Machine Learning Approaches to Credit Application Analysis:

Clustering, Classification, and Regression

ST3189 – Machine Learning
Student Number: 230753127
03.04.2025

Abstract 3

1. Introduction 4

2. Data and Methodology 4

2.1. Dataset Overview 4

2.2. Data Preprocessing 5

2.3. Analytical Approach 5

3. Results and Analysis 6

3.1. Clustering Results 6

3.2. Regression Results 9

3.3. Classification Results 11

4. Conclusion 12

Abstract

This study investigates the use of machine learning methods for analyzing data from actual credit applications. Finding trends among applicants and developing predictive models that can predict application approval results and other important characteristics are the goals.

Anonymized credit application records are included in the dataset, which comes from the UCI Machine Learning Repository. It includes both continuous and categorical variables, some of which have missing values. Whether an applicant was accepted (+) or denied (-) is indicated by the target variable. The dataset's characteristics make it possible to apply techniques for classification, regression, and unsupervised learning.

To divide candidates into relevant categories, unsupervised learning techniques such as Principal Component Analysis (PCA) and KMeans clustering were employed. A continuous variable thought to signify financial strength was predicted using regression analysis. To forecast the acceptance status of an application, classification models like Random Forest and Logistic Regression were created.

Metrics relevant to each task, including as accuracy, ROC-AUC, R2 and RMSE, were used to assess model performance. The findings demonstrate that machine learning models have the potential to be applied to credit scoring systems and can successfully differentiate applicants based on hidden patterns.

1. Introduction

When making financial decisions, credit rating is essential, especially when determining an applicant's creditworthiness. Machine learning provides methods that can increase the effectiveness and precision of these evaluations in a world that is becoming more and more data-driven. In order to find insights, identify applicant groupings, and forecast important outcomes like approval decisions, this study explores the application of machine learning approaches to credit application data.

Anonymized credit application records from the UCI Machine Learning Repository made up the dataset used in this investigation. Numerous numerical and categorical characteristics that characterize the circumstances and profile of each applicant are included. Whether the credit application was accepted or rejected is indicated by the target variable. Although the specific meaning of each variable has been removed for confidentiality, the dataset structure still provides a realistic and useful foundation for analysis.

This report's objective is to use a variety of machine learning techniques to tackle three important analytical tasks: clustering applicants into meaningful groups based on similarities (unsupervised learning), predicting a continuous financial variable assumed to represent income or available credit (regression), classifying applications as approved or rejected (classification).

Care is taken throughout the report to clearly and business-focusedly describe the methodologies and results. Performance measures are prioritized, but so are interpretability, useful insights, and the findings' applicability to credit risk assessment and decision-making.

2. Data and Methodology

2.1. Dataset Overview

The UCI Machine Learning Repository is the source of the dataset utilized in this study. It includes credit card application records that have been anonymized. With 15 input variables (designated A1–A15) and one target variable (designated A16), each row represents a single application and shows whether the credit was accepted (+) or denied (-).

A combination of categories and numerical characteristics make up the input variables. Although the variables' true meanings have been obscured for privacy reasons, they represent common characteristics used in credit evaluations, like income level, work status, or credit history. Imputation techniques were used to manage missing values in certain features (e.g., filling missing numerical values with the median).

A brief summary of the data:

Total observations		690
Features	6 numerical	9 categorical
Target	«+» = approved	«-» = rejected

2.2. Data Preprocessing

Before applying any machine learning techniques, three preprocessing steps were conducted. The first step is imputation (missing values in numerical columns were filled using the median, while categorical columns used the most frequent value). The second one is encoding (categorical features were transformed into numerical format using one-hot encoding, which creates binary columns for each category). And the last step is scaling (numerical variables were standardised to ensure equal contribution to distance-based algorithms such as clustering and K-Nearest Neighbours).

2.3. Analytical Approach

A combination of supervised and unsupervised learning approaches was used to meet the project's goals. To investigate underlying patterns in the data without depending on the goal variable, the investigation started with

unsupervised learning. Principal Component Analysis (PCA) was used to reduce dimensionality, which helped to find the most informative components and visualize the data in two dimensions. The candidates were then divided into discrete groups according to similar feature characteristics using KMeans clustering. The possible existence of client segments that would not be apparent through conventional research was revealed by this exploratory stage.

In order to forecast a continuous variable (A15), which is thought to reflect an applicant's financial health, such as income or available credit, regression modeling was used in the second portion of the research. The following three regression models were taken into consideration: Random Forest, Ridge Regression, and Linear Regression. These were chosen because they provide a balance between prediction strength and interpretability. The average prediction error and overall explanatory strength of each model were assessed using the Root Mean Squared Error (RMSE) and R2 score.

Lastly, categorization models were created to forecast whether a loan application would be approved (A16). Because it replicates the actual lending decision-making process, this assignment has special practical significance. A number of methods, such as Random Forest, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Logistic Regression, were tested. Standard evaluation criteria like accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) were used to compare the models. A comprehensive evaluation of each model's ability to differentiate between accepted and denied applications was made possible by these criteria.

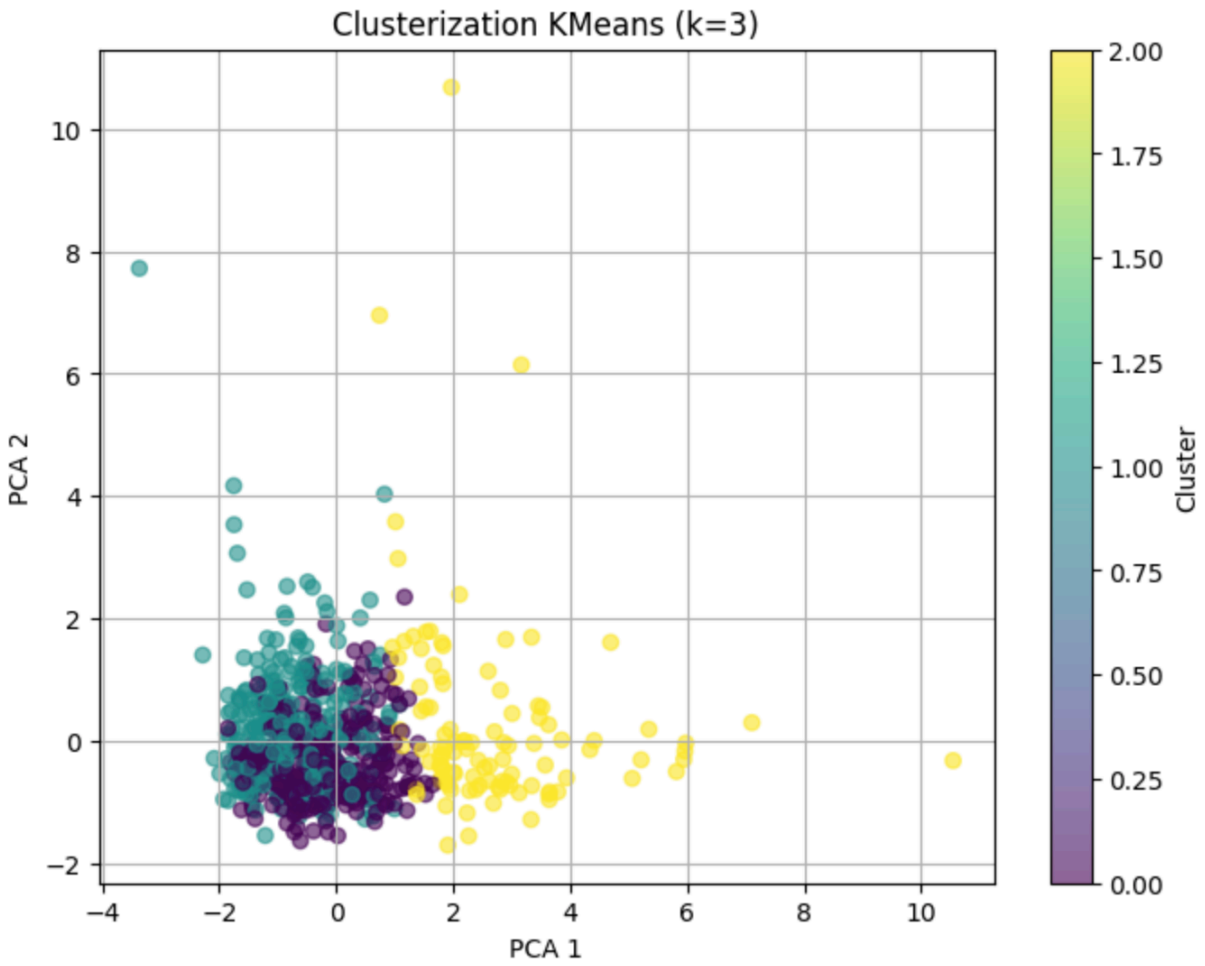
3. Results and Analysis

3.1. Clustering Results

Principal Component Analysis (PCA) was used to reduce dimensionality before KMeans clustering was used to find natural groupings within the application population. The elbow approach, which recommended three separate clusters, was used to determine the ideal number of clusters.

The average values of a number of continuous variables varied noticeably within each cluster. For instance, the greatest values for variables like A2 and A15 were found in Cluster 2, which would suggest that the applicants had more robust financial profiles. Cluster 1, on the other hand, had a higher number of applicants with possibly less desirable traits and lower average values.

The visualisation below presents the clustering results in two-dimensional PCA space, revealing distinct groupings:



Below is the table with Cluster interpretation:

Cluster 0	
Characteristic	**Value**
A15	733 - average value
A14	131
A11	2.09
A8	1.25
A3	4.61
A2	25.93

This cluster shows average values for most features. Can represent typical clients, without strong extremes. For example: stable, but not super-rich, middle-income, standard query.

Cluster 1	
Characteristic	**Value**
A15	353 - lower than in cluster 0
A14	263 - 2 times higher than the others
A11	0.53 - the lowest
A8	1.54
A3	2.84
A2	32.09 - slightly above average

This cluster represent clients with low A15 (possibly lower income/ balance) but high A14 (possibly application amount or debt load). A11 (possibly "turnaround time" or "stability") - low, it may indicate higher risk customers. This cluster may include clients that are beginning borrowers or temporary requests from less stable category.

Cluster 2	
Characteristic	**Value**
A15	3576 - maximum
A14	130

A11	8.17 - 4 times higher than cluster 0
A8	6.81 - significantly greater
A3	10.21 - maximum
A2	46.29 - maximum

This is the richest or largest segment - it leads by all signs. This cluster may represents clients with with high yield/credit volume, high activity. These are the most reliable clients or VIP-clients with high turnover.

In conclusion, such segmentation can be useful for subsequent decisions on individual approach to clients, adjustment of approval criteria and risk management.

3.2. Regression Results

Variable A14, which is thought to reflect a continuous financial indication like income, payment amount, or credit line, was predicted using regression analysis. The objective was to ascertain whether accurate numerical forecasts for this value could be made using the applicant's profile.

The Random Forest Regressor, Ridge Regression, and Linear Regression were the three models that were put to the test. These models were selected because they provide a compromise between regularization, non-linear pattern recognition, and simplicity. In order to address missing data, numerical features were scaled and imputed using median values, and all categorical variables were encoded using one-hot encoding prior to training.

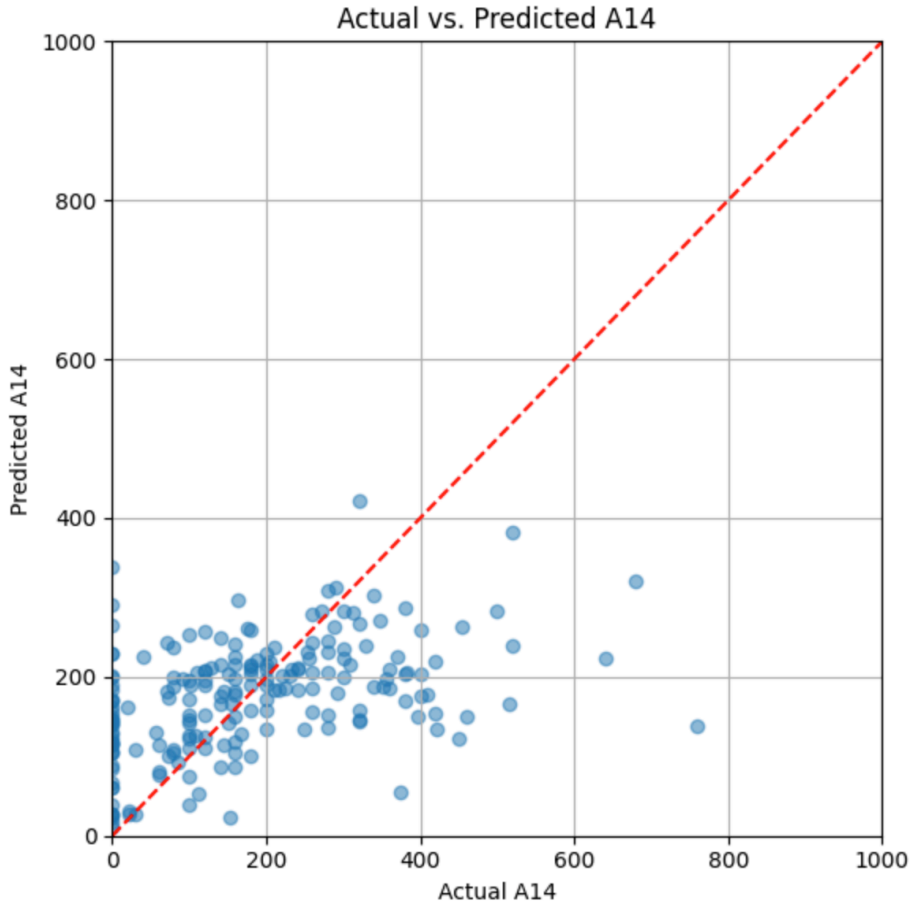
Model	**RMSE**	**R^2 Score**
Linear Regression	211.8	-0.03
Ridge Regression	210.2	-0.02
Random Forest	194.32	0.13

With the lowest root mean squared error (RMSE) and a moderate positive R2 score of 0.13, the Random Forest model performed better than the others,

suggesting that it was able to capture a small amount of the volatility in the target variable. However, both linear models showed negative R2 values, indicating that they performed worse than a straightforward mean-based prediction, and overall model performance was limited.

These findings imply that there is insufficient predictive signal in the current feature set to reliably estimate A14. The anonymized nature of the dataset or the subjectivity and noise present in the target variable could be the cause of this. Notwithstanding this, the exercise is still useful for emphasizing how crucial feature quality, data representation, and model evaluation are to predictive modeling.

The plot below illustrates the relationship between the actual values of A14 and the values predicted by the best-performing model — the Random Forest Regressor.



Perfect prediction would occur if every point fell precisely on the red diagonal line. The majority of the points in this instance, however, diverge significantly off the line, indicating the model's weak capacity to accurately predict the target variable.

Additionally, the plot indicates that the model does marginally better at forecasting mid-range A14 values, whereas the variance in forecasts rises for higher or lower values. This finding validates that A14 is only weakly predictable based on the features that are currently available and is in line with the quantitative performance measures.

3.3. Classification Results

As stated by the binary target variable A16, where + indicates approval and - indicates rejection, the final job involved predicting the result of a credit application. This represent a common classification problem that arises in the fields of financial risk and credit score.

Four machine learning models were tested 4 models: Logistic Regression, as a widely used baseline model, Random Forest, to capture non-linear interactions, Support Vector Machine (SVM), effective in high-dimensional spaces and K-Nearest Neighbours (KNN), as a simple, instance-based approach.

Every model was trained on preprocessed data that had numerical features scaled and categorical variables encoded using one-hot encoding. Thirty percent of the dataset was put aside for evaluation, while the remaining dataset was divided into training and test sets.

Accuracy, precision, recall, F1-score, and ROC-AUC were used to evaluate each model's performance. These metrics provide a fair assessment of predictive quality, especially when it comes to binary classification tasks.

With an accuracy of 87% and an area under the ROC curve (AUC) of 0.92, Random Forest outperformed the other tested models. This shows how well

the model can differentiate between applications that are accepted and those that are denied, which makes it a solid contender for real-world implementation in decision-support systems.

These findings are further supported by the ROC curves, which demonstrate that Random Forest routinely performs better than the other models over a range of classification criteria. Furthermore, a feature importance analysis showed that characteristics like A2 and A8 were among the most significant predictors of decisions regarding credit approval.

These results show that machine learning methods, especially tree-based ensembles, can effectively extract significant patterns to help credit risk assessment even while the dataset is anonymized.

4. Conclusion

In this study, data from actual credit applications was analyzed using a variety of machine learning approaches. Potential client segments with unique financial profiles were revealed by identifying significant clusters of applications through unsupervised learning. Regression research demonstrated the difficulties of predicting continuous financial values utilizing anonymized and incomplete datasets, notwithstanding its limited predictive performance.

All things considered, the experiment demonstrates the potential and constraints of machine learning in financial settings. Although more detailed or richer data is needed for certain tasks in order to achieve high precision, the models were able to identify patterns that correspond with actual credit decision-making procedures. Adding domain-specific features or investigating ensemble approaches suited to unbalanced datasets and risk tolerance levels could result in even greater advancements.