INSTITUTE FOR ADVANCED COMPUTING AND SOFTWARE DEVELOPMENT AKURDI, PUNE

Documentation On

"Automate the business loan approval system for central bank using historical data of borrowers"

PG-DBDA September 2023

Submitted by- Group No: 08

Roll No.	Name:
239538	Ruturaj Kore
239539	Shivamkumar Surwase

Mrs. Priti Take	Mr. Rohit Puranik
Project Guide	Centre Coordinator

Abstract

In this article, a large and rich dataset from the U.S. Small Business Administration (SBA) and an accompanying assignment designed to teach statistics as an investigative process of decision making are presented. Guidelines for the assignment titled "Should This Loan Be Approved or Denied?," along with a subset of the larger dataset, are provided. For this case-study assignment, students assume the role of loan officer at a bank and are asked to approve or deny a loan by assessing its risk of default using logistic regression. Since this assignment is designed for introductory business statistic courses, additional methods for more advanced data analysis courses are also suggested.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to everyone who has contributed to the completion of our project.
First and foremost, we would like to thank our project guide **Mrs. Priti Take** madam for their constant guidance and support throughout the project. We extend our sincere thanks to our respected Centre Co-Ordinator, **Mr. Rohit Puranik sir**, for allowing us to use the facilities available.
We would also like to express our appreciation to the faculty members of our department for their constructive
feedback and encouragement. Their insights and suggestions have helped us to refine our ideas and enhance
the quality of our work.
Furthermore, we would like to thank our families and friends for their unwavering support and encouragement throughout our academic journey. Their love and support have been a constant source of motivation and inspiration for us.
Thank you all for your valuable contributions to our project,

Ruturaj Kore (239538)
Shivamkumar Surwase (239539)

Table of Contents

Introduction

The Small Business Administration (SBA) was founded in 1953 to assist small businesses in obtaining loans. Small businesses have been the primary source of employment in the United States. Helping small businesses help with job creation, which reduces unemployment. Small business growth also promotes economic growth. One of the ways the SBA helps small businesses is by guaranteeing bank loans. This guarantee reduces the risk to banks and encourages them to lend to small businesses. If the loan defaults, the SBA covers the amount guaranteed, and the bank suffers a loss for the remaining balance.

There have been several small business success stories like FedEx and Apple. However, the rate of default is very high. Many economists believe the banking market works better without the assistance of the SBA. Supporter claim that the social benefits and job creation outweigh any financial costs to the government in defaulted loans.

Product Scope

To develop a robust and data-driven decision-making system that leverages company demographic data to accurately assess and determine the eligibility of businesses for loans, ultimately aiding the Small Business Administration (SBA) in making informed lending decisions and promoting responsible financial support for small enterprises.

Data Preprocessing and Cleaning

Data Cleaning:

Data cleaning involves handling missing values, outliers, standardizing data, converting categorical variables, correcting inconsistencies, removing duplicates, addressing data integrity issues, and documenting changes, ensuring dataset reliability for accurate predictive modeling.

1. **Handling Missing Values:**

- Identify any missing values in the dataset, which could arise due to incomplete records or data entry errors.
- Decide on an appropriate strategy for handling missing values, such as imputation (replacing missing values with a calculated estimate), deletion of records with missing values, or treating missing values as a separate category.

2. **Dealing with Outliers:**

- Detect outliers in the data, which are data points that significantly deviate from the rest of the observations.
- Decide whether outliers are genuine data points or errors and take appropriate action, such as correcting erroneous values or transforming/extending the model to accommodate outliers if they are legitimate.

3. **Standardizing and Normalizing Data:**

- Standardize or normalize numerical features to ensure that they are on the same scale, which can improve the performance of certain machine learning algorithms.
- Standardization involves transforming the data to have a mean of 0 and a standard deviation of 1, while normalization scales the data to a fixed range (e.g., 0 to 1).

4. Correcting Inconsistent Data:

- Identify and correct any inconsistencies or errors in the data, such as misspelled names or inconsistent formatting.
- This may require manual inspection or the use of automated data validation techniques.

5. Removing Duplicates:

- Identify and remove any duplicate records in the dataset to avoid skewing the analysis or model training process.
- Duplicate records may arise due to data entry errors or multiple entries for the same entity.

6. Addressing Data Integrity Issues:

- Ensure data integrity by validating the relationships and dependencies between different variables in the dataset.
- Check for logical inconsistencies or contradictions that could affect the accuracy of the analysis or predictions.
- Keep track of all changes made during the data cleaning process, including the reasons for each change and any assumptions or decisions made.
- This documentation ensures transparency and reproducibility of the data preprocessing steps.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. Here's a concise explanation:

Exploratory Data Analysis involves visually and statistically exploring datasets to understand their distributions, relationships, and key features, aiding in the identification of patterns, anomalies, and potential insights.

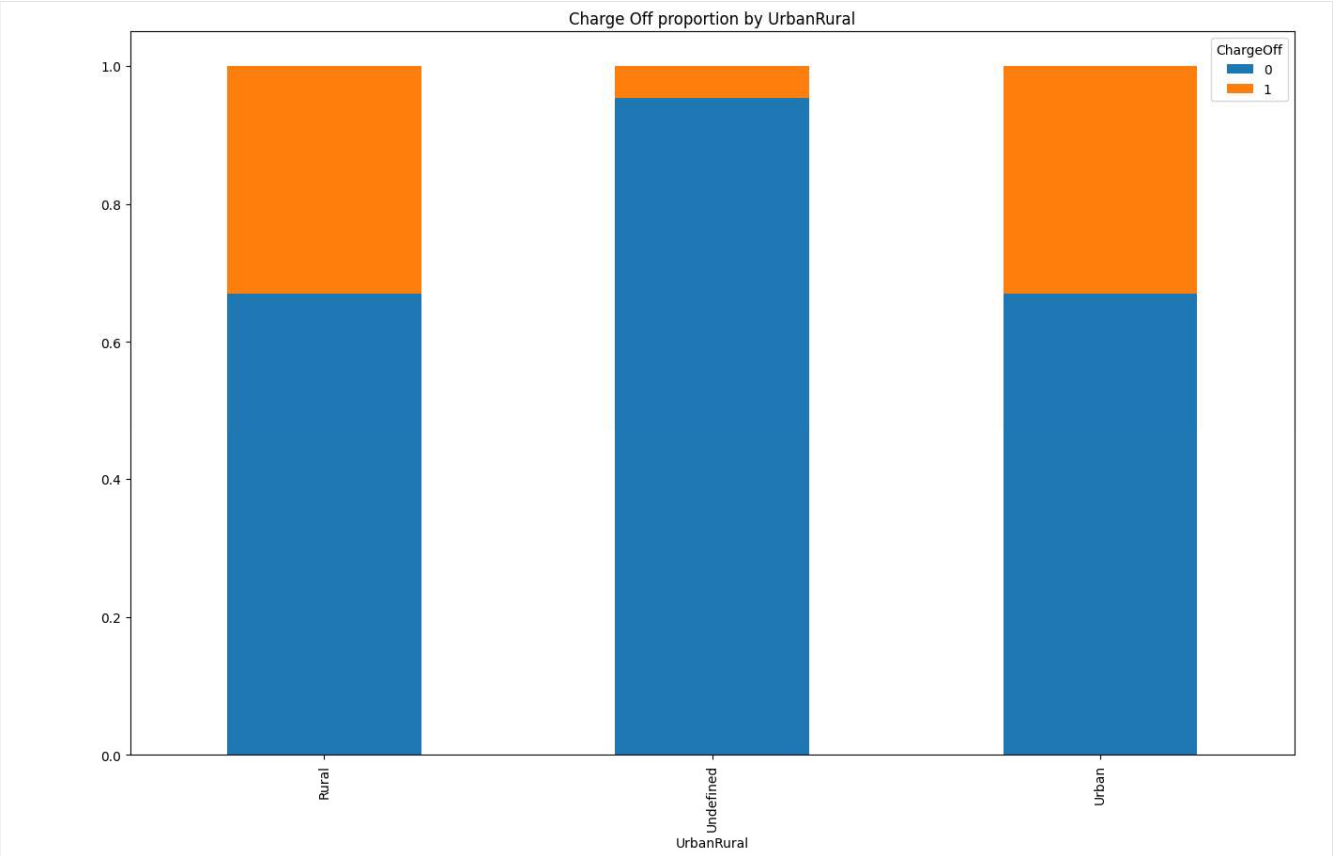

Figure 1 Column UrbanRural has strong relationship with ChargeOff.

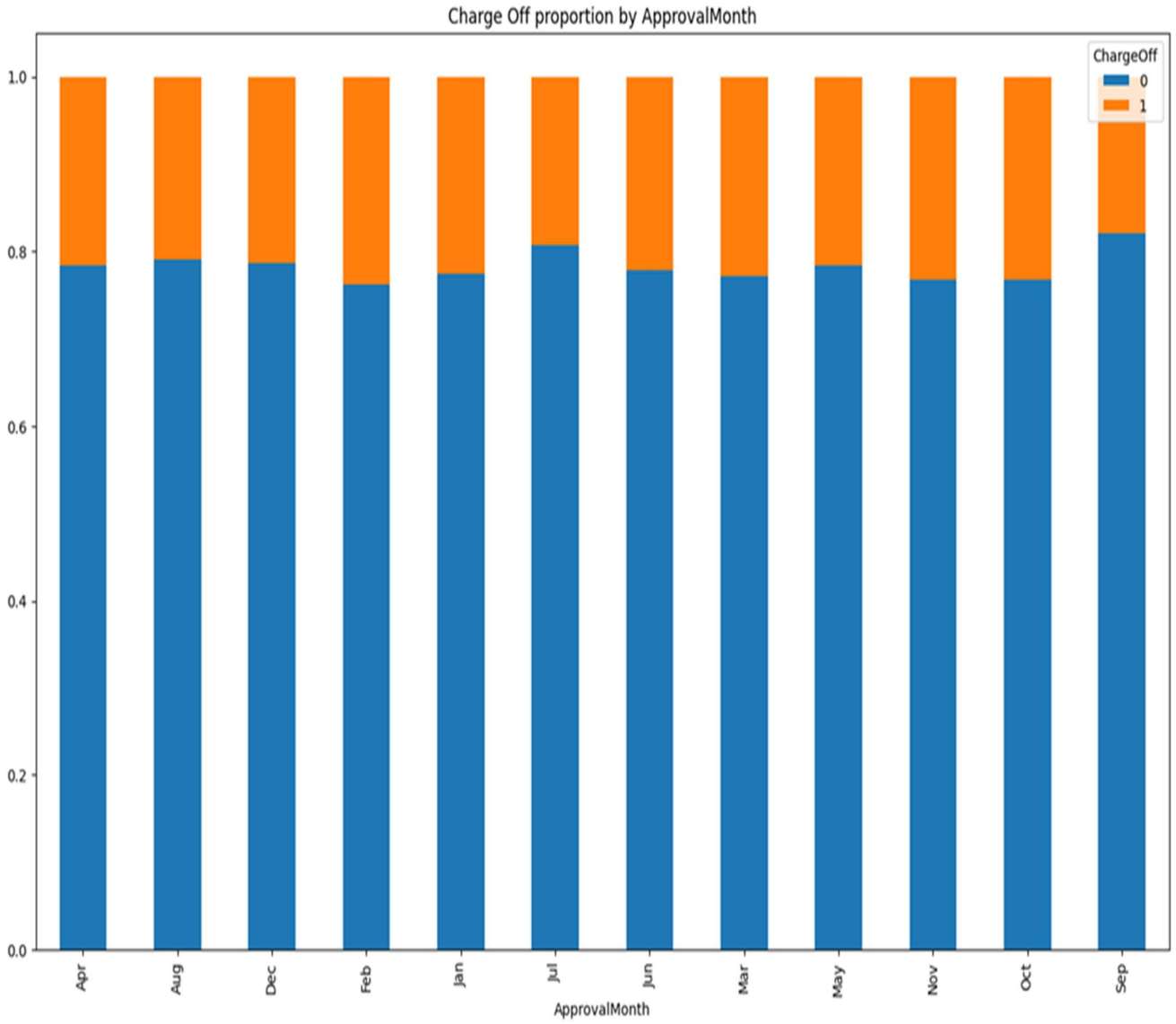

Figure 2 Column Approval Month has strong relationship with ChargeOff

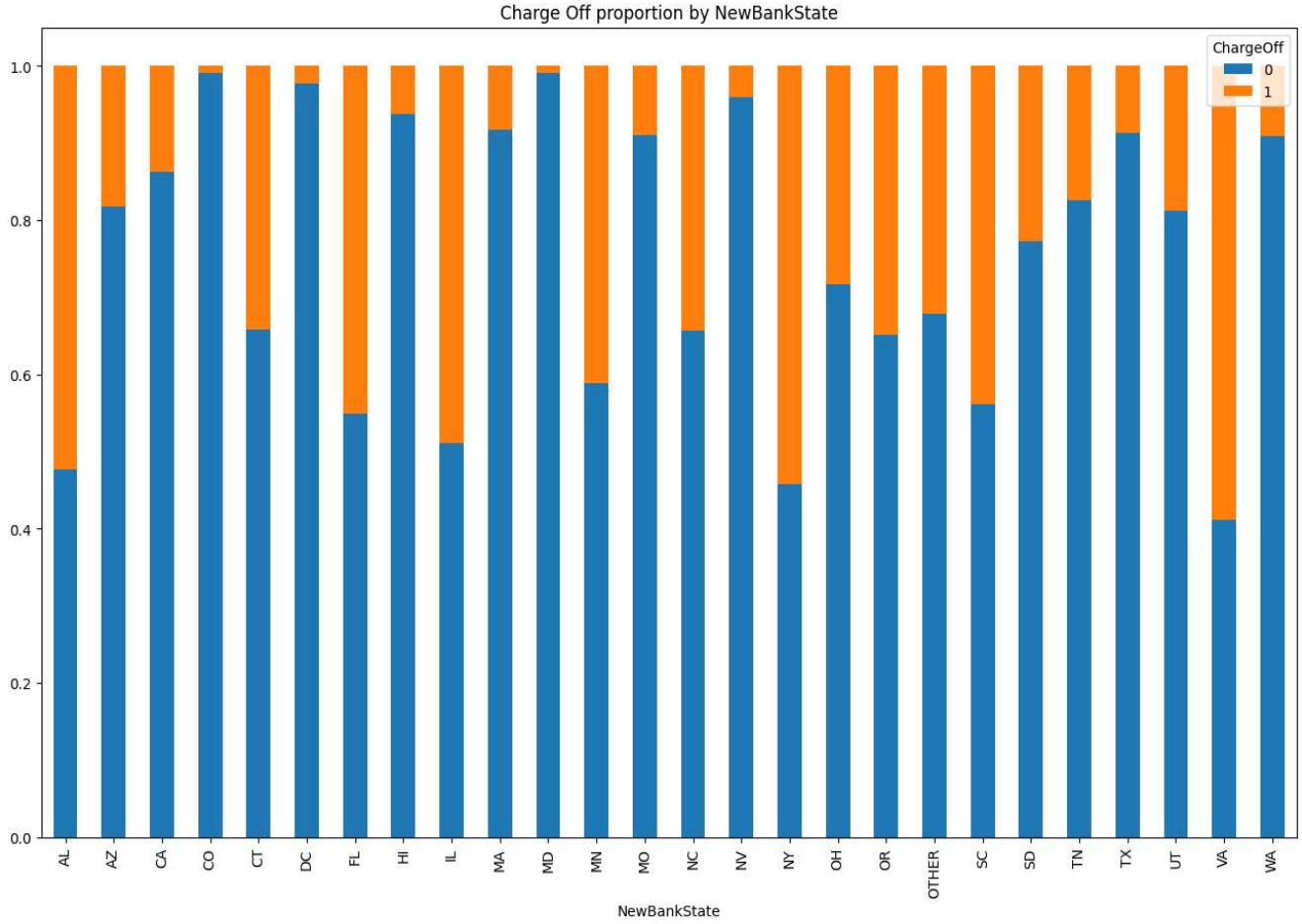

Figure 3 .Column NewBankState has strong relationship with ChargeOff.

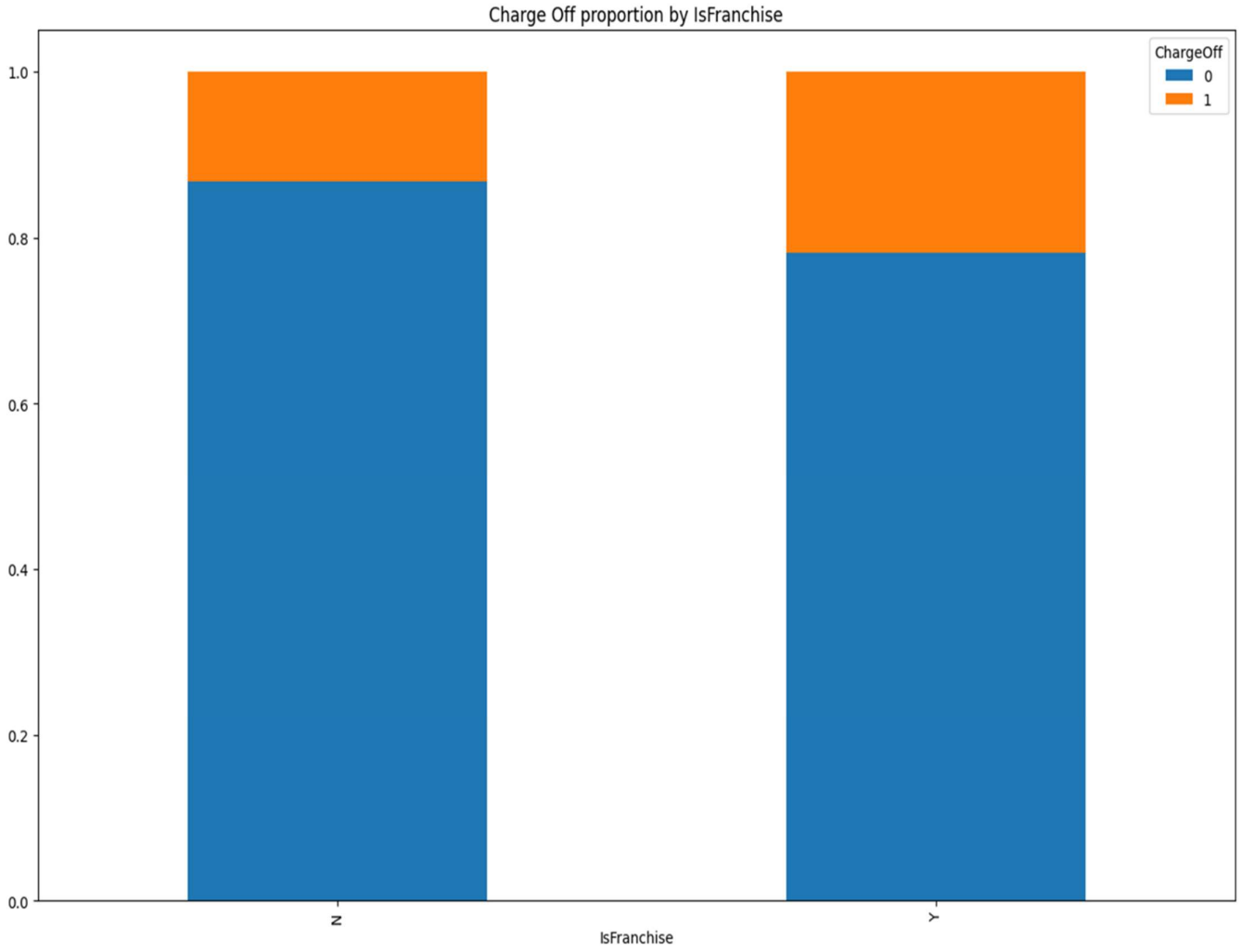

Figure 4.Column IsFranchise has strong relationship with ChargeOff.

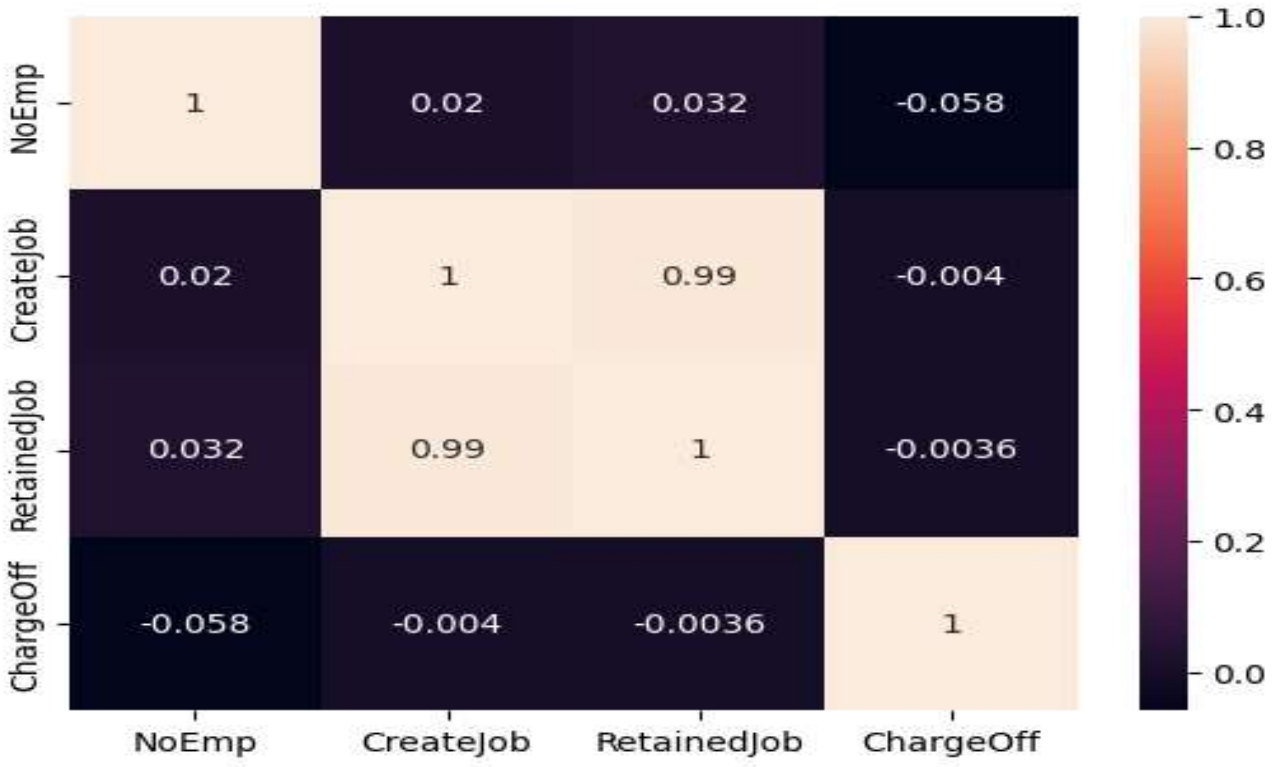

Figure 5.Correlation Between Independent Variables

Model Building

For the current machine learning model, we aim to minimize the model's prediction of companies that will actually Charge Off but are predicted as Paid in Full (PIF). Hence, we will focus on Recall Score.

"Charge Off" is a condition where a lender determines that a debt is unlikely to be collected and treats it as a loss on their financial records. This would result in losses for the borrowers.

According to the data we currently have, the Charge Off rate stands at 72.4%. This means that if a company has a debt of USD 1,000,000, it cannot pay USD 724,000 to the lender.

Strategy

In the modeling step, I will use a Pipeline to help me build the model. First of all, I will divide the data into two categories:

1. Categorical Variables
2. Quantitative Variables

A categorical variable is data that takes category or label values. On the other hand, quantitative variables take numerical values and represent some kind of measurement.

There are 13 columns that I have listed as categorical variables. Since each column values has no order with each other, I will encode them using **One-Hot Encoding**.

The rest of the columns should be listed in the other category, and since each column has a significantly different value range from each other and highly skewed, I will try to scale them using **Robust Scaler**.

Since there are no missing values, I don't need an imputer in this process.

The models that I will use is **Logistic Regression** and **Random Forest Classifier**. We will compare both of them and choose the model which give us the best result.

1. Logistic Regression Pipeline:

Logistic regression is a type of regression analysis used for predicting the probability of a binary outcome (e.g., yes/no, 1/0). In this step, the logistic regression model is trained on the preprocessed dataset. The model learns the relationship between the independent variables (features) and the dependent variable (target) by estimating coefficients that minimize the error between predicted and actual outcomes.

Once the logistic regression model is trained, it needs to be evaluated to assess its performance. Common evaluation metrics for classification tasks include accuracy, precision, recall, F1-score, and ROC-AUC score. These metrics provide insights into different aspects of the model's performance, such as its ability to correctly classify instances from different classes and its robustness to class imbalances.

```
one_hot_cols = X.select_dtypes(include='object').columns
numeric_cols = X.select_dtypes(exclude='object').columns
logit = LogisticRegression(solver='liblinear', random_state=2023)
smote = SMOTE(random_state=2023)

logit_pipe_num = Pipeline([
('scaler', RobustScaler()),
])

# for all object columns
logit_pipe_cat = Pipeline([
('onehot', OneHotEncoder(drop='first')),
])

# transforming all columns
logit_transformer = ColumnTransformer([
('pipe_num', logit_pipe_num, numeric_cols),
('pipe_cat', logit_pipe_cat, one_hot_cols)
])

# combine all pipeline
logit_pipe_combine = Pipeline([
('transformer', logit_transformer),
('rfe', RFE(logit)),
('resampling', smote),
('logit', logit)
])
```

2. Random Forest Pipeline¶:

A Random Forest pipeline involves a series of steps to prepare data, train a Random Forest model, and evaluate its performance.

Initialize a Random Forest model with hyperparameters such as the number of trees, maximum depth of trees, etc,Fit the model to the training data.
Optionally, perform hyperparameter tuning using techniques like cross-validation or grid search to find the best parameters for the model.

```
rfc = RandomForestClassifier(max_depth=7, min_samples_split=10, random_state=2023)
smote = SMOTE(random_state=2023)

rfc_pipe_num = Pipeline([
    ('scaler', RobustScaler()),
])

# for all object columns
rfc_pipe_cat = Pipeline([
    ('onehot', OneHotEncoder(drop='first')),
])

# transforming all columns
rfc_transformer = ColumnTransformer([
    ('pipe_num', rfc_pipe_num, numeric_cols),
    ('pipe_cat', rfc_pipe_cat, one_hot_cols)
])

# combine all pipeline
rfc_pipe_combine = Pipeline([
    ('transformer', rfc_transformer),
    ('rfe', RFE(rfc)),
    ('resampling', smote),
    ('rfc', rfc)
])
```

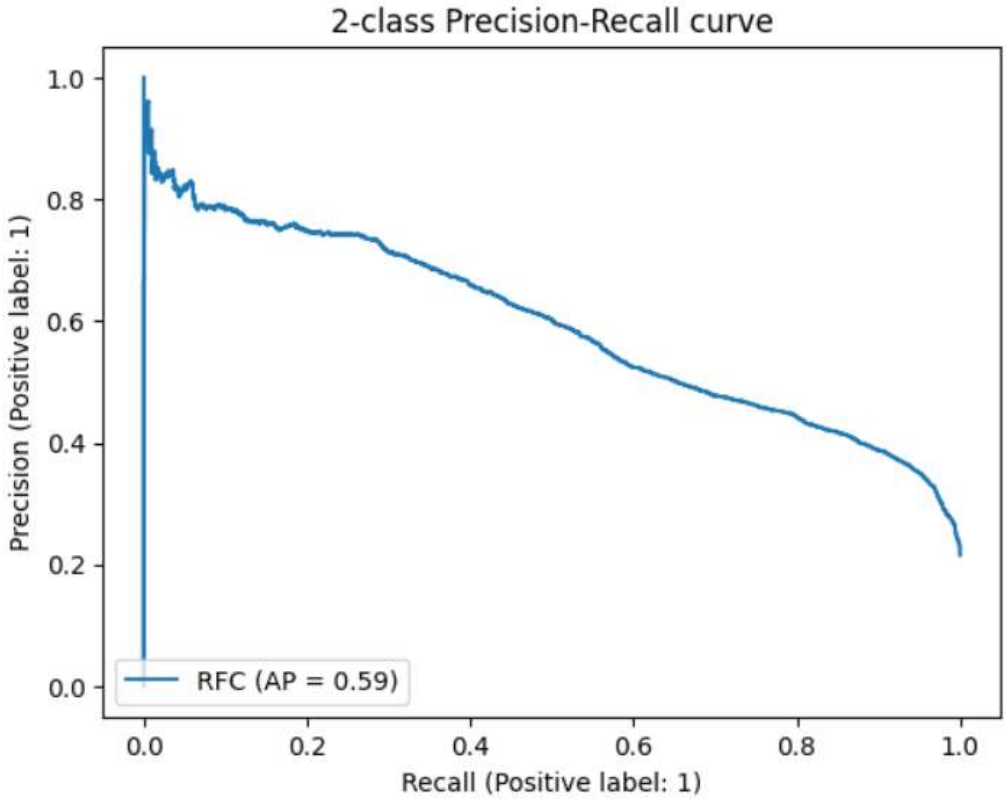

Figure 10 The confusion matrix of logistic regression model

Requirements Specification

4.1 Software Requirement:

- Windows/Mac/Linux
- Python-3.9.1
- Anaconda/Spyder
- Python Extension for VS Code
- Libraries:
  - Numpy 1.18.2
  - Pandas 1.2.1
  - Matplotlib 3.3.3
  - Scikit-learn 0.24.1
  - Flask 1.1.2
- Any Modern Web Browser like Google Chrome
  - To access the web application written in Flask

Conclusion

Up to this point, we have been able to improve the Recall to 90%. This means that out of 100 companies applying for loans and actually being potential defaulters, we can accurately predict 90 of these companies. The remaining 10 companies, we mispredict as non-defaulters, and we continue to lend money to them.

If we lend 100,000 USD to these 10 companies and, based on our previous analysis, the average charge-off rate is around 72.4%, then each company will Charge Off approximately 72,400 USD, resulting in a total default amount of 724,000 USD.

On the other hand, there is a note we should also consider. In addition to recall, we also have a precision score of 39%. This means that if we accurately predict CHGOFF (True Positive) for 90 cases, we predict CHGOFF (False Positive) for about 231 companies that are actually capable of paying in full (PIF).

SBA can further develop this model depending on SBA's needs. Because the current objective of this project is to minimize losses from Charge Off, increasing recall is the right step.

However, without using machine learning, we have limitations in assessing whether a company will default or not. At best, we can predict 50% accurately. In the same context, this would result in a loss of 3,620,000 USD from defaulting companies.

Here we can see that by using Machine Learning, we can reduce the risk of losses when providing

References

Min Li, Amy Mickel & Stanley Taylor
https://amstat.tandfonline.com/doi/full/10.1080/10691898.2018.1434342?scroll=top&needAccess=true
US Small Business Administration (2015), History retrieved August 22, 2015 from https://www.sba.gov/about-sba/what-we-do/history
Assessment and Instruction in Statistics Education College Report 2016," Available at http://www.amstat.org/education/gaise