IRS Office of Research, Applied Analytics and Statistics

COVID-19 related support

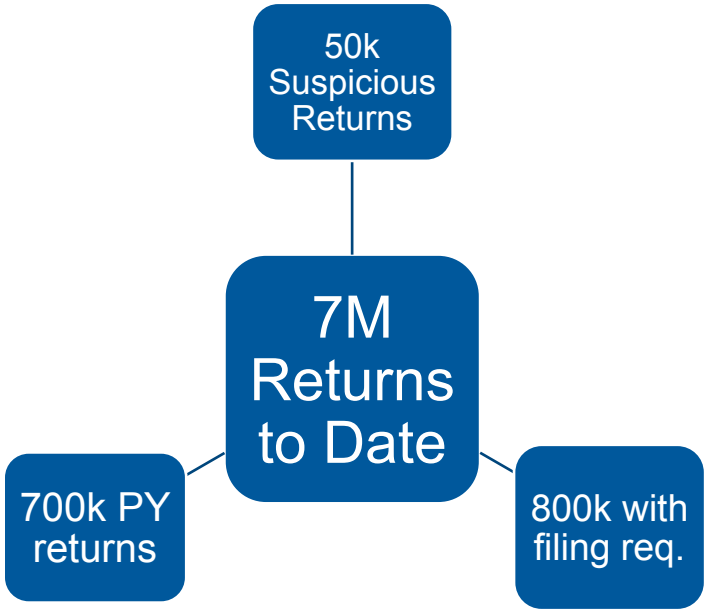
Federal Committee on Statistical Methodology 2020 Fall Conference

Detecting Suspicious Economic Impact Payment (EIP) Returns

August 17, 2020

EIP Overview

Economic Impact Payments (EIP) returns present challenges for ID Theft detection. These returns include little information and filers generally have no consistency or third-party data to help validate legitimate returns.



Total Return Population

- About 7M EIP returns filed to date and almost half were filed by mid-April
- Between 5k and 20k EIP returns come in each day since filing season ended

Unnecessary Filing

- 700k EIP returns had filed a TY 18 return and thus did not need to file an EIP return
- 800k EIP returns had a filing requirement based on third party data

Specific Suspicious Activity

- IRS has explored new methods of aggregating returns using various linking factors and developing new features to protect against IDT

Finding a Signal

Given the minimal amount of information, we employed a new approach to detect ID Theft

- Most EIP returns should have few connections to other returns - If they share linking factors with many returns it raises suspicion
- The number of groups, types of groups, and characteristics of the groups helped inform selections
- Characteristics included previous ID Theft selection activity, reject rates, leads from various sources, and developed features such as e-mail characteristics and zip to IP distance
- Once we had enough selections, we used those as targets in a selection model to help automate the selection process

Some clusters created false signals that required manual review to detect

- Homeless shelters, rehab facilities, convents, and other non-profits had groups of returns that looked suspicious but were likely not IDT
- These returns shared the same address, e-mail and, in some cases, bank account, but appeared to be serving populations that perhaps had no address or bank

Bad Cluster Example

Device	IP	IP-Zip Dist.	Bank Account	Bank	E-Mail	Address
1A1A1…	1.1.1.1	605.9	1234	Bank A	ag123@mail.com	56 Elm
1A1A1…	1.1.2.1	608.3	1234	Bank A	ag126@mail.com	57 Elm
1A1A1…	1.1.2.3	610.2	1234	Bank A	ag133@mail.com	56 Elm
1A1A1…	1.1.1.3	601.3	1238	Bank A	ag136@mail.com	62 Elm
1A1A1…	1.1.1.5	605.3	1238	Bank A	ag155@mail.com	52 Elm

- Returns are clustered by Device and, in this case, seem to be using a dynamic IP address limiting clustering capability for IP address
- All returns are using an IP address that seems to be far away from the taxpayer's zip code
- Returns are going to different accounts but associated with the same bank
- E-mail addresses have a consistent structure across the cluster
- Addresses are very similar across the cluster

Good Cluster Example

Device	IP	IP-Zip Dist.	Bank Account	Bank	E-Mail	Address
2A1A1…	2.1.1.1	3.9	2234	Bank A	jerry@help.org	22 Pine
2A1A1…	2.1.2.1	3.9	2234	Bank A	jerry@help.org	22 Pine
2A1A1…	2.1.2.3	3.9	2234	Bank A	jerry@help.org	22 Pine
2A1A1…	2.1.1.3	3.9	2234	Bank A	jerry@help.org	22 Pine
2A1A1…	2.1.1.5	3.9	2234	Bank A	jerry@help.org	22 Pine

- Returns are clustered by Device and, in this case, seem to be using a dynamic IP address limiting clustering capability for IP address
- Small IP to Zip distance
- Returns are going to the same account
- Same address used across the cluster
- Same e-mail used across the cluster, but the domain is a .org

Manual Investigation and Models

At this point, systematic selection of EIP returns is not quite possible

- The previous examples show we can use clustering to find suspect activity, but differentiating between good and bad can be difficult
- There are bad clusters that look just like the good cluster example and bad clusters that are different from both examples
- Manual investigation allows for a better distinction between good and bad among suspicious returns

Learning from selections may lead to a systematic approach

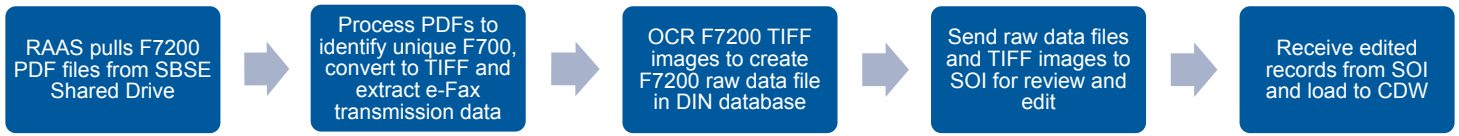
- Over the first six weeks of EIP, teams met daily to discuss findings, emerging patterns and questionable clusters
- These discussions helped identify feature generation possibilities and combinations of features highly associated with ID Theft
- Using early selections as targets, a modeling approach has been implemented which overlaps with manual selections by about 80%

Data Delivery and Support

August 17, 2020

*RAAS was asked to provide data for the **Form 7200 – Advanced Payment of Employer Credits Due to COVID-19.** This information was needed to validate the accuracy of these claims and to support a "true-up" process when the corresponding Form 94X is filed. This was a challenge due to image quality issues arising from E-FAX submission process.*

RAAS DMD Imaging Workflow For Form 7200



Lessons Learned

- E-Fax solution used as submission vehicle resulted in poor image quality making OCR process difficult
- Form 7200 fillable PDF form did not enforce field formats (text, currency, date, phone number, TIN). Field formatting would have improved OCR results
- Existing technologies such as use of bar codes or option to attach completed forms to an e-mail routed to designated e-mail address would have improved data quality and produced significantly shorter processing times
- The vast majority of F7200 submissions required extensive editing to perfect the data set

RAAS ANALYTICS

RAAS has developed reusable natural language processing, computer vision, and information extraction capabilities that address common IRS data challenges

RAAS Capability Key: Natural Language Processing | Computer Vision | Information Extraction

01: Unstructured Text

Unstructured text exists in documents of varying types, often organized by case or collection of documents relating to a taxpayer

02: Scanned Documents & Images

Taxpayers can submit paper versions of forms, which are then scanned and stored as images along side with traditional PDFs

03: Optimization for Volume

IRS has vast amounts of unstructured data that require optimized computing to systematically process and standardize data for analysis

04: Entity Resolution

IRS data refers to taxpayers in a variety of formats across data sources; resolving references is key to connect disparate data sources

05: Excessive Form Text

IRS documents regularly contains boilerplate or standard language that is not otherwise meaningful toward understanding its content or outcome

06: Embedded Tables

Tables are a common data structure within IRS text documents, and are frequently non-standard in layout and location within a document

07: Addendums

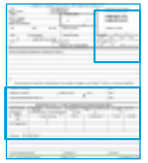
IRS forms often contain addendums with spillover information, which is still pertinent but frequently much less structured than the form itself

08: Form Version & Year Variability

IRS forms are frequently revised across and within years, which leads to inconsistent field presence, spacing, and other related challenges

09: Form Content Extraction

IRS forms are inconsistently transcribed, with specific fields often needed to be identified, extracted, and joined with structured data



Considering broad needs across the IRS to identify, extract, manipulate, and analyze text, Command Line Tools and User Interfaces will allow the IRS to leverage advances from RAAS text analysis efforts to maximize the ability to transfer code pipelines and workflows between projects

10 Command Line Tools

A command line tool allows end users to interact with different capabilities from a consolidated utility

11 User Interfaces

A user interface facilitates wider adoption of developed CLI tools by lowering barriers to use from system commands to be point and click

Form 7200 – Advance Payment of Employer Credits Due to COVID19

SBSE, CI, OTA and RAAS delivery partners were looking to obtain data elements from Form(s) 7200 filed via e-FAX to identify relief payments made, identify potential fraudulent activity and to reconcile with Q2 Form(s) 941 and 941 Schedule R

Business Problem

Due to Campus shutdowns due to COVID19 and the implementation time-frame for the relief provision, taxpayer were instructed to file Form(s) 7200 by accessing and completing a fillable PDF form and transmitting to the IRS via e-Fax. There is a need for the data from these forms to complete the "true-up" reconciliation process, track relief payments made.

Technical Problem

Although the Forms were to be filed via PDF fillable forms, the forms produced did not enforce any data entry standards nor did they leverage computational functionality that would have improved data quality and reduced taxpayer errors in completing the forms. Additionally, instructions for filing the forms were to e-FAX the form to a number provided. E-FAX transmission does not produce quality images for OCR.

To overcome these issue the it was necessary to:

- Employ Kofax solution for data extraction
- Leverage SOI/WI resources to identify and correct OCR issues

Highlights at a glance

22,316 pages processed, across over 13,938 forms filed to date

Produced data extract for e-FAX transmission

Estimated $349M in relief payments issued to 4,842 taxpayers

Technical Capabilities

Text: Page level text extraction, text classification, text characteristics

Computer-Vision: Optical character recognition, form box location

Processing Pipelines: N/A

Capabilities:

01	02	03	04	05	06	07	08	09	10	11

Statistical Tables update for EIP

August 17, 2020

Update on Statistical Tables for Economic Impact Payments (EIP)

The Statistics of Income Division is currently engaged with the Office of Tax Analysis on preparing tables to be released presenting data on the Economic Impact Payments [add link to landing page for tax stats]

- Tables will be released on specific landing page in the Individual area of TaxStats on irs.gov
- Tables will be released monthly
- Currently working with OTA on language and definitions in tables for clarity of what presented
- Three tables planned

- **Table 1** classified by size of Adjusted Gross income (11 different size classes)
- Information on number of payments and amounts for:
  - Total EIP
  - EIP paid electronically
  - EIP paid by paper checks
  - EIP paid by debit card

- **Table 2** classified by State
- Information on number of payments and amounts for:
  - Individuals with and without qualifying children
  - Total EIP
  - EIP paid electronically
  - EIP paid by paper checks
  - EIP paid by debit card

- **Table 3** classified by marital status
- Information on number of payments and amounts for:
  - Total EIP