SYRIATEL CUSTOMER CHURN PREDICTION

By: Jeniffer Njeri Gatharia

Business Objective

- As we all know customer retention is the heart of every thriving business. Dissatisfaction in the service quality of product can facilitate a customer to seize doing business with you and probably move to a competitor with better. This would lead to the business making losses.
- The goal of this project therefore is to develop a model that helps us predicts which customers are likely to churn so that we can take proactive steps to retain them and leverage the business.

Business Problem

- Currently we lack the ability to accurately identify customers at risk of churning which means we often miss opportunities to intervene before they leave. As a result, we face challenges in maintaining our customer base leading to potential revenue loss and increased costs associated with acquiring new customers. We are unable to allocate resources effectively and implement the required customer retention strategies as a result. This project aims at leveraging data and machine learning to help predict churn and solve our problem.

Questions our project aims to answer:

- What is the current churn % rate.
- What features/attributes do the customers who churn have.
- What strategies can we implement to increase customer retention.

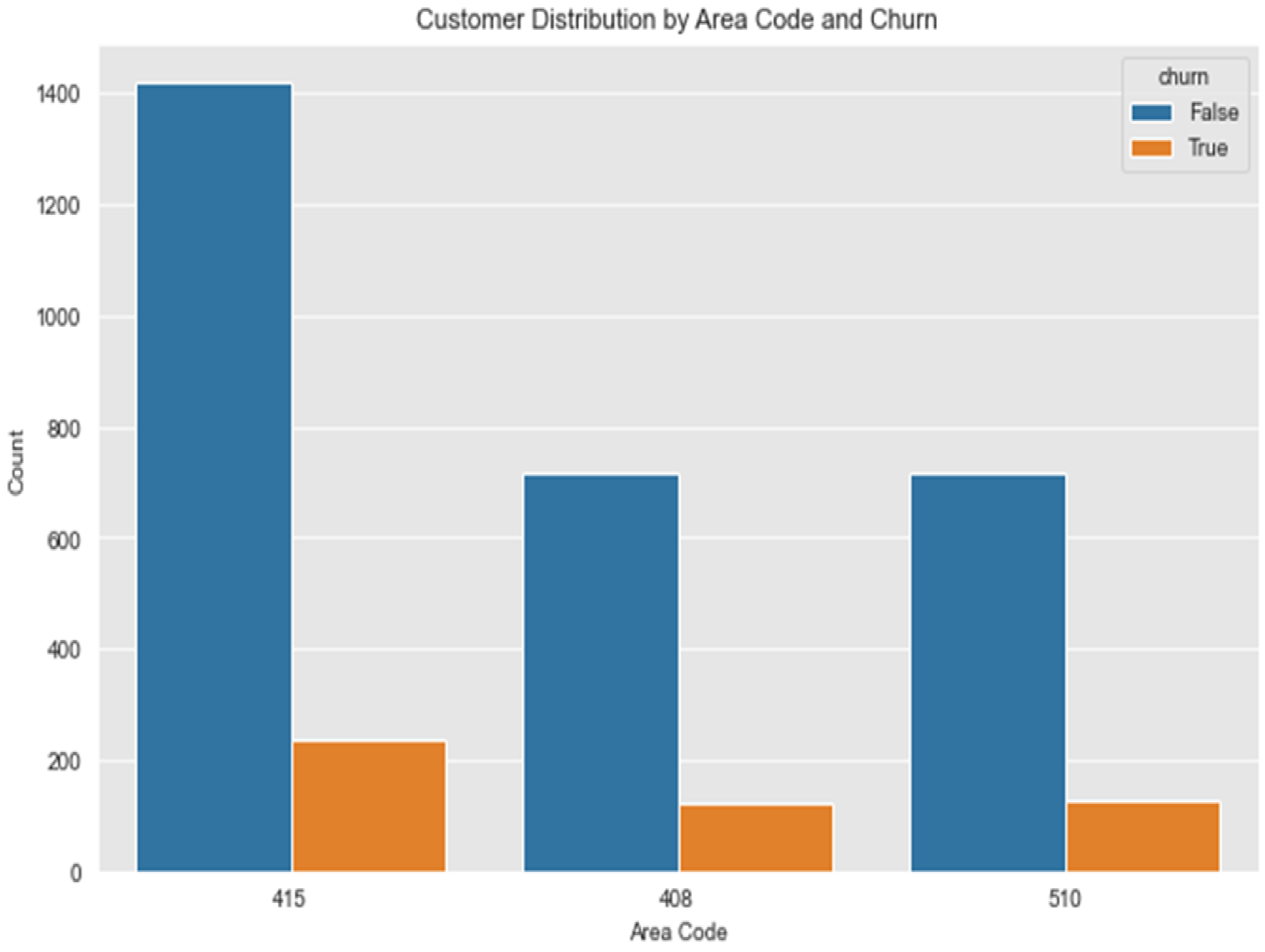
Data Understanding

- Our project utilizes the SyriaTel dataset, which was downloaded from Kaggle. The dataset contains 3,333 records and 21 features.

- We analyzed customer data, including demographics, service usage in terms of calls, charge and minutes all round the clock(day, evening and night) customer support interactions and the various packages that the service comes along with such as the international plans and voice main plans to identify patterns that may indicate a higher risk of churn.

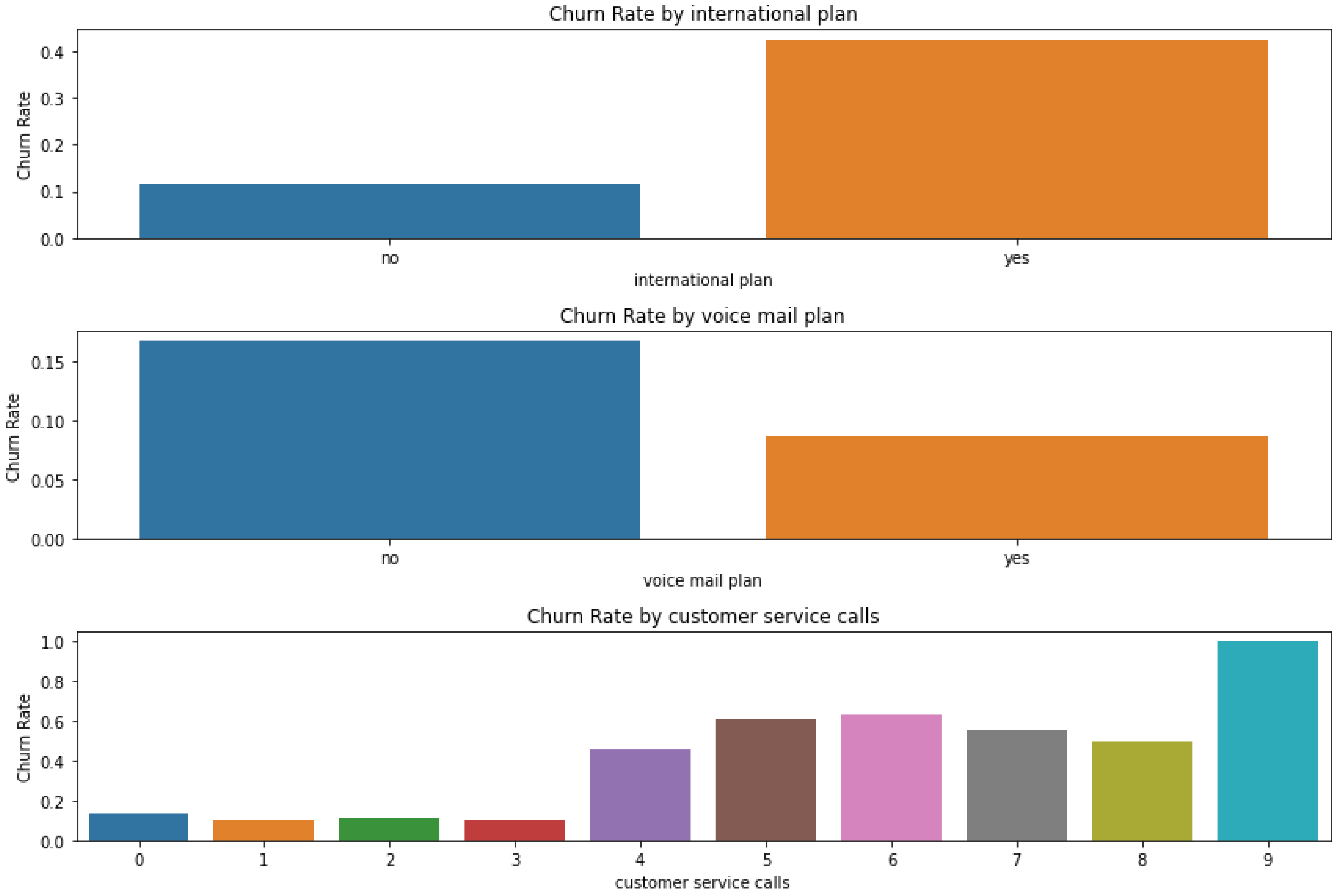
Data Finding 1:

Our analysis revealed that our customer data spans three geographical areas, identified by area codes 415, 408, and 510. Area code 415 not only has the highest number of customers but also the highest churn rate, followed by area code 510. Area code 408 has the lowest churn rate among the three.



Data Finding 2:

Our analysis shows that customers who are most likely to churn share some common characteristics: they tend to have an international plan, do not subscribe to a voicemail plan, and have a history of contacting customer service. These patterns suggest potential areas where we can focus our efforts to improve customer satisfaction and retention.


Churn Rate by international plan

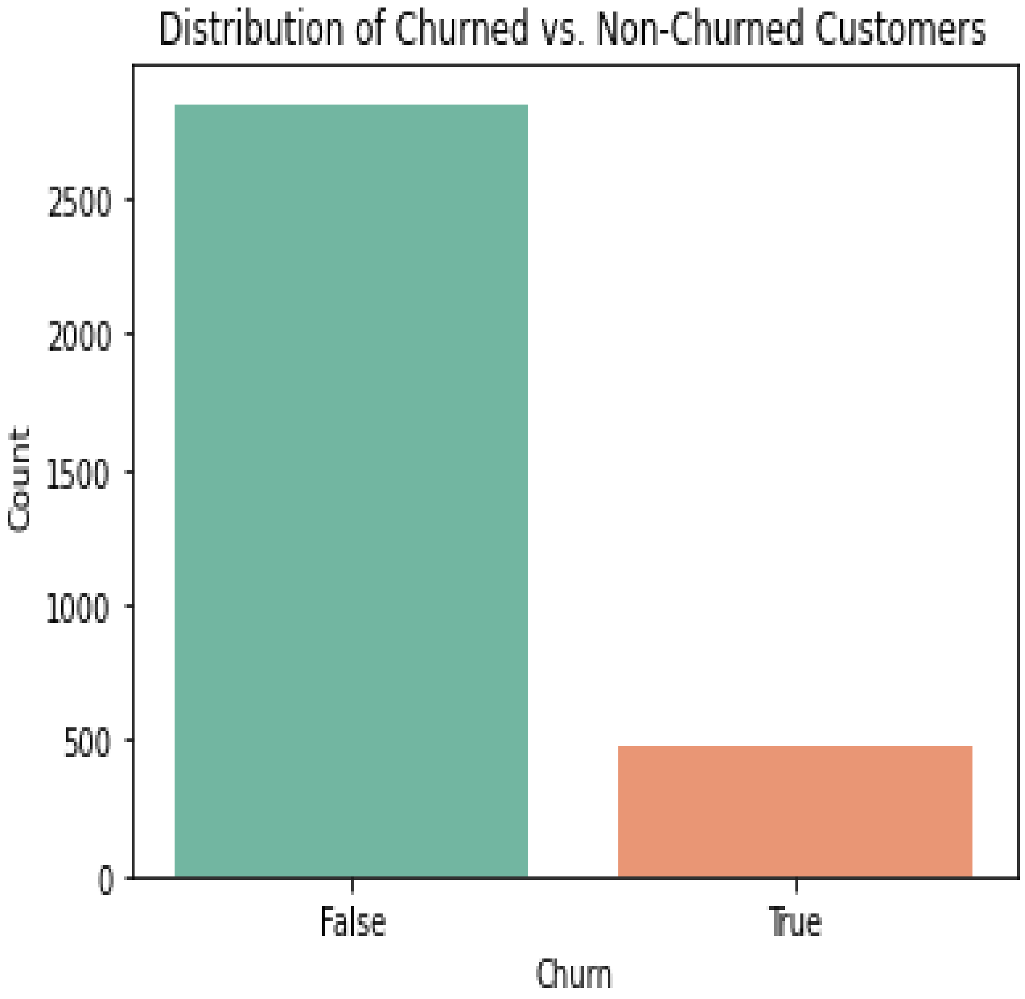
Churn Rate by voice mail plan

Churn Rate by customer service calls

Data Finding 3:

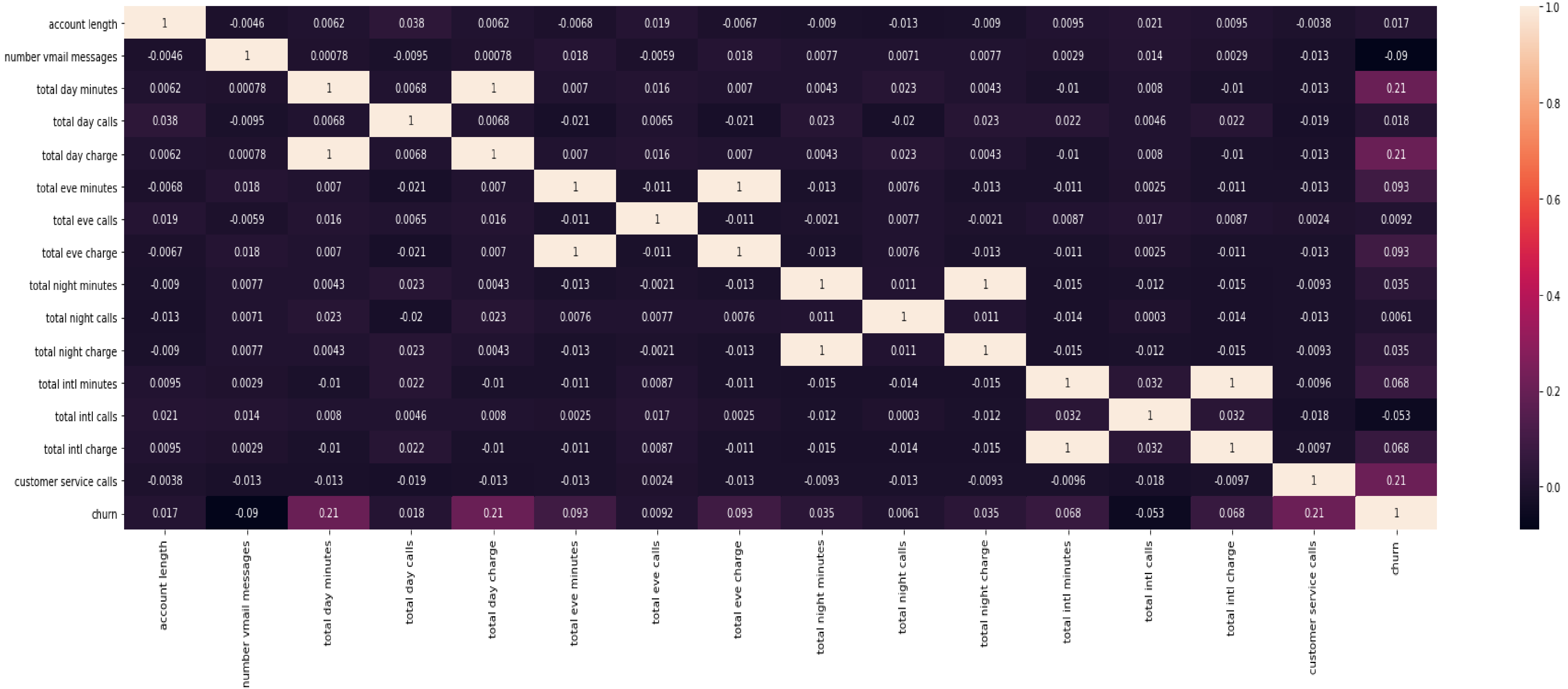
Our analysis of the target variable reveals a significant class imbalance in our dataset. Out of the total customers, 85.5% did not churn, while only 14.5% did, resulting in 483 customers who churned versus 2,850 who did not. This kind of imbalance is typical in churn datasets and highlights the challenge of accurately predicting customer churn.

- False is the no-churn while True represents churn.



Data Finding 4:

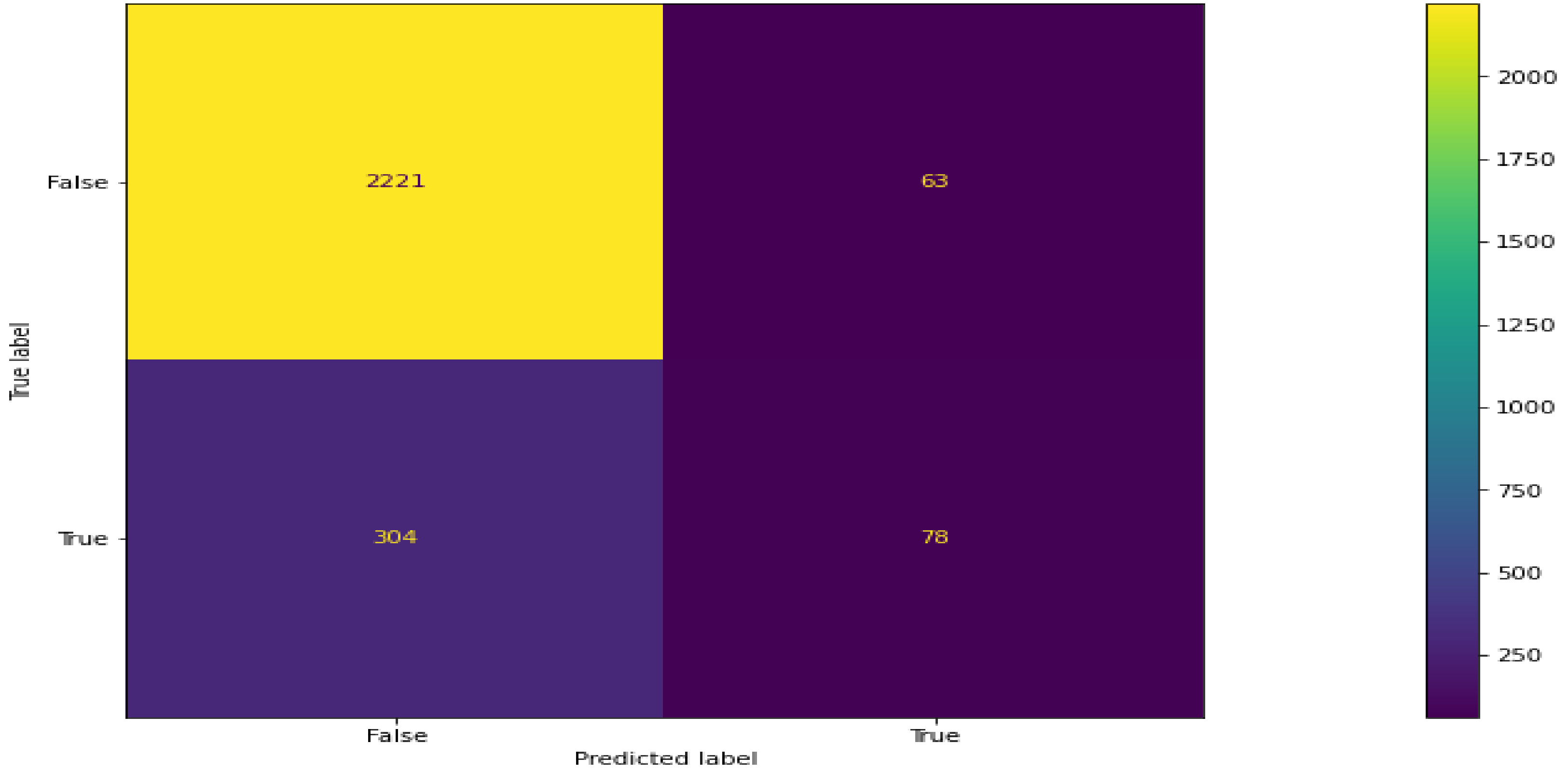
Our analysis identified high correlations between several variables, such as 'total day charge' and 'total day minutes', among others. This issue, known as multicollinearity, can make it difficult to determine the unique impact of each variable and may lead to overfitting in models like Logistic Regression. To address this, we plan to use techniques like regularization, ensuring that our models are both reliable and easy to interpret.



	account length	number vmail messages	total day minutes	total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls	churn
account length	1	-0.0046	0.0062	0.038	0.0062	-0.0068	0.019	-0.0067	-0.009	-0.013	-0.009	0.0095	0.021	0.0095	-0.0038	0.017
number vmail messages	-0.0046	1	0.00078	-0.0095	0.00078	0.018	-0.0059	0.018	0.0077	0.0071	0.0077	0.0029	0.014	0.0029	-0.013	-0.09
total day minutes	0.0062	0.00078	1	0.0068	1	0.007	0.016	0.007	0.0043	0.023	0.0043	-0.01	0.008	-0.01	-0.013	0.21
total day calls	0.038	-0.0095	0.0068	1	0.0068	-0.021	0.0065	-0.021	0.023	-0.02	0.023	0.022	0.0046	0.022	-0.019	0.018
total day charge	0.0062	0.00078	1	0.0068	1	0.007	0.016	0.007	0.0043	0.023	0.0043	-0.01	0.008	-0.01	-0.013	0.21
total eve minutes	-0.0068	0.018	0.007	-0.021	0.007	1	-0.011	1	-0.013	0.0076	-0.013	-0.011	0.0025	-0.011	-0.013	0.093
total eve calls	0.019	-0.0059	0.016	0.0065	0.016	-0.011	1	-0.011	-0.0021	0.0077	-0.0021	0.0087	0.017	0.0087	0.0024	0.0092
total eve charge	-0.0067	0.018	0.007	-0.021	0.007	1	-0.011	1	-0.013	0.0076	-0.013	-0.011	0.0025	-0.011	-0.013	0.093
total night minutes	-0.009	0.0077	0.0043	0.023	0.0043	-0.013	-0.0021	-0.013	1	0.011	1	-0.015	-0.012	-0.015	-0.0093	0.035
total night calls	-0.013	0.0071	0.023	-0.02	0.023	0.0076	0.0077	0.0076	0.011	1	0.011	-0.014	0.0003	-0.014	-0.013	0.0061
total night charge	-0.009	0.0077	0.0043	0.023	0.0043	-0.013	-0.0021	-0.013	1	0.011	1	-0.015	-0.012	-0.015	-0.0093	0.035
total intl minutes	0.0095	0.0029	-0.01	0.022	-0.01	-0.011	0.0087	-0.011	-0.015	-0.014	-0.015	1	0.032	1	-0.0096	0.068
total intl calls	0.021	0.014	0.008	0.0046	0.008	0.0025	0.017	0.0025	-0.012	0.0003	-0.012	0.032	1	0.032	-0.018	-0.053
total intl charge	0.0095	0.0029	-0.01	0.022	-0.01	-0.011	0.0087	-0.011	-0.015	-0.014	-0.015	1	0.032	1	-0.0097	0.068
customer service calls	-0.0038	-0.013	-0.013	-0.019	-0.013	-0.013	0.0024	-0.013	-0.0093	-0.013	-0.0093	-0.0096	-0.018	-0.0097	1	0.21
churn	0.017	-0.09	0.21	0.018	0.21	0.093	0.0092	0.093	0.035	0.0061	0.035	0.068	-0.053	0.068	0.21	1

Modeling

- We tested several models including Logistic Regression and Decision Trees since this is classification problem to determine which provided the best balance of accuracy and interpretability.
- Our logistic regression baseline model had this confusion matrix.



Modeling Approach

Logistic Regression

This model was used due to churn and no churn binary nature.

The Baseline model was trained with imbalanced data.

Confusion Matrix interpretation:

- **True Positive (TP):** 78 churned customers correctly predicted.
- **False Negative (FN):** 304 customers who churned but weren't predicted to.
- **False Positive (FP):** 63 non-churned customers incorrectly predicted to churn.
- **True Negative (TN):** 2221 non-churned customers correctly predicted.

On the Iterative model We used a technique called SMOTE to adjust the data so that it equally represents customers who are likely to leave and those who are likely to stay. This helps the model make fairer predictions. We applied regularization to prevent overfitting on the training data.

Decision Trees

Decision trees was chosen because they are flexible and easy to understand when making decisions.

We started with a basic model using consistent settings to establish a reference point.

The first iterative model had adjusted settings by hand to improve the models performance.

The second iterative model used a systematic approach to find the best settings which led to better accuracy and more predictions.

Evaluation

Logistic Regression

Baseline Model:

- This model mostly predicts for customers who won’t churn but struggles to identify those who will.
- This model misses many potential churners making it less useful for preventing customer loss.

Iterative Model:

- The model is better at catching potential churners which is critical for effective retention strategies.
- Even though the precision is slightly lower, the model’s improved ability to catch potential churners makes it more valuable for identifying and addressing at-risk customers.

Decision Trees

Baseline Model:

- This model has a strong performance in identifying customers who are likely to stay.
- This model struggles with identifying customers who might leave, though it still catches some.

Iterative Model With Grid Search:

- The model now does a much better job overall. It’s more accurate and reliable in predicting both customers who will stay and those who might leave.
- The model is better at telling the difference between customers who are likely to stay and those who might leave, which helps us target our retention efforts more effectively

Model Of Choice

- The iterative model with the grid search turned out to be the best option. It performs well in accurately predicting both customers who will stay and those who might leave.
- Given that it’s crucial for us to correctly identify both loyal and at-risk customers while managing resources effectively the this model is a better choice for making smart business decisions about customer retention and marketing strategies.

Possible Limitations in Real Life

The model may not identify all potential churners effectively, which means we might miss some opportunities to take action and retain those customers.

Mitigation Strategies

Modify the decision thresholds used by the model to classify churners. Lowering the threshold may help identify more potential churners but could increase false positives.

Recommendations and Future Investigations

- **Customer Service Calls Investigation**: Dig deeper to understand why some customers need to contact customer service frequently. Increase the number of positive customer interactions especially for those with frequent service calls. Consider offering more personalized support, quicker resolution times and proactive check-ins for customers showing signs of dissatisfaction. Improved support can reduce churn likelihood.
- **International Plan Churn Investigation**: Since some of the customers with international plans are leaving it is imperative to dig deep to understand whether the customer dissatisfaction is as a result of the service or the charges. If charges, consider adjusting the pricing plans or offering more flexible payment options. Transparent communication about pricing and the value provided will also help mitigate churn.
- **High Churn States Analysis**: Look into the states where many customers are leaving to identify any patterns or reasons for the high churn rates. Could be a competitor or even the level of engagement from a particular region. This will enable the business tailor fit the best retention strategies.
- The model specifically predicts churn, not other metrics like overall customer satisfaction or product engagement. While reducing churn can indirectly affect these metrics the model itself doesn't provide direct insights into them. It would therefore be limiting to use this model to rate the level of customer satisfaction or product engagement.

Thank you!!!