CSE422

Lab Project

Dataset: Telco_Customer_Churn

Rezaul Mostofa, ID: 23301511

Salman Munawar Hossain, ID: 22201227

Section: 26

Table of Contents

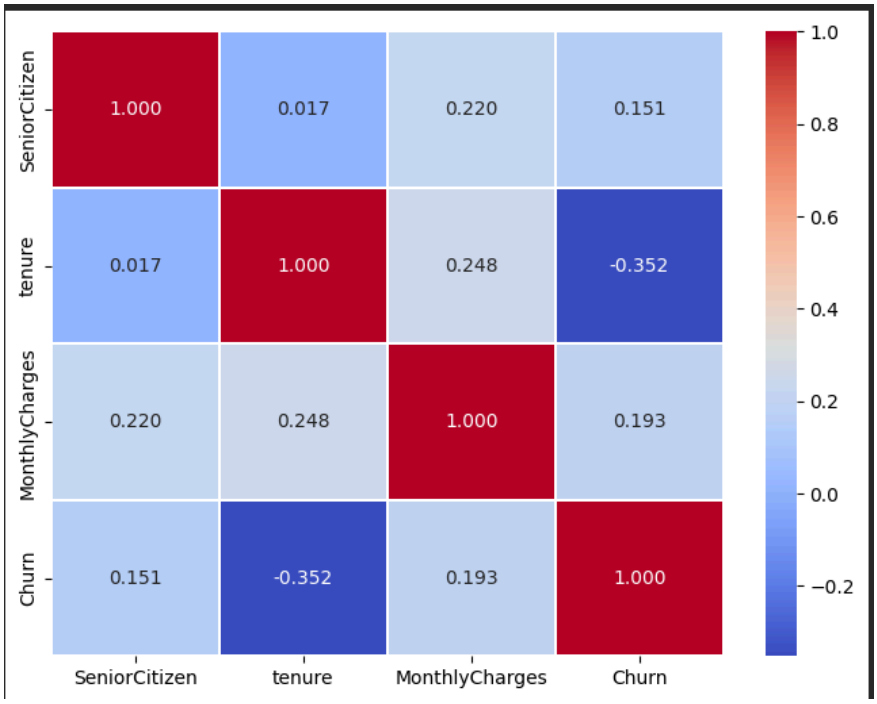
1. Introduction: Our dataset for the project is Telco_Customer_Churn. Customer. Churn refers to customers leaving a service provider. Predicting churn is vital because retaining customers is cheaper than acquiring new ones. This project aims to build machine learning models to predict customer churn using the given dataset.

2.Dataset Description:

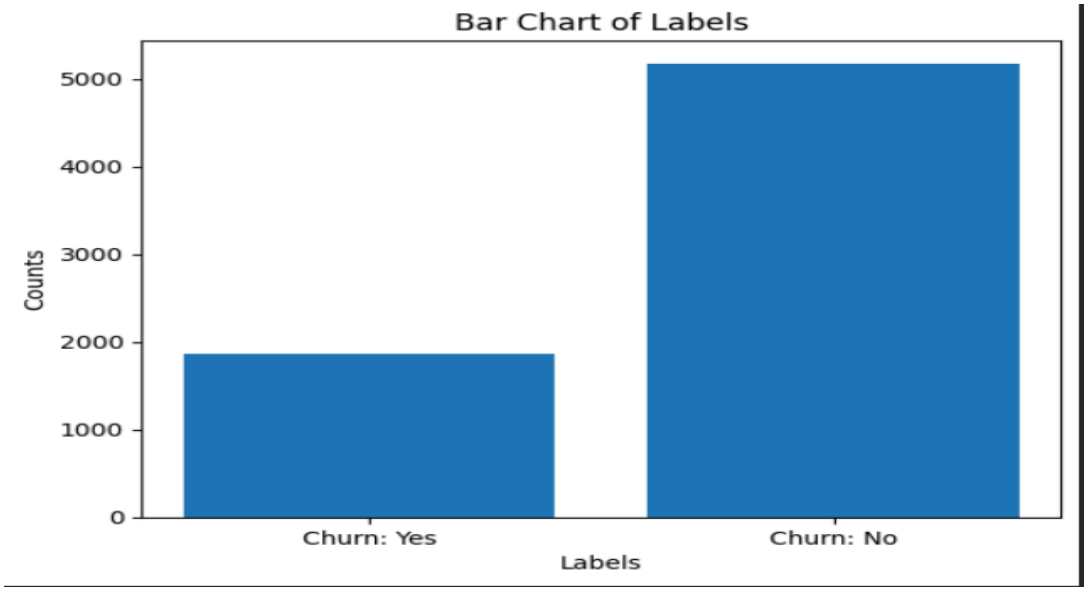
There are 21 features,7043 data points. It is a classification problem because the output variable has categorical values (Yes/No) . It has feature types of categorical and numerical. We need to encode categorical values because ML models need numerical values.

Correlation Heatmap:

The correlation heatmap shows how numerical features are related to churn. Features such as tenure and MonthlyCharges show correlation with churn. This helps to understand which features are more important for predicting customer churn.



Imbalanced Dataset: Unique classes do not have an equal number of instances. There are more No than Yes.



Exploratory Data Analysis (EDA)

Exploratory analysis revealed the following insights:

- Customers with month-to-month contracts have a higher churn rate.
- Short-tenure customers are more likely to churn.
- Customers subscribing to fewer services show a higher tendency to churn.

3. Dataset Pre-processing

3.1 Identified Faults

- Presence of invalid values in the TotalCharges column
- Categorical variables unsuitable for direct model input
- Features with different numerical scales

3.2 Pre-processing Techniques (Solutions)

- Invalid entries in TotalCharges were converted to numeric values and missing values were imputed using the median.
- The customerID column was removed as it had no predictive value.
- Categorical variables were encoded using One-Hot Encoding.
- Feature scaling was performed using StandardScaler to normalize numerical values.

4. Dataset Splitting

The dataset was split into training and testing sets using an 80:20 ratio. Stratified sampling was applied to preserve the class distribution of the target variable in both sets, ensuring fair model evaluation.

5. Model Training and Testing

The following supervised learning models were implemented:

5.1 Logistic Regression

Logistic Regression was used as a baseline model. It demonstrated good overall performance and provided well-calibrated probability estimates for churn prediction. Accuracy is 0.81.

5.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a supervised, distance-based classification algorithm that assigns class labels based on the majority vote of the k nearest data points. The algorithm relies on distance calculations, making feature scaling an important preprocessing step in this project. Based on the classification results, the KNN model achieved an overall accuracy of 0.75, with moderate performance in identifying churn customers.

5.3 Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem and assumes independence among input features. In this project, Gaussian Naive Bayes was applied due to the numerical nature of the dataset and its computational efficiency. According to the results, the Naive Bayes model achieved an overall accuracy of 0.66, showing comparatively lower classification performance than KNN.

5.4 Neural Network (MLPClassifier)

A Neural Network model was implemented using a multilayer perceptron architecture.We used smote because it generates synthetic samples for the minority class, helping the Neural Network learn balanced decision boundaries without bias toward the majority class. Despite extensive tuning and class imbalance handling using resampling techniques, the Neural Network achieved moderate performance. This highlights the limitations of neural networks on structured tabular datasets.Accuracy of this model was 0.76.

5.5 K-Means Clustering

K-Means clustering was applied as an unsupervised learning method to identify customer segments. The clusters revealed distinct patterns based on tenure and service usage, providing additional insights into customer behavior.

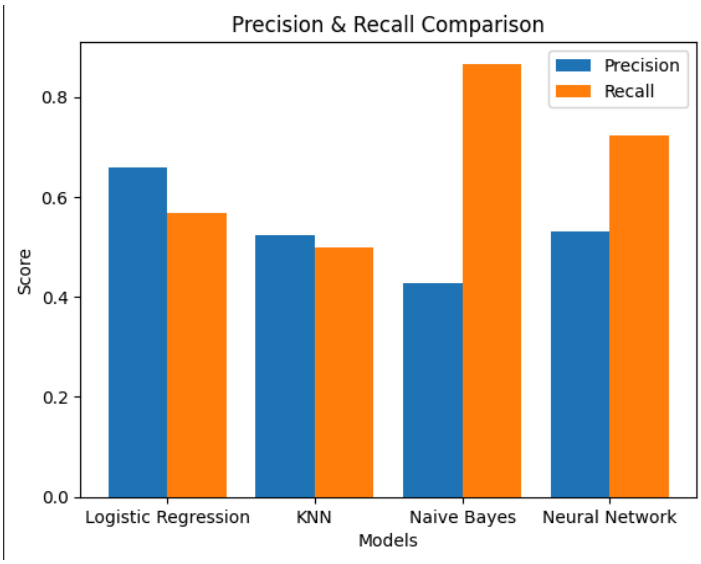
6. Model Comparison and Evaluation

The models were evaluated using the following metrics:

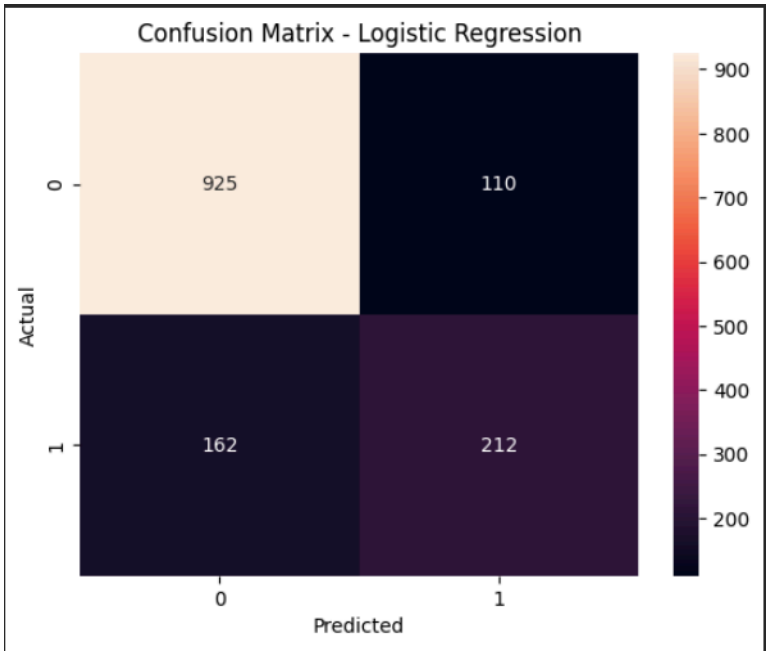
- Accuracy

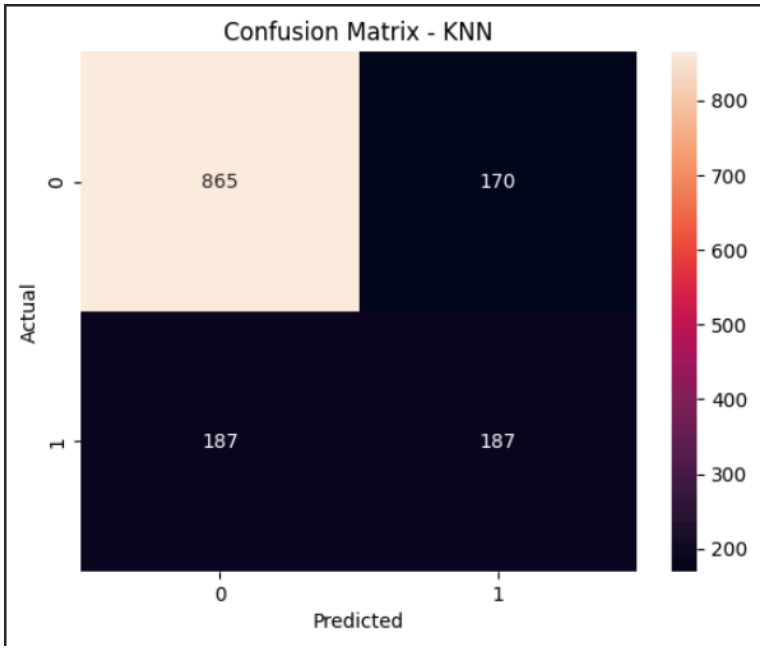
```
Logistic Regression
              precision    recall  f1-score   support

           0       0.85      0.89      0.87      1035
           1       0.66      0.57      0.61       374

    accuracy                           0.81      1409
   macro avg       0.75      0.73      0.74      1409
weighted avg       0.80      0.81      0.80      1409

KNN
              precision    recall  f1-score   support

           0       0.82      0.84      0.83      1035
           1       0.52      0.50      0.51       374

    accuracy                           0.75      1409
   macro avg       0.67      0.67      0.67      1409
weighted avg       0.74      0.75      0.74      1409

Naive Bayes
              precision    recall  f1-score   support

           0       0.92      0.58      0.71      1035
           1       0.43      0.87      0.57       374

    accuracy                           0.66      1409
   macro avg       0.67      0.72      0.64      1409
weighted avg       0.79      0.66      0.67      1409

Neural Network
              precision    recall  f1-score   support

           0       0.88      0.77      0.82      1035
           1       0.53      0.72      0.61       374

    accuracy                           0.76      1409
   macro avg       0.71      0.75      0.72      1409
weighted avg       0.79      0.76      0.77      1409
```

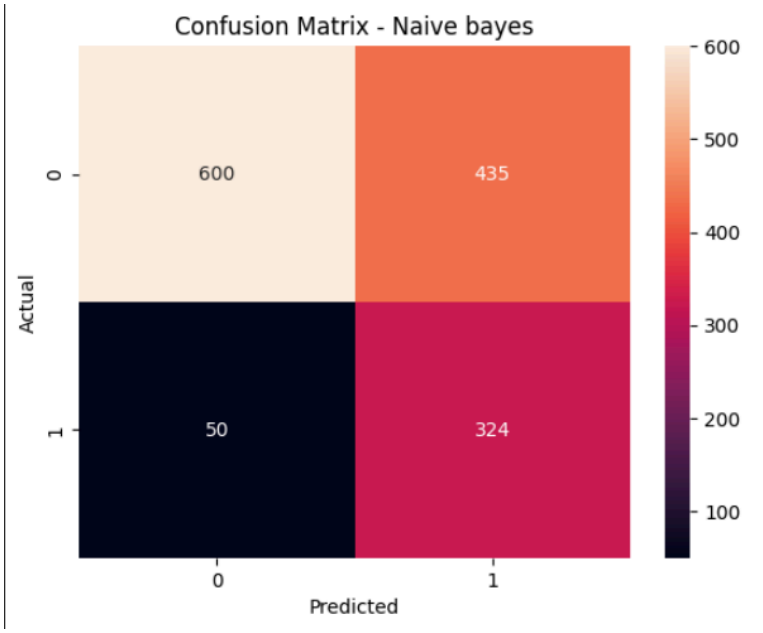
- Precision
- Recall

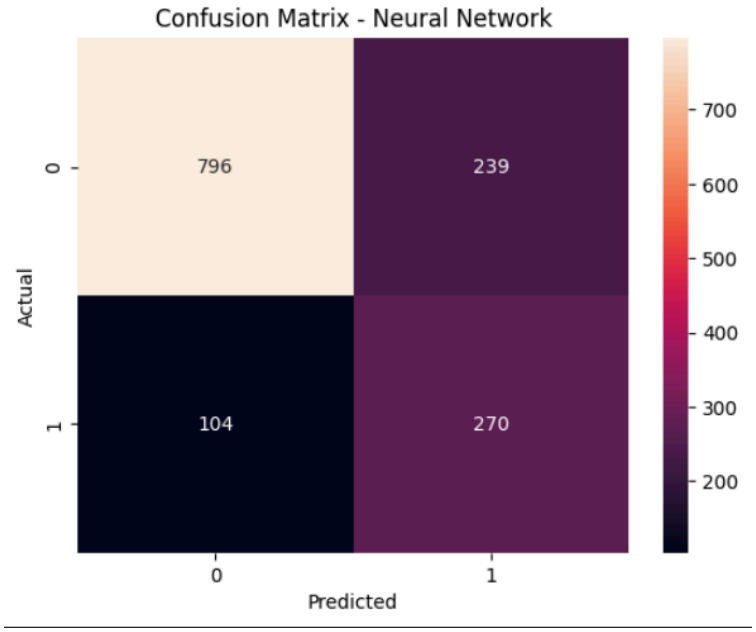


- Confusion Matrix

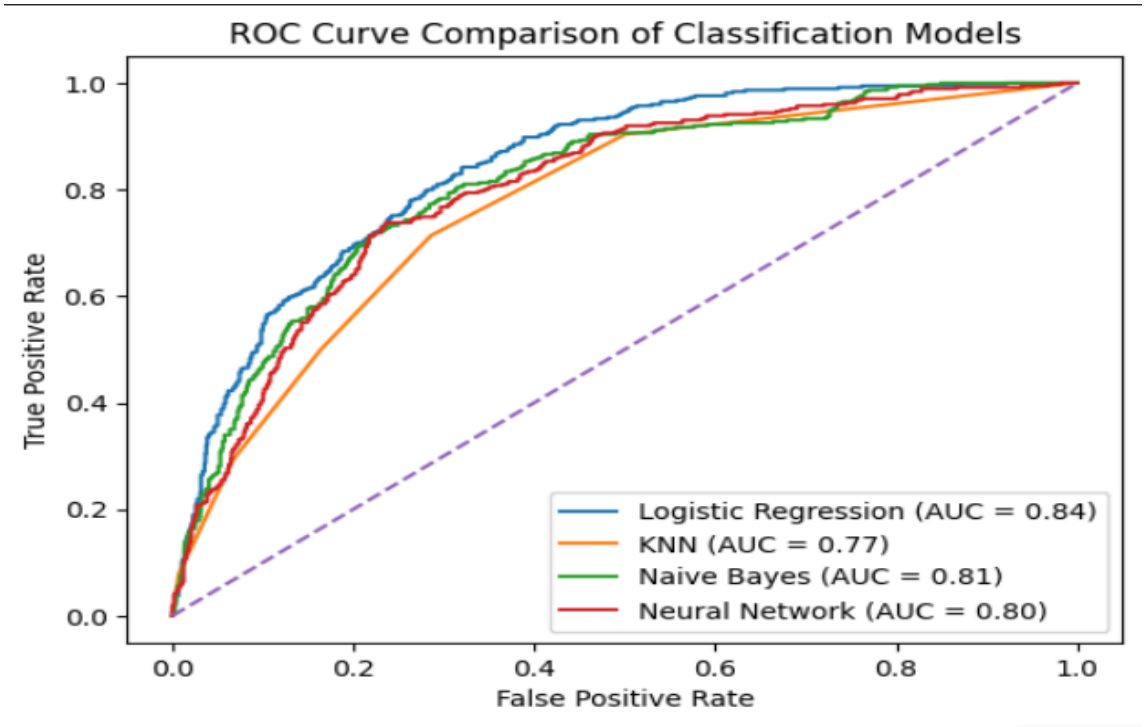



Confusion Matrix - KNN
865
170
187
187
0
1
Actual
0
1
Predicted
800
700
600
500
400
300
200


Confusion Matrix - Naive bayes
600
435
50
324
0
1
Actual
0
1
Predicted
600
500
400
300
200
100



- ROC Curve and AUC



The comparison showed that Logistic Regression outperformed all other models in terms of AUC score. Neural Network & Naive Bayes performed well, while the KNN showed lower performance due to the structured nature of the dataset.

7.Conclusion

This project explored customer churn prediction using multiple machine learning models. Proper data preprocessing and careful model selection played a crucial role in improving predictive performance.

Among all the models, Logistic Regression emerged as the most effective classifier, achieving the highest AUC score and demonstrating robustness in handling complex, non-linear relationships. Although a Neural Network was implemented, its comparatively lower performance reinforces the understanding that ensemble tree-based models are often better suited for tabular datasets.

One of the key challenges faced during this project was handling class imbalance and optimizing model performance. The insights gained from this study can help telecom companies proactively identify customers at risk of churning and implement targeted retention strategies.