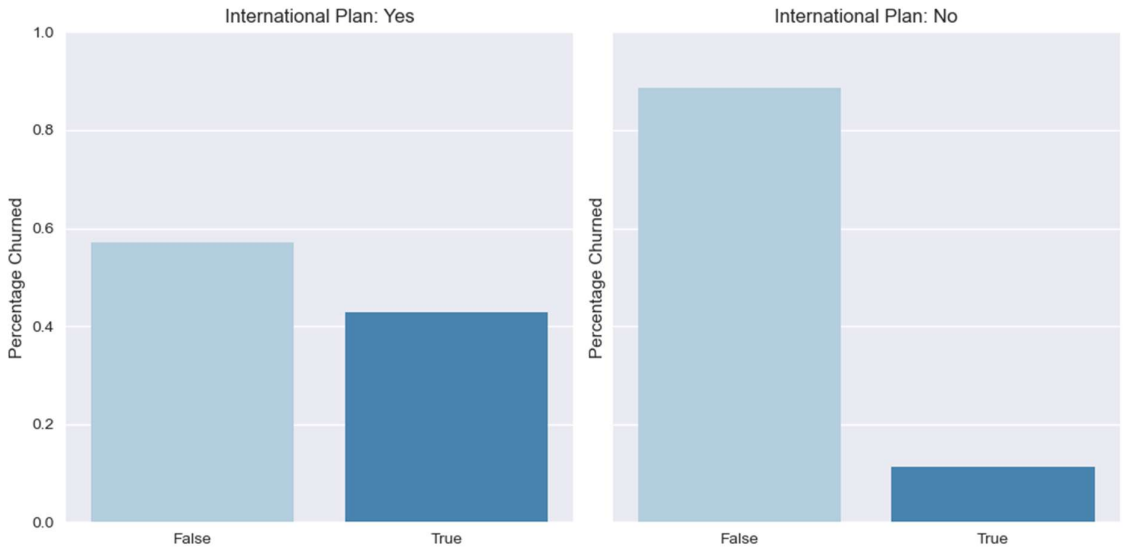
imperative in preparing our dataset for robust analysis. The transformation of the raw dataset into a meticulously structured DataFrame laid the groundwork for a comprehensive and insightful analysis of customer churn dynamics within SyriaTel's telecom domain.

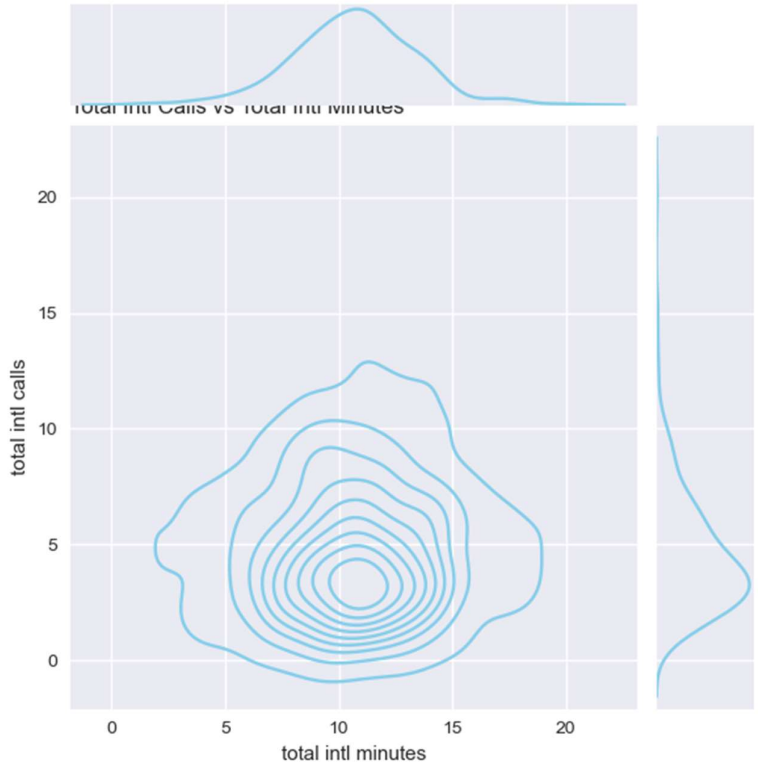
Data analysis

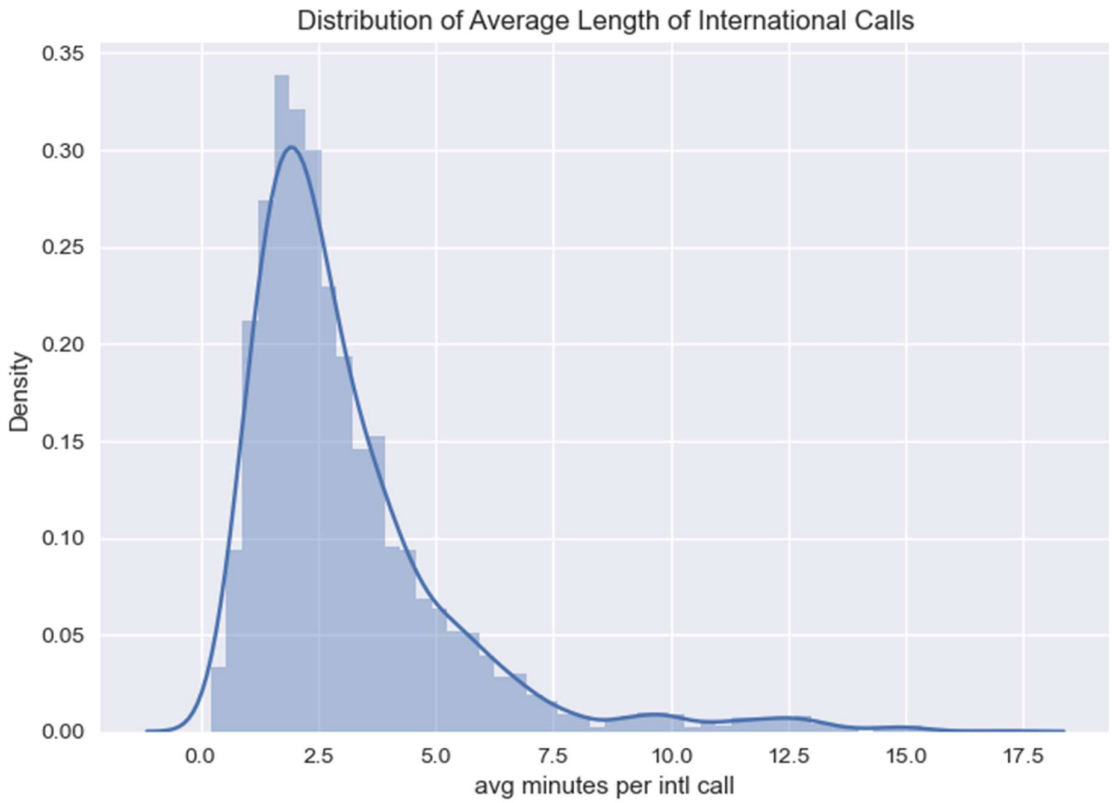


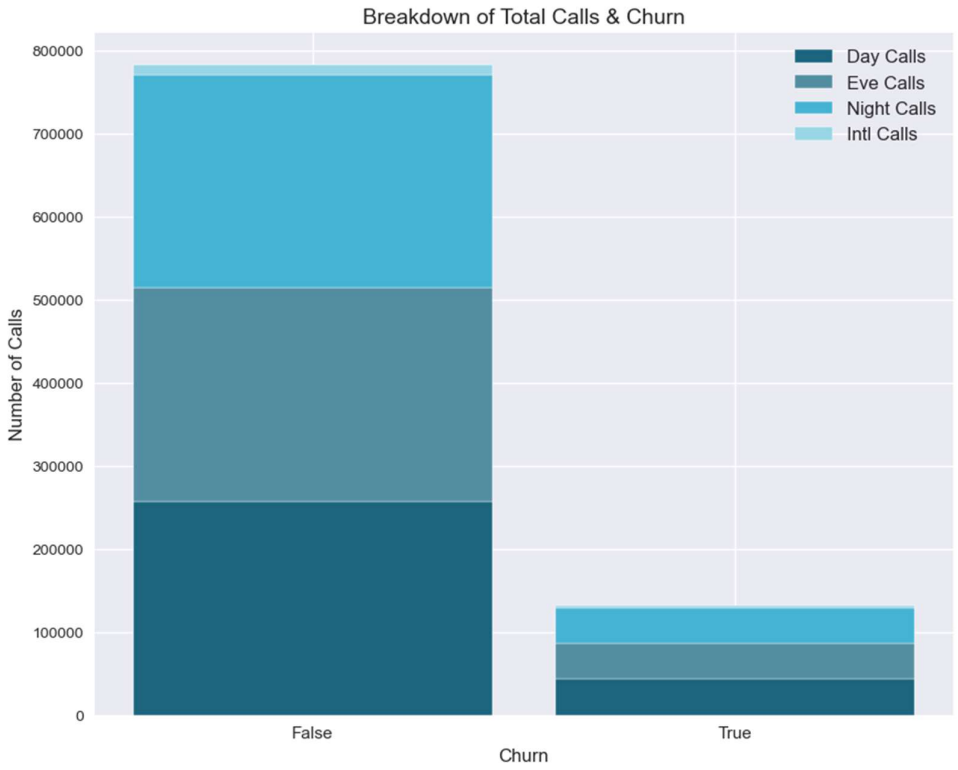

Distributions of Customer Service Calls and Churn
This customer is very loyal!
customer service calls
0
2
4
6
8
False
True
churn

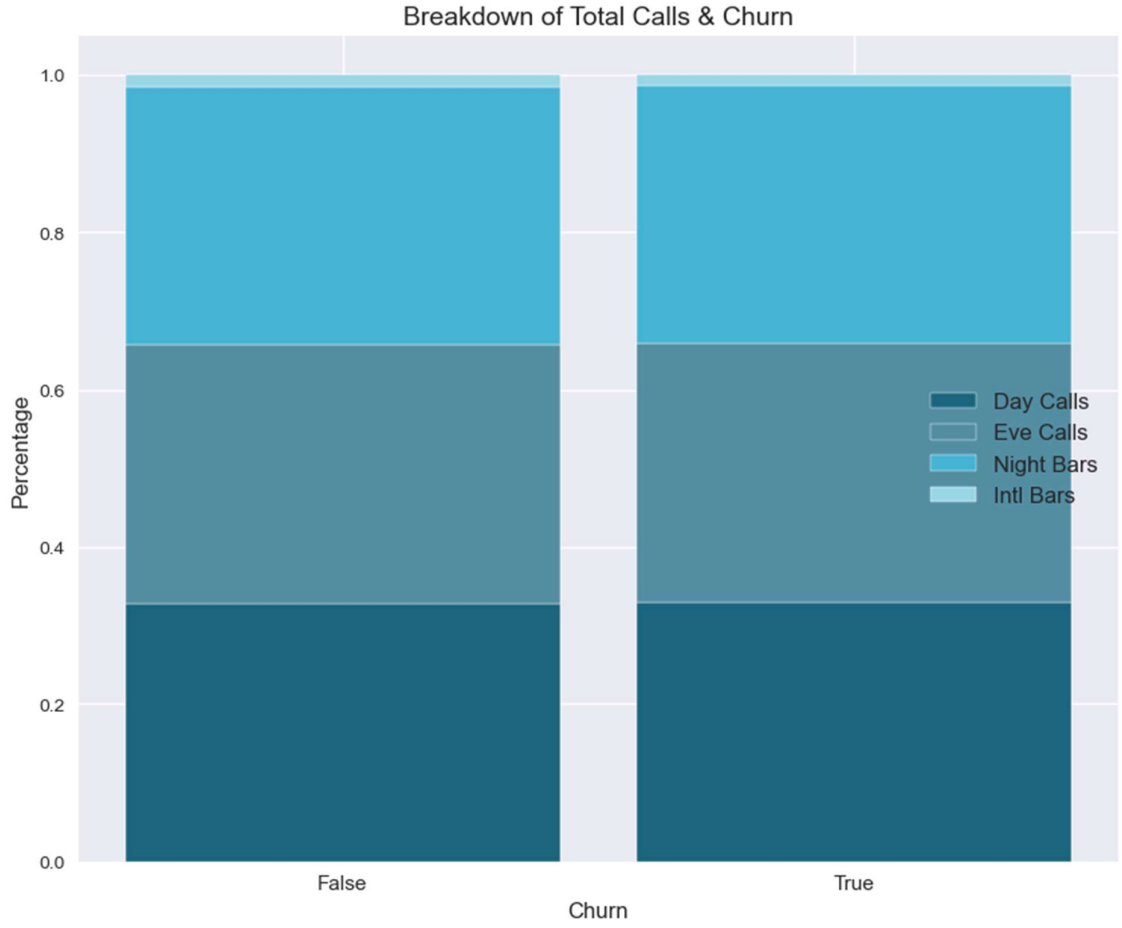

Calls at Different Times of Day
total day calls
total eve calls
total night calls
50000
100000
150000
200000
250000
Number of Calls

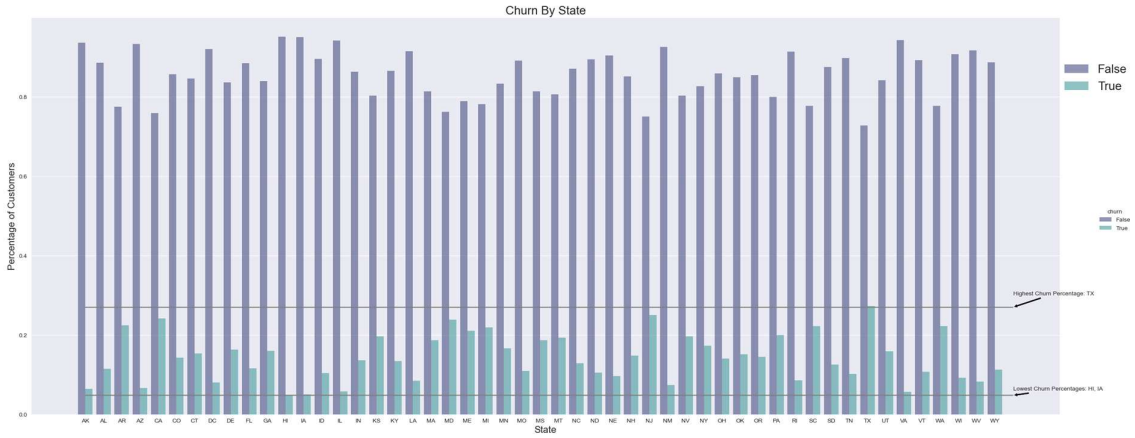

International Plan: Yes
International Plan: No
Percentage Churned
1.0
0.8
0.6
0.4
0.2
0.0
False
True

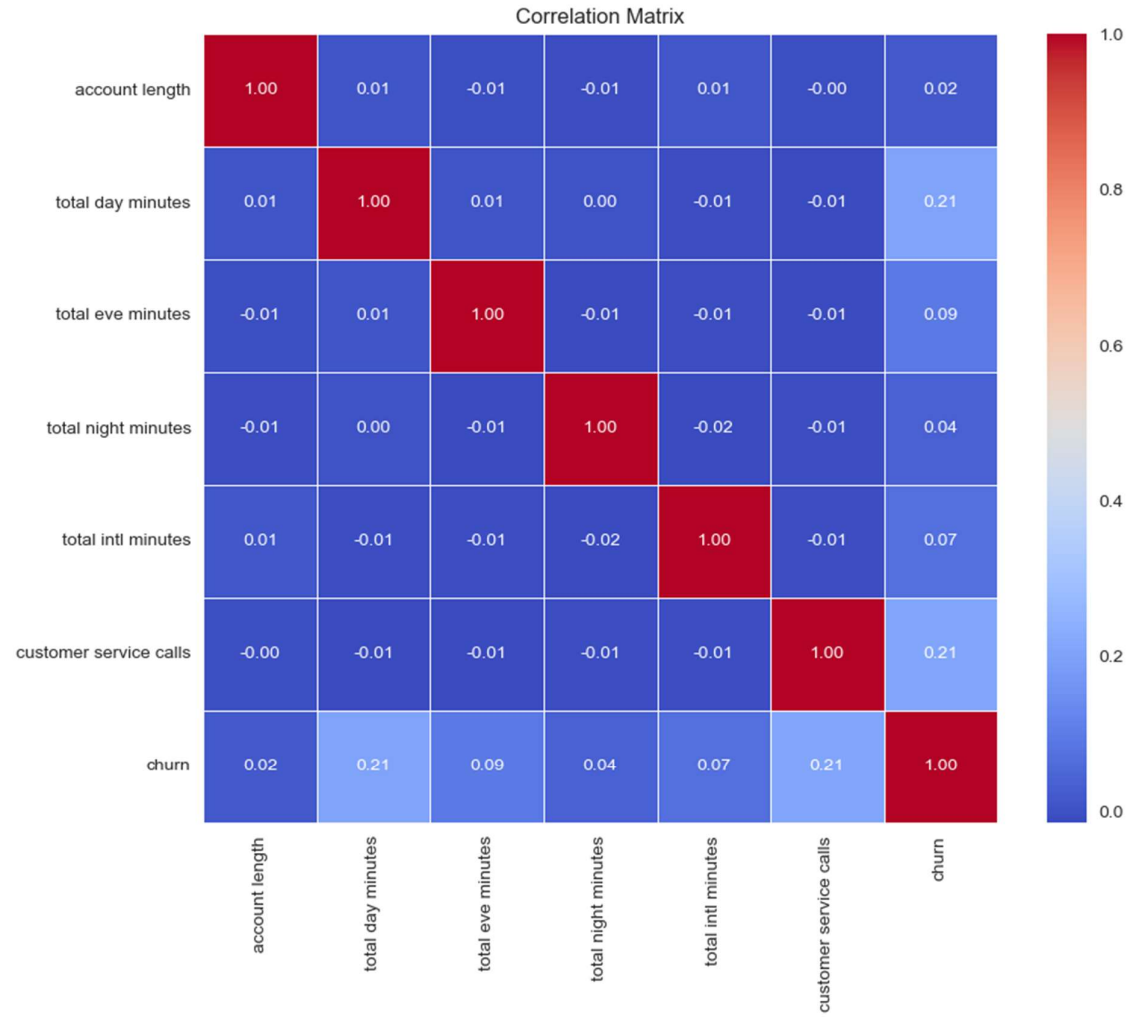

total intl calls
total intl minutes
20
15
10
5
0
0
5
10
15
20


Distribution of Average Length of International Calls
Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
avg minutes per intl call


Breakdown of Total Calls & Churn
Day Calls
Eve Calls
Night Calls
Intl Calls
Number of Calls
800000
700000
600000
500000
400000
300000
200000
100000
0
False
True
Churn


Breakdown of Total Calls & Churn
Day Calls
Eve Calls
Night Bars
Intl Bars
Percentage
1.0
0.8
0.6
0.4
0.2
0.0
False
True
Churn





Positive Correlation (Values closer to 1):

'total day minutes' has a positive correlation of approximately 0.21 with 'churn.' This suggests that customers with higher total day minutes are slightly more likely to churn.

'customer service calls' has a positive correlation of approximately 0.21 with 'churn.' This indicates that as the number of customer service calls increases, the likelihood of churn also increases.

Negative Correlation (Values closer to -1):

'number vmail messages' has a negative correlation of approximately -0.09 with 'churn.' This implies that customers with more voice mail messages are slightly less likely to churn.

Weak Correlations (Values close to 0):

Most other correlations are relatively weak, indicating a limited linear relationship.

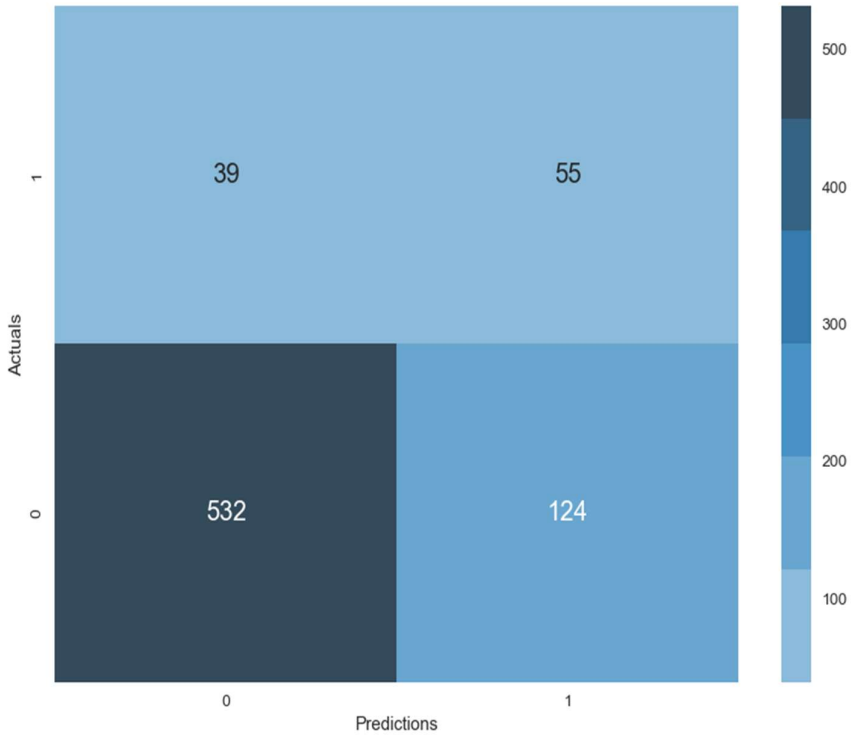
Modeling

In the modeling phase, I began by preprocessing the training dataset, saving it as 'training_set.csv' for future use. To address the issue of class imbalance, I implemented the Synthetic Minority Over-sampling Technique (SMOTE), oversampling the minority class to achieve a more balanced distribution. Subsequently, I considered a variety of classification models, including Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, Gaussian Naive Bayes, and Support Vector Machines (SVM), to identify the most suitable model for the task. To streamline the process and ensure reproducibility, I set up a pipeline that incorporated the necessary steps, such as data transformation, resampling, and model training.

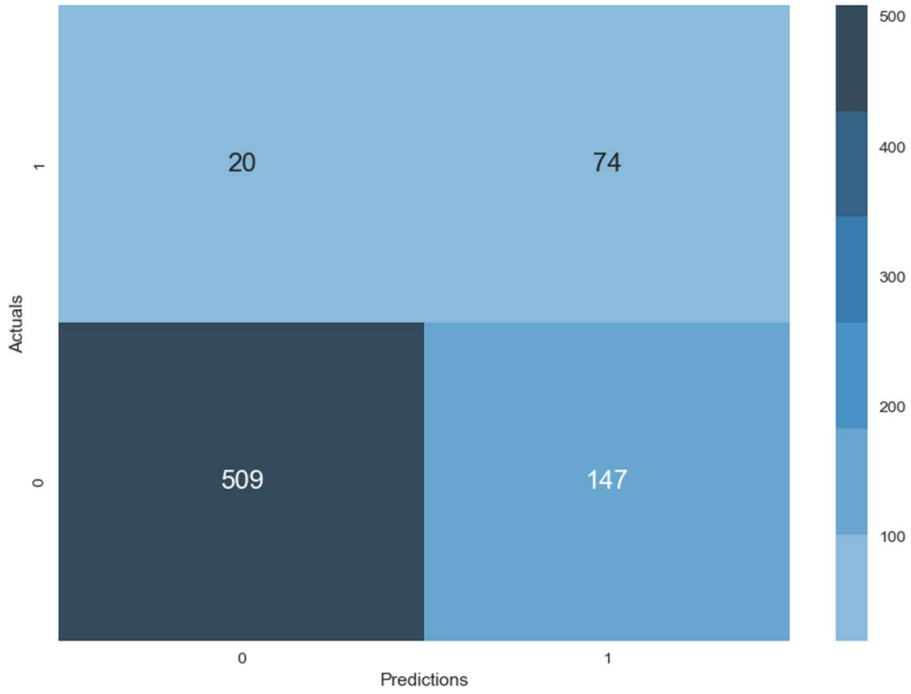
A key step involved tuning the models through a grid search, exploring various hyper parameter combinations to optimize their performance. Additionally, I assessed feature importance using the Gradient Boosting model and visualized the results through a bar chart. This analysis provided insights into which features played a crucial role in the predictive performance of the model.

Furthermore, I decided to create a new model using only the top 8 features based on their importance, aiming to evaluate whether this reduced set of features could maintain or enhance model performance while potentially saving computational resources. This comprehensive approach, encompassing data preprocessing, addressing class imbalance, model selection, hyper parameter tuning, and feature importance analysis, aimed to deliver a robust and effective solution for the classification task at hand.

```
Model: SVC()
Training Recall: 0.46680606377417666
Testing Recall: 0.5851063829787234
```

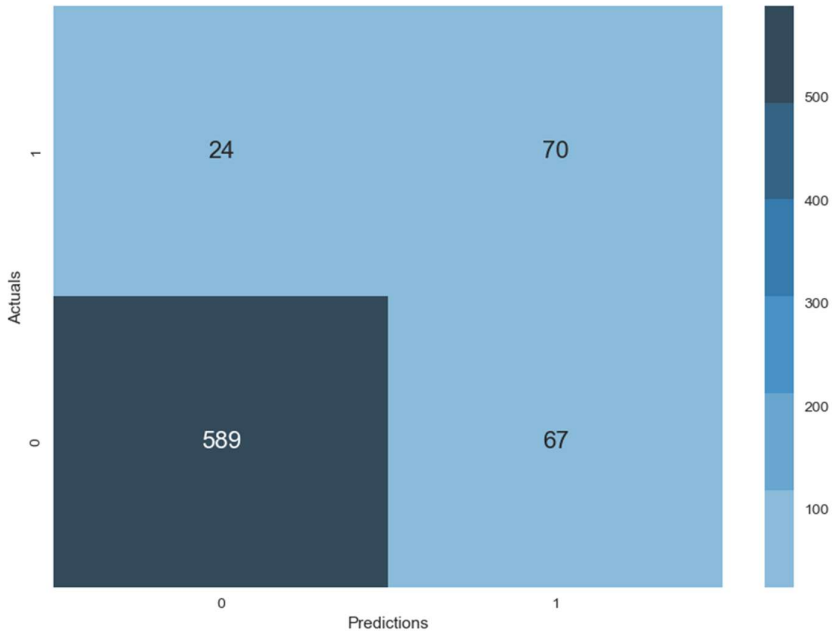


Model: GaussianNB()
Training Recall: 0.7522216414009409
Testing Recall: 0.7872340425531915

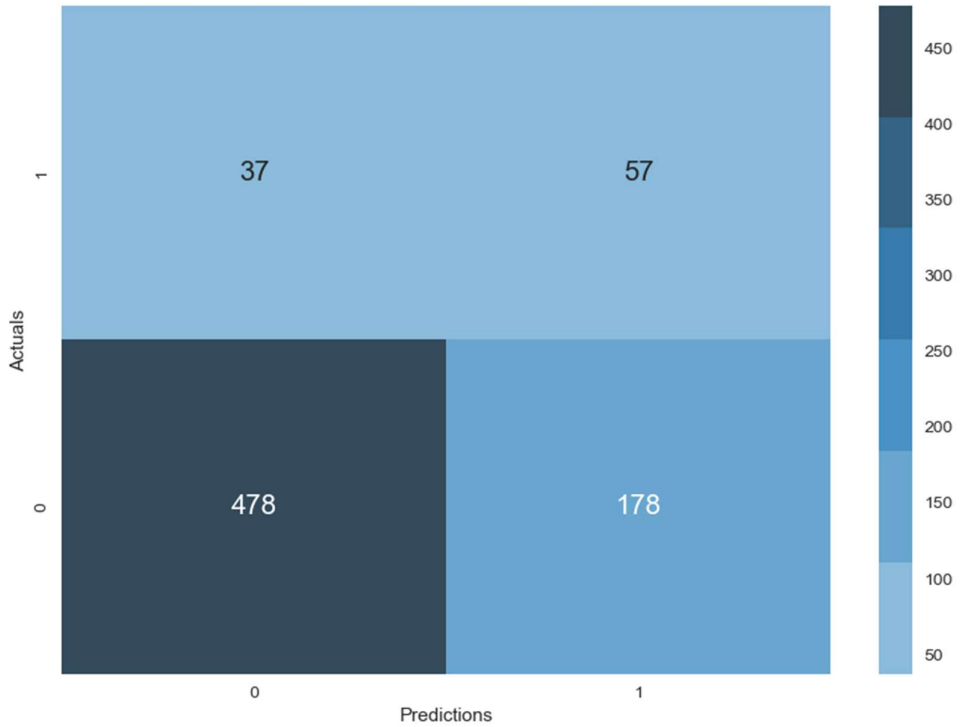


Model: GradientBoostingClassifier()

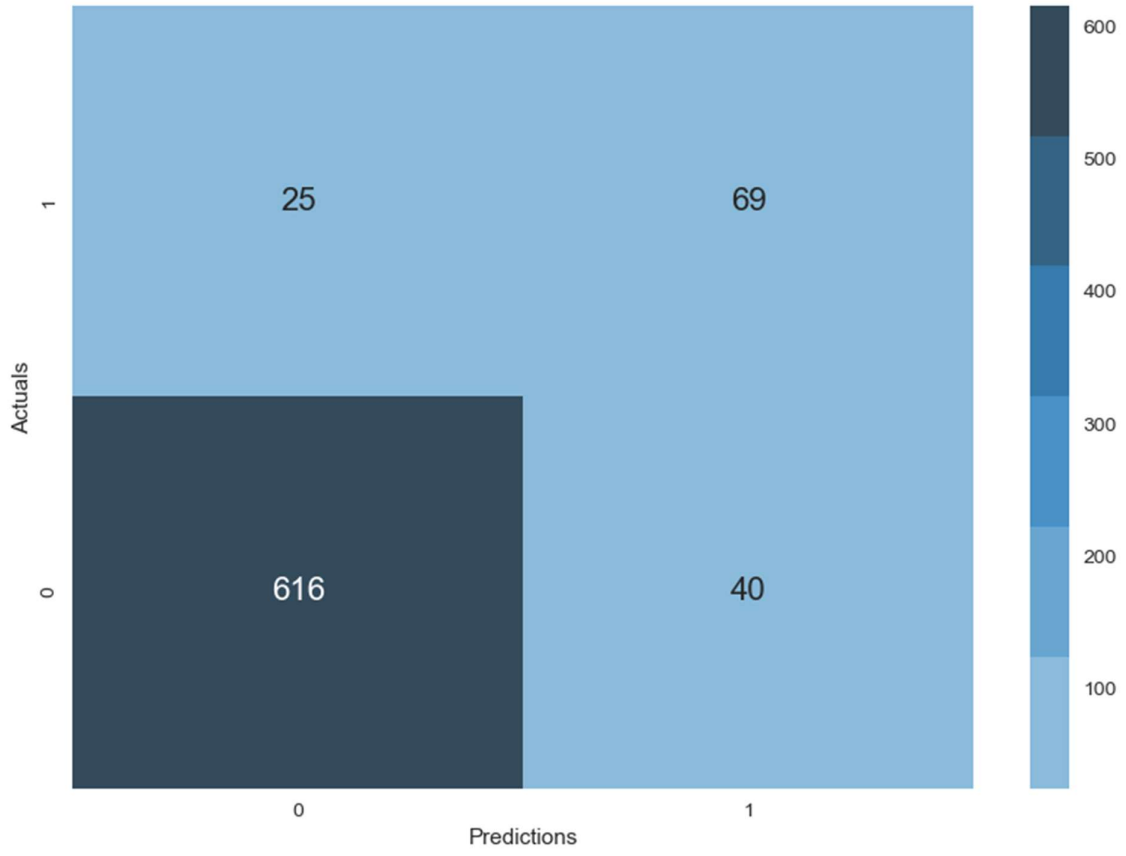
Training Recall: 0.7553580763199164
Testing Recall: 0.7446808510638298

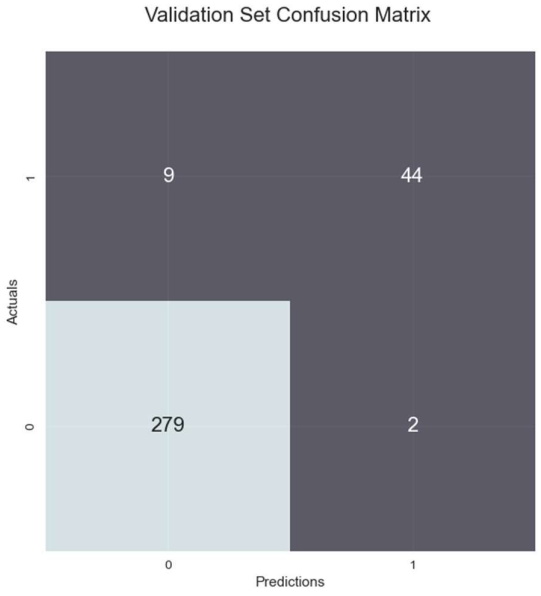


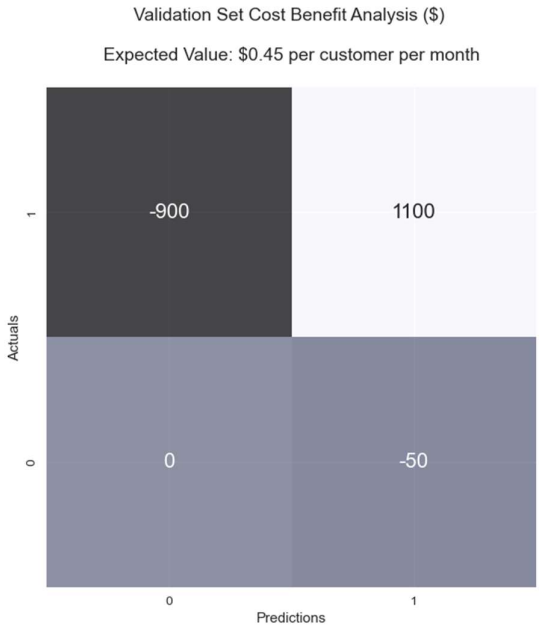
Model: KNeighborsClassifier()
Training Recall: 0.9790904338734971
Testing Recall: 0.6063829787234043



```
Model: RandomForestClassifier()
Training Recall: 1.0
Testing Recall: 0.7340425531914894
```







Based on this cost benefit analysis, our expected value from this strategy is 45 cents per customer per month. That may not seem like much, but for millions of customers it would add up. The good news here is that with this model predicting churn, we are not LOSING money! We can see the breakdown of each cost and benefit multiplied by the number of TP, TN, FP, FNs on the confusion matrix above.

Conclusion

To sum up, phone calls to customer service appear to be one of the most significant markers of client attrition. Additionally, we observe increased churn in some states, albeit it's unclear from the data why some states have a higher churn rate. Additionally, it is evident that a higher percentage of consumers with foreign plans churn than those without, possibly due to dissatisfaction with their plans.

Recommendation

1. We currently have a 14.5% churn rate for the training data set. Examining customer support calls, we may observe that the probability of churning rises in tandem with the quantity of calls. In particular, the chance of a customer churning rises from roughly 10% to 50% after at least 4 customer service calls.

2. Calls to customer care do not by themselves ensure that a consumer will leave. Actually, most of the clients who DID NOT churn only placed one or two customer support calls. It's crucial to remember that the majority of DID churners made one to four customer support calls. Consequently, a client's likelihood of churning should be shown if they have made more than three calls to customer support.

Recommendation: In light of these results, I would advise reviewing our customer service procedure. Offering a bigger incentive or discount to clients who contact customer support more than three times could be beneficial.

3. It is evident that the utilization of day, night, and international calls was nearly same for both the consumers who did not churn and the ones who did. Regardless of whether the consumer has an international plan or not, the charges for foreign minutes are the same (27 cents per minute). It's also noteworthy that a higher proportion of international plan subscribers than non-international plan subscribers experienced consumer attrition. It's probable that consumers with international plans who churned did not think paying for the international plan was worthwhile because of this comparable cost for international calls.

>>>>These results lead me to suggest adjusting the international minute rates. Customers with international plans should be able to make international calls at a lower cost than those without one.

4. It is evident that certain states have far greater churn than others. Texas has the most state-level turnover (27%), compared to all other states. California, Maryland, and New Jersey too have greater turnover rates (above 23%). Hawaii and Iowa have the lowest churn rates (less than.05%)

5. The variation in churn between states may be caused by several factors. One explanation might be the dearth of rivals in more remote locations like Hawaii and Iowa. There may be numerous other significant companies in the industry in states like Texas, California, or New Jersey, giving our clients more options if they decide to leave. Another factor can be the poor quality of service in some

>>>>>>>In light of these results, I advise investigating competitors in high-churn jurisdictions such as Texas, California, New Jersey, and others to see whether they are running any initial promotions that could force some of our clients to leave. In order to determine whether any dead zones are causing the higher rates, I also advise examining the cell signal in these states with higher churn.

Obtain additional information on rivals in states with greater churn.

Obtain additional cell signal data across the United States to identify trends in states with greater churn rates.

Examine voicemail logs to determine whether it could be a useful predictor.

Thank you

Deployment