Predicting Loan Defaulters

End Project 2

Description

Data Analysis is the process of creating a story using the data for easy and effective communication. It mostly utilizes visualization methods like plots, charts, and tables to convey what the data holds beyond the formal modeling or hypothesis testing task.

Domain: Finance

Read the information given below and also refer to the data dictionary provided separately in an excel file to build your understanding.

Problem Statement

Financial institutions incur significant losses due to the default of vehicle loans. This has led to the tightening up of vehicle loan underwriting and increased vehicle loan rejection rates.
The need for a better credit risk scoring model is also raised by these institutions. This warrants a study to estimate the determinants of vehicle loan default.

There is 1 dataset data that have 41 attributes.
You are required to determine and examine factors that affected the ratio of vehicle loan defaulters. Also, use the findings to create a model to predict the potential defaulters.

Approach:

1. Data Preliminary analysis:

- Perform preliminary data inspection and report the findings as to the structure of the data, missing values, duplicates, etc.
- Variable names in the data may not be in accordance with the identifier naming in Python. Change the variable names accordingly.
- The presented data might also contain missing values, therefore exploration will also lead to devising strategies to fill in the missing values. Devise strategies to do so whilst exploring the data.

2. Performing EDA:

- Provide the statistical description of the quantitative data variables
- How is the target variable distributed overall?
- Study the distribution of the target variable across the various categories such as branch, city, state, branch, supplier, manufacturer, etc. What are the different

employment types given in the data? Can a strategy be developed to fill in the missing values (if any)? Use pie charts to express how different types of employment defines defaulter and non-defaulters.
- Has age got something to do with defaulting? What is the distribution of age w.r.t. to defaulters and non-defaulters?
- What type of ID was presented by most of the customers as proof?
- Explain the factors in the data that may have an effect on ratings e.g. No. of cuisines, cost, delivery option, etc.

3. Performing EDA and Modelling:

- Provide the statistical description of the quantitative data variables
- How is the target variable distributed overall?
- Study the distribution of the target variable across the various categories such as branch, city, state, branch, supplier, manufacturer, etc.
- What are the different employment types given in the data? Can a strategy be developed to fill in the missing values (if any)? Use pie charts to express how different types of employment defines defaulter and non-defaulters.
- Has age got something to do with defaulting? What is the distribution of age w.r.t. to defaulters and non-defaulters?
- What type of ID was presented by most of the customers as proof?
- Explain the factors in the data that may have an effect on ratings e.g. No. of cuisines, cost, delivery option, etc.

Project Task: Week 1

Importing, Understanding, and Inspecting Data :

1. Perform preliminary data inspection and report the findings as the structure of the data, missing values, duplicates, etc.
2. Variable names in the data may not be in accordance with the identifier naming in Python so, change the variable names accordingly
3. The presented data might also contain some missing values therefore, exploration will also lead to devising strategies to fill in the missing values while exploring the data

Performing EDA:

1. Provide the statistical description of the quantitative data variables
2. Explain how is the target variable distributed overall
3. Study the distribution of the target variable across various categories like branch, city, state, branch, supplier, manufacturer, etc.
4. What are the different employment types given in the data? Can a strategy be developed to fill in the missing values (if any)? Use pie charts to express the

different types of employment that define the defaulters and non-defaulters.
5. Has age got anything to do with defaulting? What is the distribution of age w.r.t. to the defaulters and non-defaulters?
6. What type of ID was presented by most of the customers for proof?

Project Task: Week 2

Performing EDA and Modeling:

1. Study the credit bureau score distribution. Compare the distribution for defaulters vs. non-defaulters. Explore in detail.
2. Explore the primary and secondary account details. Is the information in some way related to the loan default probability?
3. Is there a difference between the sanctioned and disbursed amount of primary and secondary loans? Study the difference by providing appropriate statistics and graphs.
4. Do customer who make higher number of enquiries end up being higher risk candidates?
5. Is credit history, that is new loans in last six months, loans defaulted in last six months, time since first loan, etc., a significant factor in estimating probability of loan defaulters?
6. Perform logistic regression modeling, predict the outcome for the test data, and validate the results using the confusion matrix.

Dashboarding:

1. Visualize the data using Tableau to help user explore data to have a better understanding
2. Demonstrate the variables associated with each other and factors to build a dashboard

You can download the Data Dictionary from here - https://www.dropbox.com/sh/l7zopm0e20idnn9/AADbL0y5mig7tGRtB4D0oMq9a?dl=0