**ADMN5016 Assignment: Proof of Concept for Machine Learning Application**  
  
Venkata Pravalika, Chinthakunta

Venkata Pradeep, Gutta

School of business | St. Lawrence college

ADMN5016: Applied Artificial Intelligence and Machine Learning

Prof. Joseph Santarcangelo

April 17, 2023

**About Dataset:**

The current data set includes details of the 500 people who have opted for loan. Also, the data mentions whether the person has paid back the loan or not and if paid, in how many days they have paid. In this project, we will try to draw few insights on sample Loan data.

Please find the details of dataset below which can help to understand the features in it.

Loan\_id: A unique loan (ID) assigned to each loan customers- system generated

Loan\_status: Tell us if a loan is paid off, in collection process - customer is yet to payoff, or paid off after the collection efforts

Principal: Principal loan amount at the case origination OR Amount of Loan Applied

terms: Schedule (time period to repay)

Effective\_date: When the loan got originated (started)

Due\_date: Due date by which loan should be paid off

Paidoff\_time: Actual time when loan was paid off, null means yet to be paid

Past\_due\_days: How many days a loan has past due date

Age: Age of customer

Education: Education level of customer applied for loan

Gender: Customer Gender (Male/Female)

**Project Proposal:**

**Project Title:** Loan Data Analysis and Prediction

**Project Overview:**

This project aims to analyze loan data and build a model to predict whether a borrower is likely to default on their loan or not. The loan data has features like loan ID, loan status, principal, terms, effective date, due date, paid-off time, past due days, age, education, and gender. The data can be obtained from Kaggle.

**Project Tasks:**

1. Data Collection: Collecting loan data from Kaggle.
2. Data Cleaning: Cleaning the data by removing missing values, duplicates, and outliers.
3. Exploratory Data Analysis: Analyzing the loan data to identify patterns and trends, such as loan amount, loan duration, interest rates, and loan purpose.
4. Visualization: Creating visualizations using libraries such as matplotlib and seaborn to present the findings of the exploratory data analysis.
5. Feature Engineering: Creating new features from the existing data, such as debt-to-income ratio, loan-to-value ratio, and payment-to-income ratio.
6. Data Preprocessing: Preprocessing the data by scaling, encoding categorical variables, and splitting the data into training and testing sets.
7. Building a Model: Building a predictive model to determine whether a borrower is likely to default on their loan or not using machine learning algorithms such as logistic regression, decision tree, random forest, or XGBoost.
8. Model Evaluation: Evaluating the performance of the model using metrics such as accuracy, precision, recall, and F1-score.

**Tools and Technologies:**

Python, pandas, numpy, matplotlib, seaborn, scikit-learn, Flask.

**Expected Outcomes:**

By the end of this project, you will have a better understanding of loan data analysis, data cleaning, exploratory data analysis, feature engineering, and machine learning algorithms. You will also have a working predictive model to determine whether a borrower is likely to default on their loan or not based on the given features.

**Problem application solves:**

The loan data analysis and prediction project aim to solve the problem of predicting loan defaults by building a predictive model using machine learning algorithms. Loan defaults can cause financial losses to lenders and borrowers, and it can also affect the credit score of the borrower.

By predicting loan defaults in advance, lenders can take preventive measures to minimize the losses and avoid the potential financial risk.

Moreover, this project can also help lenders to identify the factors that are most correlated with loan defaults, such as credit score, age, education, and loan duration. By analyzing these factors, lenders can develop loan approval criteria that can reduce the risk of loan defaults in the future. This project can also benefit borrowers by providing them with insights into the factors that affect loan approvals and by helping them to make informed decisions when applying for loans.

Overall, this project can help both lenders and borrowers to mitigate the financial risks associated with loan defaults and improve the loan approval process.

**Performance of Logistic Regression model:**

The classification report for the logistic regression model on the validation data shows that the model's overall accuracy is 0.65, meaning that it correctly predicts the loan status for 65% of the validation data.

The precision, recall, and F1-score for each loan status class are as follows:

* For loan status 1 (paid-off), the precision is 0.53, recall is 0.51, and F1-score is 0.52. This indicates that the model correctly predicted 53% of the loan status 1 instances out of all instances predicted as loan status 1 (precision), and it correctly identified 51% of all loan status 1 instances in the validation data (recall). The F1-score is the harmonic mean of precision and recall, and it represents a balance between the two metrics.
* For loan status 2 (Collection), the precision is 0.73, recall is 0.65, and F1-score is 0.69. This indicates that the model correctly predicted 73% of the loan status 2 instances out of all instances predicted as loan status 2, and it correctly identified 65% of all loan status 2 instances in the validation data.
* For loan status 3 (Collection\_paidoff), the precision is 0.68, recall is 0.78, and F1-score is 0.73. This indicates that the model correctly predicted 68% of the loan status 3 instances out of all instances predicted as loan status 3, and it correctly identified 78% of all loan status 3 instances in the validation data.
* The macro-average F1-score, which is the unweighted average of F1-scores across all classes, is 0.65. The weighted-average F1-score, which takes into account class imbalance in the validation data, is also 0.65. The macro-average and weighted-average precision and recall are also equal to the overall accuracy since the classes are balanced in this case.

Overall, the classification report indicates that the logistic regression model has **moderate performance** in predicting the loan status on the validation data. There may be room for improvement by using more advanced algorithms or optimizing the model's hyperparameters.

**Performance of Decision Tree model:**

This classification report provides evaluation metrics for a decision tree model on the validation data. The model was trained to classify instances into three classes (1, 2, and 3).

The precision, recall, and F1-score are reported for each class, as well as the overall macro and weighted averages.

* The precision for class 1 is 0.57, which means that out of all instances predicted as class 1, 57% were actually class 1. The recall for class 1 is 0.60, which means that out of all actual class 1 instances, 60% were correctly classified as class 1. The F1-score for class 1 is 0.58, which is the harmonic mean of precision and recall.
* For class 2, the precision is 0.78, recall is 0.64, and F1-score is 0.71. This means that the model is better at identifying instances of class 2 than class 1 or class 3, and has higher precision than recall for class 2.
* For class 3, the precision is 0.70, recall is 0.78, and F1-score is 0.74. This means that the model is also accurate at identifying instances of class 3, and has higher recall than precision for class 3.
* The overall accuracy of the model on the validation data is 0.68, which means that 68% of the instances were correctly classified by the model.
* The macro average of precision, recall, and F1-score takes the average of these metrics across all classes, giving equal weight to each class. The macro average precision is 0.68, recall is 0.68, and F1-score is 0.68.

The weighted average of precision, recall, and F1-score takes the average of these metrics across all classes, but weights each class by its support (the number of instances in that class in the validation data). The weighted average precision is 0.68, recall is 0.68, and F1-score is 0.68. This shows that the model is performing **relatively well across all classes**, and is not biased towards any particular class.

**Interpreting the confusion matrix**

[51 13 19]

The first row of the matrix shows the performance for the first class. Out of 83 instances of this class, the classifier correctly predicted 51 instances as belonging to this class, but it incorrectly predicted 13 instances as belonging to the second class, and 19 instances as belonging to the third class.

[24 54 6]

The second row shows the performance for the second class. Out of 84 instances of this class, the classifier correctly predicted 54 instances as belonging to this class, but it incorrectly predicted 24 instances as belonging to the first class, and 6 instances as belonging to the third class.

[14 5 64]

The third row shows the performance for the third class. Out of 83 instances of this class, the classifier correctly predicted 64 instances as belonging to this class, but it incorrectly predicted 14 instances as belonging to the first class, and 5 instances as belonging to the second class.

​