Customer Churn Prediction with Machine Learning

1. Introduction

This project is a machine learning project that aims to predict customer churn for a telecom company. Customer churn refers to the number of customers who leave a company's products or services during a given time period. Predicting customer churn is crucial for businesses as it helps them retain their customers and increase revenue.

2. Objective

The primary objective of this project is to develop a machine learning model capable of accurately predicting customer churn for a telecom company. By identifying customers who are likely to discontinue the service, the company can implement targeted retention strategies, improve customer satisfaction, and reduce revenue loss. This project focuses on leveraging customer demographic data, account information, and service usage details to build and optimize a predictive model that distinguishes between churned and retained customers.

3. Dataset Used

The dataset used for this project can be found here.
https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling

The dataset contains information on customer demographics, account information, and service details. The target variable is a binary variable indicating whether the customer has churned or not.

4. Methodology

The project was divided into the following steps:

Data Cleaning and Preprocessing: The dataset was cleaned and preprocessed to remove missing values, outliers, and other inconsistencies
Exploratory Data Analysis: Various exploratory data analysis techniques were used to understand the relationship between different variables and the target variable.
Feature Engineering: New features were created to improve the performance of the machine learning models.
Model Selection: Several machine learning models were trained and tested to select the best performing model.
Hyperparameter Tuning: The hyperparameters of the selected model were tuned to optimize the performance.

Model Evaluation: The final model was evaluated on a test dataset to measure its performance.

5. Model Chosen

The **Random Forest Classifier** was chosen to address overfitting seen with the **Decision Tree** model. It combines multiple decision trees, improving **generalization** and reducing the risk of overfitting. Additionally, it handled the **data imbalance** effectively, making it a more robust choice for predicting customer churn.

6. Performance Metrics

Confusion Matrix:

```
[[1574   42]
 [ 203  181]]
```

Accuracy Score: 0.8775
245 / 2000

KappaScore is: 0.53008239861288

Classification Report:

Class	Precision	Recall	F1-Score	Support
0 (No Churn)	0.89	0.97	0.93	1616
1 (Churn)	0.81	0.47	0.60	384

Overall Metrics:

Metric	Accuracy	Macro Avg	Weighted Avg	Support
Score	0.88	0.76	0.86	2000

7. Results

The final model decision tree model showed us overfitting problem. Hence the randomised search cv on random forest classifier gave us better accuracy which is 87 percent and wrong predictions made by the model are 243/2000 and grid search cv gave us 87 percent accuracy and wrong predictions are 246/2000.

8. Challenges & Learnings

Challenges:

1. **Overfitting:** The decision tree model overfitted the training data.
   - **Solution:** Switched to **Random Forest** to improve generalization.
2. **Data Imbalance:** The dataset had more non-churned customers, leading to biased predictions.
   - **Solution:** Applied **class weighting** and **resampling** techniques.
3. **Hyperparameter Tuning:** Finding optimal hyperparameters for Random Forest was time-consuming.
   - **Solution:** Used **Grid Search CV** and **Randomized Search CV** for better tuning.
4. **Evaluation Metrics:** The model had low recall for churned customers.
   - **Solution:** Focused on **precision**, **recall**, and **F1-score** instead of just accuracy.

Learnings:

1. **Model Selection:** Choosing the right model (e.g., Random Forest over Decision Tree) is key for reducing overfitting and improving accuracy.
2. **Importance of Balanced Metrics:** Focusing on precision and recall for imbalanced datasets helps in better performance for the minority class.

9. Conclusion

The machine learning model developed in this project can be used by the telecom company to predict customer churn and take appropriate actions to retain customers. The model can also be improved further by incorporating additional data sources and refining the feature engineering process.