Credit Risk Scoring Project

Dataset Overview

The dataset comprises 1000 entries with 20 variables:

- 18 categorical variables
- 2 numerical variables
- 2 binary variables indicating whether an applicant is classified as "Good" or "Bad"

Objective

The primary objective of this project is to produce a scorecard that accurately determines the creditworthiness of applicants using the provided dataset. The credit scorecard will help financial institutions evaluate the likelihood of a borrower defaulting on a loan. By leveraging statistical techniques and machine learning models, we aim to create a robust tool that supports informed lending decisions. This analysis is carried out using the R programming language and a suite of R packages specialized for data analysis and predictive modelling.

Analysis Workflow

Splitting the Data-Set

- **Data Cleaning:**
  - Reviewed for missing values.
  - Identified and omitted 12 missing values in the "Purpose" column.
  - Final dataset: 988 entries.
- **Data Subsetting:**
  - Split data into two subsets:
    - Subset1: 532 entries.
    - Subset2: 456 entries.
  - Ensured proper alignment with the remaining rows post-cleaning.

Establishing the Training and Validation Sets

- **Principles:**
  - Split each subset into training and validation sets using the head() and tail() functions.
  - Training set: Top 70% of the rows.
  - Validation set: Bottom 30% of the rows.
- **Splitting Process:**
  - Verified total split values between "good" and "bad" align with original subsets.
- **Training and Validation Sets:**
  - Subset1:
    - Training: 209 "Good" / 163 "Bad".
    - Validation: 88 "Good" / 72 "Bad".

    - Subset2:
        - Training: 278 "Good" / 41 "Bad".
        - Validation: 118 "Good" / 19 "Bad".
- **Importance:**
    - Training sets help build the model.
    - Validation sets test the model's effectiveness on unseen data.

Variable Selection & Binning Processing

- **Variable Selection:**
    - Used Information Value (IV) to determine the predictive power of variables.
    - Selected variables for models based on IV scores:
        - Model 1: Duration, History, Savings, Age.
        - Model 2: Age, Amount, Duration, Purpose.
- **Binning Process:**
    - Applied Weight of Evidence (WoE) binning using woebin() function from the "scorecard" package.
    - Implemented bins into the dataset with woebin_ply().
    - Ensured consistent binning in validation sets.

Scorecard Generation

- **Model Training:**
    - Trained logistic and linear regression models using selected variables.
    - Generated scorecards using the scorecard() function.
- **Scorecard Application:**
    - Applied scorecards to the dataset with scorecard_ply().
    - Example models:
        - Logistic Regression.
        - Linear Regression.

Model Evaluation

- **ROC Curve:**
    - Compares and evaluates True Positive Rates (TPR) against False Positive Rates (FPR).
    - Demonstrates model accuracy through sensitivity (TPR) and specificity (True Negative Rate).
- **Credit Risk Scoring:**
    - Determines effectiveness of a scorecard in identifying “Good” customers.
- **Gini Coefficient:**
    - Measures area between the Lorenz curve and the line of equality.
    - Values range from 0 to 1; higher values indicate better predictive accuracy.
- **Kolmogorov-Smirnov (KS) Statistic:**
    - Takes maximum absolute difference in the CDFs of “Good” and “Bad” outcomes.
    - Higher KS-scores indicate better model predictive capabilities.
    - Suggested ranges:
        - 28-35: Average separation.
        - 35-45: High separation.

        - 45+: Very high-quality application scorecard.
- **Model Application:**
    - Applied scorecard models to validation sets.
    - Used `predict()` to generate predictions on validation sets.
    - Employed `prediction()` function from pROC package for comparison with actual outcomes.
    - Used `performance()` function to measure TPR against FPR and plot ROC curves.
- **Model Performance Comparison:**
    - Model 1 (checking attributes 1 and 2):
        - Linear Regression: Gini = 0.46, KS = 0.33.
        - Logistic Regression: Gini = 0.46, KS = 0.33.
        - Both models demonstrate above average separation and accuracy.
    - Model 2 (checking attributes 3 and 4):
        - Linear Regression: Gini = 0.19, KS = 0.19.
        - Logistic Regression: Gini = 0.21, KS = 0.21.
        - Both models show weak predictive accuracy and reliability.
- **Conclusion:**
    - Model 1 (linear and logistic regression) performs well with high Gini and KS scores.
    - Model 2 shows poor performance and is less reliable for distinguishing between “Good” and “Bad” applicants.

Usage

To run the analysis, use the R scripts provided in the repository. Ensure that the necessary R packages (`scorecard`, `pROC`) are installed.

Installation

1. Clone the repository.
2. Ensure you have R and RStudio installed.
3. Install required packages:

```
R
Copy code
install.packages("scorecard")
install.packages("pROC")
```

4. Run the analysis scripts in order.

Contact

For any questions or issues, please contact [Shaun Commee] at [shauncommee@hotmail.co.uk].