Overview & Data Understanding

Background Information

Customer churn, the phenomenon of customers discontinuing their relationship with a business, is a significant challenge faced by many industries. Understanding the factors that influence customer churn is crucial for businesses to implement strategies to retain their valuable customers

Problem Statement

To predict which customers are likely to churn and identify the key factors contributing to customer churn.

Objectives

1. Data Exploration: Understand the dataset, identify relevant features, and handle missing values and outliers.
2. Model Development: Build and train machine learning models to predict customer churn.
3. Model Evaluation: Evaluate the performance of the models using appropriate metrics.
4. Feature Importance Analysis: Identify the most important factors influencing customer churn.

Metrics of Success:

- Accuracy: Proportion of correct predictions.
- Precision: Proportion of true positive predictions among all positive predictions.
- Recall: Proportion of true positive predictions among all actual positive cases.
- F1-Score: Harmonic mean of precision and recall.
- ROC-AUC Score: Measures the model's ability to distinguish between positive and negative classes.

Data Understanding

Data Source: Kaggle

Data Description:

The dataset contains 10 rows and 21 columns

Here is the list of the 21 columns and what they mean

```
state**: the state the user lives in
**account length**: the number of days the user has this account
**area code**: the code of the area the user lives in
**phone number**: the phone number of the user
**international plan**: true if the user has the international plan, otherwise
false
**voice mail plan**: true if the user has the voice mail plan, otherwise false
**number vmail messages**: the number of voice mail messages the user has sent
**total day minutes**: total number of minutes the user has been in calls during
the day
**total day calls**: total number of calls the user has done during the day
**total day charge**: total amount of money the user was charged by the Telecom
company for calls during the day
**total eve minutes**: total number of minutes the user has been in calls during
the evening
**total eve calls**: total number of calls the user has done during the evening
**total eve charge**: total amount of money the user was charged by the Telecom
company for calls during the evening
**total night minutes**: total number of minutes the user has been in calls
during the night
```

```
**total night calls**: total number of calls the user has done during the night
**total night charge**: total amount of money the user was charged by the Telecom company for calls during the night
**total intl minutes**: total number of minutes the user has been in international calls
**total intl calls**: total number of international calls the user has done
**total intl charge**: total amount of money the user was charged by the Telecom company for international calls
**customer service calls**: number of customer service calls the user has done
**churn**: true if the user terminated the contract, otherwise false
```

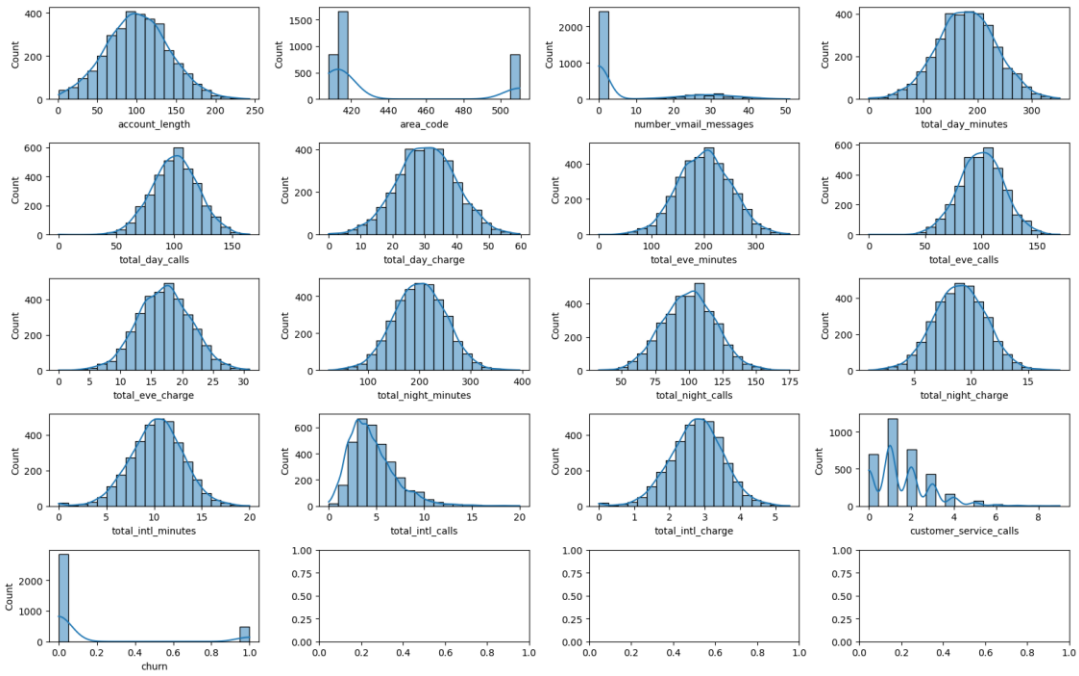
DATA PREPARATION AND ANALYSIS

Perform data cleaning by checking for missing values, outliers, null values and removing them so that they cannot affect our models

Data Analysis

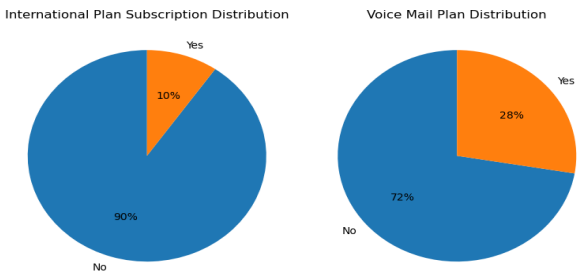
We perform a few visualization on the numerical and categorical data

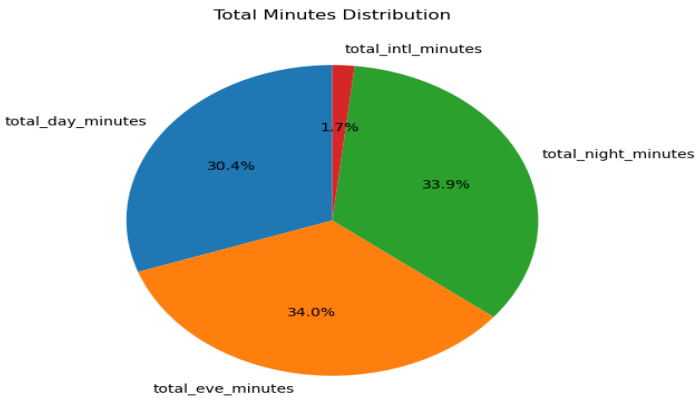
Here are the examples



We also visualized the categorical data

Here are a few examples





MODELING

Models Used:

Logistic Regression (Baseline Model):

Logistic Regression was selected as the baseline model because of its simplicity and effectiveness for binary classification tasks. It assumes a linear relationship between the features and the target variable.

Decision Tree:

This model was chosen for its interpretability and ability to handle non-linear relationships between features and the target variable.

Random Forest:

Random Forest was included as an ensemble method to improve prediction accuracy and reduce overfitting by combining multiple decision trees.

Justification for Models:

The models were selected to ensure a balanced approach, starting from a simple baseline to more advanced techniques. Logistic Regression provides a benchmark for comparison. Decision Trees allow for better feature interaction modeling, while Random Forest introduces ensemble learning to boost performance.

Metric of Success:

The models were evaluated based on accuracy, precision, recall, and F1 score to ensure balanced performance for the problem at hand.

EVALUATION

1.Logistic regression

```
Accuracy: 0.8515742128935532
Precision: 0.5294117647058824
Recall: 0.1782178217821782
F1-score: 0.26666666666666666
```

```
AUC: 0.574974635272714
```

While the Logistic Regression model has high accuracy, its low recall and F1-score suggest poor performance in identifying positive cases. This may be due to class imbalance or insufficient feature representation in the data

2.random forest

```
Accuracy: 0.9280359820089955
F1-score: 0.6962025316455697
ROC AUC: 0.7705104432704755
```

```
F1-score being 69.6% is moderately good performance. AUC of 77% is okay for the imbalanced data. This gives a generally good performance of the model.
- Having an accuracy of 92.9%, performs better than that of Logistic Regression Model, 85.3%
```

3.Decision trees

```
      Accuracy: 0.9220389805097451
F1-score: 0.74
ROC AUC: 0.844251828009656
```

Random Forest could be a more reliable choice for deployment due to its better generalization ability. The final decision should depend on additional metrics (e.g., precision, recall, and F1-score) and validation on unseen data to confirm the robustness of the models and has the highest accura

RANDOM FOREST IS A MORE RELIABLE CHOICE

RECOMMENDATIONS

```
      Enhance Call Quality: Invest in infrastructure and technology to improve call quality, ensuring a better customer experience.

- Customer Service Improvement: Focus on enhancing customer service by reducing response times, increasing efficiency in issue resolution, and offering personalized support.
```

```
- Tailored Plans for International Subscribers: Design attractive plans and
offers specifically targeted at international subscribers to increase
satisfaction and reduce churn.
- Proactive Retention Strategies: Implement proactive measures such as targeted
promotions, loyalty rewards, and personalized communication to retain at-risk
customers.
- Regular Analysis: Continuously monitor customer behavior and churn patterns,
regularly updating models and strategies to adapt to changing market dynamics.
```

NEXT STEPS

Deployment: Deploy the selected model into a production environment to make real-time predictions.

Data Collection: Continuously collect more data to improve the model's accuracy and reliability.

Model Refinement: Regularly retrain and fine-tune the model to adapt to changing customer behavior and market trends.