Predicting Bank Customer Churn with Machine Learning

In this report, I present the results of a comprehensive analysis aimed at predicting bank customer churn using machine learning techniques. The process involved five key steps, each contributing to the understanding of the dataset and the development of predictive models.

Step 1: Importing Modules and Loading the Dataset

I initiated the analysis by importing essential Python libraries, including Pandas, NumPy, CSV, and Matplotlib. Subsequently, we loaded the dataset, denoted as "Churn_Modelling.csv," into a Pandas DataFrame. This step was crucial to access and manipulate the dataset effectively.

Step 2: Exploring the Dataset

Exploration of the dataset was pivotal to gaining insights into its structure and characteristics. I employed several Pandas functions to accomplish this, such as **head()** to preview the initial rows, **describe()** to obtain descriptive statistics of numeric columns, and **duplicated()** to check for duplicates. These initial observations provided an essential foundation for the subsequent steps.

Step 3: Exploratory Data Analysis

In the third step, I performed exploratory data analysis (EDA) on specific columns of interest. Utilizing a custom module called **col_checker**, I examined columns related to customer attributes, such as credit score, age, and gender, among others. This process allowed me to understand the distribution and characteristics of these features.

Step 4: Data Processing

Data processing was a crucial phase in preparing the dataset for machine learning. I used encoding techniques, specifically one-hot encoding for the "Geography" column with three unique values and label encoding for the "Gender" column with two unique values. The encoded values were then added back to the dataset, resulting in a processed DataFrame named "df2." Additionally, I calculated feature correlations with the target variable "Exited" and visualized them to identify potential predictors of churn.

Step 5: Model Building and Evaluation

The final step involved building and evaluating machine learning models for churn prediction. I split the data into features (X) and the target variable (y), standardized the features, and then trained three regression models: Linear Regression, Decision Tree Regression, and Random Forest Regression. These models aimed to predict the likelihood of customer churn based on various features. Evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared, were calculated for each model.

Furthermore, I trained three classification models: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. These models categorized customers into churners or non-churners. Evaluation metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC were computed to assess model performance.

Results - Regression Models

Linear Regression:

- *Mean Absolute Error (MAE):* 0.283
- *Mean Squared Error (MSE):* 0.139
- *R-squared (R2):* 0.156

The relatively high MAE and MSE values indicate that this model doesn't fit the data well. The low R-squared value (0.156) suggests that only a small portion of the variance in churn can be explained by the features, making it a suboptimal choice for predicting customer churn in this context.

Decision Tree Regression:

- *Mean Absolute Error (MAE):* 0.218
- *Mean Squared Error (MSE):* 0.218
- *R-squared (R2):* -0.317

The MAE and MSE values, although lower than Linear Regression, are still relatively high. However, the R-squared value of -0.317 indicates that this model performs poorly, as it is unable to capture the underlying patterns in the data.

Random Forest Regression:

- *Mean Absolute Error (MAE):* 0.211
- *Mean Squared Error (MSE):* 0.108
- *R-squared (R2):* 0.345

Random Forest Regression offers the best performance among the regression models. The lower MAE and MSE values signify improved accuracy in predicting churn probabilities. Moreover, the R-squared value of 0.345 indicates that Random Forest Regression can explain a larger proportion of the variance in customer churn compared to the other regression models.

Result - Classification Models

Logistic Regression:

- *Accuracy:* 0.811
- *Precision:* 0.643
- *Recall:* 0.212
- *F1-score:* 0.319
- *ROC-AUC:* 0.590

The model achieved an accuracy of 0.811, indicating that it correctly classified approximately 81.1% of the cases. The precision score of 0.643 suggests that when it predicts churn, it is correct about 64.3% of the time. However, the low recall of 0.212 indicates that it misses a substantial portion of actual churn cases. The F1-score, which balances precision and recall, is 0.319. The ROC-AUC score of 0.590 represents the area under the Receiver Operating Characteristic curve, which quantifies the model's ability to distinguish between churners and non-churners.

Decision Tree Classifier:

- *Accuracy:* 0.788
- *Precision:* 0.493
- *Recall:* 0.530
- *F1-score:* 0.510
- *ROC-AUC:* 0.693

Decision Tree Classifier is another classification model that achieved an accuracy of 0.788. It displayed a precision of 0.493, indicating that it correctly classified approximately 49.3% of churn cases when predicting churn. The model's recall of 0.530 implies that it successfully identified 53.0% of actual churn cases. The F1-score of 0.510 suggests a balanced performance between precision and recall. The ROC-AUC score of 0.693 indicates good discriminatory power.

Random Forest Classifier:

- *Accuracy:* 0.860
- *Precision:* 0.794
- *Recall:* 0.448
- *F1-score:* 0.573
- *ROC-AUC:* 0.709

Random Forest Classifier, a powerful ensemble model, achieved the highest accuracy of 0.860 among all models. It exhibited a high precision of 0.794, indicating that it correctly identified approximately 79.4% of churn cases when predicting churn. However, the lower recall of 0.448 implies that it missed some actual churn cases. The F1-score of 0.573 suggests a balance between precision and recall, making it a strong performer. The ROC-AUC score of 0.709 signifies excellent discriminatory ability.

Conclusion

The results highlight the superiority of classification models, particularly the Random Forest Classifier, in predicting customer churn in the banking sector. These models demonstrated high accuracy, precision, and F1-scores, indicating their ability to classify customers into churners and non-churners effectively.

For regression models, the Random Forest Regression outperformed Linear Regression and Decision Tree Regression, with lower MAE and MSE and a higher R-squared value, making it a better choice for predicting the likelihood of churn accurately.

This analysis provides valuable insights into model performance and offers a foundation for banks to implement strategies aimed at reducing customer churn. Further refinements and the incorporation of additional features could enhance predictive accuracy and lead to more effective churn mitigation efforts in the future.