Telecom Customer Churn Data

Milestone: Project Report

Group 21

Parth Deshmukh
Malav Makadia
Gaurav Ramoliya

deshmukh.par@northeastern.edu
makadia.m@northeastern.edu
ramoliya.g@northeastern.edu

Percentage of Effort Contributed by Parth Deshmukh: 34%

Percentage of Effort Contributed by Malav Makadia: 33%

Percentage of Effort Contributed by Gaurav Ramoliya: 33%

Signature of Parth Deshmukh: *Parth Deshmukh*

Signature of Malav Makadia: *Malav Makadia*

Signature of Gaurav Ramoliya: *Gaurav Ramoliya*

Submission Date: 22 March 2024

Problem Setting:

The dataset provided pertains to customer churn in a telecommunications company, specifically in California. Customer churn refers to the phenomenon where customers discontinue their services with the company. The dataset includes a variety of information about the customers, their demographics, services subscribed to, billing details, and the reasons behind their churn.

Additionally, the dataset includes information on the population of different zip code areas. This demographic data can be utilized to understand the context in which the company operates, allowing for more targeted strategies in customer retention.

Problem Definition:

In this analysis, the primary focus is on predicting and understanding customer churn within a telecommunications company. The first objective involves developing a predictive model to identify customers who are at risk of churning. This entails identifying key indicators or features that contribute significantly to churn, providing insights that can inform proactive retention strategies.

Additionally, the analysis delves into the impact of various services subscribed to by customers, such as phone service, internet, and streaming, on churn rates. By exploring specific combinations of services that correlate with higher churn rates, the company can tailor its service offerings or retention efforts accordingly.

Furthermore, the relationship between billing details, including monthly charges, payment methods, and paperless billing, and customer churn is examined. This investigation aims to determine whether billing methods or amounts play a significant role in churn, helping the company refine its billing practices to potentially mitigate churn rates.

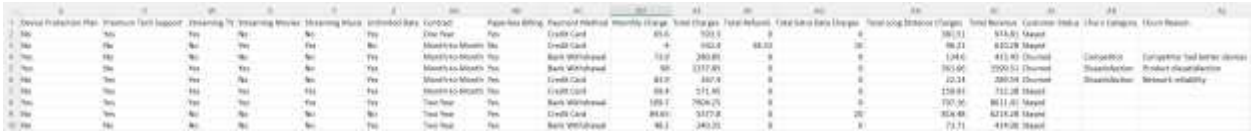
Lastly, the analysis delves into the reasons for churn provided by customers, both at a high-level category and specific reason level. Understanding the most common reasons for churn and addressing them effectively can aid in reducing churn rates and improving customer retention efforts. Additionally, revenue analysis is conducted to quantify the financial impact of churn, identifying strategies to minimize revenue loss associated with customer attrition.

Data Sources: https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics/data

Data Description:

Rows: 7043
Columns: 38

Dataset contains information of customers such as customer id, gender, age, address, married or not, services they use and total cost.



Data Preprocessing:

Handling missing values: We have handled missing values with average values of the columns. We have categorical values with the most frequent values in the columns.

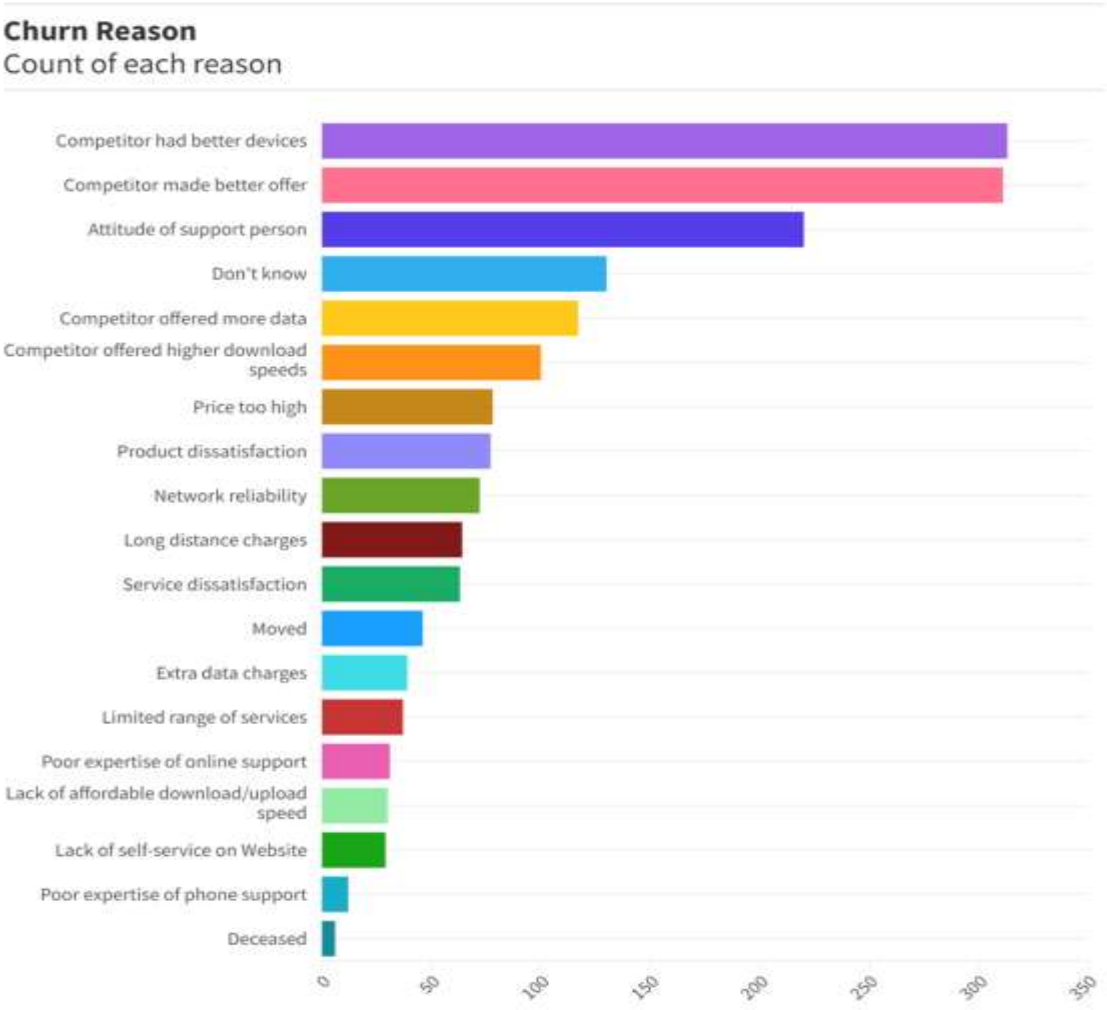
Duplicates: There are no duplicates in the dataset.

Categorical values: We have converted categorical values of columns having yes/no to 1/0 with the help of label encoding.

Data Visualization:

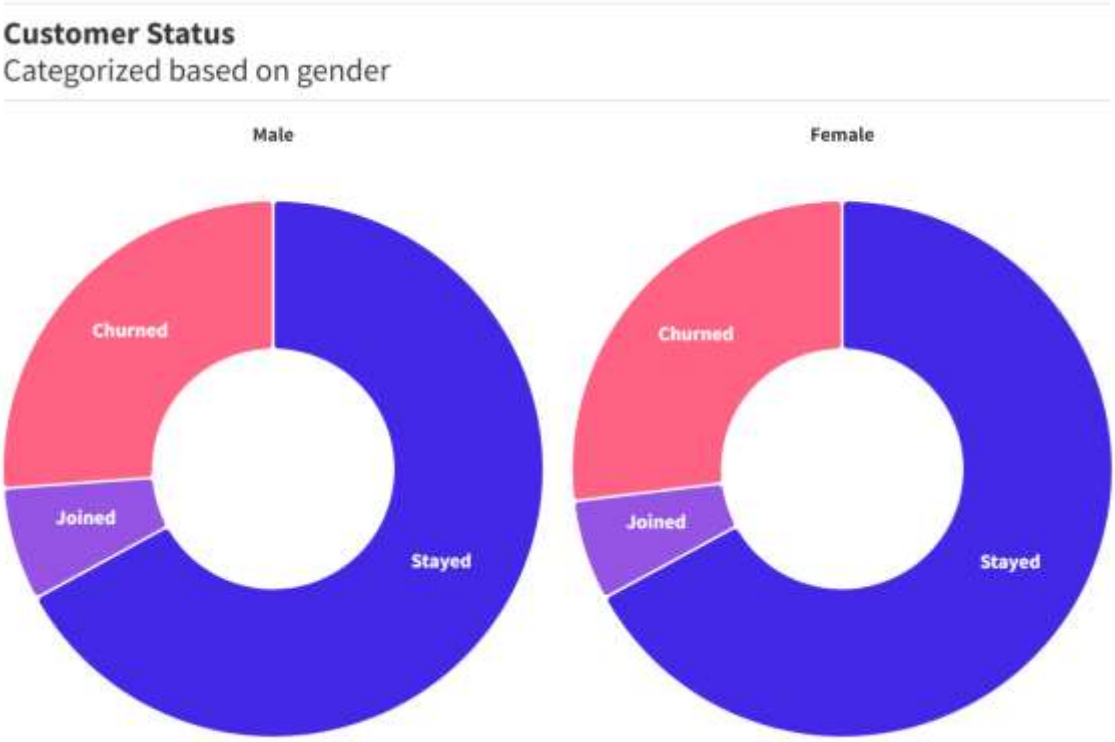
Following are the visualizations from the dataset to understand the trends and to gain the insights from the dataset.

1) **Understanding the Churn Reason:**
   Exploring the prominent reasons behind customer churn with an interactive visualization. Dive into the key factors driving customer attrition. Below is interactive visualization, the link for it is provided below the image of that visualization.



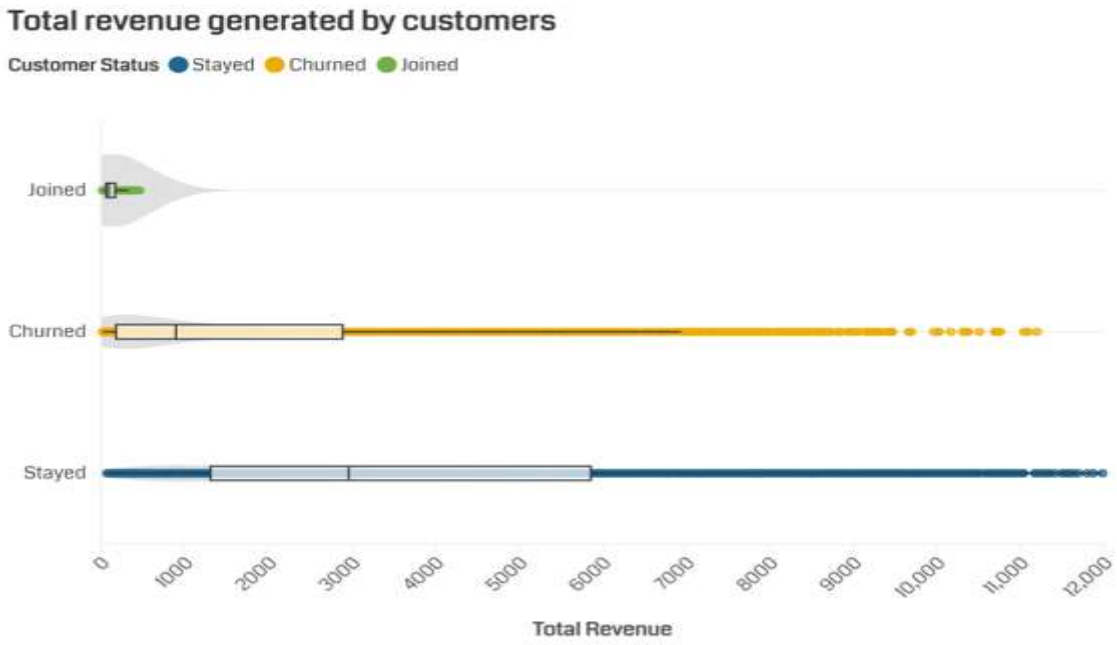
Click here for interactive visualization.
Link: https://public.flourish.studio/visualisation/16822273/

2) **Customer Status Categorization:**
Gaining insights into customer demographics by categorizing them based on gender and their status. A pie chart illustrates the distribution, revealing intriguing patterns. Below is interactive visualization, the link for it is provided below the image of that visualization.



Click here for interactive visualization
Link: https://public.flourish.studio/visualisation/16832837/

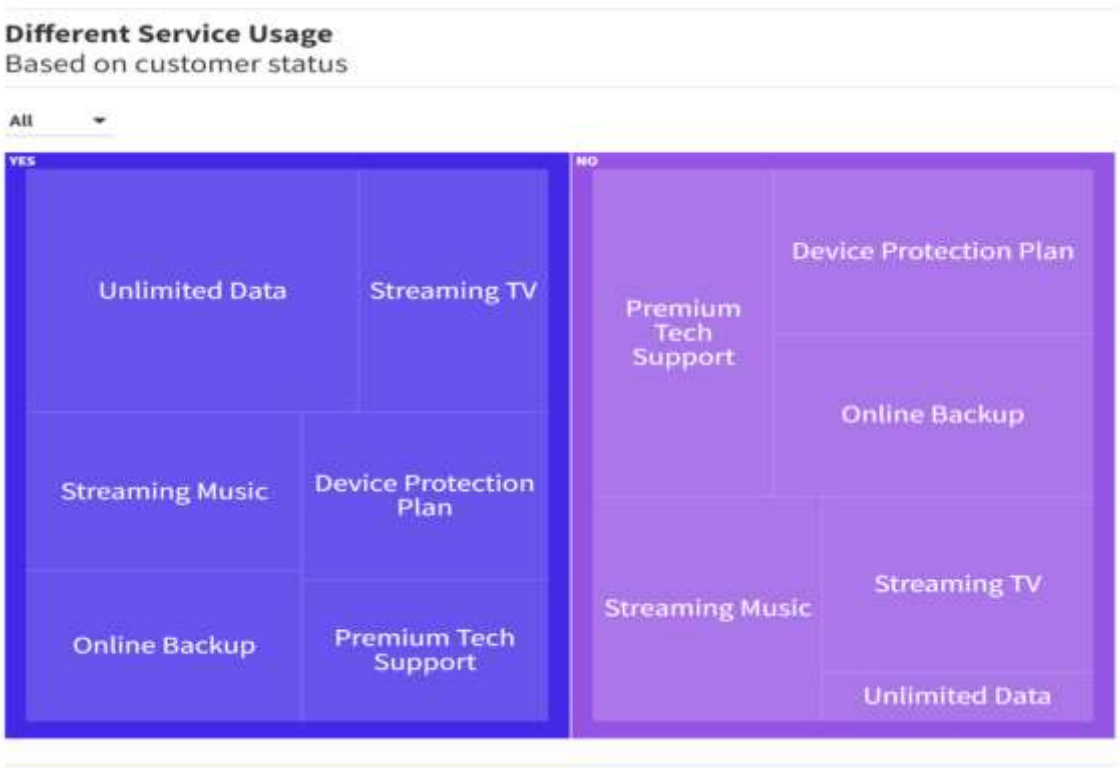
3) Total Revenue generated from all customer:



Delving into the financial landscape by examining the total revenue generated from churned, retained, and new customers. A box plot provides a clear overview, aiding in strategic decision making. Below is interactive visualization, the link for it is provided below the image of that visualization.

Click here for interactive visualization
Link: https://public.flourish.studio/visualisation/16832669/

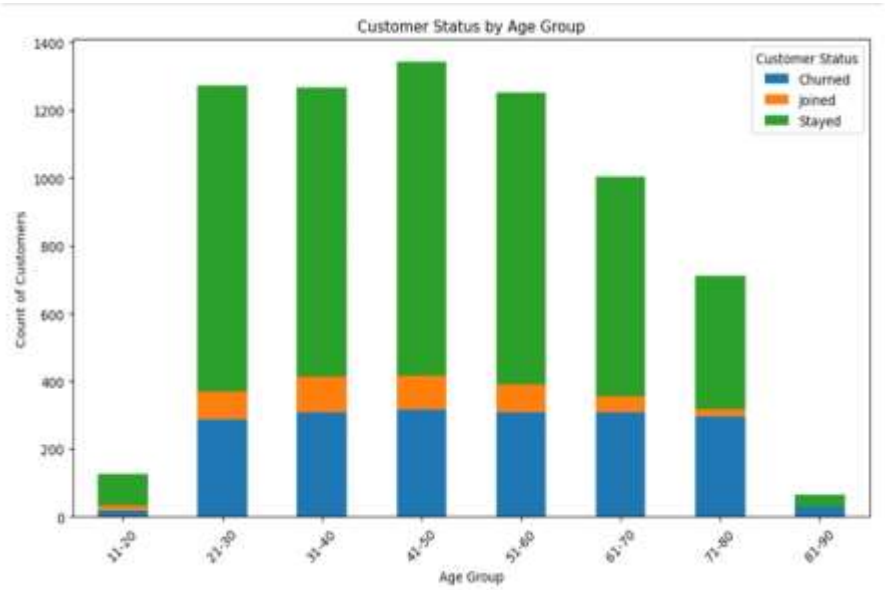
4) Services used by the customer:



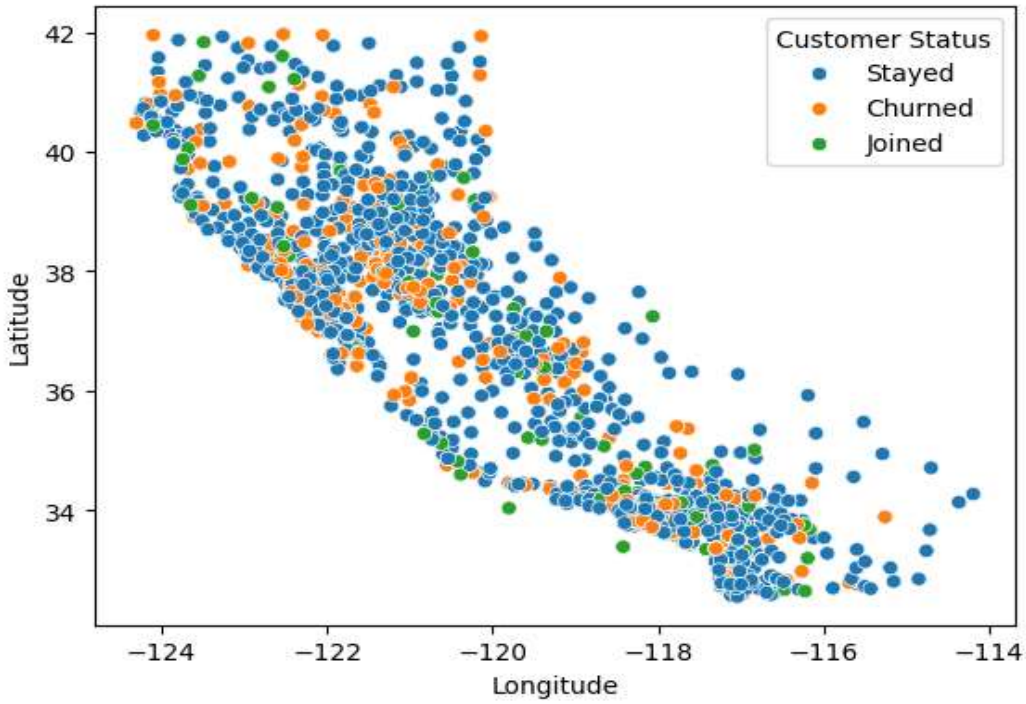
Uncovering the breadth of services utilized by customers, beyond network and internet access. Explore a tree map visualization showcasing service adoption trends. Below is interactive visualization, the link for it is provided below the image of that visualization.

Click here for interactive visualization
Link: https://public.flourish.studio/visualisation/16833063/

5) **Customers Status by Age group:**
Understanding the distribution of customers across different age groups through a comprehensive bar chart analysis. Gain valuable insights into age-based customer segmentation.



6) **Customers location:**
Visualizing the geographic distribution of customers within California, offering valuable insights into regional demographics. Explore customer concentration and potential market opportunities.



Data Mining Models:

1. **Logistic Regression:**

Interpretability: Logistic regression provides coefficients for each feature, making it easy to interpret how each feature influences the probability of churn.
Assumption of Linearity: It assumes a linear relationship between the features and the log odds of churn. This may not capture complex nonlinear relationships present in the data.

- Advantages:

Interpretable: Easy to understand the impact of each feature on the churn prediction.
Efficient: Computationally inexpensive and can handle large datasets.
Provides probabilities: Outputs probabilities of churn, which can be useful for decision-making.

- Disadvantages:

Assumes linear relationship: May not capture complex nonlinear relationships between features and churn.
Limited expressiveness: Unable to capture interactions between features effectively.

2. **Decision Trees:**

Nonlinear Relationships: Decision trees can capture nonlinear relationships between features and the target variable, which is beneficial for churn prediction tasks.
Interpretability: Decision trees provide a clear decision-making process, making it easy to understand the factors leading to churn.

- Advantages:
  Nonlinear relationships: Can capture complex nonlinear relationships and interactions between features.
  Interpretable: Easy to visualize and understand the decision-making process. Handles mixed data types: Can handle both numerical and categorical features without requiring preprocessing.
- Disadvantages:
  Prone to overfitting: Decision trees can create overly complex models that perform well on training data but generalize poorly to new data.
  Instability: Small changes in data can lead to significantly different trees, impacting model reliability.

3. **Random Forests:**

Ensemble Learning: Random forests combine multiple decision trees, reducing overfitting and improving generalization performance.

Feature Importance: Random forests provide a measure of feature importance, helping to identify which features are most influential in predicting churn.

- Advantages:
  Reduces overfitting: Aggregating multiple decision trees reduces overfitting compared to individual trees.
  High predictive accuracy: Often provides better generalization performance compared to decision trees.
  Feature importance: Can provide insights into which features are most influential in predicting churn.

- Disadvantages:
  Lack of interpretability: Random forests are less interpretable compared to single decision trees.
  Computationally intensive: Training and predicting with random forests can be slower than simpler models like logistic regression.

4. **Gradient Boosting Machines (GBM):**

Sequential Improvement: GBM builds decision trees sequentially, each focusing on the mistakes of its predecessors. This leads to higher predictive accuracy compared to standalone decision trees.
Robustness: GBM is robust to outliers and noise in the data, making it suitable for real-world datasets like telecom churn data.

- Advantages:
  High predictive accuracy: GBM sequentially builds trees, correcting errors of previous trees, leading to high accuracy.
  Handles missing data: Can handle missing values in features without requiring imputation.
  Feature importance: Provides insights into feature importance for churn prediction.
- Disadvantages:
  Prone to overfitting: Like other ensemble methods, GBM can overfit if not properly tuned.
  Computationally expensive: Training GBM can be time-consuming, especially with large datasets.

5. **Support Vector Machines (SVM):**

Effective in High-Dimensional Space: SVMs work well in high-dimensional feature spaces, making them suitable for churn prediction tasks with many features.
Kernel Trick: SVMs can use different kernel functions to capture nonlinear relationships between features and the target variable.

- Advantages:
  Effective in high-dimensional space: SVMs can handle datasets with many features, such as telecom customer churn data.
  Versatile: Can use different kernel functions to capture complex relationships between features.
- Disadvantages:
  Sensitivity to parameters: Performance heavily depends on the choice of kernel and regularization parameters.
  Computationally expensive: Training SVMs can be slow, especially with large datasets.

6. **Neural Networks:**

Complex Pattern Recognition: Neural networks can capture complex patterns in the data, potentially leading to higher predictive accuracy.
Data and Compute Intensive: Training deep neural networks requires large amounts of data and computational resources. However, they can be effective for churn prediction tasks with sufficient data.

- Advantages:
  Captures complex patterns: Neural networks can learn intricate patterns in data, potentially improving predictive performance.
  Automatic feature learning: Can automatically learn relevant features from raw data, reducing the need for manual feature engineering.
- Disadvantages:
  Requires large datasets: Neural networks require a large amount of data to avoid overfitting.
  Computationally expensive: Training neural networks can be computationally intensive, especially deep architectures.

7. **Gradient Boosted Trees (XGBoost, LightGBM):**

Speed and Performance: These implementations of gradient boosting algorithms are optimized for speed and performance, making them suitable for large-scale churn prediction tasks.
Regularization: XGBoost and LightGBM provide various regularization techniques to prevent overfitting, such as tree pruning and column subsampling.

- Advantages:
  High performance: XGBoost and LightGBM are optimized for speed and performance, often outperforming other models.
  Handles missing data: Can handle missing values without requiring imputation.

Feature importance: Provides insights into feature importance for churn prediction.

- Disadvantages:
  Complexity: Tuning hyperparameters and understanding the model internals can be challenging.
  Computationally expensive: Training and predicting with gradient boosted trees can be resource-intensive.

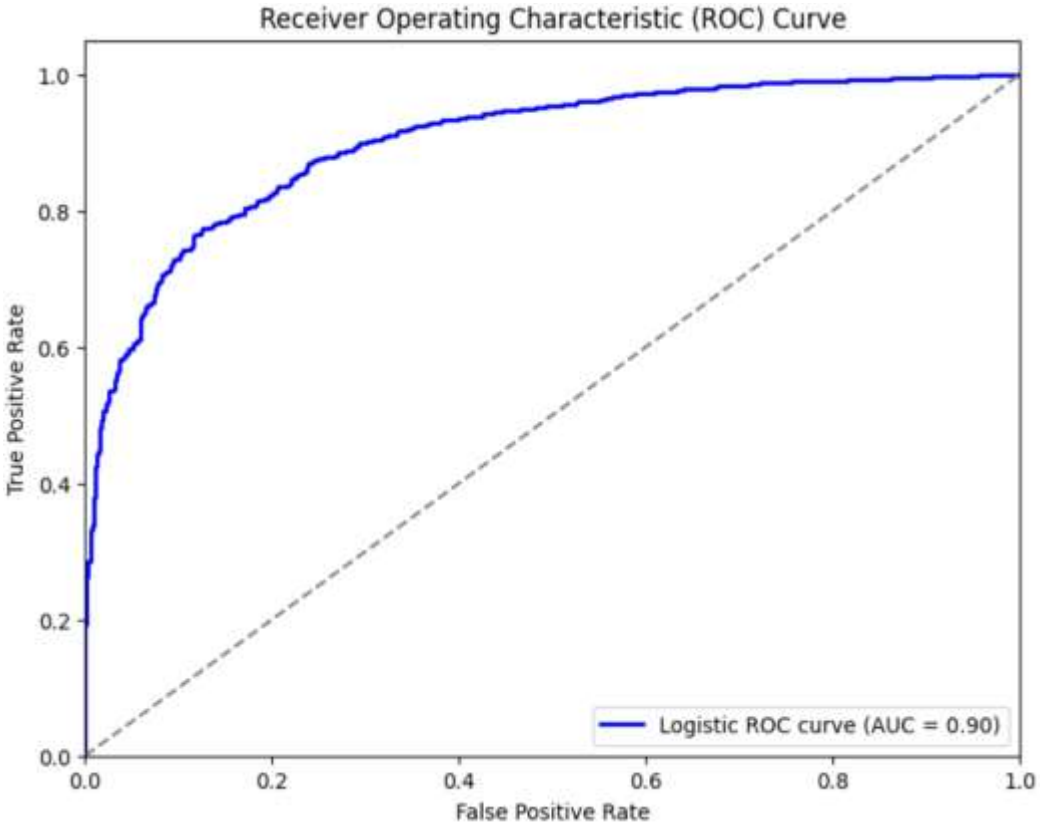
Performance Evaluation:

1) Logistic Regression:

	precision	recall	f1-score	support
0	0.73	0.71	0.72	561
1	0.89	0.90	0.89	1416
accuracy			0.84	1977
macro avg	0.81	0.80	0.81	1977
weighted avg	0.84	0.84	0.84	1977

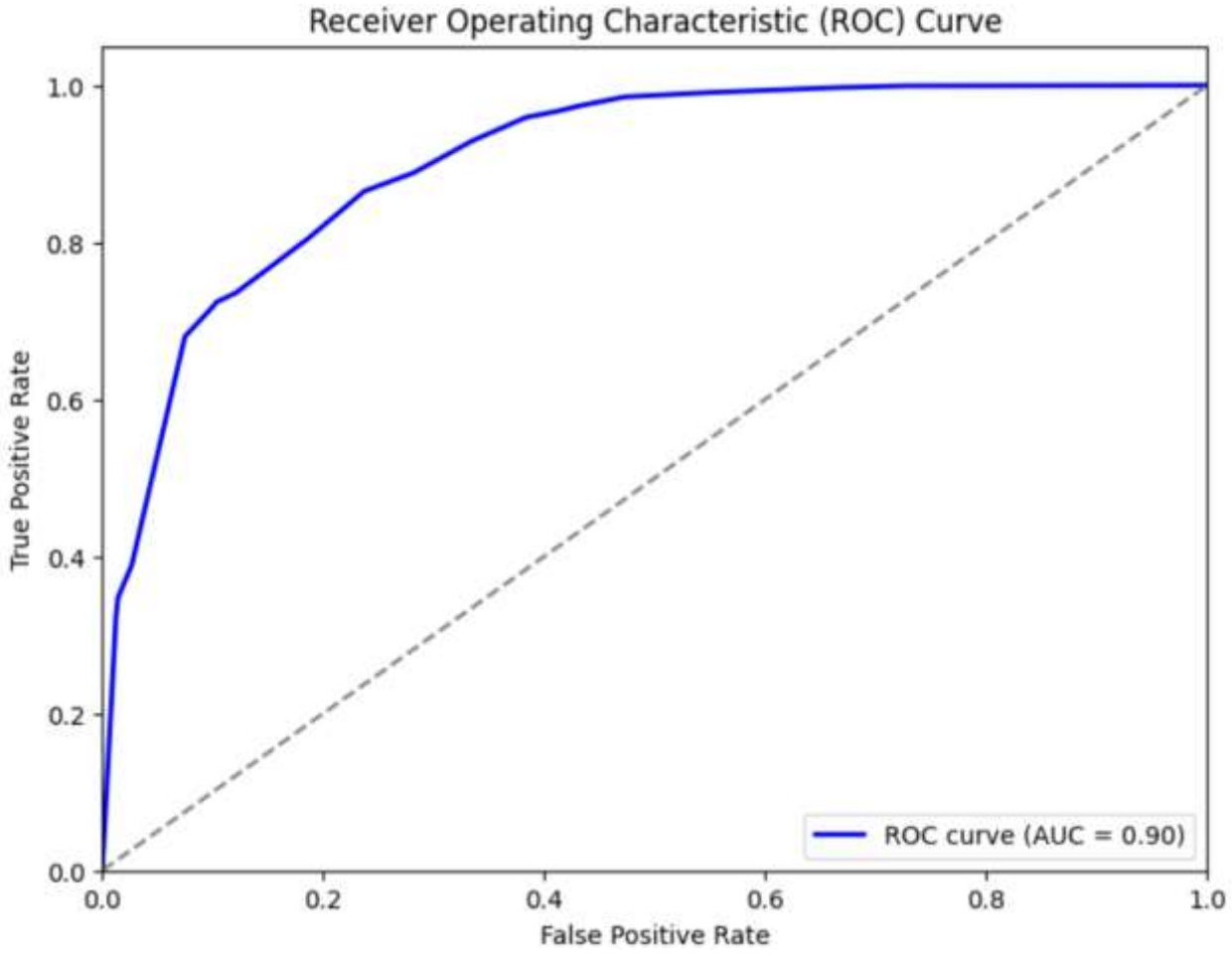
Confusion matrix (Rows actual, Columns predicted):

	0	1
0	396	165
1	143	1273



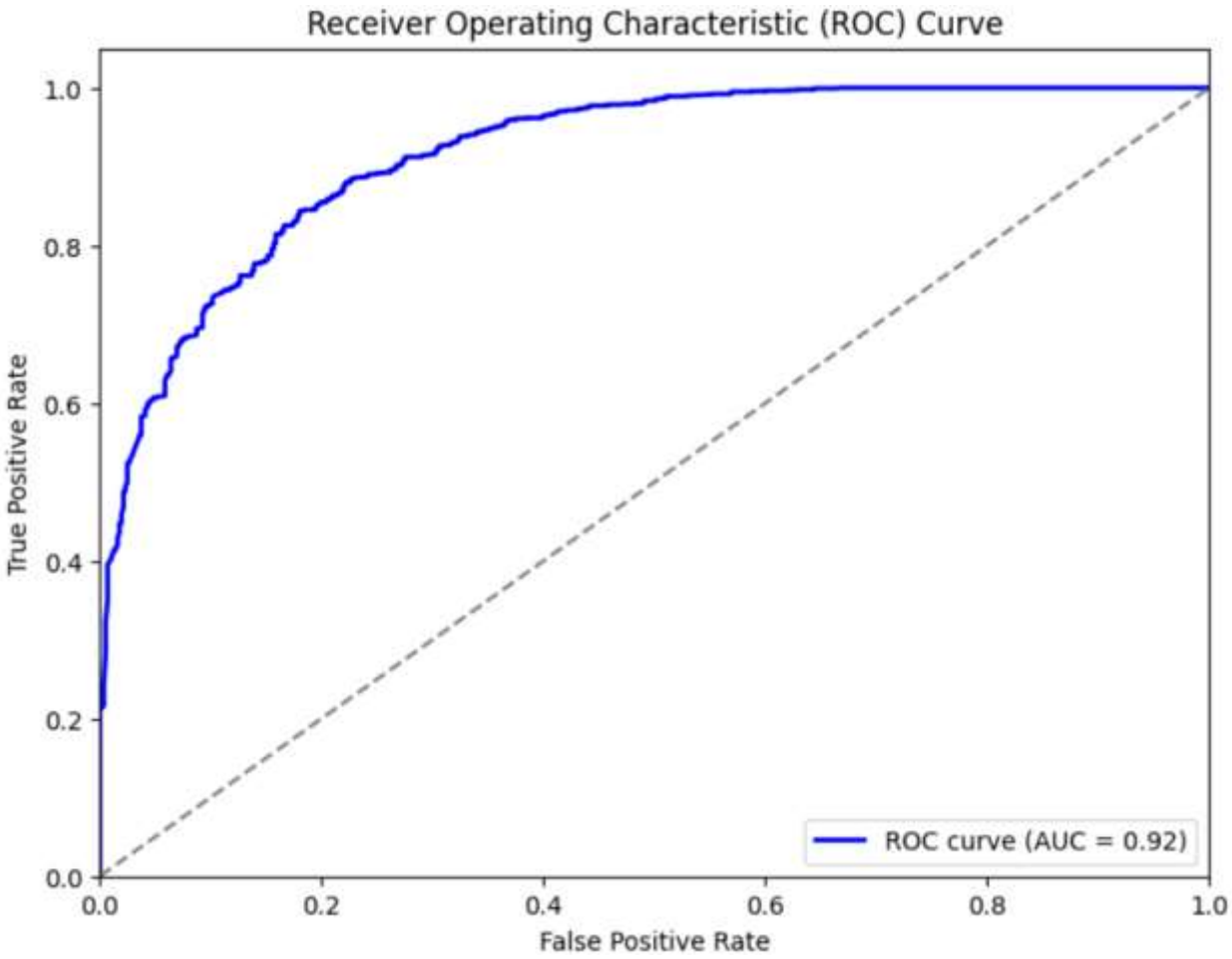
2)Decision Trees:

```
Classification report:
              precision    recall  f1-score   support

           0       0.86      0.62      0.72       561
           1       0.86      0.96      0.91      1416

    accuracy                           0.86      1977
   macro avg       0.86      0.79      0.81      1977
weighted avg       0.86      0.86      0.85      1977

Confusion matrix (Rows actual, Columns predicted):
     0     1
0  346   215
1   58  1358
```



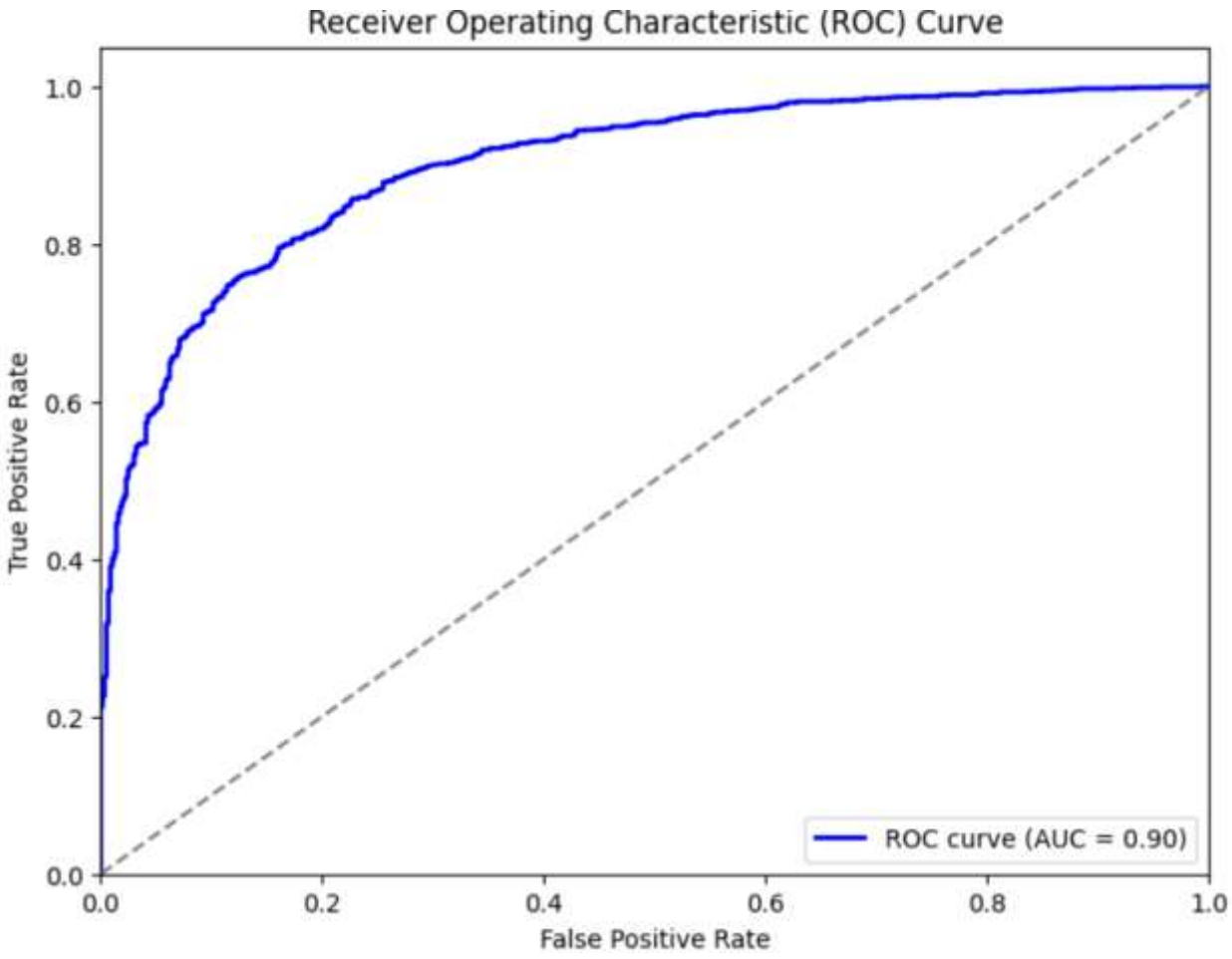
3)Random Forest:

```
Classification report:
              precision    recall  f1-score   support

           0       0.83      0.65      0.73       561
           1       0.87      0.95      0.91      1416

    accuracy                           0.86      1977
   macro avg       0.85      0.80      0.82      1977
weighted avg       0.86      0.86      0.86      1977

Confusion matrix (Rows actual, Columns predicted):
     0     1
0  365   196
1   75  1341
```



4)Support Vector Machine:

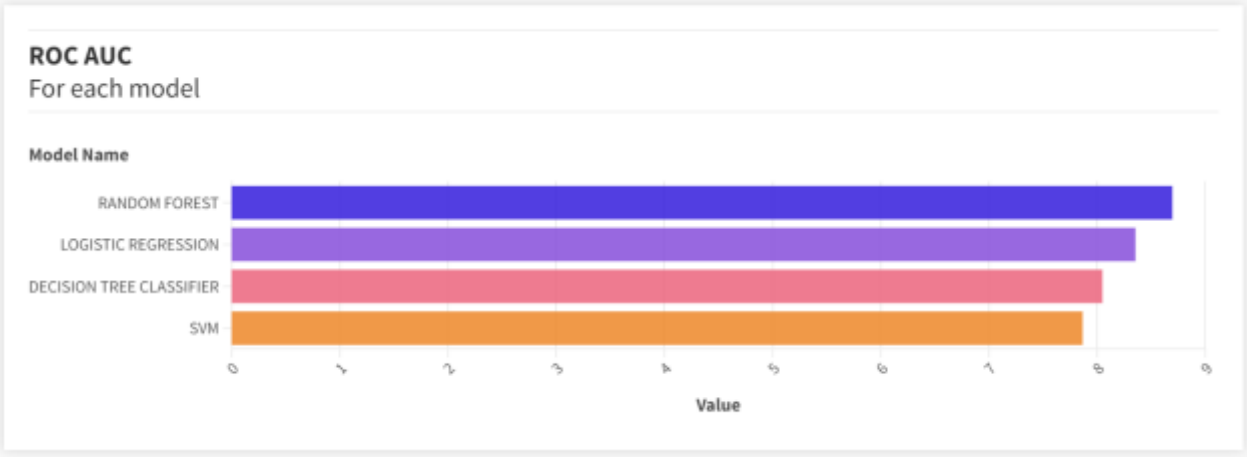
```
Classification report:
              precision    recall  f1-score   support

           0       0.73      0.72      0.72       561
           1       0.89      0.89      0.89      1416

    accuracy                           0.84      1977
   macro avg       0.81      0.80      0.81      1977
weighted avg       0.84      0.84      0.84      1977

Confusion matrix (Rows actual, Columns predicted):
     0     1
0  402   159
1  152  1264
```



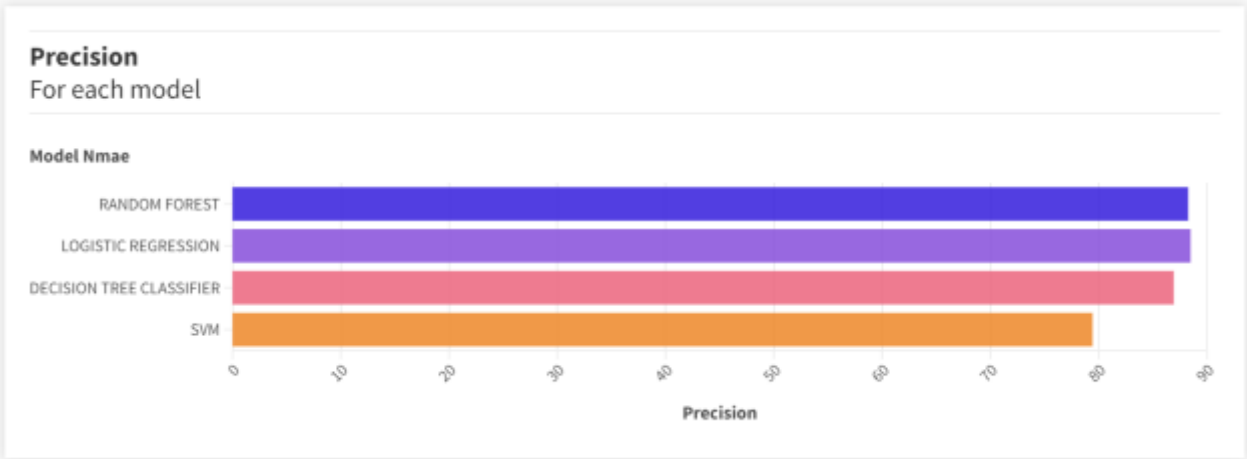
Summary of all evaluations:

ROC Curve Interpretation: All four Receiver Operating Characteristic (ROC) curves exhibit remarkable similarity, boasting an Area Under the Curve (AUC) of 90, signifying strong predictive performance. Notably, the Random Forest model demonstrates a marginally superior AUC of 92, further accentuating its efficacy in classification tasks. This consistent high performance across all models underscores their robustness and reliability in discriminating between classes, thereby affirming their suitability for practical applications

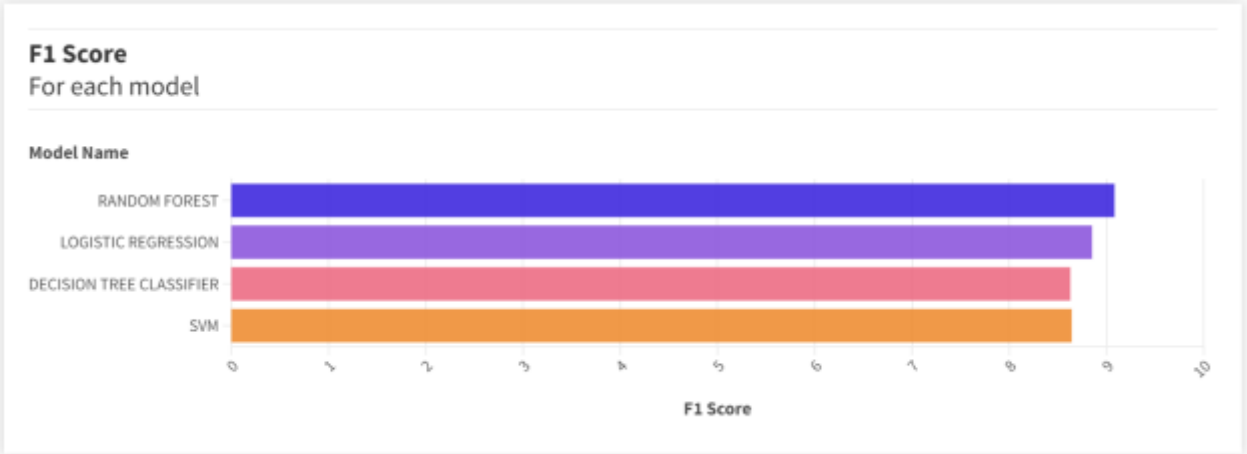
	Algorithm	ROC AUC	Accuracy	Precision	f1 Score
2	Random Forest	86.95	92.18	88.27	90.82
0	Logistic Regression	83.54	90.11	88.49	88.51
3	Decision Tree Classifier	80.49	76.67	86.94	86.28
1	Kernel SVM	78.66	85.89	79.44	86.42



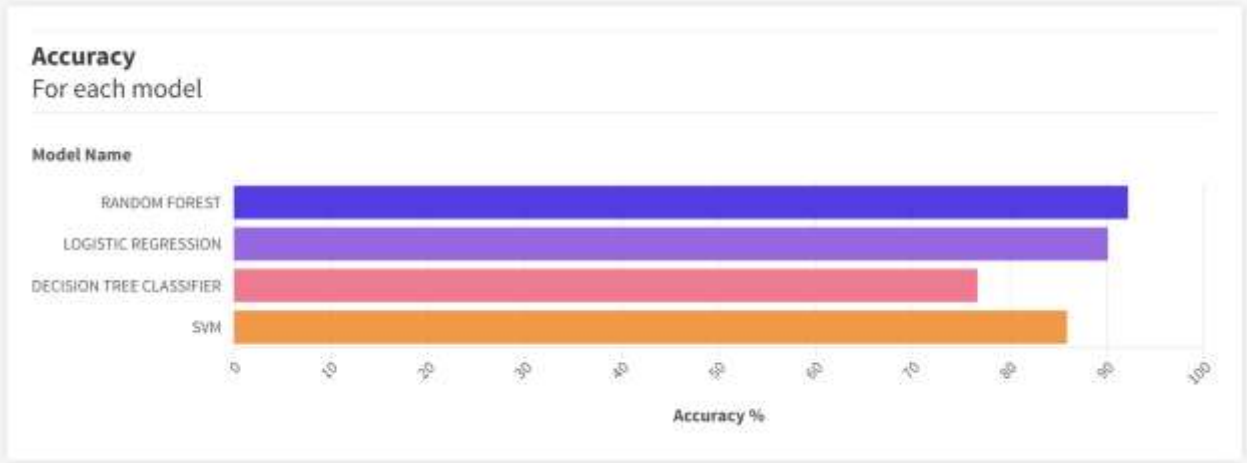
Link: https://public.flourish.studio/visualisation/17442610/



Link: https://public.flourish.studio/visualisation/17442773/



Link: https://public.flourish.studio/visualisation/17442695/



Link: https://public.flourish.studio/visualisation/17442741/

Out of four models, Random Forest model performed better than other three models with highest accuracy of 92.87 among the four models. Hence, we decided to use Random Forest model for predicting churn.

References:

Kaggle dataset:

https://www.kaggle.com/code/zakriasaad1/customer-churn-prediction-on-telecom-dataset

Exploratory Data Analysis (EDA) in Python:

https://towardsdatascience.com/exploratory-dataanalysis-eda-python-87178e35b14

Label Encoding:

https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd