WGU C951

Task 3

Machine Learning: Loan Prediction

Jazmin Rodriguez Arevalo

Student ID #: 011371511

10/13/2024

Table of Contents

[A. Project Overview 3](#_Toc181136463)

[A.1. Organizational Need 3](#_Toc181136464)

[A.2. Context and Background 3](#_Toc181136465)

[A.3. Outside Works Review 4](#_Toc181136466)

[A.4. Solution Summary 7](#_Toc181136467)

[A.5. Machine Learning Benefits 7](#_Toc181136468)

[B.1. Scope 8](#_Toc181136469)

[B.2. Goals, Objectives, and Deliverables 9](#_Toc181136470)

[B.3. Standard Methodology 10](#_Toc181136471)

[B.4. Projected Timeline 11](#_Toc181136472)

[B.5. Resources and Costs 12](#_Toc181136473)

[B.6. Evaluation Criteria 13](#_Toc181136474)

[C. Machine Learning Solution Design 13](#_Toc181136475)

[C.1. Hypothesis 13](#_Toc181136476)

[C.2. Selected Algorithm 14](#_Toc181136477)

[C.2.a Algorithm Justification 14](#_Toc181136478)

[C.2.a.i. Algorithm Advantage 15](#_Toc181136479)

[C.2.a.ii. Algorithm Limitation 15](#_Toc181136480)

[C.3. Tools and Environment 15](#_Toc181136481)

[C.4. Performance Measurement 16](#_Toc181136482)

[D. Description of Data Sets 16](#_Toc181136483)

[D.1. Data Source 16](#_Toc181136484)

[D.2. Data Collection Method 17](#_Toc181136485)

[D.2.a.i. Data Collection Method Advantage 17](#_Toc181136486)

[D.2.a. ii. Data Collection Method Limitation 17](#_Toc181136487)

[D.3. Quality and Completeness of Data 17](#_Toc181136488)

[D.4. Precautions for Sensitive Data 18](#_Toc181136489)

[References 19](#_Toc181136490)

# A. Project Overview

This project aims to develop a machine learning (ML) system that predicts loan approval for applicants. By analyzing various applicant details and financial information, the ML model will identify critical indicators of loan eligibility. This should help loan officers make decisions faster and more accurately, reducing the need for manual review. Overall, the system aims to save time and improve decision-making. With faster loan eligibility, loan officers can help future homeowners get into their dream house or get loans for any needs.

## A.1. Organizational Need

Our organization is a new online lender and bank that needs a loan eligibility model to help them process more applications and improve accuracy. Currently, loan officers manually assess applicants' financial details, which can be time-consuming and prone to human error. A machine learning model can facilitate the process by analyzing historical data and predicting whether a loan applicant will likely repay the loan. This will reduce manual errors, improve decision-making efficiency, and ensure consistent results.

## A.2. Context and Background

The financial industry relies heavily on accurate loan approval systems to minimize risks while offering services to as many people as possible. There is a balance between risk and improving the number of borrowers. Our organization currently manually assesses loan applications and determines eligibility, which has led to errors and giving loans that resulted in financial losses. Like many others, our organization uses rule-based systems, which need more flexibility when encountering new patterns in customer data. One borrower might not qualify due to specific rules but can pay the loan, and the risk is minimal. By introducing machine learning, the lender can utilize large datasets to build models that recognize complex patterns and improve overall decision-making. The use of predictive models in finance has grown significantly, as seen in many modern banks and lending institutions. This will ultimately make the lender more competitive and improve customer satisfaction.

## A.3. Outside Works Review

Three outside works related to loan eligibility were used to further research loan eligibility machine learning models. They are used to help develop a model to offer loans efficiently.

**Work 1: “Explainable prediction of loan default based on machine learning models” by Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L.**

This research article talks about how machine learning has been used in loan risk assessment, exploring the effectiveness of each algorithm. It compares several algorithms—logistic regression, decision trees, XGBoost, and LightGBM—to determine the most effective. The findings show that LightGBM and XGBoost outperform the other methods. However, the study highlights a significant issue known as the "black box", where models may provide accurate predictions but lack transparency. To address this, the authors use a technique called Local Interpretable Model-agnostic Explanations (LIME) to clarify how specific factors impact predictions.

This research is beneficial for developing a machine learning model for loan eligibility because it identifies the best-performing algorithms and emphasizes the importance of model explainability. By understanding which borrower characteristics are critical in assessing loan repayment risk, decision-makers can create more informed and reliable loan approval processes. The article also points out the limitations of LIME, such as high computational costs and local interpretability issues, but suggests that combining predictive and explainable models can lead to more effective risk management strategies in the future. It provides a solid foundation for building a transparent and trustworthy loan eligibility assessment system.

**Work 2: “An Innovative Approach to Predict Loan Eligibility of a Customer in Bank by Comparing Random Forest Algorithm over Logistic Regression in terms of Accuracy” by Sandeep, C. V., & Devi, T.**

This article compares the Random Forest Algorithm (RT) with Logistic Regression and finds a slightly better performance with RT. It highlights the importance of accurate and efficient credit risk assessment in the financial sector, as factors like bankruptcy and family obligations are common challenges for lenders. By leveraging real-time data and analyzing trends, banks can make more informed decisions, potentially reducing the rate of defaults.

The research finds that Random Forest performs slightly better than Logistic Regression, with a mean accuracy of 70.5% compared to 69.5%. This improvement, though modest, suggests that Random Forest's ability to handle complex data patterns and non-linear relationships can provide an advantage in predicting loan eligibility. Combining multiple decision trees, Random Forest's ensemble approach helps create a more robust and accurate model, especially in real-time data and varied customer profiles.

This relates to loan eligibility, as it provides a structured approach to building a predictive model that banks could use to facilitate loan approval processes. Using a dataset with relevant customer attributes, like the Kaggle dataset utilized in the study, the organization can apply and compare these algorithms to determine the best fit.

**Work 3: “A Deep Neural Network (DNN) based classification model in application to loan default prediction” by Bayraci, S., & Susuz, O.**

his article discusses the use of neural networks in predicting loan defaults, highlighting the strengths of deep learning methods in handling large, complex datasets. It emphasizes the importance of banks distinguishing between good and bad customers to manage risk effectively. While traditional methods like logistic regression and support vector machines have been standard in credit risk assessment, recent studies suggest that machine learning techniques can improve predictive accuracy. However, these methods often need to capture valuable confidence levels in their predictions. The article explores using Deep Neural Networks (DNN) with multiple hidden layers to evaluate loan clients' risk profiles, comparing their performance with traditional methods.

Using two datasets from a Turkish commercial bank, the study shows that the DNN model significantly enhances credit scoring performance, particularly in larger datasets. It outperforms logistic regression and support vector machines but may not provide significant advantages in smaller datasets, where simpler models can be more effective. Implementing deep learning algorithms can be complex and requires careful tuning of parameters. Overall, the study highlights the potential of deep learning to enhance credit scoring.

## A.4. Solution Summary

Our organization is a relatively new lender in comparison to large banks. It is around nine years old and is still growing. As the articles show, many models can be adopted, but the complexity and time it takes to establish the algorithms can be costly. Logistical regression is the best choice for this organization. There is a reason why logistic regression is the standard. DNN and RT do perform better, but it is a small amount. As for LightGBM and XGBoost, although they are better than logistical regression and even decision tree models, the complexity of the models can hinder the fast development of a loan eligibility model. Logistical regression still improves the efficiency of loans and reduces human error. The computation for logistic regression is inexpensive and has a high speed, which our organization needs.

## A.5. Machine Learning Benefits

Logistic regression is a statistical machine-learning technique that predicts the probability of an event occurring based on a given data set. The logistic regression model is more effective in loan approval than manually assessing financial details and a loan officer deciding loan eligibility. The statistical model will lead to better prediction accuracy than manual assessments because there is less room for errors. The automation will reduce the time loan officers spend on manual review, speeding up the loan approval process. Many large and small organizations use logistic regression models for loan eligibility, so it is evident that the model can easily handle an increasing volume of loan applications as the organization grows. The model works best with smaller data sets, but it still works with more extensive data sets. In conclusion, the main benefits are effectiveness, improved accuracy, and scalability.

**B. Machine Learning Project Design**

## B.1. Scope

**In Scope:**

* Data Preparation: Cleaning data and preparing for logistic regression model
* Model Development: Implementing logistic regression model
* Accuracy Assessment: Determine the accuracy of the model using performance metrics
* Integration: Integrating the organization's website to give immediate loan eligibility
* Documentation: Document model development

**Out of Scope:**

* Continuous monitoring and maintenance after deployment (beyond initial launch support)
* This is out of scope because continuous maintenance would require a change in the deadline, and it is unnecessary for the overall loan model.
* Comparisons of other models
* Handling fraud: The model will not determine if the information given by the borrower is accurate.

**Project Members:**

* Software developers
* Project Lead
* Data Scientist

## B.2. Goals, Objectives, and Deliverables

Goals

* Improve the efficiency of the loan approval process.
* Increase prediction accuracy to reduce approval errors.
* Reduce processing time by reducing manual processing.

Objectives

* Achieve a prediction accuracy of 85%.
* Increase throughput by ensuring the system can process at least 150 applications per hour
* Reduce loan processing time by 50%.

Deliverables

* Develop a Logistic Regression Model: The model must be trained and validated.
* A dashboard displaying loan predictions: There must be a web application where stakeholders can input data and receive predictions of loan eligibility.
* Documentation for model deployment: There must be a detailed documentation of model performance and user instructions.

## B.3. Standard Methodology

CRISP-DM (Cross-Industry Standard Process for Data Mining) will be implemented in this project:

**• Business Understanding:**

The objective is to define project goals in terms of loan eligibility and understand business constraints. To implement, there will be meetings with stakeholders to define objectives and expectations and the most important criteria of the model’s performance will be identified.

**• Data Understanding:**

The objective is to collect data required for project. To implement, historical loan application data will be gathered and perform exploratory data analysis (EDA) to understand and interpret the data. This will help identify patterns and detect outliers as well as find correlations.

**• Data Preparation:**

The objective is to prepare the data for the model. To implement, the data set will be cleaned to ensure data quality, the data will be split into training and test sets and convert categorical variables into binary format.

**• Modeling:**

The objective is to develop the logistic regression model. To implement, the model will be trained with the training dataset and tunning parameters to enhance performance.

**• Evaluation:**

The objective is to assess the model’s performance and validity. To implement, the model will be evaluated by using metrics such as accuracy, ROC curves, and precision. There will also be cross-validation.

**• Deployment:**

The objective is to deploy the model. To implement, the model will be implemented int the web application for user interaction and the necessary infrastructure for hosting will be added.

## B.4. Projected Timeline

November 5, 2024 – The proposal is accepted.

December 29, 2024 – The objective is defined and data is prepared for the model.

February 3, 2025 – Model is developed and evaluated.

March 21, 2025 – Deployment, testing and documentation are done.

March 26, 2025 – Delivered. The project is approved and launched.

**Sprint Schedule**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Tasks** |
| 1 | November 1, 2024 | November 5, 2024 | Project Approval |
| 2 | November 6, 2024 | November 13, 2024 | Stakeholder meetings for business understanding |
| 3 | November 14, 2024 | December 29, 2024 | Data preparation: collection and cleaning of data |
| 4 | December 30, 2024 | January 20, 2025 | Model Development |
| 5 | January 21, 2025 | February 3, 2025 | Model evaluation |
| 6 | February 4, 2025 | February 28, 2025 | Deployment |
| 7 | March 3, 2025 | March 14, 2025 | Testing |
| 8 | March 15, 2025 | March 21, 2025 | Documentation |
| 9 | March 22, 2025 | March 26, 2025 | Final Approval and Launch |

## B.5. Resources and Costs

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Software | Python and libraries | $0 |
| Web hosting | AWS | $1000 |
| GPU Server | These are used for faster model training. | $1600 |
| Labor | A project lead, data scientist, and software developer are needed to finish the project. We will assume that the project lead is $80 an hour and will need 20 hours, the data scientist will be the same, and the software engineer will do most of the work in 150 hours at a rate of $65. | $12,950 |
| Miscellaneous | These are cost like server maintenance. | $600 |
|  | **Total** | $16,150 |

## B.6. Evaluation Criteria

Describe the criteria used to evaluate and measure the success of the completed project.

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Accuracy | Model prediction accuracy of 85% or higher |
| Throughput | Applications processed per hour must be 150 or more |
| Loan Processing Time | Reduce loan processing time by 50% |

# C. Machine Learning Solution Design

## C.1. Hypothesis

The proposed hypothesis for this loan prediction project is: "Applicants with certain financial and demographic profiles, such as stable income, good credit history, and secure employment, are more likely to be eligible for loans." By analyzing historical loan data, a logistic regression model can be developed to predict loan eligibility with at least 85% accuracy. The logistical regression model will be trained with historical loan data, including applicant features like income, credit score, loan amount, and employment status, to test the hypothesis. The model accuracy will be tested with cross-validation to confirm that the model's predictions align with actual loan approvals and rejections.

## C.2. Selected Algorithm

The chosen algorithm for this project is Logistic Regression, a supervised learning algorithm well-suited for binary classification tasks like loan approval. Logistic regression is known for its simplicity, interpretability, and efficiency.

### C.2.a Algorithm Justification

*Explainable prediction of loan default based on machine learning models* states, "This type of computation is inexpensive, its speed is high, and the required storage resources are low." This is talking about the logistics regression model. It is a simple model that still has high speeds. Although there are faster models, the complexity will make the project more time-consuming and expensive. More hours would be dedicated to making other models, such as Deep Neural Networks or LightGBM, making the project over budget. Our organization is still new and cannot allocate too much of its budget for a model that will not significantly improve the accuracy. Logistic regression is also well suited for the binary outcome (approved and rejected). Overall, the model is chosen because it fits the project's scope.

### C.2.a.i. Algorithm Advantage

The number one advantage of the logistical regression model is it simplicity. The model is easy to interpret and provides probability scores that make it ideal for explaining to stakeholders and accessing the risks for each borrower.

### C.2.a.ii. Algorithm Limitation

One of the significant limitations is that logistical regression assumes linear relationships between features and outcomes, which might affect its effectiveness without a robust linear relationship. Although there is this limitation, many large institutions still rely on logistical regression because it provides adequate effectiveness.

## C.3. Tools and Environment

**Programming Language:** Python, with libraries such as Scikit-Learn for model development, Pandas for data manipulation, and NumPy for numerical operations.

**Operating System:** Any operating system that supports Python can be used

**Development Environment:** Jupyter Notebook, facilitating code development and data exploration.

**Deployment:** Amazon Web Services (AWS) for hosting the model

**Database:** MySQL stores loan data and tracks new loan applications.

The organization will use the same website that is already used, React, but they will integrate the model.

**Minimum Hardware Requirements:** At least 8GB of RAM (16GB is preferred for efficient processing), a six-core or higher processor for CPU, 256GB SSD Storage with additional storage if there is a large dataset, a dedicated GPU for faster training time, and 100GB of disk overhead for initial setup, but 200GB is preferred for a scalable model

## C.4. Performance Measurement

There will be a training phase and a baseline accuracy to measure performance. During training, performance will be monitored using cross-validation on the training dataset. Metrics like accuracy, precision, recall, and ROC curves will provide insights into how well the model fits the data. The baseline accuracy is a way to measure how well the model performs by comparing it to a theoretical baseline. The baseline will be that all applicants are eligible. To determine accuracy, the logistical regression model must exceed the accuracy of the baseline. The qualitative metrics have an accuracy of 85% or higher. Another is the throughput. The model should process a minimum of 150 applications per hour. The last metric is that the loan processing time must be reduced by at least 50% compared to the old manual processing. The time will be compared and alkalized to ensure the model performance meets all requirements.

# D. Description of Data Sets

## D.1. Data Source

The data will be extracted from the organization's historical loan data. The data set will include age, income, employment status, loan amount, credit score, and loan outcome (approved or rejected). Since the organization is relatively new, public datasets are open access from public finance research institutions if they need more data. The second option is a last resort. If possible, the organization must try to stick to its data.

## D.2. Data Collection Method

SQL queries or API requests will extract the data from the organization's database. The data can also be collected in batch mode to pull historical data periodically. Additional data points may be collected from other relevant third-party sources to enhance the dataset.

### D.2.a.i. Data Collection Method Advantage

Since most of the data comes from the organization, it will be relevant and reflect the specific loan approval standards of the organization. This makes the model exclusive to the organization.

### D.2.a. ii. Data Collection Method Limitation

The main limitation is that the data needs to represent future applicants. The data does not reflect real-time conditions, so a lack of recent data may affect the model's accuracy. The organization is also relatively new, so the data is limited and needs to be more diverse. To fix this problem, data can be added from third parties if the data is not accurate enough.

## D.3. Quality and Completeness of Data

To prepare the data, it is cleaned. The ETL process extracts the data, transforms it by cleaning it, and loads it to the model. The data is formatted so that all fields align with the logistic regression algorithm. This is where categorical variables are converted into binary format. Some fields may be irrelevant, so they will be removed to improve efficiency. The missing values will be identified and filled in with appropriate data to handle missing data. If the data cannot be located, the missing values will be filled in with the median or mean, depending on the field. If the field is not required, it can also be removed. Outliers will be removed to prevent disproportionally influencing the model. Dirty data will be corrected, and null values will be filled in without introducing bias. Personally identifiable information (PII) is not needed to determine eligibility so that it can be anonymized or removed. The only information that might be needed is the address, which can be just the state instead of the whole address. The data is rechecked to ensure consistency, and the cleaning process is done.

## D.4. Precautions for Sensitive Data

Sensitive data is present in the dataset, so policies will be established to ensure data security and privacy. There will be access control of the data, meaning only authorized personnel are involved with the project. Data encryption will prevent unauthorized access, and PII will be removed or anonymized to comply with data protection laws. There will also be employee agreements and training, during which employees must sign a non-disclosure agreement to ensure confidentiality and establish consequences of data misuse. Before signing the agreement, employees are trained to reduce accidental exposure and understand data privacy policies. The organization's primary sensitive data will be anonymized or removed, so most data privacy worries are before it is anonymized or removed.

# References

Bayraci, S., & Susuz, O. (2019). A Deep Neural Network (DNN) based classification model in application to loan default prediction. Theoretical and Applied Economics, 4(621), 75–84.

Sandeep, C. V., & Devi, T. (2022). An Innovative Approach to Predict Loan Eligibility of a Customer in Bank by Comparing Random Forest Algorithm over Logistic Regression in terms of accuracy. Journal of Pharmaceutical Negative Results, 13, 1741–1747. <https://doi.org/10.47750/pnr.2022.13.S04.210>

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. Data Science and Management, 6(3), 123–133. <https://doi.org/10.1016/j.dsm.2023.04.003>