**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA 2 – Capstone Project Proposal |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Gustavo Lopes & William da Cunha Braghin |
| **Student Number:** | sba22021 & sba22052 |
| **Assessment Due Date:** | 17/12/2023 |
| **Date of Submission:** | 17/12/2023 |

**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. |

SUMMARY

[1. PROJECT OVERVIEW 3](#_Toc153738995)

[2. /DATA UNDERSTANDING 4](#_Toc153738996)

[3. DATA PREPARATION 6](#_Toc153738997)

[4. MACHINE LEARNING IMPLEMENTATION 8](#_Toc153738998)

[5. CONCLUSIONS 10](#_Toc153738999)

[6. GITHUB REPOSITORY 11](#_Toc153739000)

[7. REFERENCES 12](#_Toc153739001)

## PROJECT OVERVIEW

The Enhancing Credit Risk Assessment through Machine Learning project aims to improve the credit risk assessment process by leveraging machine learning techniques. This project is driven by the increasing need for more accurate and efficient methods of evaluating an individual's creditworthiness.

Therefore, by utilizing the Credit Risk Dataset available on Kaggle.com, on which containing 12 key variables and 32,581 observations, the project seeks to develop a predictive model that enhances the precision and reliability of credit risk predictions.

Main Objectives:

* Enhancing Accuracy: Improve the accuracy of credit risk assessments through the implementation of machine learning algorithm.
* Risk Minimization: Minimize the risk of default by developing a model that effectively identifies creditworthy individuals.
* Optimizing Decision-Making: Provide financial analysts and credit officers with a powerful tool for making informed and data-driven credit approval decisions.

## /DATA UNDERSTANDING

For this project we are using the dataset Credit Risk Dataset available on Kaggle.com.

That dataset is composed of 12 Variables and 32,581 observations on which each row/observation represents data from individual whose applied for a loan for different purposes (.ie: Medical, Educational, and Home Improvement reasons).

Following there is a break-down of the Variables and it’s interpretation:

person\_age: Represents the age of the individual applying for the loan. Understanding age dynamics can provide insights into the correlation between age and creditworthiness.

person\_income: Represents the annual income of the individual. Examining income levels allows for a thorough assessment of the borrower's financial capacity.

person\_home\_ownership: Type of home ownership of the individual.

* rent: The individual is currently renting a property.
* mortgage: The individual has a mortgage on the property they own.
* own: The individual owns their home outright.
* other: Other categories of home ownership that may be specific to the dataset.

person\_emp\_length: Reflects the employment length of the individual in years. Employment stability can influence creditworthiness.

loan\_intent: Represents the intent behind the loan application, providing context for understanding borrower motivations.

loan\_grade: Indicates the grade assigned to the loan based on the creditworthiness of the borrower. The grading system is as follows:

* A: The borrower has a high creditworthiness, indicating low risk.
* B: The borrower is relatively low-risk, but not as creditworthy as Grade A.
* C: The borrower's creditworthiness is moderate.
* D: The borrower is considered to have higher risk compared to previous grades.
* E: The borrower's creditworthiness is lower, indicating a higher risk.
* F: The borrower poses a significant credit risk.
* G: The borrower's creditworthiness is the lowest, signifying the highest risk.

loan\_amnt: Denotes the loan amount requested by the individual, offering insights into the financial needs of borrowers.

loan\_int\_rate: Specifies the interest rate associated with the loan, a crucial factor in assessing the overall cost of credit.

loan\_status: Represents the loan status, where 0 indicates non-default (successful repayment), and 1 indicates default (failure to repay according to terms).

* 0: Non-default - The borrower successfully repaid the loan as agreed, and there was no default.
* 1: Default - The borrower failed to repay the loan according to the agreed-upon terms and defaulted on the loan.

loan\_percent\_income: Illustrates the percentage of income represented by the loan amount, aiding in understanding the debt burden relative to income.

cb\_person\_default\_on\_file: Indicates the historical default record of the individual as per credit bureau records, distinguishing between:

* Y: The individual has a history of defaults on their credit file.
* N: The individual does not have any history of defaults.

cb\_preson\_cred\_hist\_length: Represents the length of credit history for the individual, a key factor in evaluating creditworthiness based on past financial behavior.

## DATA PREPARATION

The data preparation phase is a critical step in ensuring the dataset's readiness for machine learning model development. Therefore in this phase, our attention was given to handling missing values in two key variables: person\_emp\_length and loan\_int\_rate.

Variable: person\_emp\_length

A careful examination revealed that the variable person\_emp\_length had null values, accounting for approximately 2.75% of the dataset. To maintain the integrity of the data while mitigating the impact of missing values, it was decided to drop the rows where person\_emp\_length was null. This approach was deemed appropriate due to the relatively small percentage of missing values, ensuring minimal data loss and preserving the overall representativeness of the dataset.

Variable: loan\_int\_rate

For the variable loan\_int\_rate, which represents the interest rate associated with the loan, missing values were observed. Given that the percentage of missing values in this variable was not negligible, a strategy was needed to impute these missing values effectively. The decision was made to use the mean of the existing values in loan\_int\_rate for imputation.

Using the mean for imputation it is also a reasonable choice when dealing with continuous numerical variables like interest rates. It ensures that the imputed values are representative of the overall distribution of interest rates in the dataset. This method is particularly suitable when the variable follows a normal or near-normal distribution, as the mean provides a central tendency measure that aligns with the data's central tendency.

By using these strategies, the data preparation phase addresses missing values judiciously, allowing for a more complete and robust dataset for subsequent machine learning model development.

Age Group Analysis:

In addition to this, our data preparation included a detailed analysis of age groups and targeted filtering for outlier removal.

The project incorporates an insightful age group analysis function, age\_group, designed to categorize individuals into specific age brackets and quantify the distribution within each range. The function iterates through predefined age bins, revealing the count of individuals falling within the respective brackets. This analysis provides a nuanced understanding of the dataset's age distribution, shedding light on potential trends and patterns relevant to credit risk assessment.

The output of this analysis, when applied to our dataset, provides valuable insights into the distribution of individuals across different age brackets, facilitating a more nuanced understanding of the demographic composition.

Age Filtering for Outliers:

Recognizing the importance of maintaining a relevant and realistic age range for credit risk assessment, the dataset undergoes targeted filtering to remove potential outliers. The code snippet below illustrates the removal of rows where the 'person\_age' exceeds 90, a decision likely made to exclude extreme age values that might not align with typical credit profiles.

The resulting dataset after this filtering now reflects a more realistic age distribution, with the maximum age capped at 84 and the minimum age at 20. This targeted filtering ensures that the dataset remains focused on a representative age range conducive to accurate credit risk assessment.

## MACHINE LEARNING IMPLEMENTATION

In the machine learning implementation phase, two different models were employed: Random Forest and Support Vector Machine (SVM). The primary focus was on assessing their effectiveness in predicting credit risk based on the features provided in the dataset.

**Data Splitting:** To facilitate model development and evaluation, the dataset is split into training and testing sets. The training set, comprising 80% of the data, is used to train the machine learning model, while the remaining 20% serves as the test set for assessing model performance on unseen data. This division ensures an unbiased evaluation of the model's generalization capabilities.

**Model Training:** The project employs two machine learning algorithms Random Forest and Support Vector Machine (SVM). Each algorithm is trained on the designated training set, learning patterns and relationships within the data.

**Evaluation Metrics:** Model performance is assessed using relevant evaluation metrics, such as accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of the model's ability to correctly classify creditworthy and non-creditworthy individuals, minimizing both false positives and false negatives.

**Fine-Tuning:** To optimize model performance, hyperparameter tuning is conducted. This involves systematically adjusting model parameters to achieve the best possible results. Techniques such as grid search or random search are employed to explore different combinations of hyperparameter values, enhancing the model's predictive accuracy.

**Model Comparison:** The developed models are compared based on their performance metrics, allowing for the selection of the most effective model for credit risk assessment. Considerations include both the model's ability to identify creditworthy individuals and its generalization to new, unseen data.

**Target Variable meaning:**  
0: Non-default - The borrower successfully repaid the loan as agreed, and there was no default.

1: Default - The borrower failed to repay the loan according to the agreed-upon terms and defaulted on the loan.

**Random Forest Results:**

The Random Forest model achieved an impressive accuracy of 92.46%. This suggests that the model can correctly classify borrowers into default and non-default categories in approximately 92.46% of cases. The classification report provides additional insights into precision, recall, and F1-score for both classes (0: Non-default, 1: Default).

* Precision for class 0 (non-default) is high at 92%, indicating a low rate of false positives.
* Precision for class 1 (Default) is also commendable at 93%, suggesting accurate identification of individuals at risk of default.
* Recall for class 0 is very high (99%), indicating the model's ability to capture a large proportion of non-default cases.
* Recall for class 1 is lower (71%), indicating that the model may miss some instances of actual defaults.
* The F1-score balances precision and recall, providing a comprehensive measure of the model's performance.

The confusion matrix visually represents the model's classification results, showing that it misclassified 69 instances of non-default as default and 409 instances of default as non-default.

The ROC curve illustrates the trade-off between true positive rate and false positive rate, with an area under the curve (AUC) of 0.96, indicating a high discriminatory ability of the model.

**Support Vector Machine Results:**

The Support Vector Machine (SVM) model achieved an accuracy of 90.09%. While slightly lower than the Random Forest model, it still demonstrates a reasonable level of predictive performance. The classification report reveals:

* High precision for class 0 (non-default) at 90%.
* Precision for class 1 (Default) is also good at 93%.
* High recall for class 0 (99%).
* Lower recall for class 1 (60%).

The confusion matrix shows 59 instances of misclassified non-default as default and 569 instances of misclassified default as non-default.

The ROC curve for SVM has an AUC of 0.94, indicating good discriminatory power.

## CONCLUSIONS

Comparing the results of the Random Forest and SVM models, it is evident that the Random Forest model outperforms the SVM model in terms of accuracy, recall for class 1 (Default), and the AUC of the ROC curve. The Random Forest model's ability to capture a higher proportion of true positives for defaults suggests that it may be more suitable for credit risk assessment in this context.

This comparison aligns with the well-documented versatility of Random Forests in handling complex datasets and capturing intricate patterns. The ensemble nature of Random Forests, combining multiple decision trees, often results in robust models that generalize well to unseen data.

However, it is crucial to note that model selection depends on various factors, including interpretability, computational efficiency, and the specific requirements of the credit risk assessment task.

In conclusion, the machine learning models presented in this project, particularly the Random Forest, showcase promising results in enhancing credit risk assessment. Academic research supports the use of ensemble methods for similar tasks, emphasizing the importance of exploring multiple models and metrics to make informed decisions in real-world applications.

## GITHUB REPOSITORY

Access to the GitHub repository at:

<https://github.com/sba22052/cct/blob/62acba171a5213474e0f8249fd25deddb0d5a872/CA2%20-%20Enhancing%20Credit%20Risk%20Assessment%20through%20Machine%20Learning.ipynb>

Link for lecturer invitation to the GitHub repository:

<https://github.com/sba22052/cct/invitations>

## REFERENCES

Towardsdatascience.com (2021), How to Prepare Data for Credit Risk Modeling, Available at: <https://towardsdatascience.com/how-to-prepare-data-for-credit-risk-modeling-5523641882f2> (Accessed 10 December 2023)

Investopedia.com (2023), How to Quantify Credit Risk, Available at: <https://www.investopedia.com/ask/answers/022415/what-factors-are-taken-account-quantify-credit-risk.asp> (Accessed 10 December 2023)

Corporatefinanceinstitute.com (2023), Credit Risk Analysis, Available at: <https://corporatefinanceinstitute.com/resources/commercial-lending/credit-risk-analysis> (Accessed 11 December 2023)

w3schools.com (2023), Machine Learning - Confusion Matrix, Available at: <https://www.w3schools.com/python/python_ml_confusion_matrix.asp> (Accessed 09 December 2023)

Towardsdatascience.com (2021), Understanding the Confusion Matrix from Scikit learn, Available at: <https://towardsdatascience.com/understanding-the-confusion-matrix-from-scikit-learn-c51d88929c79> (Accessed 11 December 2023)