finalProject_KanaparthiVenkata

Venkata Kanaparthi

5/16/2021

-> Identify a topic or a problem that you want to research. Provide an introduction that explains the problem statement or topic you are addressing. Why would someone be interested in this? How is it a data science problem?

1.) Introduction

Loans Default, Predicting whether the loans that may be defaulted or not. As I am a part of industry that provides loans, it is very important to atleast have 98% of recovery rate. Then only it would consider as a profit which further improves/expands the business. It can be predicted through data sciences because, we have all the information on a customer available and we can build predictive models based on that.

-> Draft 5-10 Research questions that focus on the problem statement/topic.

2.) Research questions

1.) What are the data elements that are available? 2.) What additional data elements are needed? 3.) Is un-employment a direct cause for default? 4.) Why high salaried persons also defaulted? 5.) External Factors that can cause un-employment like pandemic, recession etc

-> Provide a concise explanation of how you plan to address this problem statement.

3.) Approach

By applying regression models. Once we get to know a pattern we can either reject the loan or we can decrease the loan amount provided to the customer. This also can lead to a problem because there is a chance of rejecting a loan to a customer who might actually fall into that list but cannot be defaulted. Instead of rejecting the application, we can add some additional checks on the customer before approving the application.

-> Discuss how your proposed approach will address (fully or partially) this problem.

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

Will apply regression models on the data which has several features to help us in predicting what we need. We can use multiple regression model with predict function through which we can explain based on the various values like standardized beta, checking for multicollinearity etc

-> Do some digging and find at least 3 datasets that you can use to address the issue. (There is not a required number of fields or rows for these datasets)

-> Original source where the data was obtained is cited and, if possible, hyperlinked.

-> Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

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

DataSet 1: lending_club_loans.csv https://data.world/create-username?next=/cabudies/lending-loan-data/workspace/dataset?agentid=jaypeedevlin&datasetid=lending-club-loan-data-2007-11#

Dataset1 has all the details of a loan offered to customers and their statuses, whether it is paid off, charged off or write off. It has a total of 115 columns where there are some unnecessary fields like all_util(Balance to credit limit on all trades), total_rev_hi_lim(Total revolving high credit/credit limit), inq_fi(Number of personal finance inquiries), total_cu_tl(Number of finance trades) etc that are not useful for our analysis

DataSet 2: Unemployment.csv https://www.kaggle.com/

DataSet2 provides the unemployment rate year wise which can be used to check whether it has any dependency on the defaulted loans

DataSet 3: us_disaster_declarations.csv https://www.kaggle.com/headsortails/us-natural-disaster-declarations

Dataset3 has all the natural disasters listed year wise which can help us to identify whether this caused the loans to go default

-> Identify the packages that are needed for your project.

6.) Required Packages

ggm,ggplot2,readr,pastecs,readxl,plyr,dplyr,magrittr,purrr,stringr,QuantPsyc

-> What types of plots and tables will help you to illustrate the findings to your research questions?

7.) Plots and Table Needs

Scatter plots, histograms, line graphs

-> What do you not know how to do right now that you need to learn to answer your research questions?

8.) Questions for future steps

How to do multiple runs on different samples of data to predict how the tru population behaves