PPOL-670 Introduction to Data Science Final Project Proposal

Ryan Collins and Alex Clegg

4/17/2020

Team members:

- Ryan Collins
- Alex Clegg

Project Option:

- 1

Question of interest:

Predicting through machine learning techniques the loan size of Small Business Administration (SBA) disaster loans related to the COVID-19 Paycheck Protection Program (PPP).

Data source(s):

1) SBA Loan Date(for independent variable):

- https://www.sba.gov/about-sba/sba-performance/open-government/digital-sba/open-data/open-data-sources (can combine years 2008-2018)
- https://content.sba.gov/sites/default/files/2020-04/PPP%20Report%20SBA%204.14.20%20%20-%20%20Read-Only.pdf (would require translating into spreadsheet and would need to be tied to industry data from census below)

2) Census data (for additional predictor variables):

- https://www.census.gov/data/developers/data-sets/cbp-nonemp-zbp/zbp-api.html (provides zip code level data on payroll, number of businesses, sales and reciepts)
- https://www.census.gov/data/developers/data-sets/economic-census.html (provides additional small business metrics but at the county-wide level)
- https://www.census.gov/data/developers/data-sets/business-owners.html (provides survey data of business semtiment, and other catagorical vectors which could prove useful to model)

Summary:

As mentioned in our question of interest, our pursuit is to credibly predict SBA disaster loan size based upon historical disaster declaration loan data as well as county and zip code level census data. By combing census data such as employer payroll, employee number, sales and receipts, industry type, along with other demographic data – income levels, racial breakdown, education levels, number of businesses, etc… our hope is to get create a set of multi-variate model that provides relatively accurate predictions of loan size administered by SBA during disaster scenarios. With this model in hand, we can then generally begin to test the model on future out-of-sample data as SBA releases additional information on the program. We can also begin to articulate how this could then be used to further calculate aggregate sums which could prove useful for policy makers as they estimate total cost of programs such as PPP, which as we've seen in recent days, have had significantly more demand than thought.