Dene Stalk

BrainStation Data Science Diploma Capstone Final Report

March 2020

Executive Summary

As the COVID-19 pandemic continues to drive economic turmoil, the vast majority of traditional blue-chip companies and small organizations will continue to experience increased liquidity squeezes and concerns due to downward pressures on revenues. Managing working capital is further complicated by collection of accounts receivables uncertainty. To remain competitive, even to just remain solvent for some, cutting costs should be a top priority for every executive's cash flow strategy - especially when banks inevitably tighten up their lending.

To shore up margins through decreased expenditures without detrimental consequences to competitive advantage, an organization needs to, firstly, be forthright with their purchase order spending to lower Total Cost of Ownership[1]. This begins with an accurate understanding of what is being bought, who is buying it and who they are buying it from. As straightforward as this seems, it continues to be one of the biggest challenges organizations struggle with. This is mainly owing to inadvertently siloed spend classification processes as a by-product of different nomenclatures across regions, languages and departments.

Conventional approaches to detailed spend analysis are labour intensive and prone to human error – leading to continuous rework. This results in increased man-hours, frustration and resources as it takes an average of one year to identify and correctly classify 100,000 line items[2]. This scale of spend misclassification exists for many large non-streamlined organizations and leads to further compounding costs when disregarded. The organizations who successfully manage their spend can cut incremental costs by 5% to 10%[3].

The Information Age is well into its adolescence and innovative companies have an undeniable opportunity to improve their strategic procurement through use of machine learning. Big and small organizations, public sector and corporate alike, are wising up to the competitive advantages of being a first mover in leveraging machine learning by either developing capabilities inhouse, investing in targeted start ups or outsourcing to one of the increasing number of providers in this rapidly growing space, such as Sievo and Zycus.

Ideally an organization should tailor their spend taxonomy to capture how suppliers approach their market space to identify the best opportunities for arbitrage or leverage on volume for more favorably negotiated rates. However, this capstone took a standardized approach by using the UNSPSC Segment[4] mapped to the Government of California's 2012-2015 purchase order spend. With the goal to keep the model as transferable as possible - only supplier name, item name and item description columns were used as features. By incorporating three merged separate TF-IDF Natural Language Processing models then running it through a three classifier soft voting ensemble[5] which was able to predict 77% of line items and 92% of total spend on 65,000 rows of unseen data[6].

A few of the strengths of the model include relatively strong predictive value using only 3 free text columns and 17 categories' line items correctly classified with over 70% accuracy. However, some weaknesses I can see are possible low replicability to corporate data and 3 categories < 50% of line items correctly predicted.

Next steps include a meta-model incorporating quantity and unit price, a deeper analysis on strategic value in relation to supplier management, test transferability on new Government of California data and corporate data and creating a pyLDAvis interactive dashboard.

[1] TCO is a combination of purchasing and carrying costs
[2] Zycus – How to mint millions from your spend data
[3] BCG – Delivering on digital procurement's promise
[4] United Nations Standard Products & Service Code
[5] Soft voting models included OnevsRest Logistic Regression (4), ExtraTrees (3) and LinearSVC (0.5)
[6] See stand alone deck and code submissions enclosed for further information

Capstone Overview

Objective: Predict spend categories

Dataset: Government of California's (GoC) 2012-'15 purchase orders.

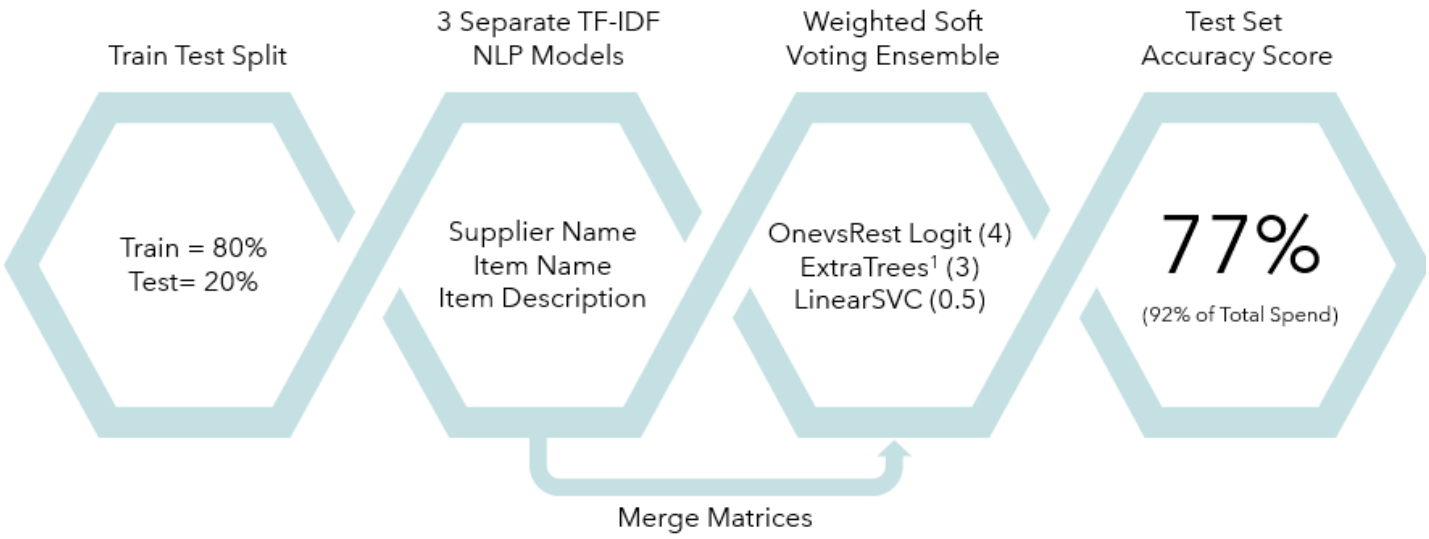
Uncleaned Dataset Description: 344k rows and 33 columns, many unique to GoC and disregarded

Cleaning Process: Deleted rows with Supplier Name = Unknown, Total Price <=0, Item Name = blank or 'confidential', bottom 20% of categories by line items placed into either Other Services or Other Goods, Segment = blank

Cleaned Data: 4 columns – Supplier Name, Item Name, Description Name and Segment

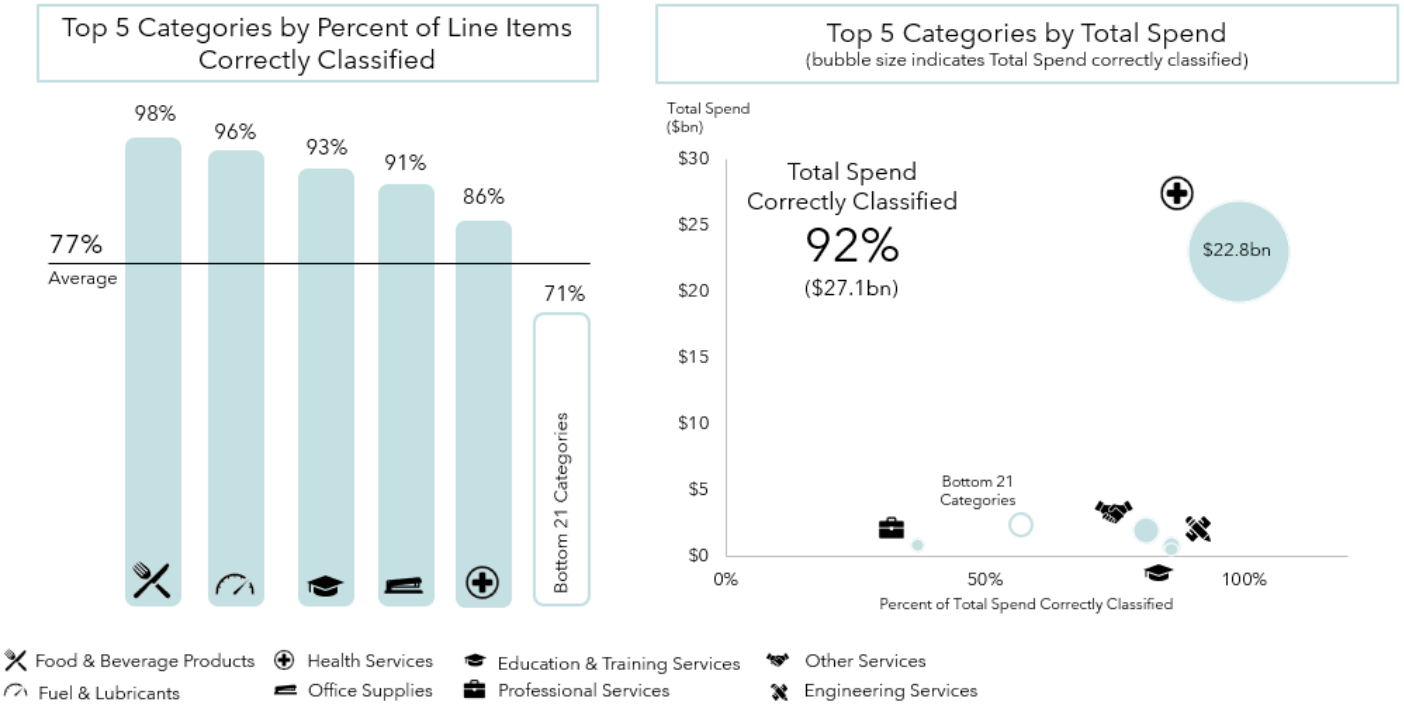
Modelling Methodology

Numerous other models where ran during the modelling phase such as Word2Vec neural networks, various stacking and voting ensembles on GCP. However, the below model delivered the best cross validation accuracy score of 76%, resulting predicting 77% of unseen data.



Results

The model strongly predicted F&B, Fuel, Education Services, Office Supplies and Health Services. Health Services' percent of spend was strongly predicted (99%) and fortunately in this case had by far the highest spend – helping drive the model predicting 92% of Total Spend.



Conclusion

The model actually exceeded my expectations on accuracy for 26 categories with 77%. However, it's applications to corporate data may be limited as a result of differences between the public sector and corperate spending focuses. Further analysis include exploring a meta-model incorporating quantity and price.

Model Strengths

- Relatively strong predictive value using only 3 free text columns
- Correctly classified the vast majority of spend
- 17 categories' line items correctly classified with over 70% accuracy
- Model is scalable

Model Weaknesses

- Not tested on corporate data
- Health Services over represented in terms of Total Spend
- 3 categories < 50% of line items correctly predicted

Further Analysis

- Meta-model incorporating quantity and unit price
- Deeper analysis on strategic value in relation to supplier management
- Test transferability on new Government of California data and corporate data
- Create an interactive dashboard