California State Polytechnic University, Pomona

Data Mining
CIS 4321
Spring 2024

Exploratory Data Analysis

Name:
Samy Benjelloun

Submitted to:
Dr. Fadi Batarseh

May 2024

Introduction

Financial inclusivity and risk management are more critical than ever, and understanding the dynamics of credit risk become important for banks and consumers. This project stems from a deep interest in technology and finance, more specifically how data-driven insights can improve decision-making in credit lending. This study is driven by the central question: "How can machine learning models predict credit risk?" By addressing these kinds of questions, the project aims to lower default rates and increase consumers' financial literacy. The insights gained from this analysis are valuable, particularly for stakeholders such as banks, credit unions, and companies that are seeking improvements in their processes. It also benefits consumers as it enhances the risk prediction models which can lead to fairer, and more accessible credit. The importance of this research goes beyond the financial sector. It also touches on economic stability and personal financial health. By identifying key predictors of credit risk and understanding the performance of various predictive models, stakeholders can implement more informed strategies that protect their interests and promote greater economic participation among consumers.

Problem Statement

Credit lending lies in the accurate assessment of risk. Financial institutions face the challenging risk of predicting which borrowers are likely to default on the loan they lend. An inaccurate prediction could lead to financial loss for the lenders or deny potentially reliable borrowers access to credit. The analytical question of this research is: "How can machine learning models predict credit risk?" This question is important as it directly influences the lenders and the borrowers. A well-constructed model could support a lender's decision-making

process using data analytics, and enabling them to issue loans with a better understanding of risk. On the other hand, for consumers, this could mean an increase in the chances of loan approval and better terms of credit.

The project seeks to explore many data analytics approaches to address this question. Since the outcome of interest is loan repayment or defaulting, which is categorical, a good technique would be a classification type of data analytics within supervised learning. These techniques include but are not limited to:

- **k-Nearest Neighbors (KNN)** algorithm is a very popular model for its simplicity and effectiveness. KNN has its foundation in the principle of proximity and similarity. It is achieved through the calculation of distances between data points, which is important in determining the nearest neighbors. It creates neighborhoods in a graph that separates graph points of people who defaulted and those who didn't.
- **Decision Tree Classier** is a method that helps identify the most significant variables affecting credit risk through an interpretable model structure. Just like KNN, Decision Trees can be used for classification and regression. The foundational concept behind decision trees is to split the data based on certain conditions, leading to an intuitive decision-making process. It is known for its simplicity, visualization, feature importance, and non-parametrics.
- **Logistic Regression** despite its name, is applied to classification tasks where it predicts discrete value. As a statistical model that estimates the probability of a binary outcome, it is particularly suited for discovering relationships between features just like the loan repayment probability. In other words, this technique works by performing linear regression on continuous features to predict a discrete feature.

- **Neural Networks (NN) (MLPClassifier)** are a foundational concept in data mining and artificial intelligence, providing powerful tools for analyzing large datasets and recognizing patterns that are not immediately apparent to humans. NN is inspired by the human brain's architecture. It is composed of an input layer, hidden layers, an output layer, and weights and activation functions.

Other models to consider include; GaussianNB, BernoulliNB, and MultinomialNB. By leveraging these models, this project aims to answer the primary question and compare the effectiveness of different algorithms in predicting credit risk. The goal is to discover which of these models will predict the creditworthy borrowers most effectively, enabling more informed lending decisions.

Understanding the Dataset

The dataset used to perform this project stems from a bank in Germany called Deutsche Bank. The CSV file includes various attributes related to applicants' credit information. The dataset contains 1,000 instances, each an individual loan applicant.

- **Numerical Attributes:**
    - **Age (int)**: Age of the applicant.
    - **Job (int)**: A numerical code representing different types of employment. 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled
    - **Credit Amount (int)**: The total credit amount applied for.
    - **Duration (int):** The term of the loan in months.
- **Categorical Attributes:**
    - **Sex (binary):** male, female

- **Housing (binary: own, rent, or free):** The applicant's housing status.
- **Saving Accounts (ordinal: little, moderate, quite rich, rich)**: Reflects the applicant's savings level.
- **Checking Account (ordinal: little, moderate, rich):** Indicates the balance levels in the applicant's checking account.
- **Purpose (nominal: car, furniture/equipment, radio/TV, etc.)**: The reason for the loan application.

- **Target Variable**
  - **Risk (Binary)**: good - not defaulting, bad - defaulting. The target variable of interest here is Risk, which indicates whether the loan was repaid (good) or defaulted (bad).
- **Challenge and Data Quality issues:**
  - **Missing Values:** 'Saving Accounts' and 'Checking Account' attributes had a lot of missing values, which could skew the model's training and predictions if not addressed properly.

saving_accounts	183 (18.3% of the dataset)
checking_account	394 (39.4% of the dataset)

  - **Imbalance Data**: preliminary analysis indicates an imbalance in the target variable with more 'good' risk outcomes than 'bad.' this imbalance can lead to biased predictions favoring the majority class.

good	700 (70% of the dataset)
bad	300 (30% of the dataset)

- **Encoding Categorical Variables:** Most machine learning models require numerical input, which necessitates the encoding of categorical variables into a numerical format to effectively train the data.

To fix these issues, the following steps were taken:

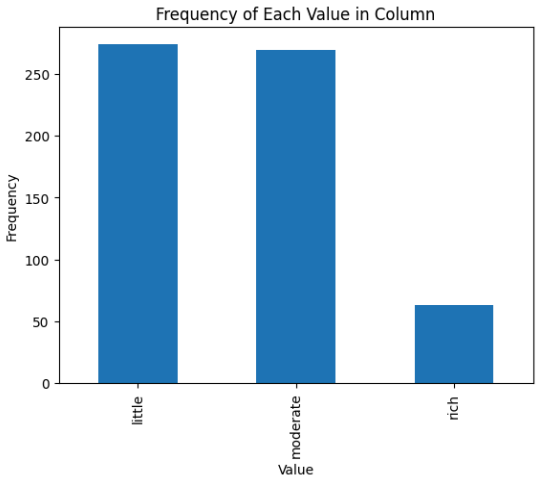
- **Handling Missing Values:** Strategies such as imputation for missing entries were taken for both accounts (saving and checking). Used the mode to fill in the missing values.

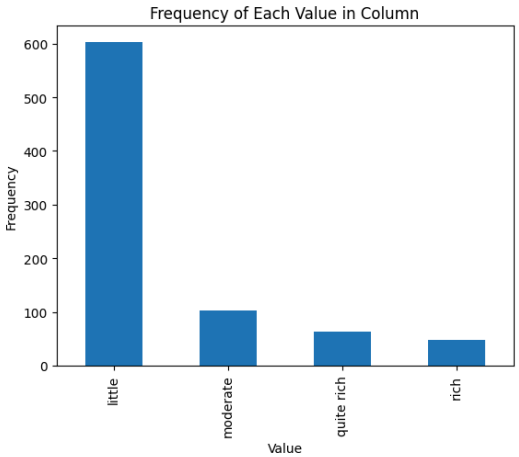
checking_account:

little	274
moderate	269
rich	63

saving_accounts:

little	603
moderate	103
quite rich	63
rich	48





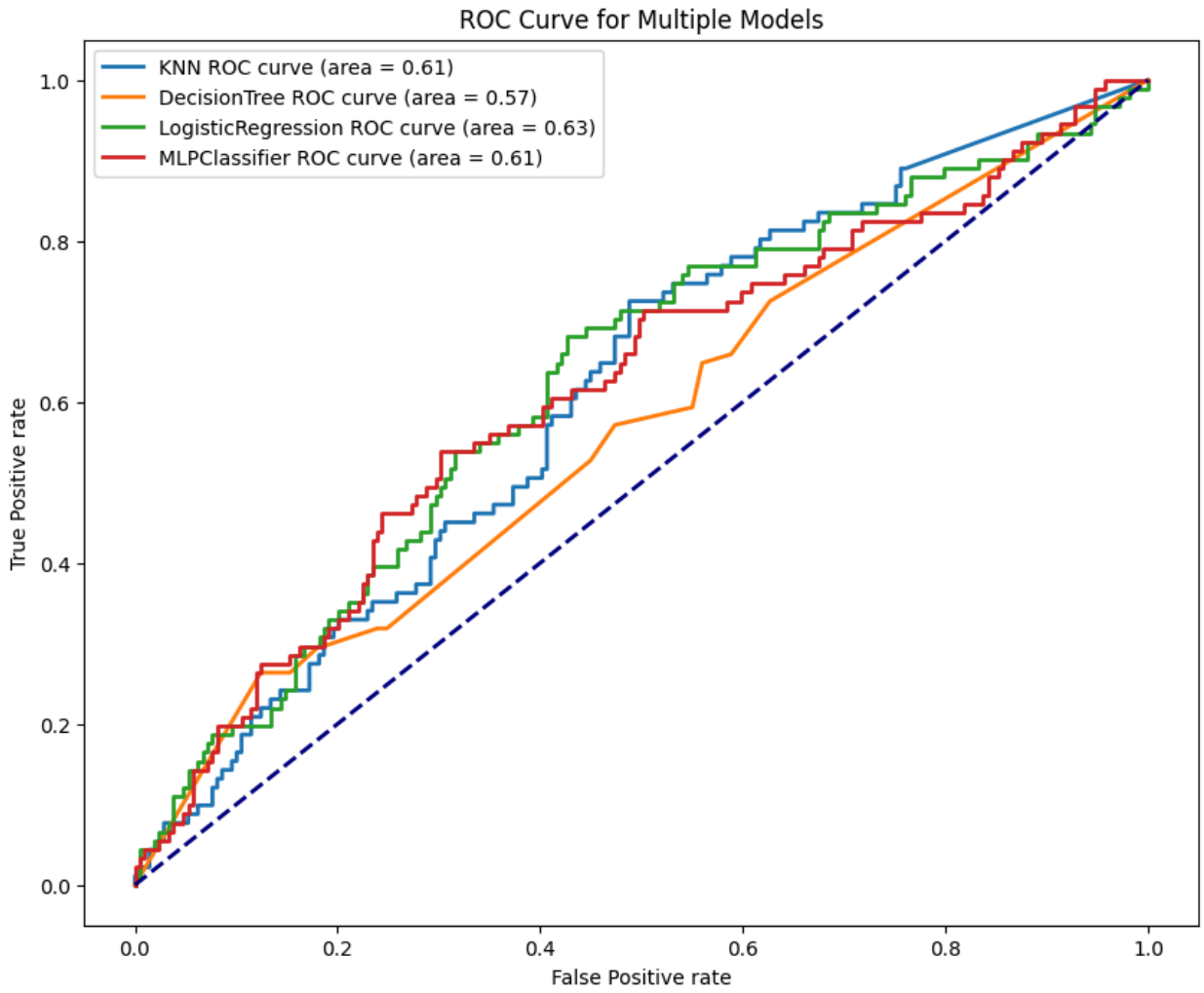
- **Addressing imbalance:** trained 70% of the dataset and tested on 30% of it
- **Encoding Technique:** Utilized One Hot Encoder from sklearn preprocessing on categorical columns to utilize all the variables for training and predicting risk.

Understanding the dataset helps with data preprocessing and informs the selection of modeling techniques that are best suited to handle this task.

Data Analytics & Results

After transforming the data which involved making all the variables numerical, inputting the missing values, and making it in a similar format, the next step was to process the data. This means setting up the feature matrix labeled as X, and the target vector labeled as y. The feature matrix (X) contains all the variables that will be used to predict the target vector (y) which is the 'risk.' The next step was to split the data between the training (70%) set and the testing set (30%). To reduce noise and ensure data integrity, Standard Scaler was used which makes the mean equal to 0 and standard deviation to 1. Finally, the data was ready to be trained, and the model built. The final algorithms that were deployed are KNN, Decision Tree, Logistic Regression, and MLP Classifier. Finally, Grid Search CV was used on X_train, and y_train to discover the best parameters to use for this dataset. These are the best parameters per model according to Grid Search CV for each model:

([('KNN', **KNeighborsClassifier**(n_neighbors=7, weights='distance')), ('DecisionTree', **DecisionTreeClassifier**(criterion='entropy', max_depth=10)), ('LogisticRegression', **LogisticRegression**(C=0.1, solver='liblinear')), ('MLPClassifier', **MLPClassifier**(activation='tanh', hidden_layer_sizes=(10, 30, 10)))])



The Receiver Operating Characteristic (ROC) is a graphical representation used to assess the performance of classification models at various threshold settings. In short, the ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR). The diagonal dashed line represents a no-skill classifier that would be correct 50% of the time. The goal here is to have a ROC curve as far as possible from 0.5 which is essentially random guessing.

- **Lines on the Graph:**
    - **KNN (blue):** The K-Nearest Neighbors model has an area under the curve of 0.61. This suggests it performs slightly better than random guessing of 0.5.

- **Decision Tree (green):** This model has an AUC of 0.57, which is very close to random guessing. This indicates the model is not performing well between classes.
- **Logistic Regression (orange):** The logistic Regression model shows an AUC of 0.63 which is the highest among the models.
- **MLP Classifier (red):** Similarly to KNN, MLP Classifier has an AUC of 0.61 which is pretty close to random guessing.

- **Overall Performance:** Unfortunately all the models had an AUC value fairly close to 0.5 indicating that there is a lot of noise in the data. Although the Logistic Regression model performed the best, an AUC of 0.63 is weak, and there is a lot of room for improvement.

Results:

	Accuracy	**Precision**	**Recall**	**F1-Score**
KNN	0.68	0.44	0.19	0.26
Decision Tree	0.67	0.43	0.26	0.33
Logistic Regr.	0.69	0.46	0.20	0.28
MLP Classifier	0.67	0.44	0.29	0.35

The classification report is a summary of the performance of a classification model. It provides important metrics that help evaluate how well the model is performing in terms of classifying different categories or classes. The report includes the following metrics:

- **Accuracy:** This metric is the overall percentage of correctly classified instances. In other words, it measures the model's overall performance.

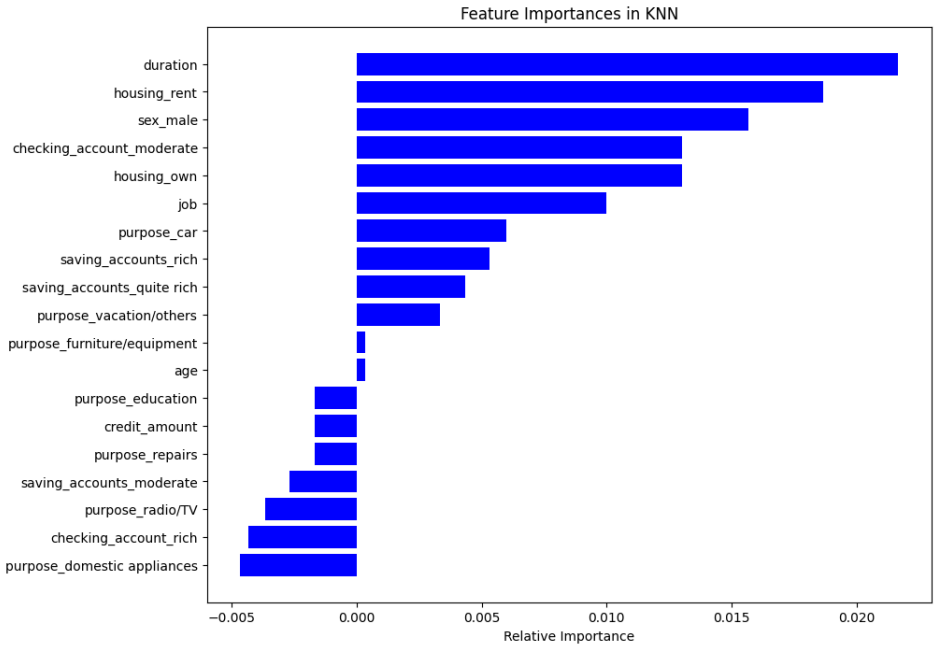
    - Although the models demonstrated similar accuracy levels, the Logistic Regression model has the highest accuracy with 69% of values correctly classified.

- **Precision:** Precision is the ratio of true positive predictions to the total number of positive predictions. It measures the accuracy of positive predictions and tells how many of the predicted positive instances were positive.
    - Again, the model that scored the highest Precision is Logistic Regression with 46%. This means that out of all the predicted positive cases class 46% of them are true positive.
- **Recall:** recall also known as sensitivity is the ratio of true positive to the total number of actual positive instances. It measures the model's ability to correctly identify all positive instances.
    - MLP Classifier correctly identified 29% of all actual positive cases which was the highest among the models used for this project.
- **F1-Score:** The F1-Score is the mean of precision and recall. It provides a balanced measure of a model's performance that considers both false positives and false negatives.
    - Again, the MLP Classifier has the highest F1-Score with 35%.

The classification report is a valuable tool for assessing the model's performance, especially in cases with imbalanced datasets. It helps understand how well a model is distinguishing between classes and where it may need improvement.

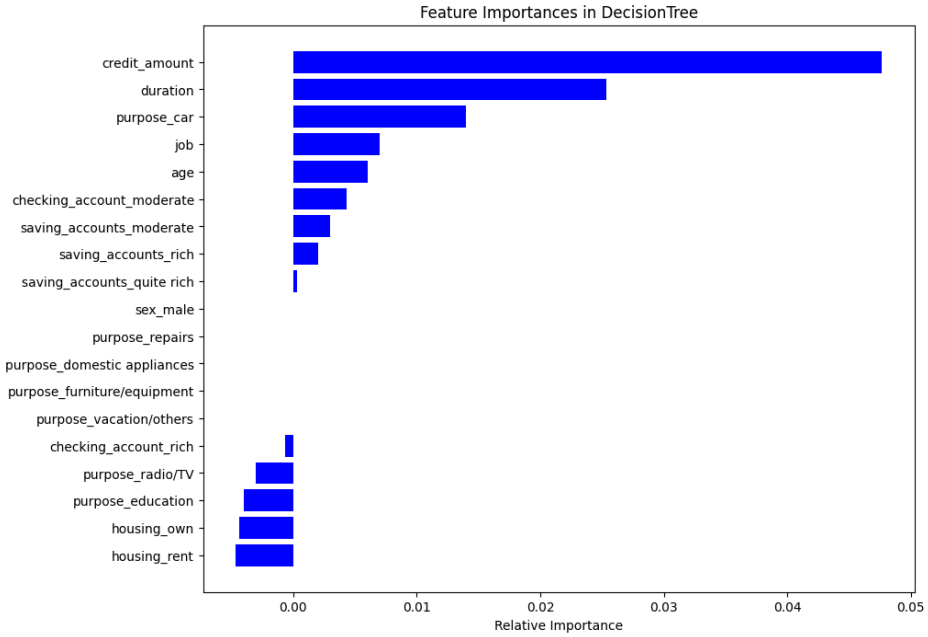
The feature importance charts show how each model weighs the impact of various features on its predictions.

K-Nearest Neighbors (KNN):



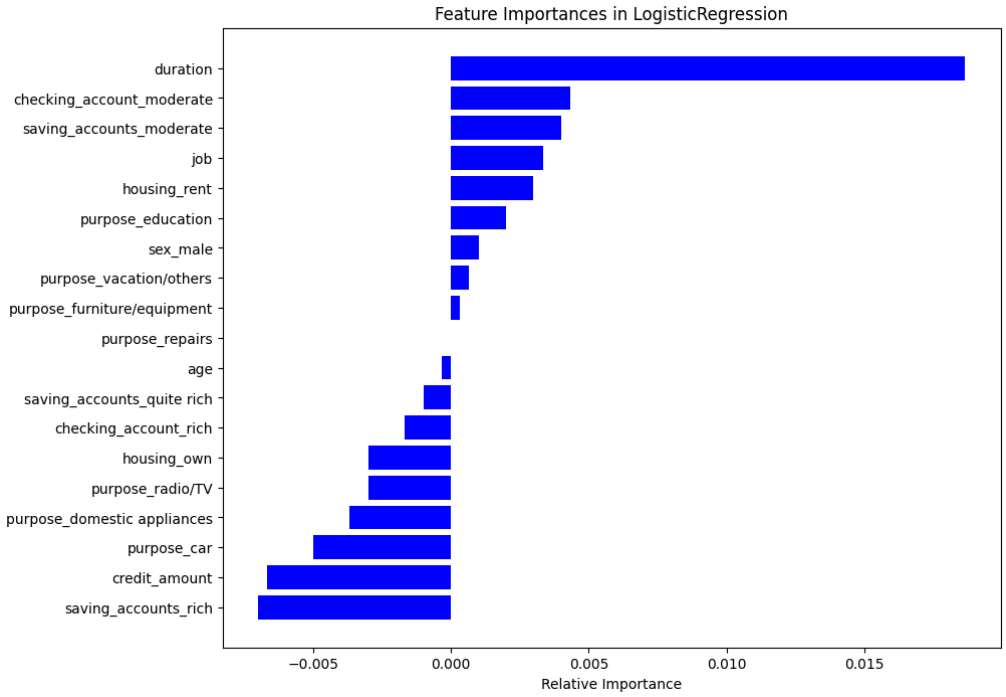
As shown in the chart above, **duration**, and **housing_rent** are the most significant features for the KNN algorithm. This suggests that factors related to the term of the loan and living situation are critical for this model.

Decision Tree:



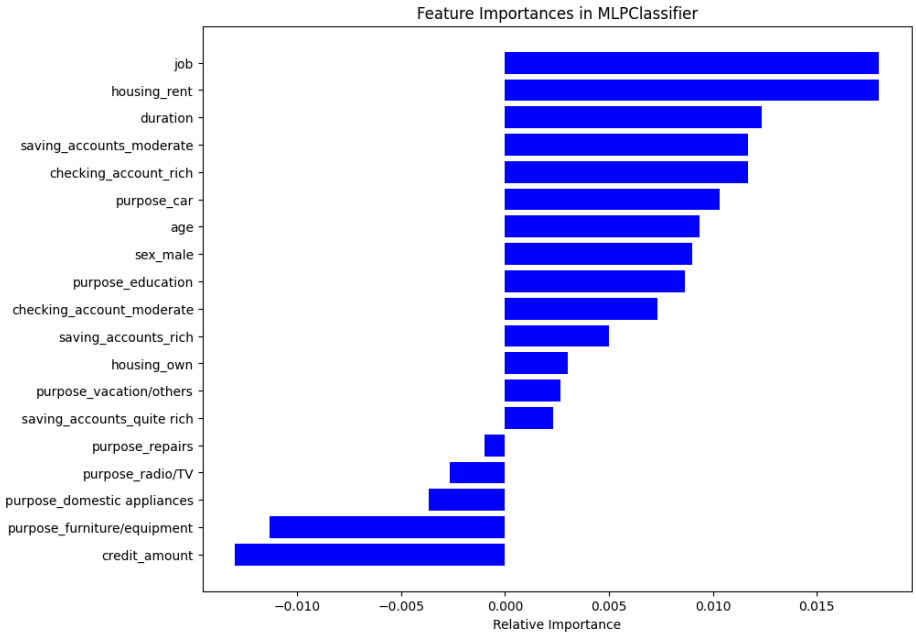
For this classifier, **credit_amount** stands out the most, likely because it is a crucial determination of credit risk. **Duration** and **purpose_car** seem to be important and suggest that car loans play a significant role in the tree's decision.

Logistic Regression:



Duration comes up again as highly influential, along with **checking_account_moderate**. Logistic Regression seems to weigh these two features heavily, which could mean they have a strong linear relationship with the outcome. On the other hand, **saving_accounts_moderate** and **job** appear to be not as important, which might be linearly separable features in the context of logistic regression.

MLP Classifier:



Job and **housing_rent** are the top two suggesting these features have complex interaction or nonlinear relationships with the outcome that the MLP can capture. With no surprise, **duration** remains consistently significant, and **saving_accounts_moderate** and **checking_account_rich** are also key factors according to the MLP model.

Conclusion

To conclude, this project embarked on an exploration to answer the question: “How can machine learning models predict credit risk?” Through the application of various machine learning techniques and data analytics processes, performance and insights were captured to answer this question. All the models performed similarly however, Logistic Regression stood out the most. Moreover, there is a lot of room for improvement, and with deeper training, we could increase the accuracy of the models. For example, creating multiple notebooks for each model and giving it more time to increase its ROC and accuracy, based on feature performance.

Reference: https://www.kaggle.com/code/garygli/predicting-credit-risk/notebook