**Data Science**

**Project Report**

**Credit Risk Assessment**

**Ali Aamir 19L-0926**

**Ammar Ahmed 19L-2271**

**Syed Ali Raza 19L-1015**

Problem Statement:  
Credit risk is the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations. Traditionally, it refers to the risk that a lender may not receive the principal and interest owed, which results in an interruption of cash flows and increased costs for collection. Correctly assessing credit risk can be a huge win to allocate finances for loan with minimum risk.

# Data Set Source:

The data set was taken from Kaggle.

# Description of the data set.

The data set contains variables relevant to predicting the credit risk of a given user. All the variables used directly or indirectly co-relate to deciding the overall probability of assessing the credit risk. The variables used in the data set are following:

|  |  |
| --- | --- |
| person\_age | Age |
| person\_income | Annual Income |
| person\_home\_ownership | Home ownership |
| person\_emp\_length | Employment length (in years) |
| loan\_intent | Loan intent |
| loan\_grade | Loan grade |
| loan\_amnt | Loan amount |
| loan\_int\_rate | Interest rate |
| loan\_status | Loan status (0 is non default 1 is default) |
| loan\_percent\_income | Percent income |
| cb\_person\_default\_on\_file | Historical default |
| cb\_preson\_cred\_hist\_length | Credit history length |

# Description of the variables used in the data set:

1. **person\_age**: Age of the borrower, which can be a factor in assessing credit risk based on age-related factors like stability, income level, and debt-to-income ratio.
2. **person\_income**: Annual income of the borrower, which is a key factor in determining their ability to repay a loan and their overall creditworthiness.
3. **person\_home\_ownership**: Indicates whether the borrower owns their home or not, which can provide insight into their financial stability and ability to manage debt.
4. **person\_emp\_length:** Employment length of the borrower, which can be used to assess their job stability and income potential in the future.
5. **loan\_intent:** The purpose of the loan, which can be used to evaluate the borrower's financial goals and their ability to repay the loan based on the intended use of the funds.
6. **loan\_grade:** A rating assigned to the loan based on the borrower's creditworthiness, which is based on their credit history, income level, and other factors.
7. **loan\_amnt:** The amount of the loan requested by the borrower, which can be used to assess their ability to repay the loan based on their income level and debt-to-income ratio.
8. **loan\_int\_rate:** The interest rate charged on the loan, which is based on the borrower's creditworthiness and the perceived risk of default.
9. **loan\_status:** Indicates whether the borrower has defaulted on the loan in the past, which can be used to assess the likelihood of future default.
10. **loan\_percent\_income:** The percentage of the borrower's income that is required to make loan payments, which can be used to assess their ability to repay the loan without defaulting.
11. **cb\_person\_default\_on\_file:** Indicates whether the borrower has a history of default on their credit report, which can be used to assess their creditworthiness and the likelihood of future default.
12. **cb\_preson\_cred\_hist\_length:** The length of the borrower's credit history, which can be used to assess their creditworthiness and the likelihood of future default based on their history of managing debt.

# Univariate Analysis Findings:

Here are the findings from the univariate analysis the variables:

**person\_age:**

The age of the borrower’s ranges from 20 to 144.The majority of the borrowers are between the ages of 20 and 40, with a peak at around 25. There are some outliers with ages above 100, which may be erroneous data points.

**person\_income:**

The income of the borrower’s ranges from 4000 to 750000. The distribution of income is positively skewed, with most borrowers earning less than 100,000. There are some outliers with extremely high incomes, which may be erroneous data points.

**person\_home\_ownership:**

There are three types of home ownership: rent, own, and mortgage. Most of the borrowers are renting their homes, followed by those who own their homes. There are fewer borrowers who have a mortgage.

**person\_emp\_length:**

The length of employment ranges from 0 to 123 years. Most borrowers have been employed for less than 10 years. There are some outliers with employment lengths above 100 years, which may be erroneous data points.

**loan\_intent:**

There are six types of loan intents: debt consolidation, credit card refinancing, home improvement, major purchase, business, and other. Most borrowers are taking out loans for debt consolidation, followed by credit card refinancing.

**loan\_grade:**

There are seven loan grades: A, B, C, D, E, F, and G. Most loans are in the C grade. There are fewer loans in the A and G grades.

**loan\_amnt:**

The loan amounts range from 500 to 35000. Most loans are for amounts between 5000 and 15000. There are some outliers with extremely high loan amounts, which may be erroneous data points.

**loan\_int\_rate:**

The interest rates on loans range from 5.42% to 23.22%. Most loans have interest rates between 10% and 15%. There are some outliers with extremely high interest rates, which may be erroneous data points.

**loan\_status:**

There are two loan statuses: 0 for paid off loans and 1 for defaulted loans. Most loans have been paid off. There are fewer loans that have defaulted.

**loan\_percent\_income:**

The loan percentage of income ranges from 0.0015 to 0.8305. The distribution of loan percentage of income is positively skewed, with most borrowers having a loan percentage of income less than 0.3. There are some outliers with loan percentages of income above 0.7, which may be erroneous data points.

**cb\_person\_default\_on\_file:**

There are two types of default on file: Y for yes and N for no. Most borrowers do not have a default on file. There are fewer borrowers who have a default on file.

# Bivariate Analysis Report

The bivariate analysis is a type of statistical analysis that investigates the relationship between two variables. In this report, we performed bivariate analysis on the provided dataset to identify the relationship between different variables.

## Heat Map

A heatmap is a graphical representation of data where the values are represented by different colors. In this analysis, we used a heatmap to visualize the correlation between the different variables. The heatmap shows the correlation coefficient values ranging from -1 to 1, where -1 represents a negative correlation, 0 represents no correlation, and 1 represents a positive correlation. The heatmap allowed us to quickly identify the variables that are highly correlated with each other.

From the heatmap, we can observe that the following variables have a strong positive correlation:

loan\_amount and person\_income

loan\_amount and loan\_percent\_income

loan\_amount and debt\_to\_income\_ratio

loan\_amount and person\_age

We also observed a negative correlation between person\_income and person\_age, which indicates that younger people tend to have lower incomes.

## Correlation Matrix

The correlation matrix is another tool used to measure the relationship between variables. The correlation matrix is a table that shows the correlation coefficient between each pair of variables. In this analysis, we used the correlation matrix to identify the variables that are highly correlated with each other.

The correlation matrix confirmed the observations from the heatmap. We observed a strong positive correlation between loan\_amount and person\_income (correlation coefficient = 0.39), loan\_amount and loan\_percent\_income (correlation coefficient = 0.25), and loan\_amount and debt\_to\_income\_ratio (correlation coefficient = 0.10).

## Boxplots

Boxplots are used to display the distribution of a categorical variable for different levels of a continuous variable. In this analysis, we used boxplots to visualize the distribution of loan\_amount for different levels of loan\_grades.

From the boxplots, we observed that the median loan\_amount is higher for higher loan\_grades, indicating that people with higher loan\_grades tend to get larger loans. However, we also observed some outliers in the boxplots for some loan\_grades, indicating that some people with low loan\_grades have also been approved for large loans.

# Multivariate analysis

From the multivariate analysis using PCA, we were able to reduce the dimensionality of the data set from 12 features to 2 principal components. The scatter plot of the first two principal components showed that there were distinct clusters in the data set. This indicated that there may be underlying patterns or groups in the data that could be further explored through clustering analysis.

We then applied k-means clustering to the data set to identify these potential groups. We determined the optimal number of clusters to be 4 using the elbow method and created a scatter plot of the data points colored by their assigned cluster. This allowed us to visually observe the separation of the data points into distinct groups.

Overall, the multivariate analysis provided valuable insights into the potential groupings and patterns in the data set, which could inform further analysis and modeling.

# Visualization

The following visualizations were created to gain insight into the loan dataset. They cover various aspects of the data, including the distribution of person income, person age vs person income, person income by home ownership, loan interest rates by loan grade, and default on file distribution. In addition, donut charts were created to visualize the distribution of person home ownership and loan intent. Finally, a pair plot was created to visualize the pairwise relationships between the variables of interest. These visualizations provide a better understanding of the loan data and can be used to inform further analysis and modeling.

**Distribution of Person Income**

The histogram shows that the distribution of person income is right-skewed, with a long tail to the right. This suggests that there are a few individuals with very high incomes, while most individuals have lower incomes.

**Person Age vs. Person Income**

The scatterplot shows a positive correlation between person age and person income, indicating that as people get older, they tend to earn more money.

**Person Income by Home Ownership**

The boxplot shows that people who own their home tend to have higher incomes than those who rent or have a mortgage. This suggests that home ownership may be a marker of financial stability.

**Loan Interest Rates by Loan Grade**

The boxplot and violin plot both show that higher loan grades (i.e., lower risk) tend to have lower interest rates. This makes sense, as lenders are less likely to charge high interest rates to borrowers who are less risky.

**Pairplot**

The pairplot shows the relationships between several variables, including person age, person income, loan amount, loan interest rate, and debt-to-income ratio. Some notable findings include a positive correlation between loan amount and loan interest rate, and a negative correlation between person income and debt-to-income ratio.

**Person Home Ownership and Loan Intent**

The donut charts show that most borrowers either rent or have a mortgage, and that most loans are taken out for medical purposes. These findings suggest that home ownership may be an important factor to consider when assessing credit risk, and that borrowers often take out loans for important life events such as medical expenses.

**Default on File Distribution**

The pie chart shows that most borrowers do not have a default on file, indicating that they have a good credit history. This is a positive finding for the lender, as it suggests that the borrower is less likely to default on the loan.

# Machine Learning Modeling and Evaluation

In this report, we evaluate three different machine learning models for the problem of predicting credit risk percentage. The goal is to determine which model performs best on this

## Models:

We will compare the performance of three different models for credit risk prediction:

* Linear Regression
* K-Nearest Neighbors Regression
* Random Forest Regression

## Model Training and Evaluation:

We will train and evaluate each model using the following steps:

Split the data into training and testing sets (80/20 split) Train the model on the training data. Evaluate the model on the testing data using various metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2) score, Mean Absolute Error (MAE), Accuracy, Precision, Recall, F1 Score, and ROC AUC depending upon the type of model being trained.

## Model Comparison:

## Linear Regression:

Linear Regression is a commonly used model for predicting continuous numerical values. In our case, we use Linear Regression to predict the percentage of credit risk associated with each borrower. We preprocess the data by one-hot encoding the categorical variables. The model is trained on the training data and evaluated on the testing data. The model achieves the following performance metrics:

* MSE: 2.2552371878306147e-05
* RMSE: 0.004748933762257181
* R^2 score: 0.9980005403431631
* MAE: 0.00293117528866775

## K-Nearest Neighbors Regression:

K-Nearest Neighbors (KNN) is a non-parametric algorithm used for both classification and regression tasks. In our case, we use KNN Regression to predict the percentage of credit risk associated with each borrower. We preprocess the data by converting the target variable to a categorical variable. The model is trained on the training data and evaluated on the testing data. The model achieves the following performance metrics:

* Accuracy: 0.9986189964707688
* Precision: 0.9985856724764876
* Recall: 0.9986189964707688
* F1 Score: 0.9985493183794802
* ROC AUC: 0.9866918221622896

## Random Forest Regression:

Random Forest is an ensemble algorithm that combines multiple decision trees to make predictions. In our case, we use Random Forest Regression to predict the percentage of credit risk associated with each borrower. The model is trained on the training data and evaluated on the testing data. The model achieves the following performance metrics:

* MSE: 1.1620242442841775e-05
* MAE: 0.0005062452048492166
* R^2 score: 0.9989697666350794

## Conclusion:

In conclusion, we have compared three different models for predicting credit risk: Linear Regression, K-Nearest Neighbors Regression, and Random Forest Regression. After evaluating the performance of each model, we found that all three models performed very well with high R-squared values, indicating that they were able to capture a significant portion of the variance in the target variable.

However, the Random Forest Regression model achieved the best results overall, with the lowest Mean Squared Error and Mean Absolute Error, and the highest R-squared value. The K-Nearest Neighbors Regression model achieved high accuracy, precision, recall, and F1 score, but its ROC AUC score was lower than that of the other models, indicating that it might not perform as well with imbalanced data.

Therefore, we decided using the Random Forest Regression model for predicting credit risk, as it provides the most accurate and precise predictions and can handle both numerical and categorical variables with ease. Additionally, Random Forest models are robust against overfitting and perform well even with noisy data.