Content

Objective

Building predictive models to identify customers who are likely to leave a telecom service is the goal of the project. Accurate churn prediction can enhance customer satisfaction and assist businesses in identifying customers likely churn early and take action to keep them.

Dataset Overview

Resampled_Training_Data.csv is the dataset that was used; it contains features related to customer information like

- Contract type
- Tenure
- Monthly charges
- Internet and phone services

Churn_1 is the targeted column, which indicates whether

- A client has churned (1)
- If the customer stayed (0)

Data Preprocessing

Target & Feature Split: The dataset was divided between target (y = Churn_1) and feature variables X.

Normalisation: In order to ensure uniform scale and better model performance, two numerical features—tenure and MonthlyCharges—were normalised using StandardScaler.

Train-Test Split: In order to maintain the target's class distribution, the dataset was divided into 80% training and 20% testing sets using stratified sampling.

Model Training and Results

Tested four different models on the same data

Logistic Regression Model

- An easy-to-understand model.
- provided poor results.
- Good for analysing simple patterns.

Random Forest Model

- combined into one ensemble.
- produced superior outcomes to those of logistic regression.
- shows performance in managing complex relationships.

KNN (K-Nearest Neighbors)

- Uses the behaviour of similar customers to predict a customer's churn.
- For huge datasets, performance was slower but still acceptable.
- Performs best when the data is properly scaled, which we made sure of by normalising it.

XGBoost

- An extremely effective boosting algorithm.
- Performed better than any other model.
- Used default settings for first training (max depth = 5, learning rate = 0.1, etc.).

Comparison of Model Accuracy

We used Python to generate and store each classification model's accuracy score in order to evaluate how well it performed. XGBoost, Random Forest, KNN, and Logistic Regression were the models used in this comparison. Each model was linked to its accuracy score using a dictionary structure, which made it simple to publish each result and iterate through the models. The code then used Python's built-in max() function to determine which model had the highest accuracy. This offered a simple and straightforward method for identifying the best model based on accuracy. After that, the model with the best performance and matching accuracy score were printed.

Model Accuracies:

- Logistic Regression: 0.52
- K-Nearest Neighbors (KNN): 0.68
- Random Forest: 0.73
- XGBoost: 0.75

Cross-Validation

- We tested the XGBoost model using 5-Fold Cross-Validation:
- They divided the data into five groups.
- Once for testing and once for training, each group was utilised.
- A positive evaluation of the model's performance was therefore provided.
- The standard deviation and mean accuracy show stable results.

Hyperparameter Tuning

In order to enhance XGBoost's performance, we tested with several settings using GridSearchCV, such as:

- Learning rate
- Total number of trees
- Tree depth maximum
- The subsampling
- Regularisation values

The model was retrained after we selected the most suitable set of parameters after testing many combinations.

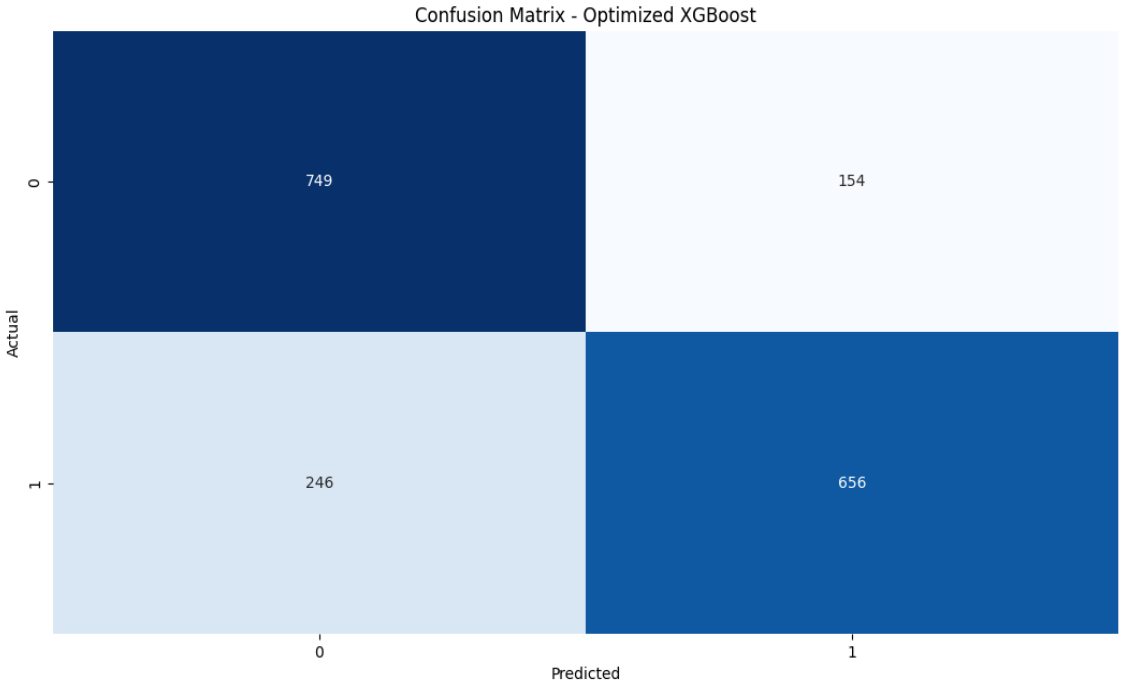
Final Model Evaluation

We tested the optimised XGBoost model. Performance was measured using:

- Accuracy

- Precision
- Recall
- F1-score

The model's ability to predict churned versus non-churned customers was displayed through a confusion matrix plot.



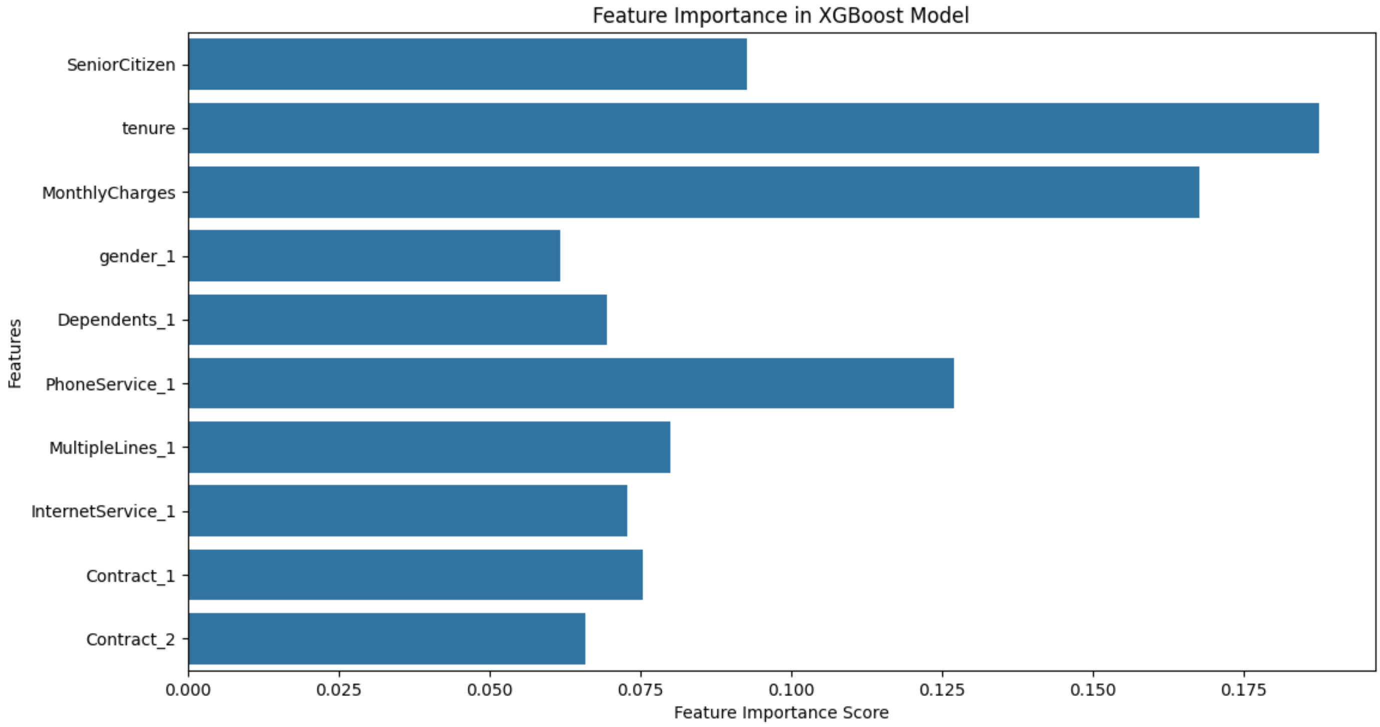
Feature Importance

Additionally, we used the built-in importance scores of XGBoost to determine which attributes had the most influence on churn prediction.

The following were among the top features:

- MonthlyCharges
- tenure
- Possibly features related to contract type and internet servicE

These were visualized using a bar chart,



In Conclusion

- XGBoost performed the best out of the four evaluated models.
- With larger datasets, KNN proved slower but still produced good predictions.
- Results from Random Forest and Logistic Regression were reasonable.
- In order to increase accuracy, proper scaling, cross-validation, and tuning were required.
- The most important customer attributes for churn prediction were also identified by the analysis.
- The outcomes can help the business in concentrating on the right customers and lowering churn through focused tactics.