Customer Churn Analysis

Alex Miningwa

TABLE OF CONTENTS

01

OVERVIEW

INTRODUCTION

- The primary objective of this project is to help the business reduce customer churn by identifying key factors that contribute to customers leaving the service. By analyzing customer data, we aim to provide actionable insights and recommendations that will enhance customer retention strategies, thereby fostering a more stable and loyal customer base.
- By employing a Random Forest classifier, we aim to uncover key factors influencing churn and generate actionable insights that can help the business implement targeted retention strategies.

- The process involves rigorous model evaluation to ensure robustness and reliability, ultimately supporting data-driven decision-making to reduce customer attrition and enhance long-term customer loyalty.

- Methodology: Data preprocessing, feature engineering, oversampling with SMOTE, and model evaluation.

02

BUSINESS UNDERSTANDING

PROBLEM STATEMENT

- Customer churn occurs when a customer stops using the company's products or services.
- The business is facing challenges with customer retention, leading to increased costs and lost revenue. Understanding the reasons behind customer churn is crucial for developing strategies that improve customer satisfaction and loyalty.
- By identifying at-risk customers early, SyriaTel can implement targeted retention strategies to reduce churn and improve customer satisfaction and loyalty.

Objective

Our goal is to develop a robust predictive model that accurately identifies customers at risk of churning, utilizing data preprocessing, feature engineering, and advanced machine learning techniques performance.

Stakeholders

- SyriaTel
- Customer Retention team
- Marketing team
- Customers
- Marketing team

Business Impact

- Revenue loss
- Increased cost
- Market share decline
- Reputation damage

03

DATA UNDERSTANDING

Dataset

The dataset used in this project is sourced from Kaggle and provides comprehensive customer information. This dataset includes a variety of attributes related to customer demographics, service usage patterns, and customer support interactions, alongside indicators of customer churn. The customer churn dataset used in this project contains 3,333 records of customer information, focusing on attributes that may influence their likelihood to churn (leave the service). The dataset includes various features such as demographic details, service usage patterns, and customer service interactions. It is split into features like State, Account Length, International Plan, and Churn, with the target variable being whether the customer has churned. This dataset serves as the foundation for training and evaluating machine learning models aimed at predicting customer churn, enabling businesses to proactively address potential customer losses. We organize and process this dataset using dataframes to ensure efficient handling and analysi

Data Analysis

- We first explored the dataset to understand its structure and contents. This involved examining descriptive statistics, checking for missing values, and identifying any outliers.
- Next, we conducted visualizations to gain insights into the distribution of key features, correlations between variables, and the proportion of churned customers.
- We also performed feature engineering to create new variables or transform existing ones to enhance the predictive power of our models. Additionally, we conducted statistical tests or exploratory data analysis to identify significant factors associated with churn.
- The data analysis phase aimed to uncover patterns and relationships within the data that could help us better understand customer churn behavior.

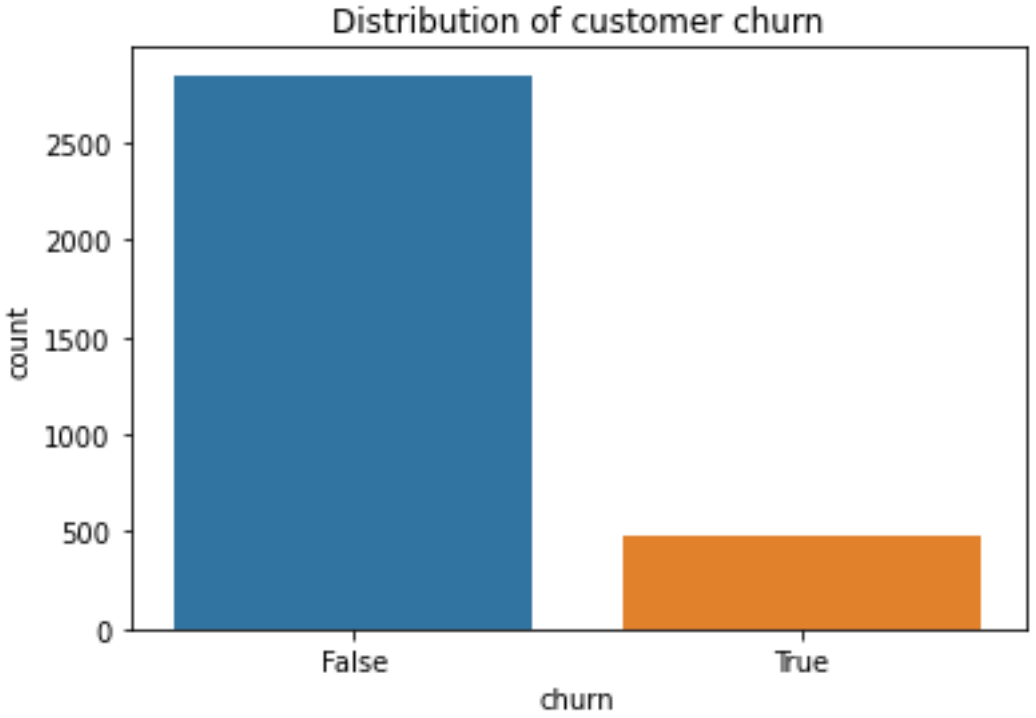
Data Cleaning

The data preparation process involved preliminary data cleaning to ensure the dataset's quality. The remaining tasks include encoding categorical variables, such as the presence of an international plan, into a numerical format suitable for machine learning models, and addressing class imbalance by applying techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the target variable's classes, ensuring the model performs well on both majority and minority classes.

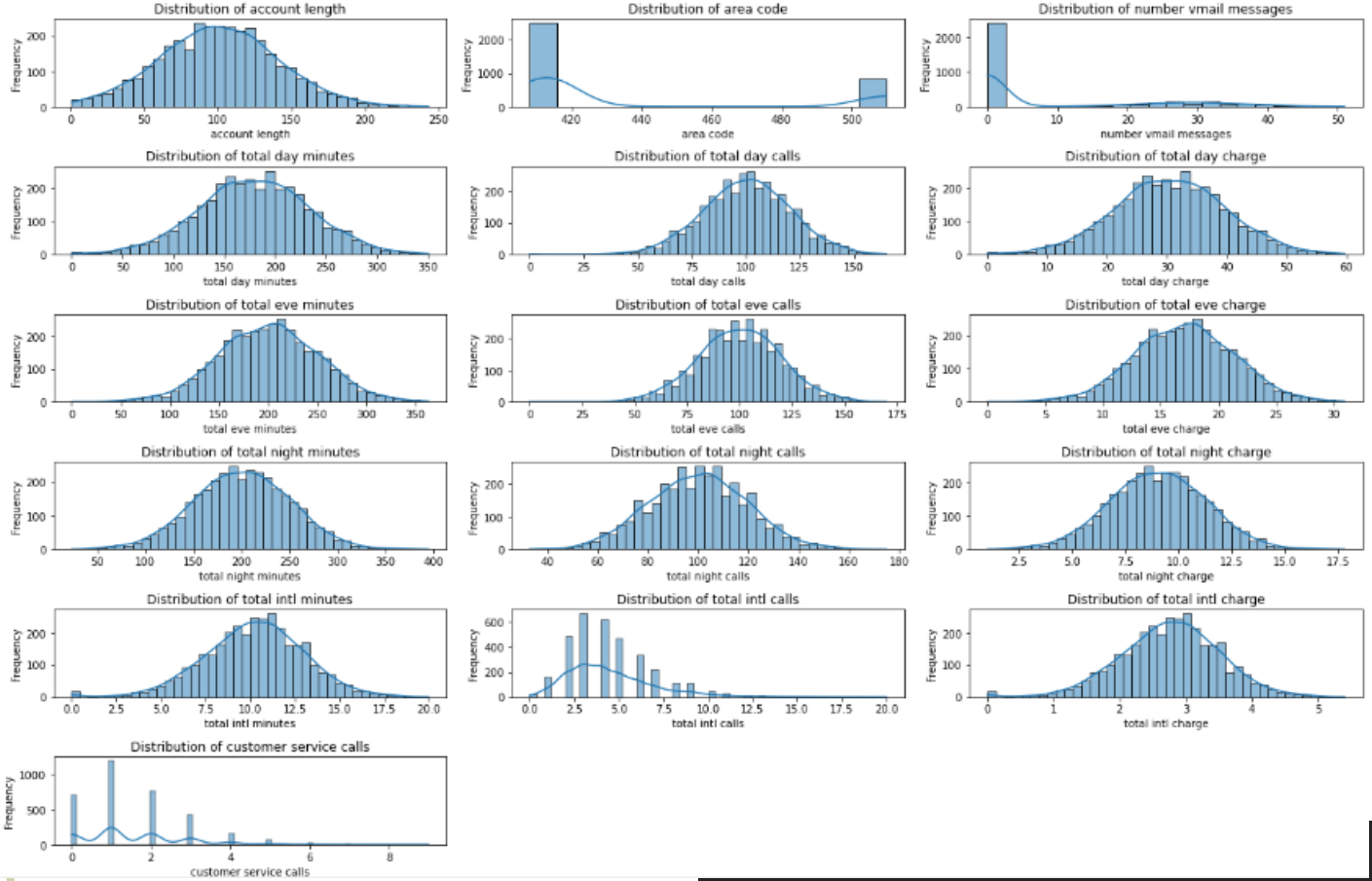
04

Visualizations

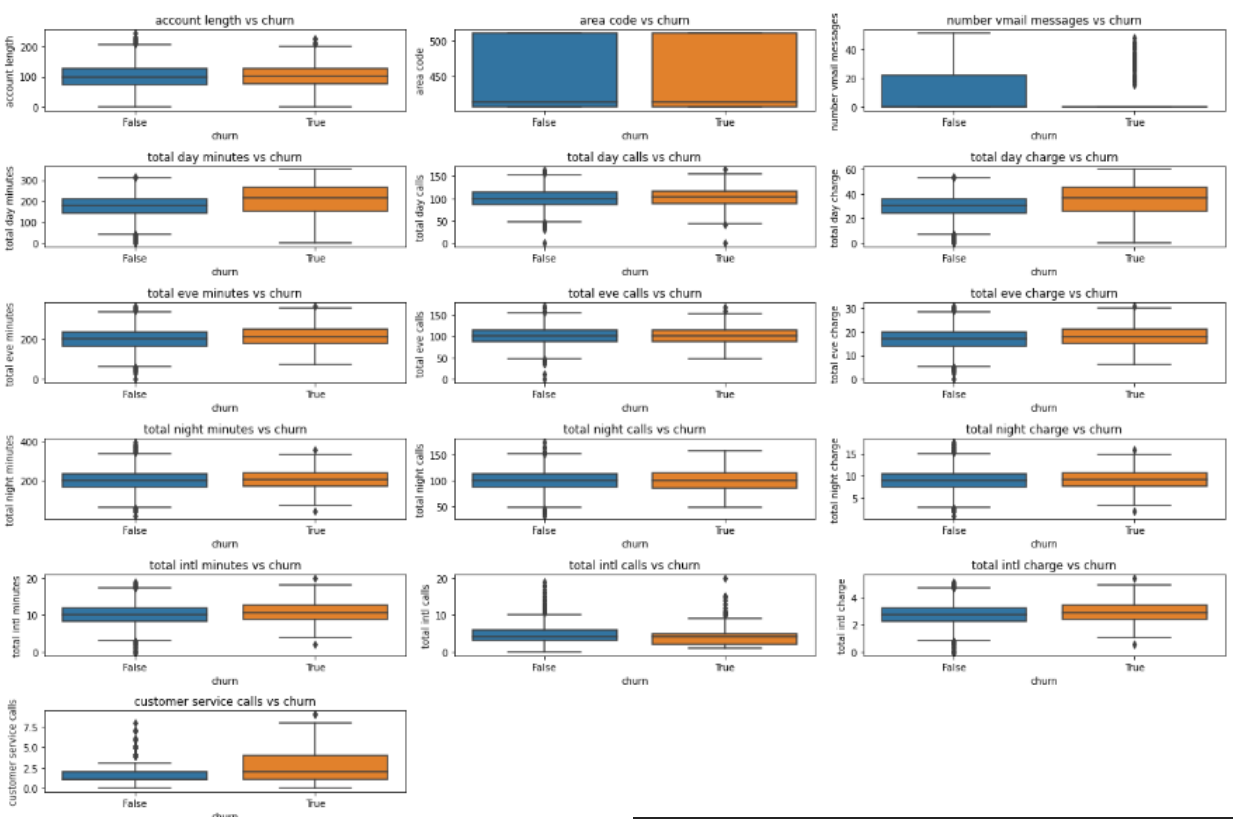
Class Imbalance



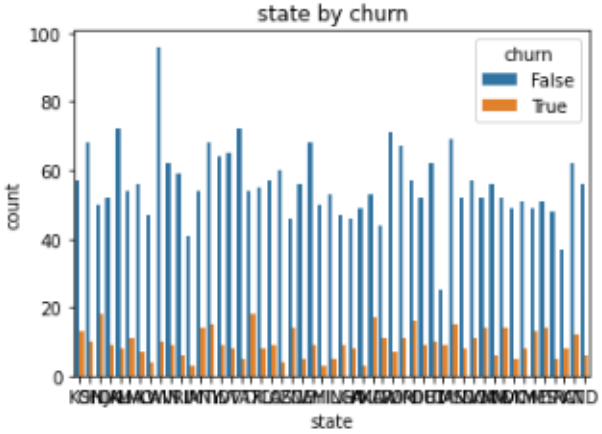
Distribution of numerical features

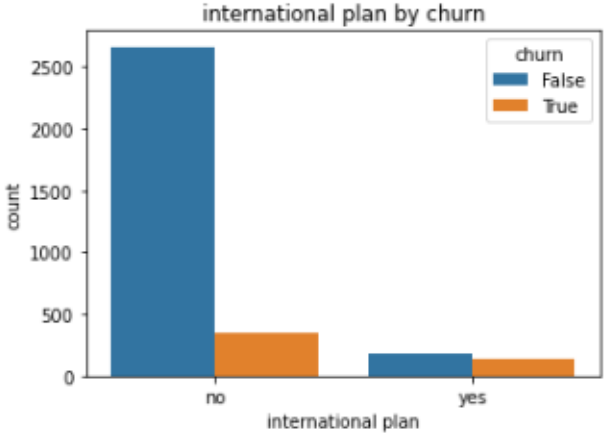


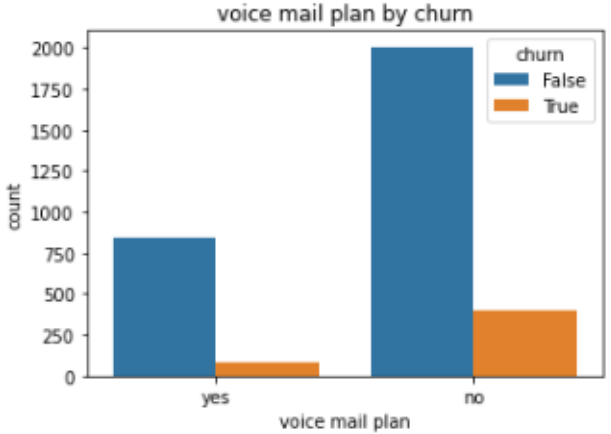
Analyzing the relationship between numerical features and the target variable



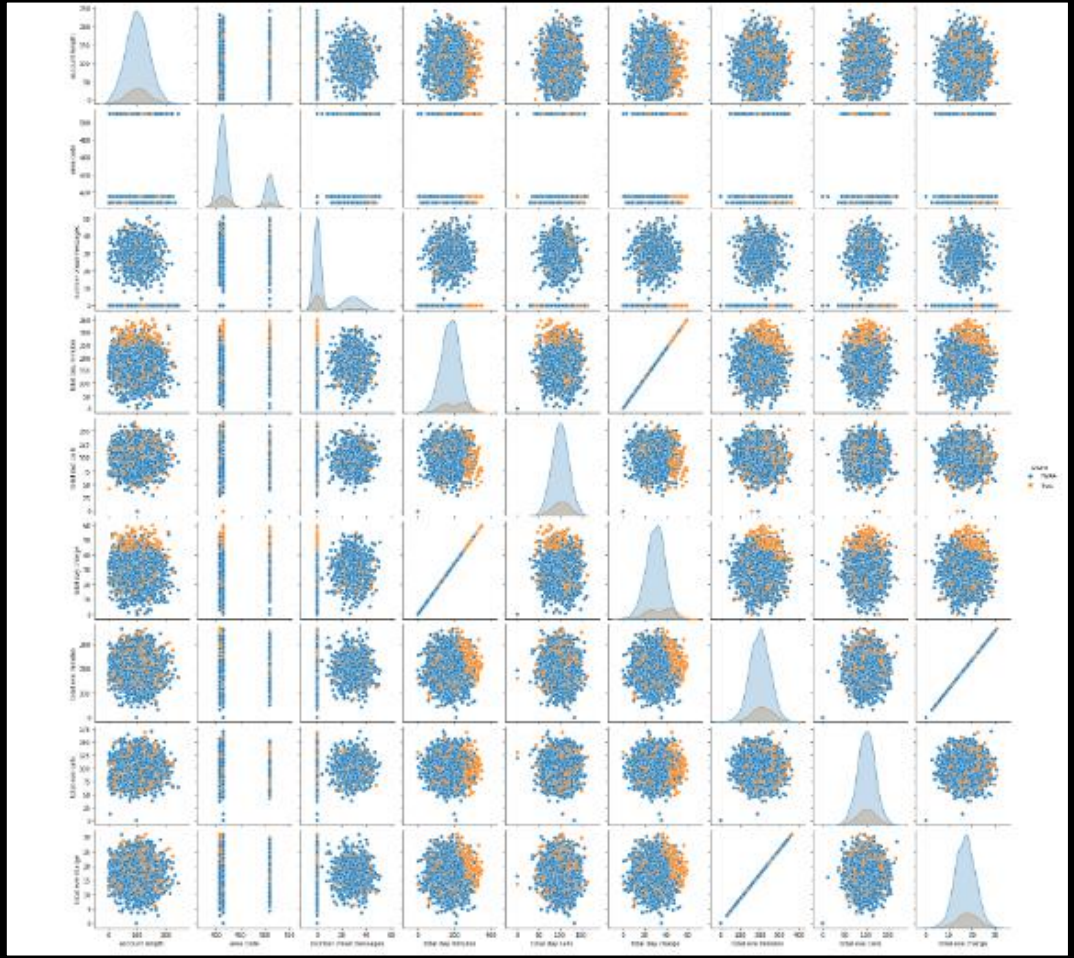
Distribution of Categorical features







Pairplot



05

MODELLING

Models Used

- Logistic Regression was the first model that was used. We trained the Logistic Regression model with a random state of 42, made predictions, and printed out the results. The accuracy of the Logistic Regression model on the test set was printed, followed by the classification report and confusion matrix.
- Next we trained a Decision Tree Classifier with a random state of 42, made predictions, and printed out the evaluation results. The accuracy of the Decision Tree model on the test set was printed, followed by the classification report and confusion matrix.
- Finally, we trained a Random Forest Classifier and printed out the outputs.

Applying SMOTE

- To address the class imbalance issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied.
- SMOTE generates synthetic samples for the minority class, balancing the distribution of classes in the dataset.

Hyperparameter Tuning

GridSearchCV was employed to find the optimal hyperparameters for the Random Forest Classifier. This process systematically searches through a grid of hyperparameters, using cross-validation to determine the best combination that maximizes model performance.

06

EVALUATION

Metrics

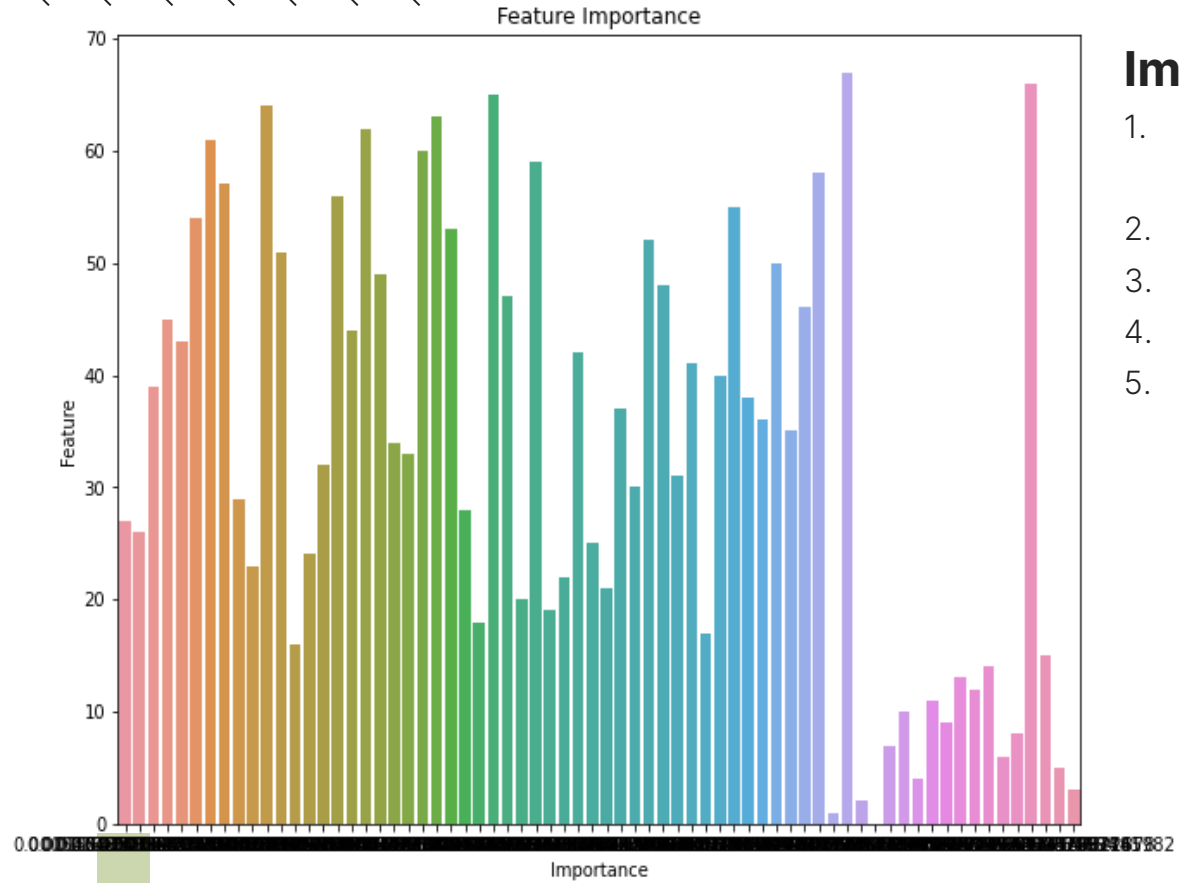
- The evaluation of the model's performance included various metrics such as Accuracy, Precision, Recall, and F1-Score. These metrics provide insights into different aspects of the model's predictive capabilities.
- To compare the three models, Logistic Regression, Decision Tree, and Random Forest, we first evaluated their performance metrics on the test set. Logistic Regression achieved an accuracy of 85.5%, demonstrating good precision (87%) for the "False" class but lower recall (19%) and F1-score (28%) for the "True" class.
- Decision Tree outperformed with an accuracy of 94.3%, showing balanced precision (95%) and recall (98%) for the "False" class and acceptable precision (88%) but lower recall (72%) for the "True" class.
- Random Forest attained an accuracy of 89.8%, with a strong precision (89%) and perfect recall (100%) for the "False" class, but lower recall (34%) and F1-score (50%) for the "True" class.
- The confusion matrices also highlight the model performances

Cont:

- We improved the random forest model using hyper parameter tuning where it showed a significant increase in its accuracy to predict customer churn.
- We got an accuracy of 94.3% and a best score of 95.4% which improved our model's ability to predict the probability of a customer to stop using a company's product due to various reasons.

07

RECOMMENDATION



Important Features

1. Total Day Minutes and Total Day Charge
2. Customer Service Calls
3. International plan
4. Total Evening Charge
5. State-Specific Insights

Recomendations

Based on the important features, these are the recommendations we gave Syriatel:

1. Syriatel should monitor high-usage customers for potential churn triggers and offer special plans or discounts to heavy users to encourage retention.
2. Improve customer service quality and responsiveness. Track frequent callers and ensure their issues are resolved promptly and satisfactorily.
3. Introduce attractive evening and off-peak hour plans or discounts to cater to users who predominantly use the service during these times.
4. Analyze and address state-specific customer behaviors and issues. Customize marketing and retention strategies to cater to regional preferences and needs.
5. Promote competitive international plans and ensure customers are aware of these options. Offer personalized discounts for frequent international callers.

Thank YOU