Assignment 3:

Credit Score Prediction using Machine Learning Models

Overview

This assignment focuses on building, training, and evaluating machine learning models to predict a customer's **credit score** (Good, Poor, or Standard). Students will use **train.csv** for model training and testing, while **vald.csv** will be used solely for validation to assess model generalisation. The assignment involves tasks such as **data cleaning, feature engineering, model selection, hyperparameter tuning**, and **model comparison**. At the end, students will provide a comprehensive analysis and recommendations based on their findings.

Datasets Overview

1. **train.csv**
    - Contains historical financial data of customers, including their credit scores (target variable: C_Credit_Score).
    - This dataset will be used for both **training and testing** the models.
2. **vald.csv**
    - Contains similar customer data but **without the target variable**.
    - This dataset will be used for **model validation**, assessing how well the trained models generalise to new, unseen data.

Assignment Tasks and Instructions

1. Environment Setup (Google Colab)

- Ensure that all the required libraries are available in the Colab environment.
- If any libraries are missing, use !pip install commands to install them.
- Import the necessary libraries, including **pandas, numpy, scikit-learn, xgboost, tensorflow, matplotlib, and seaborn**.

Task:
Make sure the environment is properly set up, and all libraries are imported before proceeding with the analysis.

---

2. Data Loading and Overview

- Load the **train.csv** and **vald.csv** datasets.
- Perform an **initial exploration** of both datasets by:
  - Displaying the first few rows.
  - Checking for **missing values** and **data types**.
  - Generating **summary statistics** for numerical features.

Task:
Provide a basic overview of the datasets, identifying any inconsistencies or missing data that need to be addressed.

---

3. Exploratory Data Analysis (EDA)

- Use **visualisations** to explore key patterns in the data:
  - Histograms to analyse distributions (e.g., income, credit history).
  - Scatter plots to understand feature relationships (e.g., income vs. debt).
  - Heatmaps to identify **correlations** between numerical features.
- Identify **outliers** and discuss any potential issues with the data.

Task:
Interpret the key insights from the visualisations. Identify which features might play an important role in predicting credit scores.

---

4. Data Cleaning and Feature Engineering

- Handle **missing values** (e.g., mean imputation or removing records if appropriate).
- Perform **feature engineering** to create new useful variables (e.g., debt-to-income ratio).
- **Encode categorical variables** such as occupation or loan type using One-Hot Encoding.
- **Scale numerical features** to standardise the data using techniques like StandardScaler.

Task:
Provide detailed interpretations of the steps taken to clean the data and the new features created. Explain why these changes are expected to improve model performance.

---

5. Model Building and Testing

- Train the following models on **train.csv**:
    1. **Logistic Regression**
    2. **Decision Trees**
    3. **Random Forests**
    4. **XGBoost**
    5. **Artificial Neural Networks (ANN)**
- **Split train.csv** into **training (80%)** and **testing (20%)** to evaluate model performance.
- Train each model on the training data and test it on the test set.

Task:
Document the results for each model, including key metrics such as **accuracy, precision, recall, F1-score, and AUC-ROC**. Interpret the performance of each model.

---

6. Hyperparameter Tuning and Cross-Validation

- Apply **hyperparameter tuning** using GridSearchCV or RandomSearchCV to optimise the models.
- Use **k-fold cross-validation** (e.g., k=5) to validate model stability and avoid overfitting.
- Apply **regularisation** where applicable (e.g., L1, L2 for logistic regression).

Task:
Interpret the improvements (if any) achieved through tuning and cross-validation. Explain which hyperparameters were most effective for each model.

---

7. Validation using vald.csv

- Use the trained models to predict outcomes for **vald.csv** (which does not contain the target variable).
- This step assesses how well the models generalise to unseen data.
- **Record the predictions** and discuss any differences between the test and validation results.

Task:
Interpret the validation results and discuss how well the models performed on unseen data. Identify any potential issues, such as overfitting or underfitting.

---

8. Model Comparison

- Create a **comparison table** to summarise the performance of all models.
    - Include metrics such as **accuracy, precision, recall, F1-score, AUC-ROC, training time, and validation time**.
- Identify the **best-performing model** and explain why it performed better than the others.

Task:
Interpret the model comparison results. Discuss the strengths and weaknesses of each model and justify the choice of the best model.

---

9. Conclusion and Recommendations

- Summarise the overall findings of the project.
- Highlight which **features were most impactful** in predicting credit scores.
- Provide **recommendations** on how the models could be improved further (e.g., additional tuning, different algorithms).
- Offer suggestions on how customers can **improve their credit scores** based on the analysis (e.g., reducing debt, paying on time).

Task:
Provide a detailed, well-organised conclusion, tying together the key insights and outcomes from the project.

Deliverables

Students must submit the following:

1. **Print of IPython Notebook (.pdf)**
2. **IPython Notebook (.ipynb)**

Data Fields and Descriptions

train.csv (Training and Testing Dataset)

This dataset contains customer financial and behavioural data along with the target variable, **C_Credit_Score**.

vald.csv (Validation Dataset)

This dataset contains the same financial and behavioural fields as **train.csv** but **does not include the target variable** (C_Credit_Score). It is used for **model validation** to assess how well the trained models generalise to new, unseen data.

No.	Field Name	Description
1	**I_ID**	Unique identifier for each transaction or record.
2	**C_Customer_ID**	Unique identifier for each customer, allowing their financial records to be grouped.
3	**M_Month**	Month of the transaction or record, used for time-based analysis.
4	**N_Name**	Full name of the customer (used for reference purposes).
5	**A_Age**	Age of the customer in years, impacting eligibility and financial behaviour.
6	**S_SSN**	Customer's Social Security Number (used for credit reporting and verification).
7	**O_Occupation**	Primary occupation of the customer, indicating earning potential.
8	**A_Annual_Income**	Total yearly income, used to assess creditworthiness and affordability.
9	**M_Monthly_Inhand_Salary**	Monthly take-home salary after tax and deductions.
10	**N_Num_Bank_Accounts**	Total number of bank accounts held by the customer across institutions.
11	**N_Num_Credit_Card**	Total number of credit cards owned by the customer, indicating credit usage.
12	**I_Interest_Rate**	Interest rate applied to loans or credit cards. Higher rates may affect repayment.
13	**N_Num_of_Loan**	Number of active loans (e.g., personal loans, car loans).
14	**T_Type_of_Loan**	Categories of loans taken (e.g., mortgage, personal loan).
15	**D_Delay_from_due_date**	Average delay (in days) between the due date and actual payment date.
16	**N_Num_of_Delayed_Payment**	Total number of instances where payments were delayed.
17	**C_Changed_Credit_Limit**	Adjustments (increase/decrease) made to the credit limit.
18	**N_Num_Credit_Inquiries**	Number of credit inquiries made about the customer, indicating financial activity.

19	**C_Credit_Mix**	Variety of credit types used by the customer (e.g., credit cards, personal loans).
20	**O_Outstanding_Debt**	Total unpaid debt across all credit facilities.
21	**C_Credit_Utilization_Ratio**	Percentage of used credit compared to the total available credit limit.
22	**C_Credit_History_Age**	Length of the customer’s credit history, indicating financial experience.
23	**P_Payment_of_Min_Amount**	Whether the customer pays only the minimum required amount each month.
24	**T_Total_EMI_per_month**	Total amount of Equated Monthly Installments (EMIs) paid across loans.
25	**A_Amount_invested_monthly**	Monthly investment amount in financial products or savings.
26	**P_Payment_Behaviour**	General pattern of the customer’s payment habits (e.g., consistent or irregular).
27	**M_Monthly_Balance**	Remaining account balance at the end of each month.
28	**C_Credit_Score**	Customer’s credit score category (**Good, Poor, Standard**), serving as the target variable.