Credit Score Classification: Model Evaluation and Interpretation

Moamen Elsharkawy 202202015

1 1. Data Preprocessing and Splitting

Data Cleaning

The dataset was loaded from `train.csv`, and high-cardinality identifier columns such as `ID`, `Customer_ID`, `Name`, and `SSN` were removed. Missing values and duplicates were addressed, and data types were inspected to determine appropriate preprocessing.

Encoding

- **Label Encoding** was used for the target variable `Credit_Score`.
- **One-Hot Encoding** was applied to nominal features including `Month`, `Occupation`, `Payment_of_Min_Amount`, and `Payment_Behaviour`.
- **Ordinal Encoding** mapped the `Credit_Mix` feature from `Bad` to `Excellent` on a scale from 0 to 3.
- **Target Encoding** was used for the high-cardinality feature `Type_of_Loan`.

Feature Engineering

The `Credit_History_Age` column, originally in the format "X Years and Y Months," was converted into total months to create a numerical feature.

Train-Test Split

The dataset was split into training and testing sets using `train_test_split` to ensure unbiased model evaluation.

2 2. Model Evaluation

Five machine learning models were trained and evaluated using accuracy on the test set. The results are shown in Table 1.

Model	**Accuracy**
Random Forest	**0.8070**
Decision Tree	0.7417
K-Nearest Neighbors	0.6600
Logistic Regression	0.5487
Naive Bayes	0.5300

Table 1: Accuracy comparison of classification models

3 3. Interpretation of Random Forest

Feature Importances

The Random Forest model's feature importances were computed based on the average decrease in impurity. The top 10 most influential features are listed in Table **??**.

SHAP (SHapley Additive Explanations)

SHAP was used to interpret model predictions globally and locally. The SHAP summary plot revealed which features had the greatest influence, and the force plot explained individual predictions by showing how each feature contributed to the final output.

Partial Dependence Plots (PDP)

PDPs were used to illustrate the marginal effect of the top two features on the model's output, helping to understand how changes in one feature affect the prediction while holding others constant.

4 4. Conclusion

The Random Forest model was the best-performing model with an accuracy of 80.70%. Through SHAP and feature importance analysis, we gained valuable insights into the factors influencing model predictions, contributing to a more explainable and trustworthy credit scoring model.