DATA*6300 (01) W24

ANALYSIS of BIG DATA
Final Project Report

Analyzing Global Financial Database & Loan Borrowing Prediction

Submitted by

Namrata Roy (1260835)

Master of Data Science

Department of Mathematics and Statistics

University of Guelph

Introduction

This project conducts an analysis of the Global Financial Inclusion (Global Findex) Database of 2014. The analysis aims to reveal data patterns, trends, correlations, and insights that can assist businesses, organizations, or researchers in making informed financial decisions. At the end of the year, this analysis can help measure a comparative financial scenario of 10 years ago and now. Additionally, it can predict loan borrowing to understand the present socioeconomic condition. This falls under the realm of predictive modeling, a crucial aspect of machine learning (ML) that uses statistical techniques to forecast future outcomes based on historical data. As the push towards data-driven decision-making gains ground across financial industries, predictive modeling becomes pivotal. This project leverages a range of ML algorithms to meet a critical need within the field, enabling more accurate outcome predictions that are foundational to strategic planning and operational efficiency.

Problem Statement

The main goal of this project is to accurately analyze the dataset and predict outcomes across the diversity and complexity of the dataset. Although this challenge is universal, it requires specific adaptations of machine learning algorithms to optimize performance in each specific domain. The project aims to explore the dataset to analyze patterns, trends, correlations, and insights and identify the most effective machine learning algorithms for predicting loan borrowing behaviour. To achieve this, the focus will be on both prediction accuracy and model efficiency.

Method

In this project, the code was written in Python. PySpark was used for exploratory data analysis (EDA) and prediction models along with other necessary libraries like pandas, numpy, matplotlib, seaborn, etc. The methodology consists of two parts, and these are:

I. Dataset and Preprocessing

The dataset used in this project was derived from the World Bank [1], comprising the financial inclusion database of the year 2014. The dataset includes around 86 features and 146688 samples from over 150 countries, with 100 indicators. The database contains a range of indicators that provide valuable information on the ownership of financial institutions and mobile money accounts. It also includes data on how these accounts are utilized for savings and payments, as well as the reasons behind such usage, such as receiving government transfers, wage payments, and agricultural payments. Additionally, the database captures information on how adults send and receive domestic remittances, their savings behaviour, and the different methods they use for savings, such as banks or informal savings clubs. It further records the sources of borrowing, such as banks, friends, or family members, as well as the purposes of borrowing, such as home purchases, school fees, and emergencies.

The data preprocessing involved several steps, including addressing missing values, encoding categorical variables using median values, and normalizing numerical features using mean values. The goal was to prepare the data for exploratory data analysis (EDA) and modeling. Additionally, to enhance the analysis, a few new categorical features were created based on other numerical features. These new features were encoded from the origin website using functions such as 'gender_status', 'education_level_status', 'domestic_remittances_status', and 'universal_status_checking_function', among others.

After the initial preprocessing, the next steps included feature selection based on correlation analysis and initial EDA insights. The aim was to refine the model inputs for optimal performance.

II. Modelling

The project explored six Machine Learning algorithms: Logistic Regression, Decision Tree, Support Vector Classifier, Naive Bayes, Random

Forest, and Gradient Boosting. Each algorithm was chosen based on its suitability for the dataset characteristics and the problem at hand—for example, Decision Trees for their interpretability and Random Forests for handling non-linear data without overfitting. Models were trained using binary classification to ensure generalizability, and performance was evaluated using accuracy, F1-score metrics, mean square error (MSE) and root mean square error (RMSE).

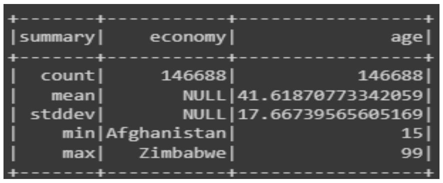
To simplify the target variable for binary classification tasks, the label feature 'borrowed' was modified to '0' and '1'. The categorical and numerical features were then separated into two variables. The 'categoricalColumns' variable represents categories such as economy and regionwb, while the 'numericCols' variable contains numerical values like age and educ. To convert the string values of each categorical column into numerical indices, a 'StringIndexer' is applied, followed by encoding the indices into a one-hot encoded vector using 'OneHotEncoder'. This process is necessary as most machine learning algorithms require numerical input. The resulting one-hot encoded categorical and numerical features are combined into a single feature vector using 'VectorAssembler', which is added to the 'stages' pipeline for preprocessed data. To scale the features and ensure equal contribution to the model's performance, 'StandardScaler' is employed and added to the 'stages'. Lastly, the dataset is split into training and testing sets using 'randomSplit' with a split ratio of 80% for training and 20% for testing, and a seed (seed=42) for reproducibility is chosen.

Results and Discussion

I. Results for Exploratory Data Analysis

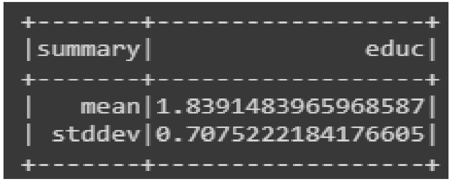
During EDA, significant patterns and insights were identified to inform the subsequent modeling phase, including distribution checks, outlier identification, and correlation analysis. The EDA was done in two parts: one is statistical analysis, and the other is graphical analysis.

1. **Statistical Analysis:** The statistical analysis shows that there were seven regions South Asia, Sub-Saharan Africa, Europe & Central Asia, High Income: OECD, Middle East, East Asia & Pacific, Latin America & Caribbean. There were more female (53.1%) participants than male (46.9%).



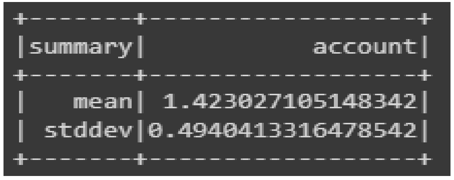
```
+-------+-----------+-----------------+
|summary|    economy|              age|
+-------+-----------+-----------------+
|  count|     146688|           146688|
|   mean|       NULL|41.61870773342059|
| stddev|       NULL|17.66739565605169|
|    min|Afghanistan|               15|
|    max|   Zimbabwe|               99|
+-------+-----------+-----------------+
```

Fig-1: Statistics of country vs age



```
+-------+------------------+
|summary|              educ|
+-------+------------------+
|   mean|1.8391483965968587|
| stddev|0.7075222184176605|
+-------+------------------+
```

Fig-2: Statistics of Education



```
+-------+------------------+
|summary|           account|
+-------+------------------+
|   mean| 1.423027105148342|
| stddev|0.4940413316478542|
+-------+------------------+
```

Fig-3: Statistics of Accounts

Fig-1 shows the statistics of the economy which represents the country vs age where the minimum value of the age is 15 and the maximum is 99. Maximum participants were from Zimbabwe and the minimum were from Afghanistan. Moreover, the mean value of age is 41.62 where the standard deviation is 17.67. On the other hand, Fig-2 and Fig-3 represent the mean and standard deviation of education and account status respectively. The mean value of educated individuals is 1.84 and individuals having an account is 1.42. The standard deviation is 0.71 and 0.49 respectively.

2. **Graphical Analysis:** In this project, five different types of graphical analysis have been performed. There are several types of graphs and charts, including pie charts, histograms, line charts, bar charts, and box

plots. In this discussion, we will highlight some of the most notable ones.

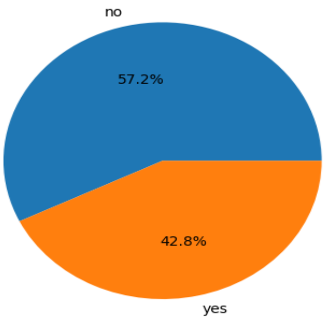


Fig-4: Loan Borrow Ratio

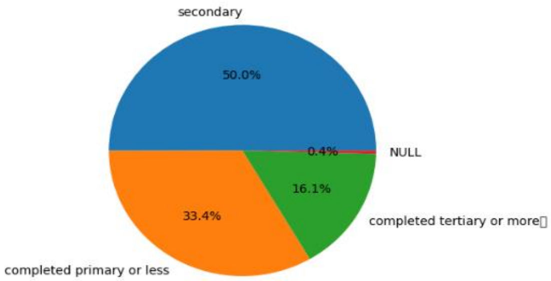


Fig-5: Education Ratio

Fig-4 shows 42.8% of people borrowed loans. Fig-5 depicts that 33.4% of people completed primary school or less and 16.1% just completed tertiary school or more. However, almost half of people completed secondary school.

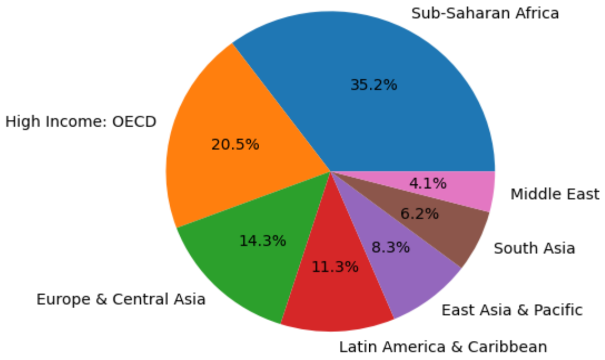


Fig-6: Region Ratio

Fig-6 shows that Sub-Saharan Africa has the highest number of participants with 35.2% and the Middle East has the lowest with 4.1%.

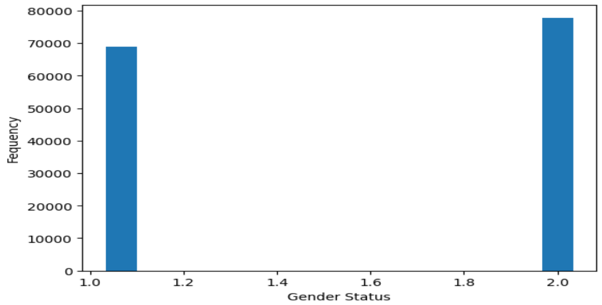


Fig-7: Gender Status

Fig+7 shows a histogram of gender status where male participants were 68866 and females were 77822.

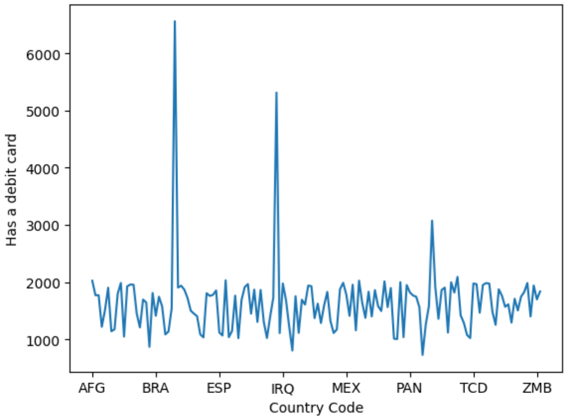


Fig-8: Total of having Debit Card against Country

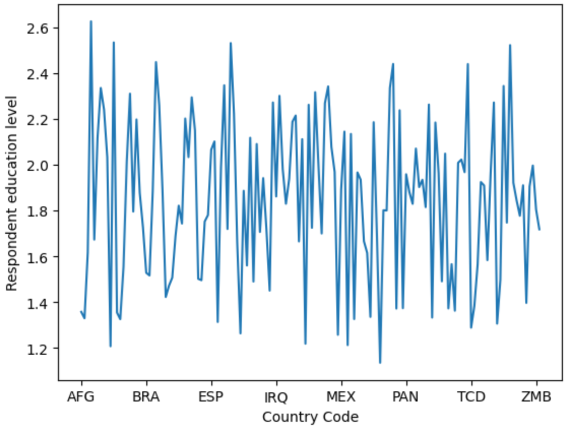


Fig-9: Mean of Education against Country

Fig-8 shows a line chart of the total number of people having debit cards against the countries. Fig-9 is a line chart of the mean of education against the countries.

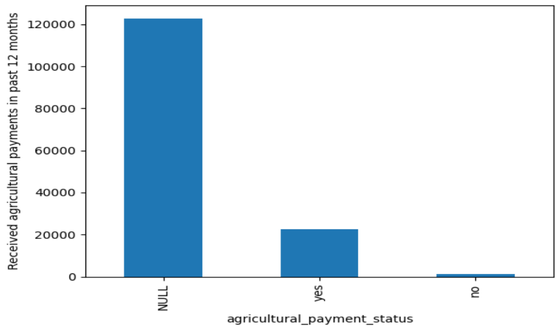


Fig-10: Agricultural Payment

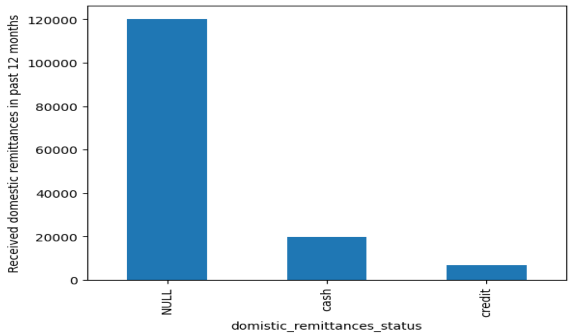


Fig-11: Received Domestic Remittances

Fig-11 shows a bar chart showing only 22615 individuals received agricultural payments, 1250 didn't receive them, and 122823 didn't respond. Fig-11 represents the domestic remittances received status, where 19712 participants received in cash, 6709 in credit and others did not respond.

II. Machine Learning Results

The results from the ML models showed varying degrees of accuracy and efficiency.

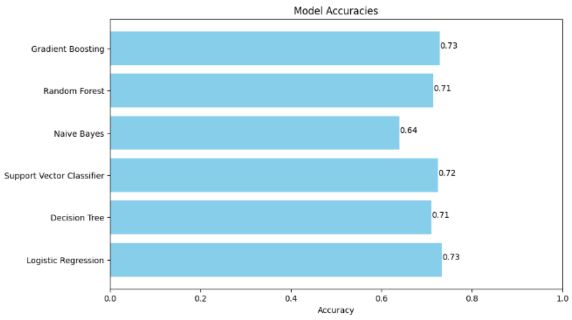


Fig-12: Accuracy Comparison

The graph Fig-12 is a horizontal bar chart titled "Model Accuracies", which displays the accuracy of six different machine learning models. Logistic Regression and Gradient Boosting appear to be the most accurate models with scores of 0.73. The Naive Bayes model has the lowest accuracy at 0.64. The other models (Decision Tree, Support Vector Classifier, and Random Forest) have similar accuracies, ranging from 0.71 to 0.72.

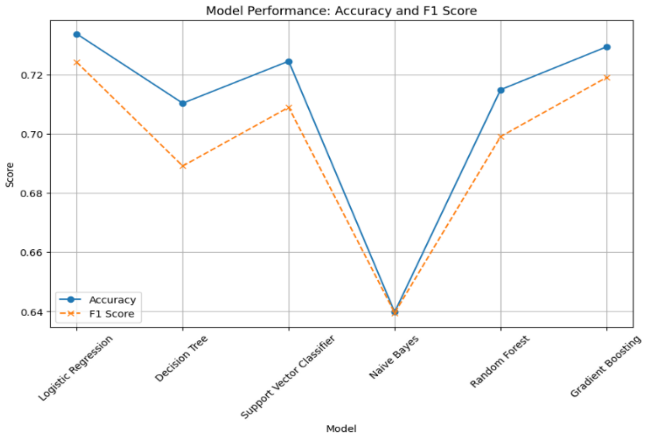


Fig-13: Model Performance: Accuracy and F1 Score

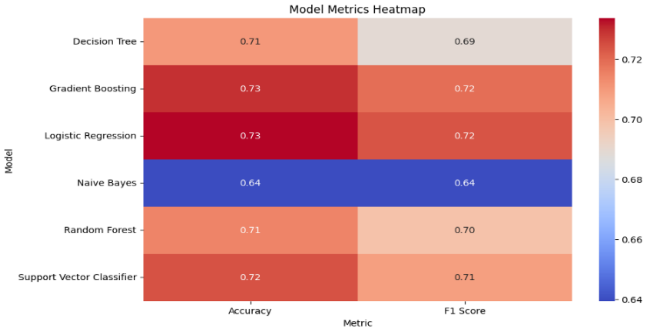


Fig-14: Model Metrics Heatmap

In Fig-13 each model is listed on the Y-axis, while the metrics, Accuracy and F1 Score, are on the X-axis. Fig-14 is the heatmap that visualizes the accuracy and F1 score for six different machine learning models. The heatmap uses a color gradient to represent the value of each metric, with red shades indicating higher values and blue shades indicating lower values. For each model-metric combination, the exact numerical score is provided within the corresponding cell. The Gradient Boosting model has the highest accuracy and F1 score among all the models, whereas the Naive Bayes model scores the lowest on both metrics.

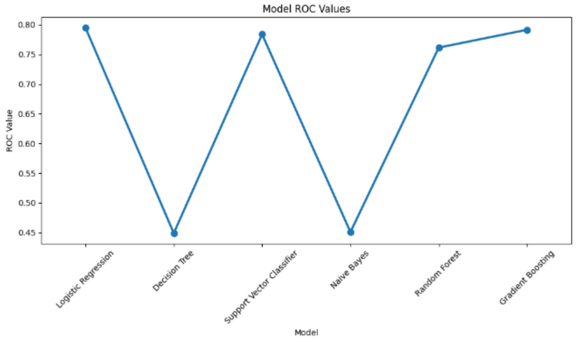


Fig-15: ROC of Six Models

Displayed on Fig-15 is a line graph entitled "Model ROC Values," depicting the Receiver Operating Characteristic (ROC) values of various machine learning models. The X-axis enumerates the models used: Logistic Regression, Decision Tree, Support Vector Classifier, Naive Bayes, Random Forest, and Gradient Boosting. Meanwhile, the Y-axis denotes the ROC value, which spans from 0.5 to approximately 0.8.

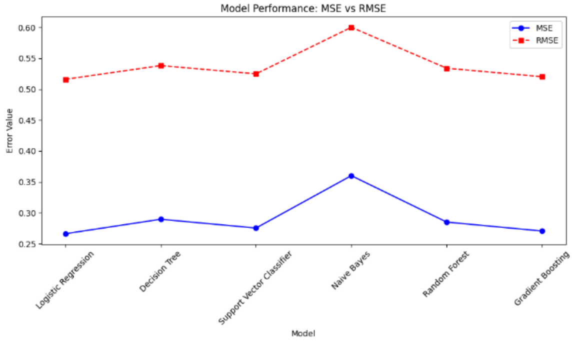


Fig-16: Model Performance: MSE vs RMSE

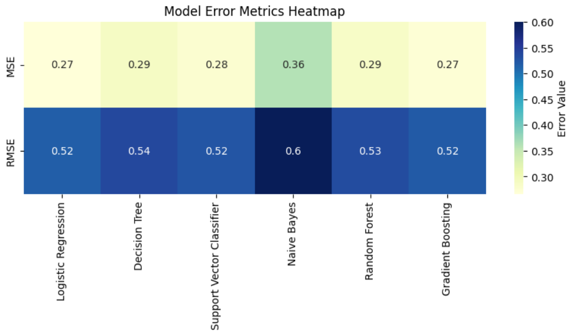


Fig-17: Model Error Metrics Heatmap

Fig-16 and Fig-17 are the line chart and heatmap of model performance based on Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) across different machine learning models. Based on the error metrics, Logistic Regression and Gradient Boosting seem to perform the best among the tested models, as those display the lowest MSE and RMSE. Meanwhile, Naive Bayes exhibits the highest MSE and RMSE, indicating that it may not perform as well as the other models. The remaining models, including Decision Tree, Support Vector Classifier, Random Forest, show moderate values for these metrics, with MSE ranging between approximately 0.28 to 0.36 and RMSE ranging from about 0.52 to 0.6.

Analyzing all the accuracy, f1 score, mse and rmse score Logistic Regression and Gradient Boosting performance is the best than other models.

Conclusion

This project aimed to analyze a dataset and understand the data patterns, as well as predict outcomes using various machine learning (ML) algorithms. The process included thorough data preprocessing, exploratory analysis, and model evaluation, demonstrating the potential of predictive modeling in solving complex problems. The study highlights the importance of selecting the appropriate ML algorithms, such as Logistic Regression and Gradient Boosting, based on the dataset and specific problem, providing valuable insights in the field of predictive analytics.

Reference

[1] https://microdata.worldbank.org/index.php/catalog/2512/get-microdata