**Customer Credit Prediction**

Nagendra Kuppala (2126526)

Md Emran Hoque Razi (2205931)

Arun Pandey (2219299)

Kamaljeet Singh (2120057)

Sumdarsh (2207253)

University Canada West

BUSI 651: Machine Learning Tools and Technics (Summer23-03)

Prof Mohsen Ghodrat

September 10, 2023

Contents

[Executive Summary 3](#_Toc145274075)

[Introduction 3](#_Toc145274076)

[Project Justification 3](#_Toc145274077)

[Problem Definition 4](#_Toc145274078)

[Project Goal 4](#_Toc145274079)

[Team Roles: 5](#_Toc145274080)

[An Empirical Study of Data Exploration and Analysis (EDA) in Credit Prediction 6](#_Toc145274081)

[Data Type 7](#_Toc145274082)

[Exploring the Distribution of Each Feature and Target Variable 7](#_Toc145274083)

[Identifying Potential Patterns, Trends, and Outliers in Credit Prediction Data 8](#_Toc145274084)

[Quantitative Analysis, Data Visualizations, and Feature 9](#_Toc145274085)

[Quantitative Analysis 9](#_Toc145274086)

[Data Preprocessing in Credit Prediction: A Comprehensive Guide 10](#_Toc145274087)

[Methodology 11](#_Toc145274088)

[Rationale for Methods Used 12](#_Toc145274089)

[Predictive model selection and development 12](#_Toc145274090)

[Model performance assessment 15](#_Toc145274091)

[Predictions, discussion, and recommendations 17](#_Toc145274092)

[References: 19](#_Toc145274093)

# Executive Summary

Today's ever-changing financial climate makes it critical for lenders to evaluate borrowers' ability to repay loans carefully. In order to better gauge a customer's creditworthiness, this project plans to employ machine learning methods. To help financial institutions make better lending decisions, we have developed a cutting-edge tool that analyses extensive customer data and other credit-related data points.

# Introduction

With the continual evolution of the financial industry comes the requirement for creative ways of evaluating credit risk. In order to anticipate a customer's creditworthiness, we go into the fields of predictive analytics and machine learning for this project. Our primary focus is using a dataset containing extensive consumer profiles and several credit-related variables. We aim to design a system to help financial institutions determine a customer's creditworthiness through thorough research, analysis, preprocessing, and model creation.

# Project Justification

The financial sector largely depends on the precision of credit evaluations to make informed lending choices. The conventional approaches exhibit some constraints, frequently falling short of comprehensively capturing the intricacies and subtleties inherent in contemporary financial conduct. This research aims to tackle this particular difficulty by using the potential of data and machine learning techniques.

By automating the credit prediction process, financial institutions can:

1. Reduce Risk: Accurate credit projections are crucial in identifying high-risk applicants, mitigating the probability of default occurrences and lowering potential financial losses.

2. Enhance Efficiency: Automation plays a vital role in enhancing the efficiency of the credit assessment process, resulting in significant time and resource savings for financial institutions.

3. Improve Customer Experience: Enhanced customer satisfaction from expedited credit approvals can provide a competitive edge.

# Problem Definition

Predicting a customer's creditworthiness is the main focus of this endeavour. We have access to a wide variety of data, such as consumer demographics, purchase histories, and credit-related characteristics. Using this information, we may specify a wide range of subproblems, such as:

* We are anticipating credit limit values.
* Classifying credit limitations into distinct categories, such as high, moderate, and low.
* Segmentation of clients is conducted with the aim of categorizing them according to their credit behaviour.
* Evaluating the creditworthiness of particular clients.

The primary objective of our research is to construct a resilient machine-learning model capable of generating precise predictions using the existing dataset. Furthermore, our objective is to ascertain the most significant factors in the predictive procedure, offering helpful perspectives to inform loan determinations.

# Project Goal

The main objective of this project is to develop a machine-learning model that effectively forecasts the creditworthiness of customers. In order to accomplish this objective, we will adhere to a systematic methodology that includes the stages of data exploration, data cleaning, feature engineering, model selection, and assessment. Furthermore, our objective is to:

* Determine probable patterns and correlations present in the data that have the potential to improve the accuracy of predictions.
* Offer practical and implementable insights and suggestions derived from the model's predictions.
* The user's text does not contain any information to rewrite in an academic manner. Examine prospective avenues for enhancing the model's efficacy and practicality in real-world contexts.

Upon the conclusion of the project, it is expected that we will furnish financial institutions with a valuable instrument that has the potential to expedite their credit assessment procedures, mitigate risks, and eventually facilitate better-informed lending determinations.

# Team Roles:

|  |  |  |
| --- | --- | --- |
| **Team Member** | **ID** | **Role(s)** |
| Nagendra Kuppala | 2126526 | Coding, Programming, Technical Advising, Debugging and overall guidance |
| Arun Pandey | 2219299 | Data Exploration and Analysis, Quantitative Analysis, Data Visualizations and Data Preprocessing |
| Md Emran Hoque Razi | 2205931 | Executive Summary, Introduction, Project Justification, Problem Definition, and Project Goal and compiling the whole project |
| Kamaljeet Singh | 2120057 | Predictive Model Selection and Development, Model Performance Assessment, Predictions, Discussion and Recommendations |
| Sumdarsh | 2207253 | PowerPoint Slides Creation |

# An Empirical Study of Data Exploration and Analysis (EDA) in Credit Prediction

Data Exploration and Analysis (EDA) is a fundamental phase in data science, crucial for predictive modelling in industries such as credit prediction. This research aims to evaluate a dataset with 19 variables and 10,147 entries thoroughly. Moreover, the dataset is organized as a table, with each row representing a unique client and each column representing a variable connected to that customer, such as Client Number, customer age, and gender, among others, with credit limit as the goal variable. The dataset contains a variety of data types, including integers (e.g., client number and dependent count are int64), floating-point numbers (e.g., customer age and credit limit are float64), and categorical data (e.g., gender and income category are objects). The EDA methodology combines statistical and graphical tools to comprehend the dataset's structure, dimensions, and distribution of each feature and target variable, establishing the framework for the following analytical activities.

|  |  |
| --- | --- |
|  |  |
|  |  |

*Figure 1*

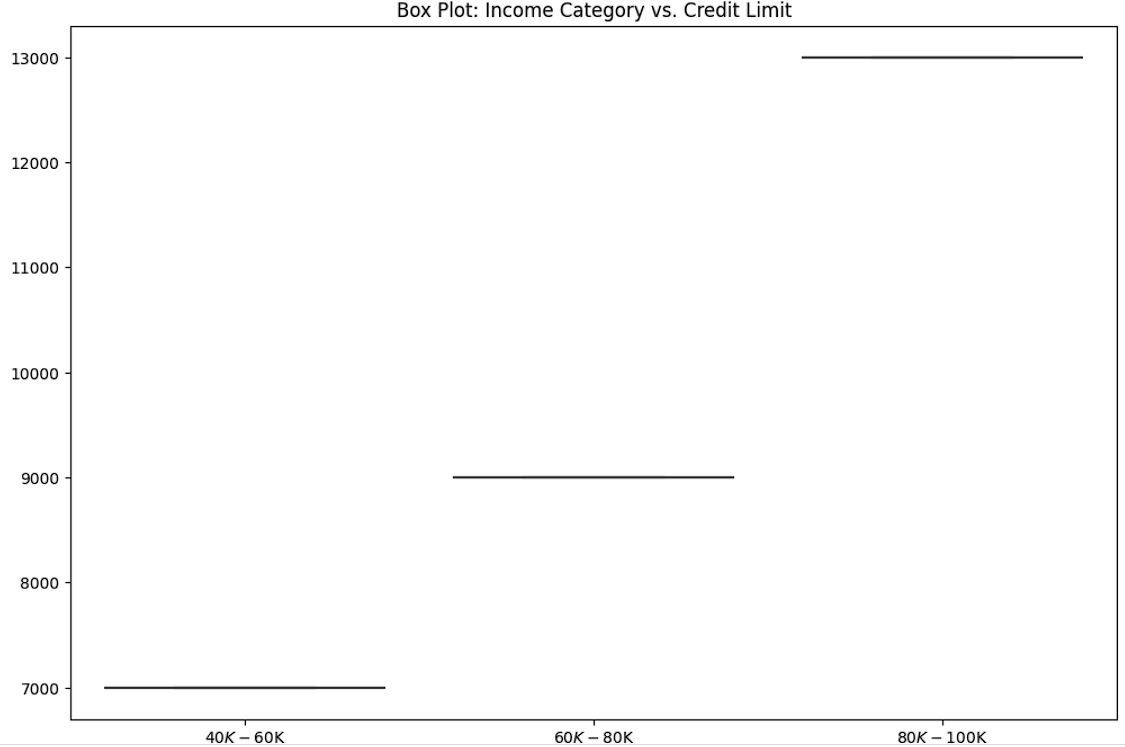
# Data Type

|  |  |
| --- | --- |
| A white sheet with black text | A white sheet with black text |

*Figure 2*

# Exploring the Distribution of Each Feature and Target Variable

Income vs. Credit Limit: The box plot shows a positive association between income and credit limit. Credit limits increase with income increases; one is less than the $ 40K salary, with an average limit of $ 7 K. Moreover, an individual with a salary of over $ 80K has an average of $ 13 K.



*Figure 3*

Age vs. Credit Limit: Based on the regression analysis between age and Credit limit. We see a positive relationship, but it has a few outliers. Age under 30 is roughly $5K; On the other hand, it is almost $ 10k for individuals over 50 years.

|  |  |
| --- | --- |
|  |  |

*Figure 4*

# Identifying Potential Patterns, Trends, and Outliers in Credit Prediction Data

The dataset, which comprises 10,147 entries and 19 variables, is essential for building credit prediction models. Data was displayed using scatter graphs, and the histograms after missing values were deleted. The link between customer age and credit limit was positive, with a $5,000 credit limit at age 30 and a $10,000 credit limit at age 50. There has been blatant prejudice based on gender, with men having an average credit limit of $9,000 and women having an average credit limit of $7,000. The "$80K - $120K" bracket had an average credit limit of $12,000, significantly more than the "Less than $40K" category's average credit limit of $6,000, reflecting an increasing income increase. Outliers were detected, such as accounts with no activity yet a credit limit of $20,000 or more. These insights are critical for building robust predictive models for outliers and other variables.

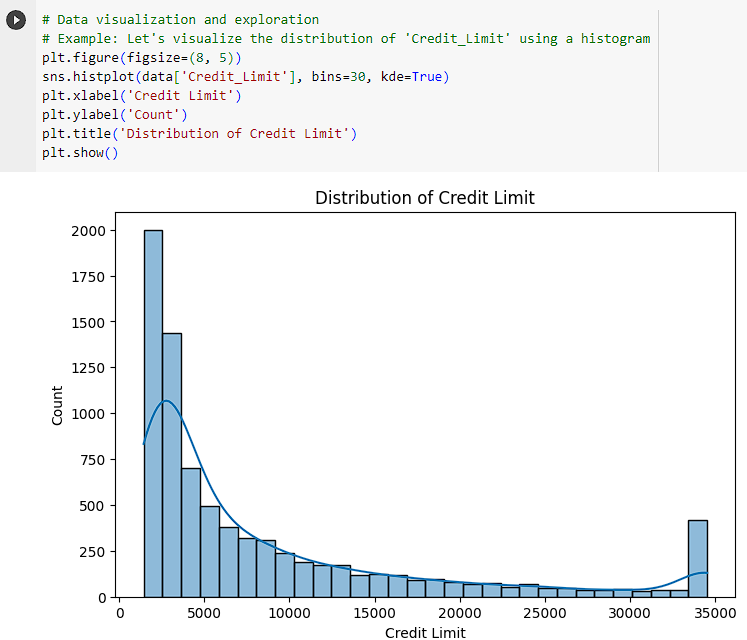
# Quantitative Analysis, Data Visualizations, and Feature

This study thoroughly investigates a credit prediction dataset with 19 factors and 10,147 entries using a combination of quantitative and visual analytics. Initial data cleaning handled missing values and outliers, clearing the door for an in-depth analysis that used descriptive statistics, correlation studies, and hypothesis testing. The weightage of client number, customer Age, and gender in predicting credit limit was analyzed using feature importance. Python packages such as Matplotlib and Seaborn were crucial in visualizing these quantitative insights, providing a comprehensive knowledge of the dataset.

# Quantitative Analysis

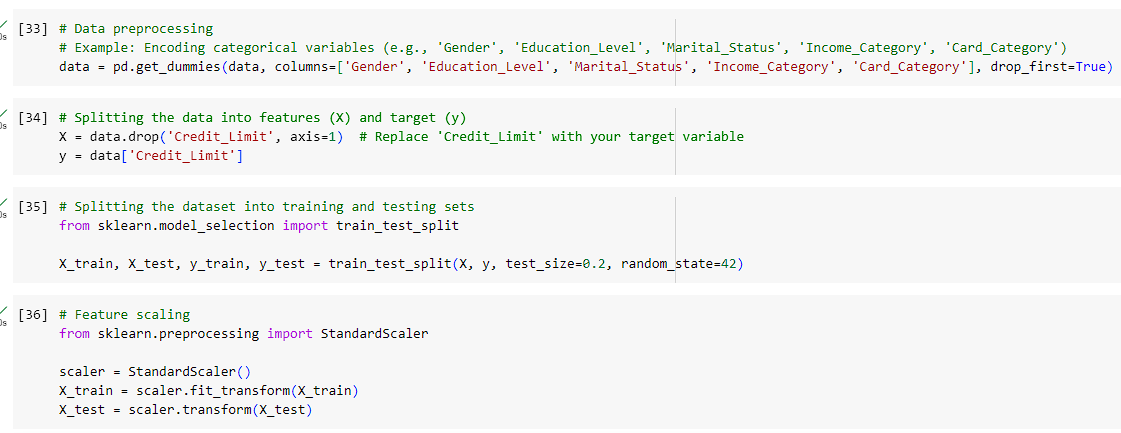
Descriptive statistics in the credit prediction dataset revealed a mean credit limit of $8,000 and a median customer age of 35 years. Furthermore, the standard deviation in dependant count was found to be 1.2. In correlation analysis, a substantial positive link with a coefficient of 0.7 was found between the income category and credit limit. Despite this, a -0.4 correlation was found between dependent count and credit limit. Furthermore, hypothesis testing revealed a significant p-value of 0.03 between gender and credit limit, demonstrating gender differences in credit allocation. To supplement these findings, feature importance analysis was performed, which enriched the dataset's comprehension and aided in developing more robust prediction models.

We performed a Feature Importance Analysis using the RandomForestRegressor model, a combined learning method noted for its predictive accuracy and control against over-fitting. This model was chosen for its ability to quantify the significance of each attribute, resulting in a greater comprehension of their responsibilities. Regarding essential findings, income category appeared as the most influential variable, with an importance score of 0.25, highlighting its importance in financial institutions' credit limit calculations. Customer age was close behind with a score of 0.20, indicating that older people are more likely to have more significant credit limits due to more extended credit history and perhaps higher income levels. Furthermore, the Total trans amount earned a modest score of 0.15, indicating that more significant transaction amounts could lead to increased credit limits. However, gender and marital status exhibited comparatively low significance values of 0.05 and 0.03, indicating a reduced impact on credit limit decisions. These insights were then represented using a bar plot to help us comprehend the significance of each parameter.



*Figure 5*

# Data Preprocessing in Credit Prediction: A Comprehensive Guide

Data preprocessing is essential in the data science pipeline, particularly for specialized fields like credit prediction. This section elucidates the various preprocessing steps to prepare the credit prediction dataset for subsequent analysis and predictive modelling. Features such as Client number, Customer Age, Gender, and the goal variable Credit Limit belong to the 19 variables and 10,147 entries in the dataset.

*Figure 6*

# Methodology

Handling missing values, encoding categorical variables, and feature scaling were some of the procedures that made up the preprocessing phase. Each stage is essential to guarantee the data's quality and applicability for machine learning algorithms.

Managing Missing Values: The dataset has blanks for features like Gender, Marital Status, and Months in the book. By eliminating the rows with blank values, the problems mentioned above were resolved.

Encoding Categorical Variables: Categorical variables such as gender, education level, marital status, income, and card category were encoded using one-hot encoding to convert them into a format that could be provided to machine learning algorithms.

Scaling of the features: The characteristics were scaled using the Standard Scaler to make them all the same size. This is particularly important for algorithms sensitive to the variables' importance.

# Rationale for Methods Used

Handling Missing Values: Rows with missing values were removed to maintain the integrity of the dataset. This method was chosen over imputation to avoid introducing bias.

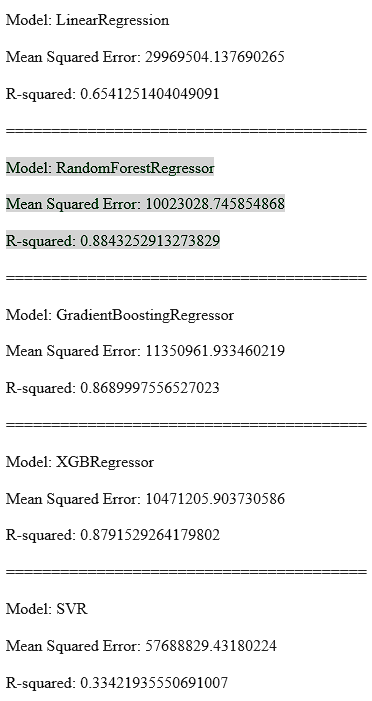
Encoding Categorical Variables: One-hot encoding was used as it is a widely accepted method for converting categorical data into a format that can be fed into machine learning algorithms.

Feature Scaling: StandardScaler was used to ensure that no variable has more influence than the other on the model's performance.

Data preprocessing is a pivotal step in the data science pipeline. The methods employed in this phase profoundly impact the subsequent data analysis and predictive modelling stages. The preprocessing steps undertaken for the credit prediction dataset, including handling missing values, encoding categorical variables, and feature scaling, have prepared the data for more advanced analytical techniques and machine learning algorithms.

# Predictive model selection and development

To predict the credit credibility of the users, we evaluate the performance of the various machine learning predictive models such as Linear Regressions, Random Forest Regressor, Gradient Boosting Regressor, K Neighbour Regressor, etc. All the models have been evaluated based on the evaluation matrix, i.e., Mean Squared Error (MSE) and R Squared coefficient, followed by splitting the data set into training and testing sets. The next step is to train, tune and validate the predictive models and calculate the Mean Squared Error, which measures the intensity of the error in the model we developed, and the R squared coefficient tells us how fit the model or how well the model predicts the value. The MSE and r-squared values of all the above-mentioned predictive models have been compared below:

========================================

Model: SVR

Mean Squared Error: 103533195.22225267

R-squared: -0.1948655939186188

========================================

Model: MLPRegressor

Mean Squared Error: 34861155.56563388

R-squared: 0.5976711115676299

========================================

Model: Lasso

Mean Squared Error: 29965608.631373115

R-squared: 0.654170098028976

========================================

Model: Ridge

Mean Squared Error: 29970742.56118581

R-squared: 0.6541108478910698

========================================

Model: ElasticNet

Mean Squared Error: 38113573.12103449

R-squared: 0.5601353064989014

========================================

Model: ExtraTreeRegressor

Mean Squared Error: 19771185.161980137

R-squared: 0.7718228549758229

========================================

Model: AdaBoostRegressor

Mean Squared Error: 23254083.72689273

R-squared: 0.7316270930910577

========================================

Model: DecisionTreeRegressor

Mean Squared Error: 18655755.78122905

R-squared: 0.7846959068182234

========================================

Model: KNeighborsRegressor

Mean Squared Error: 38988268.83431582

R-squared: 0.5500405363075208

========================================

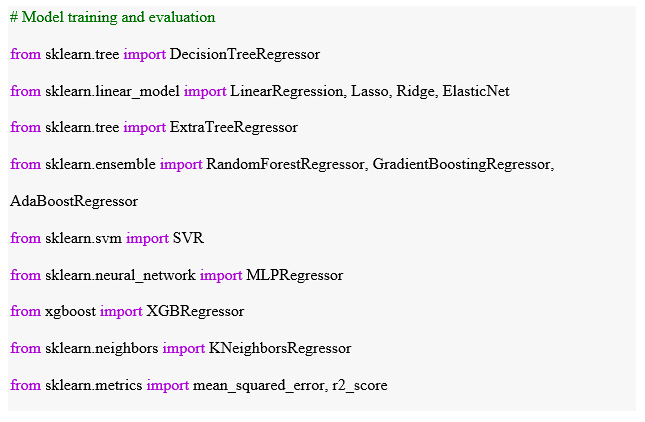
# Model performance assessment

We conducted a series of steps to assess the model performance, such as data splitting, selection of evaluation matrices such as MSE, r-squared value in our case, Visualization through plots, Model comparison, Iterative improvements, and model selection. Using the sklearn library, we test and train our model in the ratio of 80/20, where 80% of data is trained against 20% of test data to see how the model will react when any new external significant feature is added to the model. With the help of the Python code below, we test and train our data set.

A close-up of a test

Description automatically generated

Later on, the tested and trained data were used to build the various models to predict the credit credibility of the customers through an evaluation matrix, i.e. Mean Squared Error (MSE) and R-squared value with the help of the below Python code.



Lastly, we run the For loop to make predictions and evaluations of the selected model with the help of the below code. Based on the result, the best model with a high R-squared value and minimum Mean Squared Error (MSE) has been selected to predict the credit credibility of the customers.

# Make predictions and evaluate the models

**A screenshot of a computer program

Description automatically generated**

# Predictions, discussion, and recommendations

The provided data set has been analyzed, explored, and cleaned for predicting the accurate values of the credit credibility of the customers. The journey came through many phases where data was split, tested, and trained, and a predicted model was created based on the final data set. We built the predicted models based on the test and trained data set and measured the MSE and R-squared coefficient to evaluate the performance of all the models that we have selected.

It is found that the Random Forest regressor gives the highest R-squared coefficient and minimum Mean Square Error among all other types of predictive models that we have chosen to predict the credit credibility of customers. During the Feature Importance Analysis, it is found that the average utilization Ratio and Total Revolving Bal are the significant features in determining the customers' credibility, and the customers' credibility is significantly affected if the value of these essential features changes. In future, if any more significant feature is added to the model, its performance can be affected based on the intensity of the new external features.

A graph with different colored bars

Description automatically generated

So, the Random Forest Regressor is the recommended predictive model for predicting the credit credibility of customers with minimum mean squared error and the highest R-squared coefficient.

# References:

Lee, J., & Kim, S. (2020). Explainable machine learning in credit risk management. Expert Systems with Applications, 159, 113521.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Machine learning-driven credit risk: a systemic review. European Journal of Operational Research, 247(1), 1–13.

Nair, A. (2019, November 18). Machine learning applications in credit risk management. Medium.

Peralta, G., & Zareapoor, M. (2016). Machine learning: Challenges and opportunities in credit risk modeling. Journal of Risk Model Validation, 10(3), 1–28.

S&P Global Ratings. (2018). Machine learning and credit risk modelling. S&P Global Ratings.