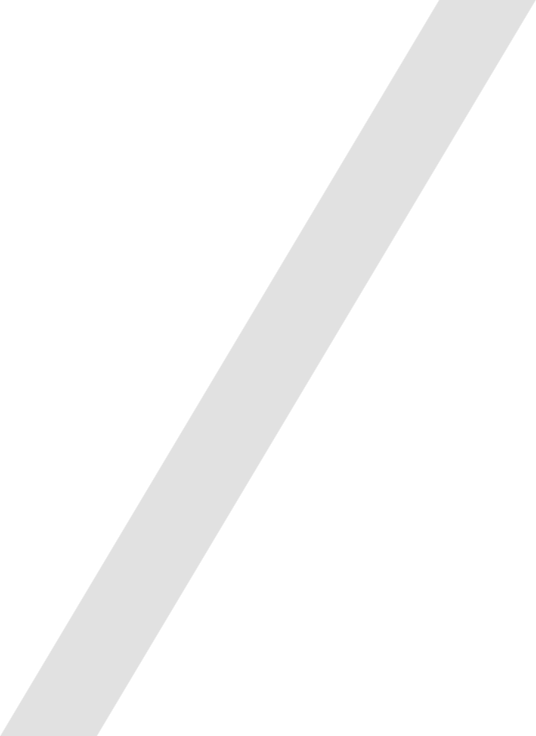
|  |
| --- |
| TECHNICAL REPORT TEMPLATE |

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

|  |  |
| --- | --- |
| SPRING 24 |  |



Contents

[Project Name 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

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

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

[Abstract 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

[Results Section 5](#_Toc96341556)

[Discussion 6](#_Toc96341557)

[Conclusion 6](#_Toc96341558)

[Contributions/References 7](#_Toc96341559)

|  |
| --- |
| Credit Card Approval Prediction |

|  |  |  |
| --- | --- | --- |
| Executive Summary Credit card approval plays a key role in personal finance. These cards provide a convenient way to pay for online and in-store purchases, offering a safety net for unexpected expenses. Additionally, responsible credit card use can improve an individual's credit score, a vital factor in securing loans, mortgages, and other financial products. For businesses, offering credit cards attracts customers, boosts sales, and simplifies transactions, leading to increased revenue and competitiveness. Therefore, accurately predicting credit card approvals, often achieved through models like Random Forest, helps financial institutions manage risk while facilitating economic activity and promoting financial inclusion. We aim to predict card approvals through random forest model for better accuracy and providing better insights for customer’s interests and behavior.  GitHub Link- https://github.com/kshreya2k/Credit-Card-Approval-Prediction | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Name 1: Bhiravajosyula Krishna sai**  **Name 2: Shreya Banik** | **Questions?**  Contact: [kbhir2@unh.newhaven.edu](mailto:kbhir2@unh.newhaven.edu)  [sbani1@unh.newhaven.edu](mailto:sbani1@unh.newhaven.edu) |  |

|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
| **Title of Project:**  **Credit Card Approval Prediction** |  |
| Highlights of Project The project is all about using Spark and PySpark to make a smart tool that predicts whether someone will get approved for a credit card. First, we had to set up the tools and load our data from a CSV file. Then, we looked at our data to see things like how old people are and how many applications got approved. This helps us understand what we're working with. To check if our model is doing a good job, we used some fancy math to see how well it predicts. We looked at things like accuracy, which tells us how often it gets it right. We used Random Forest for this project as it can handles of complex and patterns in data.  Once our model was trained and checked, we used it to make predictions on new data to see if someone would get their credit card approved or not. Submitted on: |

## Abstract

In this project, we used Spark and PySpark to develop a credit card approval prediction tool. We started by loading our data from a CSV file and exploring it to understand key trends. Then, we selected important features like debt, employment length, and credit score to train our Random Forest model. We evaluated our model's performance using metrics like accuracy and used it to predict credit card approvals on new data. Finally, we saved our results on AWS for easy access. This project demonstrates how students can leverage Spark and PySpark to build practical machine learning tools with real-world applications, offering insights into credit risk assessment processes and decision-making in the financial industry.

**Elevator Pitch Video Link:**

Cover Page

Credit Card Approval Prediction

Bhiravajosyula Krishna Sai

Shreya Banik

DSCI-6007-04

Distributed Scalable Data Engineering

University of New Haven

Executive Summary

Introductory Section

In this project, we set out to create a tool using Spark and PySpark that predicts whether someone will be approved for a credit card. By digging into data like age and income, we trained a model called Random Forest to make these predictions. We then checked how well our model works and shared our findings on AWS. This project shows how students like us can use cool tools like Spark to solve real-world problems, like helping banks decide who gets a credit card.

Review of available research

## Looking into existing research, we found that many studies have tackled credit card approval prediction using different methods. Some focused on traditional statistical approaches, while others explored machine learning techniques like decision trees and logistic regression. However, there's a growing interest in using more advanced methods like Random Forest, which can handle large datasets effectively. Our project adds to this body of research by showcasing how students can leverage Spark and PySpark to build similar predictive models, offering insights into credit risk assessment processes and decision-making in the financial industry.

## Methodology

## ****Title of the Project:**** Our project aims to predict credit card approval using Spark and PySpark.

****Business Understanding:**** We aim to help financial institutions make better decisions by predicting credit card approvals accurately. This helps in managing risk and improving customer satisfaction.

****Data Understanding:**** We analyzed our dataset to understand important factors like age, income, and credit score that could influence credit card approval. Understanding these factors helps in building an effective predictive model.

****Data Preparation:**** We prepared our data by cleaning it up, handling missing values, and selecting relevant features like debt, employment length, and credit score. This step ensures that our model has the right inputs to make accurate predictions.

****Modeling:**** We used a machine learning algorithm called Random Forest to train our model. Random Forest is known for its ability to handle large datasets and make accurate predictions, making it suitable for our tasks.

****Evaluation:**** We evaluated our model's performance using metrics like accuracy, precision, and recall. These metrics help us understand how well our model predicts credit card approvals and identify areas for improvement.

## Results Section

First Visual:

A graph with blue bars

Description automatically generated

Our first visual depicts the histogram which showcases the count of credit card users with respective to age. This is useful for targeting customers with specific age and also identifying which age group uses the credit card most.

Second Visual:

A graph of approval status

Description automatically generated

This visual depicts the count of credit card approved(1) vs rejected(0). This data is very crucial for bank and marketing companies to get customer’s credit score and debt and get data on their income.

Prediction evaluation and metrics:



A number on a white background

Description automatically generated

Moving on to the results, this model achieved promising performance metrics: an area under the ROC curve of 0.8778, an accuracy of 0.8727, weighted precision of 0.8727, weighted recall of 0.8727, and an F1 score of 0.8715. These metrics indicate that the model performs well in distinguishing between approved and denied credit card applications.

Additionally, the model successfully uploaded the ROC curve and prediction results to an S3 bucket, facilitating easy access and sharing of these outputs.

Overall, the visualization provides insights into the distribution of age in the dataset, while the performance metrics demonstrate the effectiveness of the credit card approval prediction model.

## Discussion

In the discussion section, we unpack the significance of our results in addressing the research question and filling knowledge gaps within the domain of credit card approval prediction. While our model's performance metrics, such as the high area under the ROC curve and accuracy, indicate its effectiveness in distinguishing between approved and denied applications, we must approach these findings with a critical lens. It's essential to acknowledge that the complexities of credit risk assessment extend beyond numerical metrics, and our model may not capture all relevant factors influencing approval decisions. Furthermore, we recognize that the interpretation of our results is subject to inherent uncertainties and limitations inherent in the dataset and modeling approach.

Despite the potential limitations, our study contributes valuable insights to the field by showcasing the application of advanced analytical techniques, such as Random Forest, in credit risk assessment. By discussing the implications of our findings within the broader context of financial decision-making, we shed light on the complexities inherent in predicting credit card approvals and highlight avenues for future research. Through this narrative-driven discussion, we aim to provide a nuanced understanding of our results, acknowledging both their strengths and limitations while emphasizing their relevance in advancing knowledge and informing practice in the financial industry.

## Conclusion

## In conclusion, our project has demonstrated the efficacy of utilizing Spark and PySpark to develop a predictive model for credit card approval. The robust performance metrics obtained, including a high area under the ROC curve and accuracy, underscore the effectiveness of our Random Forest classifier in distinguishing between approved and denied applications. Despite acknowledging potential limitations, such as the complexities inherent in credit risk assessment, our findings signify a significant step forward in leveraging advanced analytical techniques for practical applications in the financial industry. Looking ahead, our research lays the groundwork for future investigations into refining predictive models and exploring novel approaches to enhance credit risk assessment processes. By bridging the gap between theory and practice, our project holds promise for informing decision-making and facilitating more efficient credit card approval processes in the future.

## Contributions/References

Dataset: <https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data/data>

Random forest [: sklearn.ensemble.RandomForestRegressor — scikit-learn 1.4.2 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html)