Milestone 3

Team: Tactical

Members: Mario Aldana, Katie Briggs, Shefers Sarkar

Introduction

**Abstract**

The purpose of this study was to put into practice and examine the impact of predictive analytics for a data science project. This project was to look at past trends of tax returns for zip codes located throughout the United States and utilize the data to predict the upcoming tax cycle returns. This may be a difficult challenge due to the current year being in a quarantine pandemic and tax deadlines being so different from the previous year. According to the International Monetary Funds, forecasting the tax revenue will lead to an underestimation of the revenue decline. This is due to many countries taking tax policy and administration measures in response to the pandemic. We will take the forecasting approach with these issues in mind.

**Into/Background of the Problem**

The issue that our group is working toward is to forecast the revenue of the individual tax returns. As Benjamin Franklin wrote in a letter to Jean-Baptiste Le Roy in 1789, “this world nothing can be said to be certain, except death and taxes.” Paying taxes is considered a civic duty of our nation and the money withheld is used for government programs, government salaries, police and firefighters. There are other common resources that taxes go towards, our roads, libraries, parks, and schools. Purchasing items from a store also has a tax percentage associated with them, which drives the local and overall economy. By forecasting the tax revenue, the resources that use these funds know exactly what they can afford or have insight into the health of the economy.

**Methods**

Our predictive analysis followed the Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is a process encompassing six steps that guide us to develop and implement models from data. These steps are, in order, business understanding, data understanding, data preparation, modeling, evaluation, and deployment. During the business understanding portion we collaborated over which project avenue to take and decided that forecasting taxes was a topic that one of our group members was familiar with.

For data understanding step we viewed many datasets and decided on one from the Internal Revenue Service located at data.gov. This dataset had over 165K observations. Even with that number of attributes, some significant variables were missing or set to zero. This was discovered during the data preparation step. To deal with the missing values of zip codes, it was decided to locate and enter in these values. The values were reference from a chronological listing of zip codes located throughout the United States. Removing these variables would have caused a large portion of taxation information to have been lost and our forecast would be underestimated. Also, in this step of CRISP-DM we explored the data, conducted variable selection, handled outliers, and discovered correlations.

Our plan for modeling will be to train, test and split the data to validate which model will be appropriate to choose. We will fit the initial model and evaluate the performance. Cross- validation will be accomplished with our training data and will be tuned for hyperparameters. Our predictive analytics approach will evaluate on the validation data and for the final model we will evaluate on the test data.

**Results**

Our plan is to post our analysis results here.

**Discussion/Conclusion**

We plan to conclude on our project and our experiences here.

References:

[1] https://catalog.data.gov/dataset/zip-code-data

[2] <https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2018-zip-code-data-soi>

[3] Singha, G. (2019), Evaluating a Data Mining Model, retrieved from: <https://www.pluralsight.com/guides/evaluating-a-data-mining-model>

[4] ChristopherGS, Udemy, (2019)  How to Deploy Machine Learning Models, Retrieved from:  <https://christophergs.com/machine%20learning/2019/03/17/how-to-deploy-machine-learning-models/>

[5] Wikipedia, (2020), Death and Taxes Idiom, Retrieved from: <https://en.wikipedia.org/wiki/Death_and_taxes_(idiom)>

[6] Boyd, D, Dadayan L., (2014) Revenue Forecasting, Rockefeller Institute of Government State University of New York, Retrieved: <https://rockinst.org/wp-content/uploads/2018/02/2014-09-30-Revenue_Forecasting_Accuracy.pdf>