Identifying Fraud within the Premium Tax Credit Filing Population

Travis Williams

Introduction

*Problem*

The Premium Tax Credit (PTC) is a tax credit that is associated with the Affordable Care Act legislation. Qualifying taxpayers who sign-up for insurance with the Federal Marketplace can have their healthcare premium partially subsidized (paid for) or choose to have the subsidy paid back to them in the form of a credit when they file their tax return. In either case, taxpayers are required to fill-out tax Form 8962 and attach it to their return to ‘reconcile’ this process. As with other recently created tax credits, a large number of tax returns are filed that seek to fraudulently claim the PTC credit. Hence, this project will sought to use IRS data to predict which tax returns claiming the credit are likely to be fraudulent.

*Customer*

IRS operational areas are tasked with the prevention of fraud for specific tax credits. Fraud prevention generally involves a collaborative effort between several IRS research divisions and these operational areas. Though the IRS risk assessment offices currently have fraud filters in place, filters need to be adjusted as new tax credits are offered to the taxpaying population. Working as a research consultant, I will provide an algorithm that can be used by the IRS risk assessment office to augment their current fraud filters to include fraud detection for tax returns that claim the PTC. Improving the robustness of the filters will allow IRS operational areas to do a more effective job at selecting PTC cases to audit, meaning audit cases selected will be more likely to be fraudulent cases. Preventing fraudulent tax credits from being processed is especially important because there is very slim chance money issued for fraudulent credits can be recovered.

Data

*Description and Wrangling*

As IRS tax data is not available to the public, I used two internal IRS sources to gather data for the project: Business Objects reports and the IRS data warehouse known as the Compliance Data Warehouse (CDW). I used Business Objects to query a report of tax returns that were filed claiming the PTC credit for Tax Year (TY) 2014. I exported this report into a series of 5 separate CSV files which I then imported into Jupyter Notebook for further wrangling. In addition, I used SAS to query data on audited tax returns from the CDW. I then exported this data from SAS as a CSV file and imported it into Jupyter Notebook. Once all datasets were in Jupyter, I stacked the 5 CSV files from Business Objects into one single data frame and merged that data frame with the CDW audit dataset to create one data frame with PTC and audit data.

In the existing dataset, several binary indicator variables contained missing values, which I replaced with 0’s. I chose to do so in this case based on past experience with missing IRS data fields, but had I not had such existing knowledge, I would have consulted with someone more familiar with the data to determine if missing values were truly indicative of a ‘no’ for the indicator variable.

*Important Fields and Limitations*

Using my existing knowledge of IRS program operations, I limited the dataset to only variables I felt may be relevant to fraud within PTC. Had I not had such pre-existing knowledge, my approach would have been to consult with subject matter experts to determine fields to keep and eliminate. The resulting dataset contained a multitude of factors that could potentially impact the likelihood of a taxpayer filing a fraudulent PTC return, which I grouped into three main areas:

1. Behavioral factors- data that relates to specific actions by the taxpayer, such as including specific tax forms, calculation errors, and timeliness of return filing.

2. Monetary Factors- information such as taxpayer Adjusted Gross Income (AGI), amount of PTC claims, etc.

3. Demographic Factors- factors such as taxpayer age, location, etc.

In addition, the dataset contains a variable to show the result of the audit cases, whether the return was found to be a ‘good’ return (non-fraudulent) or a ‘bad’ return (fraudulent return). This variable was used as the target during analysis.

The dataset has two major limitations. First, the dataset is based on data from TY 2014, which is several tax seasons behind the most current IRS tax year data. TY 2014 was chosen because though PTC tax data was available was subsequent tax years, complete audit data was not. Had full TY 2017 audit data been in place, the ideal method would have been to use TY 2016 tax data to run exploratory analysis and train a model, then test the model on TY 2017 tax data. In this case, as no other data option exists to show fraudulent returns past TY 2015, I will test the model I create on TY 2015 tax data. As a consulting analyst, this may be a good opportunity for me to suggest a pilot study to have IRS operations collect more current information for a sample of PTC fraud cases for TY 2016 and TY 2017, then use that information to create a better model. As fraud schemes tend to change annually, modeling from the most current data should in theory provide the most value for operations.

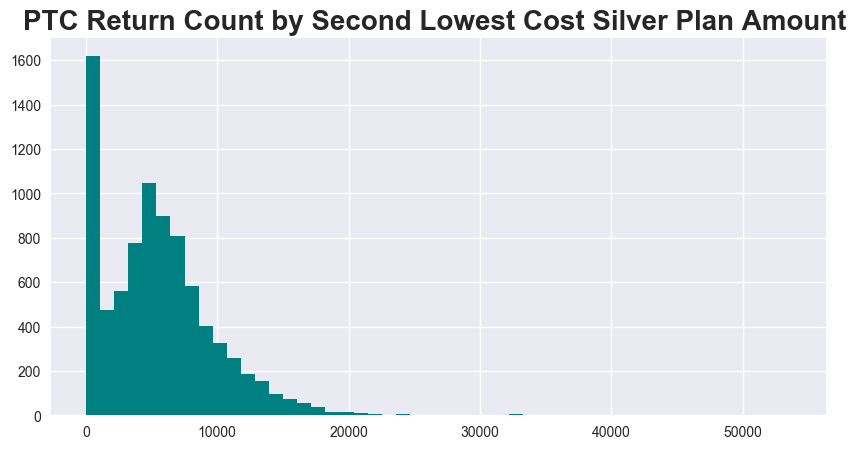
The second limitation involves the mindset behind PTC tax fraud. Quantitative data will show us the factors that may make it likely for a return to be fraudulent. What is will not provide is in depth information on the mindset behind fraud schemes as well as the systemic gaps in the process that allow fraud to perpetuate. As a consulting analyst, I’d suggest qualitative research and process auditing/mapping to augment the results of the quantitative analysis.

Method

*Exploratory Analysis*

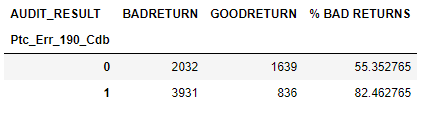
As a first step, I looked at a numerical representation of the distribution of all the variables in the dataset. As many of the continuous variables had large outliers, I chose to create histograms for several of these variables to see a visual representation of where the data began to trail off (Graph 1 below). I used this information to create new categorical break outs for relevant continuous variables in the dataset.

**Graph 1: Histogram of PTC Return Count by Second Lowest Cost Silver Plan Amount**

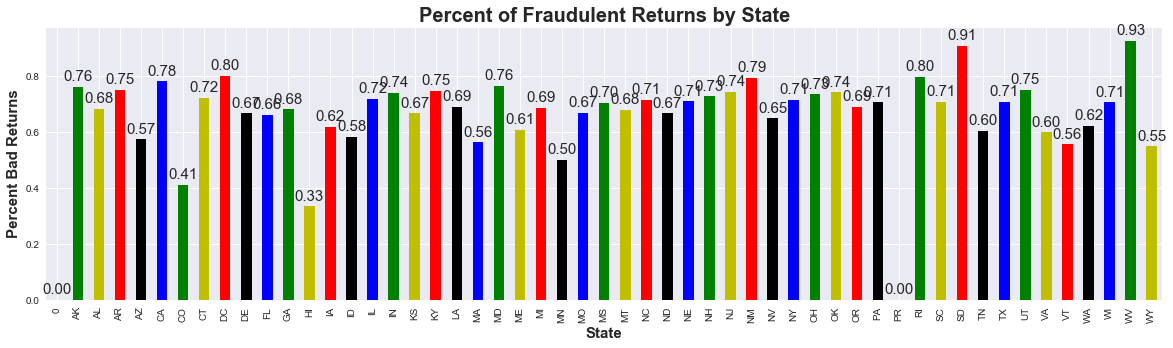
**

Once I created a series of new categorical variables, I used descriptive statistics to explore relationships between these variables and fraud. I used cross-tabulations to examine both the percent and volume of fraudulent tax returns for each variable. In addition, I produced bar graphs to display a visual representation for the percent of fraudulent returns for each variable (see Table 1 below for an example). This was particularly effective for variables in which there were a large number of groups, such as state (see Graph 2 below). A visual representation was easier to see in this case.

**Table 1: PTC Fraudulent Returns Based on Whether Taxpayer Attached Form 8962 (PTC Error Code 190)**



**Graph 2: Percent of Fraudulent PTC Returns by State**

****

***Inferential Statistics***

After using descriptive statistics to identify factors that might be good indicators of fraud, I used inferential statistics to explore which factors were statistically significant as well as the magnitude of the relationship. I used a point biserial correlation to compare continuous variables (factors such as income, etc.) to fraud (which is a dichotomous variable) and a chi-square test to determine if there was a significant difference in proportion of fraud between groups for categorical variables.

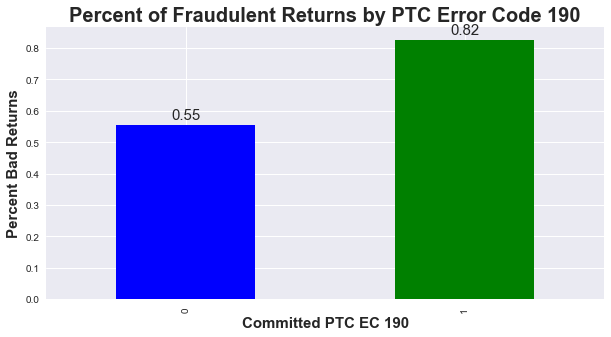
Results

After testing a variety of potentially relevant factors to determine their impact on fraudulent PTC returns, I was able to determine the following factors had the potentially highest impact on filing a fraudulent PTC return.

*Whether or not the taxpayer attached Form 8962*

To properly ‘reconcile’ the PTC on their tax return, taxpayers are required to attach a Tax Form 8926 with their return (also known as an Error (EC) 190). Taxpayers that did not attach a Form 8926 were 27% (82% fraud rate for those with PTC EC 190 vs. 55% rate for those who no PTC EC 190) more likely to file a fraudulent tax return (see Graph 3 below).

**Graph 3: Percent of Fraudulent Returns Based on Whether Taxpayer Attached Form 8962**

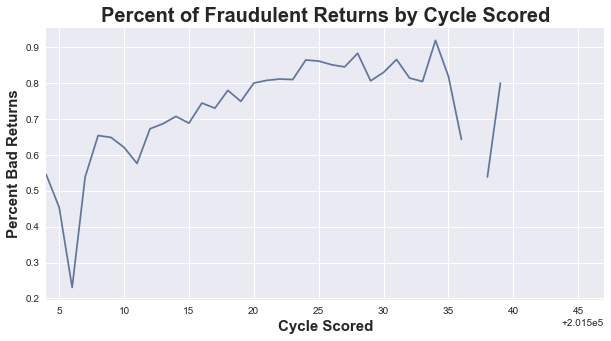
****

In addition, there was a statistically significant weak positive correlation between attaching a Form 8962 and fraud (correlation=.30, p=2.75e-169).

*Timing of when the tax return was filed*

Past research has shown timeliness of tax return filing can be indicative of fraud. Examining a line graph of the percentage of fraudulent returns by the cycle (week) the tax return was filed shows a clear increase in fraud as the tax filing season progresses (see Graph 4 below).

**Graph 4: Percent of Fraudulent Returns by Timing of Return Filing**

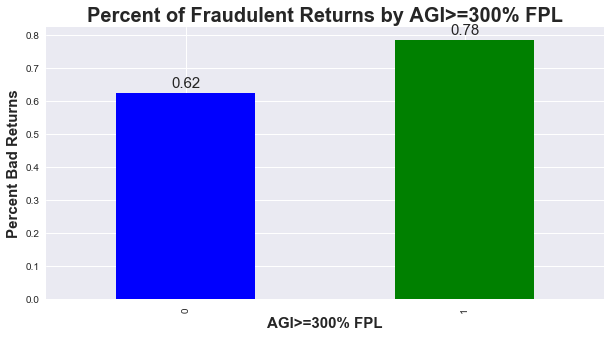
****

There was a statistically significant weak positive correlation between timing of the return filing and fraud (correlation=.32, p=1.09e-200).

*Income Compared to Poverty Level*

As the PTC is a low income credit, it is less likely taxpayers who had a high Adjusted Gross Income (AGI) compared to their poverty level calculation (poverty level calculation is based on factors such as family size) would receive the tax credit. Taxpayers with an AGI over 300% of their calculated poverty level were 16% more likely to file a fraudulent return than those under this amount (78% fraud rate for those with AGI equal to or over 300% vs. 62%% rate for those under the amount) (see Graph 5 below).

**Graph 5: PTC Fraudulent Returns Based on AGI Equal to or Over 300% of Poverty Level**

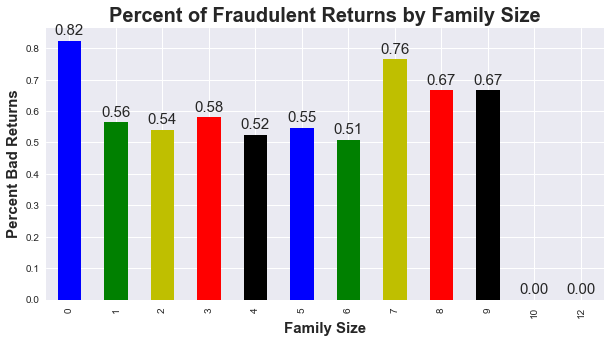


Results of chi-square show a statistically significant difference in the proportion of fraud cases based on an AGI greater than or equal to 300% of the poverty level calculation (p=1.89e-57).

*Family Size*

Examining a graph of the percentage of fraudulent returns by family size shows a difference in fraud based on family size (see Graph 6 below).

**Graph 6: Percent of Fraudulent Returns by Family Size**

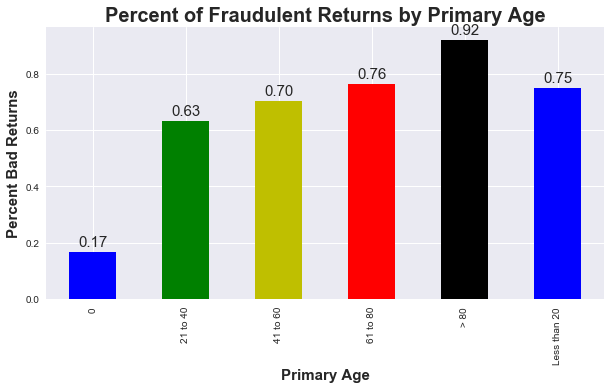
****

There was a statistically significant weak negative correlation between family size and fraud (correlation=-.25, p=1.48e-117). This was most likely attributed to 82% of returns with a family size of 0 being fraudulent.

*Age*

Examining a graph of the percentage of fraudulent returns by age of the primary person on the tax return shows a difference in fraud based on age (see Graph 7 below).

**Graph 7: Percent of Fraudulent Returns by Age**

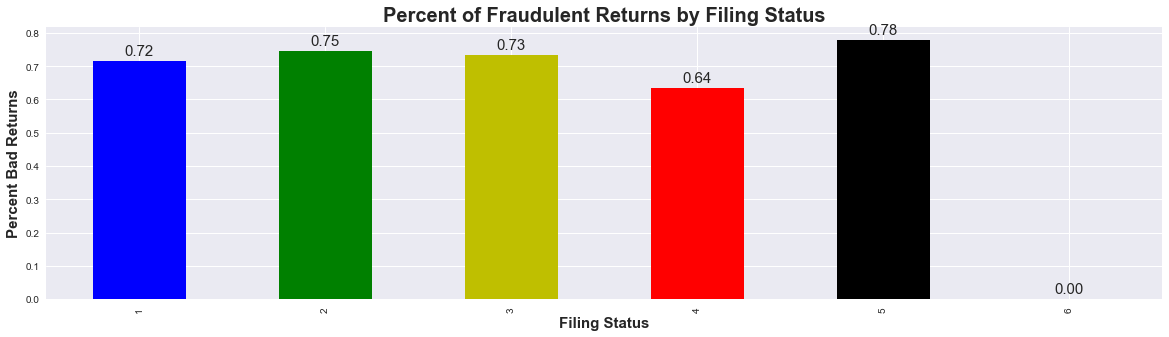
****

When removing cases in which age was listed as 0, there was a statistically significant weak positive correlation between age and fraud (correlation=.12, p=1.30e-27).

*Filing Status*

Examining a graph of the percentage of fraud shows returns filed with a filing status of 4 (filed as ‘Single’) are less likely to be fraudulent (see Graph 8 below).

**Graph 8: Percent of Fraudulent Returns by Filing Status**



Results of chi-square show a statistically significant difference in the proportion of fraud cases for filing status of ‘Single’ compared to all other groups (p=7.61e-20).

Next Steps

I will use the aforementioned variables as inputs (and fraud as the output variable) to build a model to predict fraudulent PTC cases. I will partition the TY 2014 data into a training and test set to build and test the model, and then I will import TY 2015 data to test the model on new data. If possible, I will create an algorithm that can be used for case selection.