A Credit Analysis Case Study: Exploring the Power of EDA in Mitigating Financial Risk

Angelin Dafini Dhavaraj Jeyachandra, Jiya Jha, Swetha Sathiyakumar

Class of Artificial Intelligence & Machine Learning, Lambton College in Toronto

Toronto, Canada

https://github.com/jiyajha/Data-Mining

https://github.com/dafini18/Data-Mining

https://github.com/Swetha-s95/Data-Mining

***Abstract*— This proposal aims to explore the application of exploratory data analysis (EDA) in risk analytics in the banking and financial services industry. The study will utilize machine learning techniques to analyze a publicly available credit dataset and develop a machine learning model capable of predicting borrower default likelihood. The proposal outlines the motivation, methodology, intended experiments, planning, and milestones for the credit EDA case study.**

Keywords— exploratory data analysis, risk analytics, banking, financial services, machine learning, credit dataset, borrower default, prediction model, machine learning model, case study.

I. INTRODUCTION

The credit EDA case study aims to explore the application of exploratory data analysis (EDA) in the context of risk analytics in the banking and financial services industry. The objective of the study is to develop a basic understanding of how data is leveraged to minimize the risk of financial losses associated with lending to customers.

II. DATASET

https://www.kaggle.com/datasets/dssouvikganguly/application-datacsv

The dataset above is a collection of demographic and financial data on individuals who have applied for loans from a financial institution. The dataset contains information about the loan applicant's gender, age, employment status, education, family status, income, and other relevant factors.

The dataset has a total of 307,511 records and 122 columns. The columns contain a mix of categorical and numerical data, including features such as 'NAME_CONTRACT_TYPE' (whether the loan is a cash loan or a revolving loan), 'AMT_INCOME_TOTAL' (the applicant's income), 'OCCUPATION_TYPE' (the applicant's occupation), and 'NAME_HOUSING_TYPE' (the type of housing the applicant lives in).

This dataset can be used to build predictive models to determine whether an individual's loan application should be approved or denied based on their financial and demographic characteristics. It can also be used for exploratory data analysis to gain insights into the demographics and financial status of loan applicants

III. MOTIVATION

In today's data-driven business environment, financial institutions are increasingly relying on data analytics to make informed lending decisions and mitigate financial risk. However, with the vast amount of data available, it can be challenging to identify meaningful patterns and trends that can inform decision-making. The credit EDA case study aims to address this challenge by demonstrating how EDA can be used to analyze customer behavior and minimize the risk of financial loss.

IV. METHOD

The proposed method for the credit EDA case study is to use machine learning techniques to analyze a publicly available credit dataset. The dataset will be sourced from the UCI Machine Learning Repository, which contains a variety of credit datasets for machine learning research purposes. The study will focus on exploring the data through EDA techniques such as data visualization, statistical analysis, and feature engineering. The goal is to predict the likelihood of a borrower defaulting on a loan, based on a set of key variables such as credit score, income, and debt-to-income ratio.

V. INTENDED EXPERIMENTS

The intended experiments for the credit EDA case study is to develop a machine learning model capable of accurately predicting borrower default likelihood, by utilizing various EDA techniques such as univariate, bivariate, and multivariate analysis of both categorical and numerical features. The study will focus on evaluating the effectiveness of these techniques in identifying meaningful patterns and trends within the data that can inform lending decisions and minimize the risk of financial loss.

VI. PLANNING & MILESTONES

The primary responsibility for each milestone will be shared among the group members – Angelin Dafini, Jiya Jha, and Swetha Sathiyakumar, with each member taking on specific tasks related to the milestone. The group will collaborate on all aspects of the project, including data acquisition, preprocessing, data analysis, and interpretation of results.

The credit EDA case study will be divided into the following milestones:

TIMELINE	TASK	TEAM MEMBERS INVOLVED
WEEK 1	Importing Libraries and reading the data **Description:** In this milestone, we will import the necessary libraries and read the credit dataset from the UCI Machine Learning Repository.	All
WEEK 1	Descriptive Analysis of theRaw Data **Description:** This milestone involves exploring the dataset's structure, identifying the variables' data types, and calculating summary statistics.	All
WEEK 2	Data Cleaning **Description:** In this milestone, we will identify and handle missing values, check for duplicates, and perform other data-cleaning activities.	Angelin Dafini
WEEK 2	Categorical and Numerical features **Description:** This milestone involves exploring the categorical and numerical features in the dataset.	Angelin Dafini
WEEK 3	Univariate Analysis of Categorical columns **Description:** In this milestone, we will perform a univariate analysis of the categorical features in the dataset.	Jiya Jha
WEEK 3	Univariate analysis of Numerical data **Description:** This milestone involves analyzing the numerical features' distribution, central tendency, and dispersion.	Jiya
WEEK 4	Bivariate Analysis **Description:** In this milestone, we will explore the relationship between pairs of variables in the dataset.	Swetha Sathiyakumar
WEEK 5	MultivariateAnalysis **Description:** This milestone involves analyzing the relationship between multiple variables in the dataset.	Swetha Sathiyakumar
WEEK 6	Interpretation ofresults **Description:** In this milestone, we will interpret the results of the EDA and draw conclusions about the dataset's characteristics and the likelihood of borrower default.	All

VII. CONCLUSION

The credit EDA case study aims to demonstrate the application of EDA techniques in the context of risk analytics in the banking and financial services industry. By exploring a publicly available credit dataset and developing a machine learning model to predict borrower default, the study will provide valuable insights into how data can be leveraged to minimize the risk of financial loss. The planned milestones will ensure a systematic approach to the project, and the collaboration among the group members will ensure that the project's objectives are achieved successfully.

VIII. REFERENCES

1. Kaggle Dataset: https://www.kaggle.com/datasets/dssouvikganguly/application-datacsv
2. "Python for Data Analysis" by Wes McKinney - This book provides a comprehensive introduction to data analysis using Python, including a chapter on EDA.
3. "Python Data Science Handbook" by Jake VanderPlas - This book covers a range of topics related to data science, including EDA using Python