Milestone 5

Team: Tactical

Members: Mario Aldana, Katie Briggs, Shefers Sarkar

**Abstract**

The purpose of this study is to put into practice and examine the impact of predictive analytics on a data science project. This project is 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. For example, in 2019, the deadlines for filing and paying individual tax due amounts were moved. Also, many companies have not been able to generate enough income or have had to close. This has affected both employers and workers. 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**

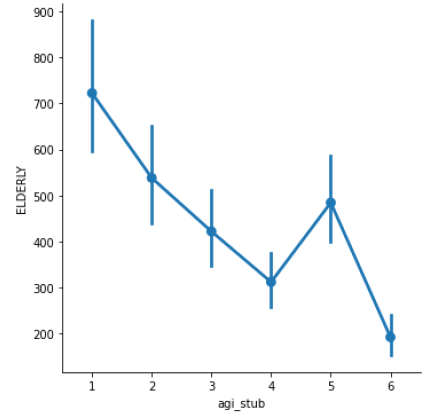
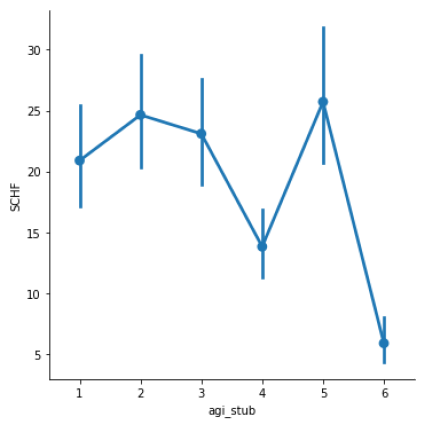
The issue that our group is working toward is to forecast the revenue of the individual tax returns. As Benjamin Franklin wrote in a letter to Jean-Baptiste Le Roy in 1789, “this world nothing can be said to be certain, except death and taxes.” Paying taxes is considered a civic duty of our nation and the money withheld is used for government programs, government salaries, police, firefighters, and so on. 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 follows 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. In the tax environment, there are two important areas of tax collection. Individual and Partnership. Our project will be working only with individuals; individuals are people. Taxpayers pay taxes to the federal government and the state government. We are going to work with federal tax returns data.

For the data understanding step, we viewed many datasets and decided on one from the IRS (Internal Revenue Service) located at irs.gov, “*SOI Tax Stats - Individual Income Tax Statistics - 2018 ZIP Code Data (SOI)*”. This dataset had over 165K observations and 153 variables. Even with that number of attributes, some significant variables were missing or set to zero, for example, the zip code variable. This was discovered during the data preparation step. To deal with the missing values of zip codes, it was decided to mark these zip codes with a tag, '*99999*'. The values were a 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. For example, correlations in continuous variables, we discovered correlations in two important variables, Farmers (SCHF) and Elderly (ELDERLY). Farmers are people who produce food in the rural sector, "Farmers are agricultural professionals who grow crops and take care of animals on farms they often own and maintain themselves" [1]. Taxpayers who derive at least two-thirds of their yearly gross income from farming or ranching activities are Farmers. On the other hand, taxpayers who are 65 or older are Elderly. Both variables, Farmer and Elderly have a similar trend. Therefore, these variables are closely related. Our decision was to combine them into one new feature.

Correlation between Farmer and Elderly features:

 Fig. 1 - Elderly Feature Fig. 2 - Farmer Feature

In Fig. 2, we can see the farmer average for each range of adjusted gross income (AGI). Each AGI range is categorized in this way:

|  |  |
| --- | --- |
| 1 | $1 under $25,000 |
| 2 | $25,000 under $50,000 |
| 3 | $50,000 under $75,000 |
| 4 | $75,000 under $100,000 |
| 5 | $100,000 under $200,000 |
| 6 | $200,000 or more |

In the Farmer Feature plot, the bars correspond to the sample size at each range of AGI, larger bars indicate a smaller sample size. In conclusion, as taxpayer have more AGI, they are less likely to be Farmers. The same happen to elderly taxpayers, as taxpayer have more AGI, they are less likely to be Elderly people. We saw this correlation between these features; therefore, we made the decision to create a new feature called *farm\_elderly*. In the dataset, we will use the following filing statuses:

*Single*: Filing status is single

*Joint*: Filing status is married filing jointly

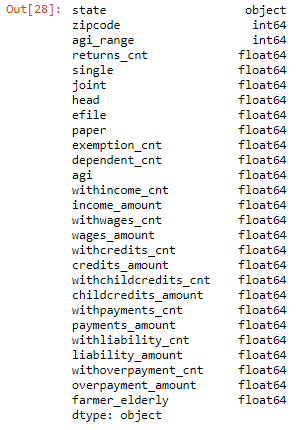
*Head of household*: Filing status is head of household

*Farmer/Elderly*: New feature when the taxpayer is filing a return as farmer or elderly.

We accomplished our plan for modeling by splitting our data into 3 sets (train, validation, and test) that validated which model will be appropriate to choose. We also accomplished an algorithm comparison that provided us with insight on which model would be best to fit. We fitted the initial model and evaluated the performance which will be discussed in the Result section. Cross-validation was accomplished with our training data and was tuned for hyperparameters. Our predictive analytics approach evaluated the validation data and for the final model and we evaluated the test data.

**Results**

Of the first 3 stages, business understanding, data understanding, and data preparation of the CRISP-DM methodology, as far as a result we have: nulls, outlier, correlation, and powerful variables identification. Also, we plotted some best variables. Final variables:



**Codebook:**

|  |  |
| --- | --- |
| state | The State associated with the ZIP code |
| zipcode | 5-digit Zip code |
| agi\_range | Size of adjusted gross income |
| 1 = $1 under $25,000 |
| 2 = $25,000 under $50,000 |
| 3 = $50,000 under $75,000 |
| 4 = $75,000 under $100,000 |
| 5 = $100,000 under $200,000 |
| 6 = $200,000 or more |
| returns\_cnt | Number of returns |
| single | Number of single returns [Filing status is single] |
| joint | Number of joint returns [Filing status is married filing jointly] |
| head | Number of head of household returns [Filing status is head of household] |
| efile | Number of electronically filed returns |
| paper | Number of computer prepared paper returns |
| exemption\_cnt | Number of individuals [3] |
| [3] Beginning in 2018, personal exemption deductions were suspended for the primary, secondary, and dependent taxpayers. However, the data used to create the “Number of individuals”—filing status, dependent status indicator, and identifying dependent information—are still available on the Form 1040. This field is based on these data. |
| dependent\_count | Number of dependents |
| agi | Adjust gross income (AGI) [7] |
| [7] Does not include returns with adjusted gross deficit |
| withincome\_cnt | Number of returns with total income |
| income\_amount | Total income amount |
| withwages\_cnt | Number of returns with salaries and wages |
| wages\_amount | Salaries and wages amount |
| withcredits\_cnt | Number of returns with total tax credits |
| credits\_amount | Total tax credits amount |
| withchildcredits\_cnt | Number of returns with child and dependent care credit |
| childcredits\_amount | Child and dependent care credit amount |
| withpayments\_cnt | Number of returns with total tax payments |
| payments\_amount | Total tax payments amount |
| withliability\_cnt | Number of returns with tax liability |
| liability\_amount | Total tax liability amount [12] |
| [12] “Total tax liability” differs from “Income tax”, in that “Total tax liability” includes the taxes from recapture of certain prior-year credits, tax applicable to individual retirement arrangements (IRA's), social security taxes on self-employment income and on certain tip income, advanced earned income payments, household employment taxes, and certain other taxes listed in the Form 1040 instructions. |
| withoverpayment\_cnt | Number of returns with total overpayments |
| overpayment\_amount | Total overpayments amount |

**Train, Test, Split:**

Our training and test data used the variables of agi, income\_amount, wage\_credits, credits\_amount and childcredits\_amount for X and our target variable was liability\_amount as Y. We conducted the ratio of 1/3 for training and 2/3 for testing. The X variables were then normalized. Normalization of the target variable was not required.

**Correlation Matrix:**

A correlation matrix was created to visualize the correlation between features.

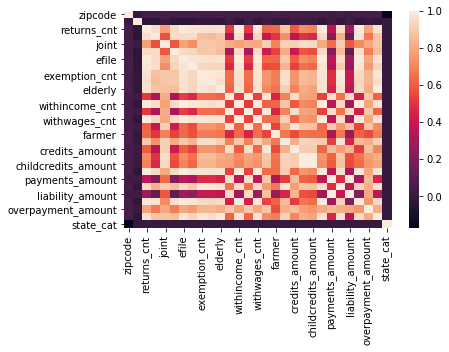


Figure 1. Correlation Matrix

To understand the correlation further we removed the highly correlated features.

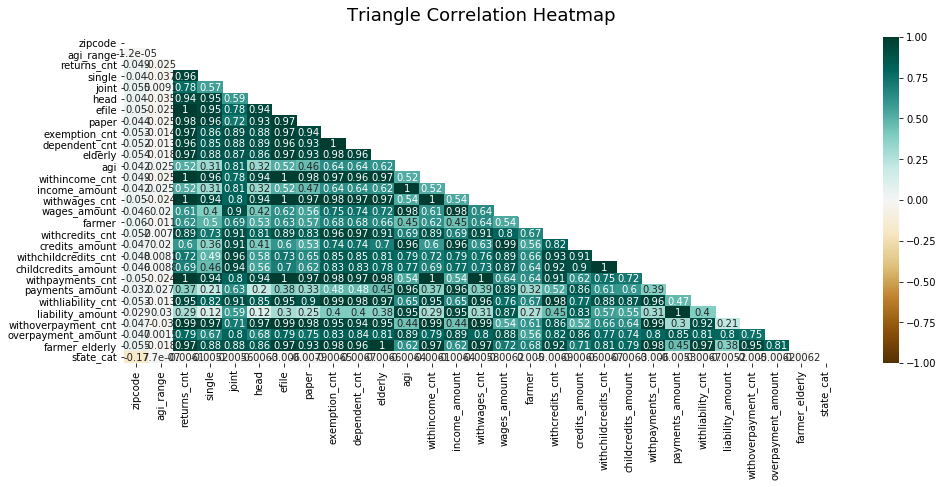


Figure 2. Triangle Correlation

**Linear Regression (Model 1):**

Our first model was linear regression, and we fitted our training and test data to make a prediction.

|  |  |  |  |
| --- | --- | --- | --- |
| **LR Coefficient** | **LR Intercept** | **LR Train Score** | **LR Test Score** |
| -1.38774090e+09 | -2004.7706918865879 | 0.990 | 0.974 |
| 1.68449827e+09 |  |  |  |
| -1.20894091e+08 |  |  |  |
| -1.53147861e+06 |  |  |  |
| -1.58688713e+07 |  |  |  |

Our training score was 99 % and the testing score 97.4%.

**Ordinary Least Squares (Model 2):**

In our second model we calculated the least squares with the response variable being agi\_range and the predictor being the liability\_amount.

|  |  |  |
| --- | --- | --- |
| **R-squared** | **Adjusted R Square** | **F-Statistic** |
| .001 | .001 | 150.2 |

With the condition number being so large, it may indicate strong mutliconllinearity. This time we tried the model on the response of payment\_amounts and the predictor remaining liability\_amount.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **R-squared** | **Adj R Square** | **F-Statistic** | **Const** | **Liability\_amount** |
| .991 | .991 | 1.915e+07 | 2457.29 | 1.02 |

**Model Comparison:**

We used a model comparison algorithm on our trained and tested data to help us decide which model should be attempted next.

Our results were:

LR: 0.086567 (0.015367) Linear Regression

LDA: 0.085075 (0.013012) Linear Discriminant Analysis

KNN: 0.088060 (0.014471) Near Neighbor

CART: 0.089552 (0.013350) Decision Tree

NB: 0.088060 (0.020789) Gaussian NB

SVM: 0.086567 (0.015367) Support Vector Machine

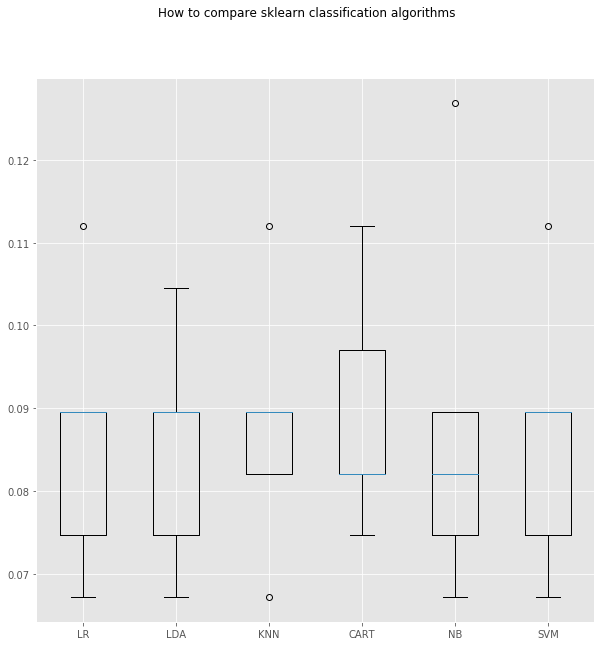


Figure 3 Comparison of Algorithms

Looking at the comparison of the algorithms a Decision Tree Classifier or Gaussian Navies Bayes may garner a closer look at predicting the liability amount.

**Gaussian NB (Model 3):**

To conduct the Naïve Bayes algorithm, the class was predicted, and the distribution was scored. After, the parameter was estimated and then charted by the liability amount by each return. No contour appeared to segregate the variables and the Navies Bayes was decided to not be a good fit.



Figure 4 Bayes Decision Boundaries

**Challenges**

Challenges presented themselves over the course of this project. For starters, several of our feature selection methods were giving different responses as to which features were strongly correlated. The Cross Validation caused computer crashes with memory issues as well as the algorithm comparison due to the largeness of the dataset.

Also, many of the models we used were new to us. Sometimes they were applied wrong, or they were not suited to the type of data we had. We attempted to use many more models than the ones mentioned here, including some Random Forests that performed very poorly, but in the end we believe we chose the best models for the job.

**Discussion/Conclusion**

We understand the tax business and preparing the data have taken considerable time. With the final variables and observations. In summary, exploratory data analysis was performed on the variables to help identify which ones might not be good predictors of taxes. Plotting the variables as histograms helped to visualize each data point and determine which ones had a normal distribution. However, this still was not enough evidence to make a solid decision on which variables to use in the analysis. Several feature selection techniques were used to help with this decision, plotting the variables and a correlation matrix. These techniques were helpful in feature selection and several variables were removed from the analysis (farmer and elderly) if their correlation value was relatively small when compared to other variables.

Our initial linear model results were good, but we wanted to include other models to ensure that we were making the most accurate predictions possible. We obtained an R2 value of 0.99 for our training subset of data. This was about a 30% improvement over the models we used.

**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>

[7] Donohoe, Ashley. How Much Money Do Farmers Make on Average Annually? March 26, 2018. Retrieved from *work.chron.com/much-money-farmers-make-average-annually-3185.html*