Cloud Spend Classification

General Services Administration | IT Category Team, Federal Acquisition Service

Theodore Jeliazkov, Data Engineering Fellow | *Cornell University, Statistics*

Keywords: Machine learning, spending analysis, keyword analysis, classification

Summary:

Spending analyses identify how much the government spends on specialized products. In many cases, the government is interested in cloud services spending. The current method for identifying cloud spend from government obligations data searches contract descriptions for cloud-related keywords to identify cloud spend. However, this approach requires manual maintenance of keyword lists and data cleaning after the keyword search has been run. Here, Theodore develops a proof-of-concept model which uses contract descriptions to classify cloud spend. The developed model is precise, and captures most cloud spend. Developing a production-level model for specialized product spend classification has the potential to decrease classification time from 5 hours to 3 seconds, and fully automate spend analyses that typically take weeks to complete.

U.S. General Services Administration
OFFICE OF INFORMATION TECHNOLOGY CATEGORY

Cloud Spend Classification

Theodore Jeliazkov - QT1D

For Government Use Only

Background

- QT1D conducts cloud spending analyses
- Identifying relevant contracts is challenging
- Keyword search is the current approach
- Keyword method has some drawbacks
  - Keyword list must be manually maintained
  - Search results need to be verified

Objective

Build a machine learning model to automate cloud contract classification.

Methodology

- Acquire FPDS data for past 5 fiscal years
- Identify cloud contracts using keyword search method
- Train model using contract descriptions and cloud labels

Results

- 99.4% accuracy
- 93% precision
- 79% recall

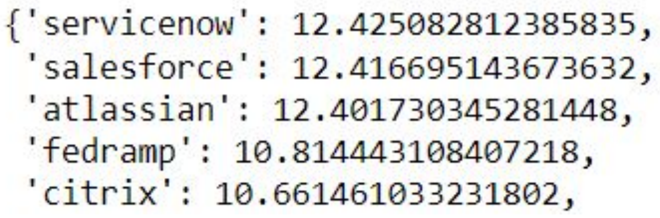
Results

- 3.5s runtime on 35 million data points

Results

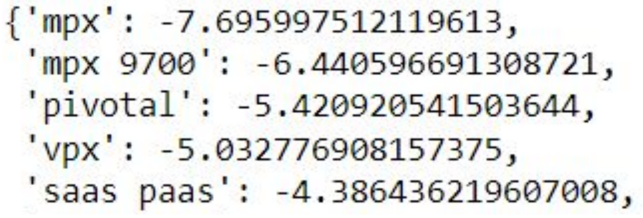
- Keyword importance values

Top 5 cloud words



```
{'servicenow': 12.425082812385835,
 'salesforce': 12.416695143673632,
 'atlassian': 12.401730345281448,
 'fedramp': 10.814443108407218,
 'citrix': 10.661461033231802,
```

Top 5 non-cloud words



```
{'mpx': -7.695997512119613,
 'mpx 9700': -6.440596691308721,
 'pivotal': -5.420920541503644,
 'vpx': -5.032776908157375,
 'saas paas': -4.386436219607008,
```

Implications

- Feedback on current keyword method
  - Keyword list can be improved
- Accuracy can be quantified
- Fully automated spend analyses
  - Model can run on FPDS updates in <1 second
  - Connect FPDS to dashboard for real-time updates