Proposal

Weiling Li

11/7/2019

Proposal for MA678 Midterm Project

Two Projects

After doing some research on the datasets, I picked two projects and finding a hard time to decide which one to use. I will illustrate the two projects 1st and then list the reason why I am interested in them and other questions that asked in the assignment.

First Project: Risk Analysis of bad debt given amount requested, interest rate and other related inputs.

Project Description

General Ideas

The dataset mainly comes from **lending club Kaggle dataset**. (There are different lending club dataset available on kaggle and I would like to pick one that is the most cleaned or even a SQLite database version such that I could upload it to MSSP server or SCC cloud computing and do analysis on a small subset and then run the full set on cluster computing)

I would like to propose a project that finds the association between Risk of bad debt(delinquency?) vs interest rate, amount requested and other related variables.

To make it more interesting, I would like to add some **national economy indicators** into the fitting model to see if there is an association of the risk and the national level economy.

Potential Variables that goes into the model

After examing the dataset, I proposed the following potential variables:(Those in brackets are potential predictors in individual level or group level, those has "?" at the end indicates that I am not so sure if I should add them into the model)

- Risk of bad debt as the outcome
- Individual level predictors
  - Amount Requested
  - Interest Rate?
  - (Credit Score)
  - (Date of loan Issued?)
- Group level predictors
  - (Credit Score quantile?)
  - State of loan issued
  - Term Requested
  - (Date of loan Issued catergorized as month)
  - (*Economy indicators* as of month)

Potential Model

Because the outcome is a probability, I would like to fit a multi-level logistic model or similar sort which would transform my inputs into probability space to compute a probability of delinquency happens.

Potential Problem

For this project. I think there are two main issues:

1. How to define a bad debt? Do we count mispayment rate? Do we put a threshold on how many months of missing payment as a sign to determine bad debt?
2. Alison pointed out that the interest rate of a loan is calculated based on the risk of not paying back at the point of issuing the loan. So assuming their model is correct, we will anticipate a highly correlated relationship between interest rate and the likelihood of bad debt and other factors might only have a really small effect on the outcome.

Future direction

Base on the comment Alison provided, I think the future direction or maybe I should reform my project(if I indeed going to take this one) as: Does the risk adjusted Interest Rate issuing to the loan and adequatly reflect the observed risk of a bad debt?

Second Proposal: Restaurant Rating & Review modeling

Project Description

General Ideas

For restaurant businesses, it is very important to find a place where your food is more likely to be loved. Yelp's dataset contains a lot of the information of existing restaurants. We could try to predict the rating and reviews and explore their relationships. However, to makes thing more exciting, I would like to restrict our dataset on restaurants in Boston and greater Boston area. And then, I would like to add an average residential building value for every zipcode and add them as random effect in the model to see if it can explain some of the deviation of the rating and reviews.

The Property value dataset mainly comes from **Property Assesment on Analyze Boston**. The dataset provides detailed Taxable value of every building in the City of Boston and some region in the greater Boston area as well.

Potential Variables that goes into the model

- Model outcome: Restaurant ratings(Or ratings for a given period?)
- Potential Individual level predictors:
  - Reviews
  - Checkin
- Potential Group level predictors:
  - Average Property Assesment Value for a given zip code?
  - zip code?
  - Genre
  - Dollar Sign(how expensive)

Potential Model

Multi-Level Linear Regression on overall ratings? or ML ordinal regression on individual responses? I haven't really decided yet

Potential Problem

For a MLLM model, is that really adequate for our inputs? each user can only choose from 1 to 5, which is turncated both ends. So in this case a turncated model maybe more adequate?

For a ML Ordinal regression, will this make the problem too complicated? In reality will people usually use machine learning instead?

Future Directions

The idea of bringing in Property value data is because of a speculation that when I choose a restaurant, not only the food, I also consider the environment and the location of the restaurant, so maybe there will be an strong association between how ppl think of the location(represented by average property value) and how people feel about the food?

So for future direction, I would suggest if the results turns out to be true. We can try to find a better way to measure how people think of the location and surroundings.

Timeline

I propose the following timeline:

1. Decided on 1 project by the end of **Nov. 11 2019(Monday)**.
2. Start data wrangling and because the data itself is too big, I might have to work with a subset of the data.(randomly select a fraction of the data) Repropose a clear variables and model selection by **Nov. 17th 2019** using a fraction of the data.
3. Run modeling on the subset of data, adjusting model and interpret the results. by **Nov. 24th 2019**.
4. Run the full analysis possibly on BU SCC bash job and interpret the results to see if that holds by **Nov. 27th 2019**.
5. Submit report by **Dec. 5th 2019**