8.3 Final Project Step 1

Sharath Chandra Tummanapally

11-9-2021

Introduction

In a personal lending industry, one of the most difficult challenges is to reduce the number of loan defaulters. In Discover Financial services where I work now, Personal loan defaults has been a significant problem especially during this COVID-19 period. In this final project, I want to forecast how we can decrease loan defaults. I am using the data I found in a website on personal loans customers. This problem would be interesting to solve because it's existing problem in loan business and eventually by predicting the defaulter, it could be beneficial to the lending organization. I am planning on building a model for this loan prediction. This prediction might help any bank to be prepared for any foreseen pandemics in future.

Research questions

1. What are the variables that impact loan defaulting?
2. How to build patterns in data set that predict potential loan defaulters?
3. What are the variables that we need to consider to create a model? Among those, what are independent/dependent variables?
4. Did Unemployment rate and interest rate impact the laon defaulting?
5. Are there any external factors that could lead a person to loan defaulting?
6. Can our model be implemented in business and successfully predict the loan defaulters?

Approach

We can create a regression model that would identify the potential loan defaulters and assess if age, profession, house ownership,income variables have impact on defaulters which in turn could help business to approve loans on consumer basis.

How your approach addresses (fully or partially) the problem.

We can apply regression models on the data to help us in this prediction. Use plots to assess the models and to detect who are at higher risk of defaulting a loan. Run correlation test between variables. Using this model, we can give the loans to customers that would have high probability of paying addresses the problem in hand.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

1. These are following datasets to work with:

- Sample Prediction Dataset.csv - This dataset is a prediction dataset should is a .csv file with sample customer id and risk_flag with values as 0 and 1 which says about whether the loan is defaulted or not.

- Test Data.csv - This dataset is the test dataset we used to apply the model created by training data. The patterns identified on training data are to be applied to the test dataset to identify potential defaulters.
- Traning Data.csv - This dataset includes a training dataset with variables like income, age, experience etc to each customer ID. Source:https://www.kaggle.com/subhamjain/loan-prediction-based-on-customer-behavior

2. These datasets explain the unemployment and employment rate in percentage for different states in India for last one year. https://www.kaggle.com/gokulrajkmv/unemployment-rate-in-india-during-covid-19/data https://unemploymentinindia.cmie.com/

3. This dataset gives us more info on the interest rate across india. https://api.worldbank.org/v2/en/indicator/FR.INR.LEND?downloadformat=csv

Required Packages

The packages that are required to address this problem: * readxl * ggm * ggplot2 * dplyr * QuantPsyc * purrr * caret

Plots and Table Needs

- Histograms
- Scatter plots
- Line graphs

Questions for future steps

Identify the limitations of the data we have. Do we need more data to work with?