A logo for college computing

Description automatically generated

**Date of Submission: 19/05/2024**

|  |  |  |
| --- | --- | --- |
| *Marluce Taciana Bora* |  |  |
| *Student Number: 2024141* |  |  |
| *Module Title: Strategic Thinking* |  |  |
| *Project to Implement Machine Learning Models for Credit Risk* |  |  |
| *Assessment Due Date: 19/05/2024* |  |  |
| *Date of Submission: 19/05/2024* |  |  |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Abstract

This report describes a comprehensive project focused on the implementation of machine learning models for credit risk assessment. The main objective is to develop predictive models capable of accurately classifying customers based on their credit risk, supporting decision-making in the lending process. Using the "German Credit Risk - With Target" dataset available on the Kaggle platform, the project followed a structured methodology that included the following steps:

* Exploratory Data Analysis (EDA): Initial exploration and visualization of data to identify patterns, trends, and potential data quality issues.
* Data Preparation: Cleaning, transforming, and preparing the data to ensure it is ready for modeling.
* Implementation of Machine Learning Algorithms: Training and evaluation of various machine learning models, including Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Decision Tree, Naive Bayes, Random Forest, Support Vector Machine, and XGBoost.
* Model Evaluation: Comparison of models based on performance metrics, with emphasis on recall, accuracy, and stability.
* Best Model Selection: Detailed analysis of the best models, Logistic Regression (LGR) and Linear Discriminant Analysis (LDA), justifying your choice based on performance and interpretability.
* Results and Conclusions: Interpretation of the results obtained, highlighting the superior performance of the Support Vector Machine (SVM) and the robustness of the LGR and LDA models.

The report concludes with a vision for future analysis and improvement, proposing the implementation of advanced techniques and the exploration of new machine learning algorithms over the next six months to further increase the accuracy of predictive models. A detailed timeline of what has been done and activities planned is provided to ensure the continued success of the project.

Contents

[1. Introduction 4](#_Toc167048331)

[2. Project Plan 4](#_Toc167048332)

[3. Business Understanding 4](#_Toc167048333)

[4. Data Understanding 4](#_Toc167048334)

[5. Exploratory Data Analysis 5](#_Toc167048335)

[5.1 Samples of analyses performed. 5](#_Toc167048336)

[6. Data Preparation 8](#_Toc167048337)

[6.1 Identification of Outliers 8](#_Toc167048338)

[6.2 Treatment of Outliers 9](#_Toc167048339)

[6.3 Coding of Categorical Variables 10](#_Toc167048340)

[6.4 Standardization 11](#_Toc167048341)

[6.5 Correlation and Heatmap Analysis 11](#_Toc167048342)

[7. Machine Learning Implementation 12](#_Toc167048343)

[8. Results 14](#_Toc167048344)

[9. Pipeline and Trait Selection 14](#_Toc167048345)

[10. Future Recommendations 14](#_Toc167048346)

[11. Future Challenges and Solutions 14](#_Toc167048347)

[12. Project Timeline and Milestones Achieved 14](#_Toc167048348)

[12.1 Initial Phase (Completed) 15](#_Toc167048349)

[12.2 Intermediate Phase (Next 3 months) 15](#_Toc167048350)

[12.3 Final Round (Last 3 Months) 15](#_Toc167048351)

[13. Conclusion 15](#_Toc167048352)

[14. Reference 15](#_Toc167048353)

# Introduction

Effective credit risk management is essential for the financial industry, particularly to assess the likelihood of default by potential customers. In an increasingly data-driven environment, the application of machine learning algorithms emerges as a promising approach to predicting borrower behavior.

This report details a comprehensive project focused on implementing machine learning models for credit risk assessment. The main objective is to develop predictive models capable of accurately classifying customers based on their credit risk, supporting decision-making in the lending process. The project follows a well-defined management methodology, with an emphasis on steps such as exploratory data analysis, data preparation, implementation of machine learning algorithms, and analysis of the results.

Within the scope of this project, I will explore the "German Credit Risk - With Target" dataset from the Kaggle platform. This dataset provides detailed information about customers of German banks, including age, income, credit history, and defaulting status on previous loans. The methodological approach aims not only to develop robust predictive models, but also to generate insights that can enhance credit risk analysis, benefiting both financial institutions and potential customers.

# Project Plan

The project follows the following plan:

* Definition of the problem and objectives.
* Understanding of data and business context.
* Exploratory data analysis.
* Data pre-processing.
* Implementation of machine learning algorithms.
* Evaluation and optimization of the models.
* Presentation of results and conclusions.
* Identification of challenges and future recommendations.

# Business Understanding

The main objective of this project is the construction of predictive models in order to improve the decision-making process in the granting of loans. Seeking to reduce the risk of default and increase the operational efficiency of financial institutions through the application of advanced machine learning techniques.

# Data Understanding

The "German Credit Risk - With Target" dataset was used, providing information on customer characteristics, credit history, and delinquency status.

A screenshot of a computer

Description automatically generated

Fig1: German Credit Risk Dataset

A graph showing a credit risk distribution

Description automatically generated

Fig 2: Credit Risk Distribution

# Exploratory Data Analysis

I performed statistical analysis and visualizations to explore the relationships between variables and identify patterns in the data. I used a variety of graphs, such as histograms, boxplots, and heat maps, to visualize the distributions and correlations between variables.

# Samples of analyses performed.

I noticed some interesting insights when looking at the age distribution by risk charts. In the first chart, showing the overlap of age distributions for customers considered "good" (green) and "bad" (red), I noticed that the age distribution for "good" customers appears to be broader and tends to be more uniform compared to the age distribution for "bad" customers. However, there is considerable overlap between the two distributions, indicating that age alone may not be a determining factor in predicting credit risk. Other factors may be influencing the risk rating.

A graph showing the age distribution

Description automatically generated

Figure 3: Age Distribution

In addition, the Age Count by Risk chart shows the count of customers in different age groups divided by credit risk. It is noted that the most represented age groups for both the "good" and "bad" groups are between approximately 20 and 40 years old. However, the proportion of "bad" customers appears to be higher in younger age groups, while the proportion of "good" customers increases in older age groups. This suggests that younger customers may be more likely to be rated as "bad," while older customers are more likely to be rated as "good." These analyses helped us better understand the behavior of the variables.

A graph of age counting by risk

Description automatically generated

Figure 4: Age Counting by Risk

The graphs indicate how credit values are distributed between the "good" and "bad" risk groups. When analyzing the distribution of credit values, I observed that the violin chart shows a higher density in smaller credit values for the "bad" risk group. This suggests that customers with lower credit are more likely to be rated as "bad." On the other hand, the boxplot reveals that the median credit value for the "good" risk group is higher, which suggests that customers with higher credit are more likely to be classified as "good".

In addition, I examined the variability and presence of outliers in credit values among risk groups. We note that the presence of many outliers in the "bad" risk group in the boxplot may indicate greater variability in the amounts of credit granted to higher-risk clients. This observation is complemented by the violin chart, which shows a wider and less concentrated distribution for the "bad" risk group. These analyses provided us with insight into how credit values relate to risk groups and helped us better understand patterns in the data.

A graph of a number of red and blue points

Description automatically generated with medium confidence

Figure 5: Distribution of Credit Amount by Risk Level

Another analysis reveals that the distribution of loans classified as "bad" is broader and has a higher density over longer durations. This suggests that loans with longer durations are more likely to be classified as high-risk ("bad").

A diagram of a diagram showing a variety of shapes

Description automatically generated with medium confidence

Figure 6: Distribution of Loan Duration by Credit Risk

Several other analyses were performed to explore complex patterns and relationships between the variables. Due to the length and complexity of the analyses, their details can be viewed in the Python project that will be submitted along with this report. The source code and companion visualizations offer a deeper, more technical view of the approaches and insights developed throughout the project.

# Data Preparation

The pre-processing steps included treatment of missing values, coding of categorical variables, standardization of numerical variables, and selection of characteristics. This ensured that the data was in an optimal state for building accurate and effective predictive models.

# Identification of Outliers

I used a function to detect outliers in each column of the DataFrame by calculating the lower and upper bounds based on the interquartile range (IQR). Values outside these limits were considered outliers.

A screenshot of a computer program

Description automatically generated

Figure 7: Function to detect Outliers.

The results indicated the presence of outliers in several columns, including "Age", "Job", "Credit amount" and "Duration".

# Treatment of Outliers

Replaces outliers with calculated thresholds to minimize their negative impact.

A screenshot of a computer code

Description automatically generated

Figure 8: Function replace outliers with limits.

After applying this function, I verified the changes in the descriptive statistics, observing a reduction in the mean and standard deviation for some variables, indicating a lower dispersion of the data.

A table of numbers and numbers

Description automatically generated

Figure 9: Descriptive statistics

\*Average age (Age) decreased slightly.

\*The standard deviation of age (Age) also decreased, indicating a smaller dispersion of the data around the mean.

\*The maximum value for age has been adjusted to 64.5, which shows that outliers have been treated.

\*The mean and standard deviation for the credit amount have been reduced, suggesting a decrease in the dispersion of credit amounts.

\*The maximum value for Duration has been adjusted to 42, indicating that utliers have been treated in this column as well.

# Coding of Categorical Variables

To allow machine learning models to work with categorical variables, I turned these variables into one-hot encoding variables.

A screenshot of a computer

Description automatically generated

Figure 10: Coding of Categorical Variables

# Standardization

The standardization of the data was important so that all numerical variables had the same scale, facilitating the performance of various machine learning algorithms. I used the RobustScaler method for this task, which is robust to outliers.

The analysis of the descriptive statistics after standardization confirms that the columns were correctly standardized, with means close to zero and standard deviations close to one.

Further detailed analyses can be viewed in the Python project that will be submitted along with this report.

# Correlation and Heatmap Analysis

Correlation analysis helps to understand the relationships between the independent variables and the target variable. I used a heatmap to visualize the correlations.

A chart with numbers and a red line

Description automatically generated with medium confidence

Figure 11: Heatmap

This visualization allowed the identification of strong and weak relationships between the variables, helping in the selection of relevant characteristics for modeling

# Machine Learning Implementation

Several algorithms were trained and evaluated, including logistic regression, decision trees, random forest, SVM, XGBoost, among others. The evaluation of the models was performed using cross-validation and performance metrics.

A screenshot of a computer code

Description automatically generated

Figure 12: Machine Learning Implementation

* Logistic Regression (LGR): 0.883371 (0.067871)

Mean recall: 88.33% Standard deviation: 6.78% Stable model with high recall, indicating that it identifies positive samples well.

* Linear Discriminant Analysis (LDA): 0.870533 (0.070737)

Mean recall: 87.05% Standard Deviation: 7.07% Good performance with a slight variation between folds.

* K-Nearest Neighbors (KNN): 0.838311 (0.075017)

Average recall: 83.83% Standard deviation: 7.50% Lower recall compared to LGR and LDA models, but still good.

* Decision Tree Classifier (CART): 0.766250 (0.084334)

Average recall: 76.62% Standard deviation: 8.43% Lower and relatively stable performance.

* Gaussian Naive Bayes (NB): 0.745714 (0.080858)

Mean Recall: 74.57% Standard Deviation: 8.08% Similar performance to CART, but with greater variation between folds.

* Random Forest (RF): 0.870740 (0.061147)

Average recall: 87.07% Standard deviation: 6.11% Good performance and reasonable stability.

* Support Vector Machine (SVM): 0.955443 (0.039990)

Mean recall: 95.54% Standard Deviation: 3.99% Excellent performance with very little variation, indicating high accuracy and robustness.

* XGBoost (XGBM): 0.844139 (0.056622)

Mean recall: 84.41% Standard deviation: 5.66% Good performance, but inferior to LGR, LDA, KNN and RF.

A graph with blue and black rectangular boxes

Description automatically generated

Figure 13: Algorithm Comparison

# Results

The results showed that some models showed promising performance in predicting credit risk, while others underperformed.

The SVM (Support Vector Machine) had the best performance in terms of average recall (95.49%) and stability (standard deviation of 3%). This means that SVM is excellent at correctly identifying positive (or "Risk\_good") samples.

Logistic Regression (88%) and Linear Discriminant Analysis (87%) performed well with moderate variation. Logistic Regression also provides coefficients that are easily interpretable, which makes it easier to understand how each characteristic contributes to credit risk prediction.

Random Forest (87%) also performed well, with a high recall and acceptable variance.

Decision Tree (76%) and Gaussian Naive Bayes (75%) had the lowest performance, indicating that they are less effective at correctly identifying positive samples.

XGBoost (84%) performed reasonably but inferior to the models mentioned above.

# Pipeline and Trait Selection

A pipeline was created for trait selection and normalization, followed by the application of the Logistic Regression model. However, the evaluation showed that the accuracy in the test set was lower than expected, reinforcing the need to evaluate other models.

# Future Recommendations

More in-depth analysis, such as feature engineering, experimentation with different algorithms, and hyperparameter optimization to further improve the models, is recommended. Additionally, it is important to consider continuous monitoring of models in production and regularly updating data.

# Future Challenges and Solutions

Some challenges faced during the project included handling missing values and adjusting model hyperparameters. Future solutions include the use of advanced data imputation techniques and the automation of hyperparameter tuning processes.

# Project Timeline and Milestones Achieved

The project timeline included steps such as exploratory data analysis, data preparation, algorithm implementation, model evaluation, and presentation of results. Key milestones achieved include the completion of the exploratory analysis, the training and evaluation of the models, and the preparation of this report.

# Initial Phase (Completed)

* Data Collection and Pre-processing
* Data collection from the "German Credit Risk - With Target" dataset.
* Data cleansing and pre-processing.
* Initial Deployment of Models
* Implementation of several machine learning models (LGR, LDA, KNN, CART, NB, RF, SVM, XGB).
* Performance evaluation of the models using cross-validation.

# Intermediate Phase (Next 3 months)

* Template Enhancement
* Fine-tuning of hyperparameters of the selected models (LGR, LDA, SVM).
* Implementation of regularization techniques and feature engineering.
* Advanced Model Exploration
* Implementation and evaluation of neural networks and deep learning.
* Tests with ensemble models (AdaBoost, Gradient Boosting, Stacking).
* Detailed Performance Analysis
* Evaluation of additional metrics for a more complete understanding of the performance of the models.
* Implementation of stratified cross-validation.

# Final Round (Last 3 Months)

* Implementation of Data Balancing Techniques
* Use of techniques such as SMOTE to deal with class imbalance.
* Interpretation and Communication of Results.
* Application of model interpretation techniques (LIME, SHAP).
* Preparation of detailed and visual reports to communicate the results.

# Conclusion

This report presents a comprehensive overview of the project to implement machine learning models for credit risk assessment, demonstrating an effective project management methodology and highlighting the key insights gained from the analysis of the data. Given that this project is planned to continue for another 6 months, several additional steps and advanced techniques will be implemented to further improve the accuracy and robustness of the models. With the planned implementation of advanced techniques and continuous improvements, it is expected to further increase the accuracy and robustness of the models, ensuring reliable and accurate results over time.

# Reference

Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Available in:: <http://archive.ics.uci.edu/ml>.

Kaggle. (s/d). Students Performance in Exams. Available in:: <https://www.kaggle.com/spscientist/students-performance-in-exams>.

OECD. (2010). PISA 2009 Results: Learning to Learn – Student Engagement, Strategies and Practices (Volume III). Paris: OECD Publishing. Available in:: <http://dx.doi.org/10.1787/9789264083943-en>