|  |
| --- |
|  |

|  |
| --- |
| Electrical & Computer Engineering & Computer Science (ECECS) |

|  |  |
| --- | --- |
| Fall 2024 |  |



Contents

[Loan Eligibility Prediction 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on:4/23/2024 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

Introduction…………………………………………….4

Review of Research………………………………….5

Methodology…………………………………………..6

Result Section………………………………………….7

Conclusion…………………………………………….14

|  |
| --- |
| Loan Eligibility Prediction |

|  |  |  |
| --- | --- | --- |
| Executive Summary In our academic undertaking, Loan Eligibility Prediction, my colleagues and I have endeavored to demystify loan approval process in light of contemporary approaches to machine learning. As students, we faced the challenge head-on: to minimize the time and effort consuming and in many cases inaccurate verdicts that characterize the current financial lending systems. In this work, we started from a dataset obtained from Kaggle that contains 4,269 data points after preprocessing. It provided an opportunity to assess its efficacy of the different algorithms such as Logistic Regression and Random Forest based on the accuracy as well as the F1-Score.  The result of our work is a web application running on AWS EC2 instance, representing the essence of the modern generation’s approach to technology and finance. The application is our proof of concept and call to a possible future where eligibility for loans is clear, efficient, and based on the available data. person at a table writing in a notebook with people around | | |
|  | | |
| **Team Members:**  **Mallikarjuna Reddy Bandi**  **Yeshwanth Labba**  **Nikitha Yarlagadda** |  |  |

|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
| Highlights of Project The result of our work is a web application running on AWS EC2 instance, representing the essence of the modern generation’s approach to technology and finance. The application is our proof of concept and call to a possible future where eligibility for loans is clear, efficient, and based on the available data. Submitted on: 12/08/2024 |  |
|  |

Introductory Section

In the landscape of financial services, the loan approval process stands as a critical gateway to economic growth and personal advancement. However, this gate is often guarded by outdated, inefficient methods prone to error and bias, leaving many potential borrowers stranded on the threshold of opportunity. Our project, "Loan Eligibility Prediction," addresses this pivotal challenge by leveraging the precision and efficiency of machine learning (ML).

Rooted in the discipline of data engineering, our initiative is not just a theoretical exploration, but a practical solution designed to transform voluminous data into actionable insights. We've tapped into a wellspring of potential in the form of a rich dataset sourced from Kaggle, which reflects real-world complexities of loan eligibility criteria. Through a meticulous process of data preprocessing, feature engineering, and model training, our team has translated this raw information into a refined tool that predicts loan eligibility with remarkable accuracy.

Our goal was to build a bridge between the technical sophistication of ML algorithms and the real-world needs of both lenders and borrowers. The resulting web application embodies a synergy of advanced analytical models—such as Logistic Regression, Decision Trees, KNN Classifier, Random Forest, and XGBoost—and a user-centric interface, all hosted within the robust framework of AWS EC2 cloud services. This strategic deployment ensures our solution's scalability and accessibility.

By harnessing the capabilities of machine learning within the realm of data engineering, we aim to dismantle the traditional barriers that have long hindered the loan approval process. The forthcoming sections will unravel the complexities of our approach, elucidating how the sophisticated interplay of algorithms and data orchestration culminates in a solution poised to transform the credit landscape.

Review of available research

Research in financial technology emphasizes the growing application of machine learning (ML) to enhance credit assessments. Studies highlight that traditional methods struggle with complex datasets, a gap that ML techniques like Random Forest and XGBoost are filling by offering more accurate predictions (Johnson et al., 2018; Khan et al., 2021).

Efficient data processing and robust pipelines are essential for ML deployment, directly influencing model performance (Lee & Yoon, 2019). The integration of these models into cloud platforms like AWS EC2 enables scalable and efficient computing, aligning with modern operational needs in finance (Green, 2020).

This literature underscores the relevance of our project's approach, combining advanced ML algorithms and cloud technology to innovate loan eligibility processes.

References:

* Johnson, R. et al. (2018). Finance Research Letters, 28, 317-323.
* Khan, S. et al. (2021). Journal of Risk and Financial Management, 14(7), 301.
* Lee, J., & Yoon, W. (2019). AI in Finance, 7(1), 13-22.
* Green, B. (2020). Computational Economics, 55(4), 1101-1119.

## 

GitHub Link: [https://github.com/DistributedandScalableEngineeringTeam9/Loan- Eligibility-Prediction](https://github.com/DistributedandScalableEngineeringTeam9/Loan-%20Eligibility-Prediction)

## Methodology

We are using the CRISP-DM approach, which focuses on the proper handling of data, and modeling. We used libraries such as Pandas and Scikit-learn for data preprocessing and training and Apache Airflow for automation of the processes. The models which were tested included Logistic Regression, Random forest, XGBoost and were judged on parameters such as accuracy, F1 score and other parameters. The deployment on AWS guarantees real time scalability and reliability.

The “Loan Eligibility Prediction” is our work that we base on the CRISP-DM framework to develop a guideline for the complex process of predictive modeling. Each phase of the methodology has been executed to ensure the development of a robust predictive system:

Business Understanding: We began our project by exploring the financial domain to appreciate the difficulties within the loan approval environment including high levels of defaults and tedious processing time. The idea was to come up with a solution that would eliminate some of these problems in order to help both lenders and borrowers.

Data Understanding: For this study, data was obtained from Kaggle containing 4,269 datapoints and 13 diverse features: The data quality and the features’ relevance were then assessing appropriately. This dataset contains some of the most important credit scores indicators, and was selected for its density and accuracy, which became the base for our models.

Data Preparation: The data engineering phase we went through ware mostly a series of cleaning and formatting processes. We also stripped the features of any outliers such as leading spaces in the column names and also encoded the categorical features. Numerical values were normalized so that the data was consistent for use in our machine learning models, which is crucial to achieve success.

Modeling: Subsequently, we set up and optimized various types of the Machine Learning models such as Logistic Regression, Decision Tree, K Nearest Neighbor Classification, Random Forest, and XGBoost using prepared data. Each of the models was selected based on its applicability to capture various patterns and interactions in the data presented to select the best model to use for our application.

Evaluation: To see if the models could correctly identify the eligibility of the loan, the models were benchmarked against a range of performance indicators—accuracy, precision, recall, and F1-Score. We used cross-validation techniques since these helps to avoid overfitting and also helps to check the sanity of the models.

Across these stages, care was taken to ensure that the data engineering activities provided a good foundation for the ML models and that there was clean transfer from raw data to the ML models. The product is a system hosted on AWS EC2 that provides highly accurate loan eligibility predictions while also supporting variable usage.

As such, our findings show the resilience of using ML for decision making. From the comparative assessment of the various algorithms, two Algorithm; Random forest and XGBoost gave the best results, which included accuracy and F1 score. In our solution, we use real-time data updates, feature importance analysis, and AWS deployment to demonstrate that our solution works well. For more understanding, feature impact is displayed as a graph while ROC curve is displayed.

## Results Section

The Loan Eligibility Prediction project gave encouraging findings that would corroborate the benefits of data engineering and machine learning in altering the financial decision making system. All of our findings are based on a robust analysis completed within a Jupyter notebook where we employed an iterative and reproducible approach to data analysis. The findings are summarized as follows:

Data Engineering Pipeline: Thus, our data pipeline was built with care and consists of the following components:

\* Data Ingestion: Using tools of Python, like Pandas, to read and analyze data sets.

\* Data Storage: Using durable storage technologies like Amazon S3 for processing big datasets at a great scale.

\* Data Processing: Using of tools that are available in Scikit-learn; for example encoding and normalization of the data.

\* Data Consumption: Flask-based web application is designed for model interaction and offers real-time data predictions as a result of the user input.

Model Deployment: Our models were deployed using AWS EC2 platform, this provided a highly reliable and elastic base. This guaranteed uninterrupted access and real time point prediction.

Data Visualization: Crucial insights were visualized using Matplotlib and Seaborn, and these visualizations can be found in the appendix section of this report. Key graphs include:

* Feature Importance Plots: Showcasing the variables with the most significant impact on loan eligibility.

A diagram of a software development process

Description automatically generated

Model Performance Metrics : A comparison of the degree of correctness, clearness, specificity, sensitivity, and F1-statistic indicators for different models.ROC Curves: Comparing the accuracy of the models in terms of true positive and false positive.

A graph of a function

Description automatically generated with medium confidence

Empirical FindEmpirical Findings: From the actual analysis, it was found that for the following:

Random Forest and XGBoost were more accurate, and produced higher F1 score compared to other models showing there aptness for predictive tasks.

\* Relative importance of the features is different, and these are the features bearing more weight; credit history, income, and loan amount are some of the significant factors to determine eligibility to the loan.A chart with numbers and a red and blue squares

Description automatically generated with medium confidence

Cross-validation scores further reflected the fact that the above models were capable of being generalized.

A graph of a graph with numbers and a number

Description automatically generated with medium confidence A diagram of a tree confusion matrix

Description automatically generatedA graph of a forest confusion matrix

Description automatically generated

Web Application Output: Web Application Output: The last result of our project is the web application which is the front-end to the raw output of our predictive model. The last two screenshots showing the interface of the applications and the real-time prediction values.

A computer screen shot of a loan application

Description automatically generated

A screenshot of a computer screen

Description automatically generated

## Conclusion

The ‘Loan Eligibility Prediction’ project best illustrates how the practice of machine learning can revolutionize the process of lending decisions. Embedded into scalable data processing and delivered on AWS, the sophisticated algorithms of the solution improve performance and customer confidence. Subsequent releases will consider live models and stream processing integrated into APIs, which will show our focus on progression.

The “Loan Eligibility Prediction ” project signifies a great advancement in the facilitation and accuracy of loans approval process. Some strategic decisions our team took to achieve the best-of-class solution include the use of sophisticated machine learning algorithms and a highly targeted data engineering workflow, which are capable of delivering accuracy and speed beyond existing state of the art.

It has emerged from our analyses that the integration of algorithms such as the Random Forest and XGBoost into a data engineering pipeline significantly improves the loan assessment process compared to conventional approaches. The final web application shown in the Appendices D involves a friendly user interface that makes real-time loan eligibility assessment possible for both the financial institutions and customers.

We understand that despite significant development, the models we used in our project are resistant to future progression in machine learning schemes, so our system anticipates and can incorporate future improvements in predictive analytics. It is more likely that in the successive versions of our application we will use real time data streams and more complex models to improve the forecasts.

June 24, 2019 The help and cooperation of the data science and engineering community has been vital for the accomplishment of this project. We are also grateful to those who offered their comments and advice and to those who contributed to the technical aspects of this process.

There are 4C candies of research directions that would follow and build upon our proposed studies. Among them are ways to modify the amount and quality of features by using additional data sources, the use of neural networks for creating deep learning predictions, and the possibilities of applying our predictive system to other industries.

Therefore, the “Loan Eligibility Prediction” project cannot be considered as the path end but rather as the start of the path that will lead to a more effective and open financial environment supported by data science tools and methods.