Government Contract Analysis and Automated Curation

by : Emily Anderson, Jordyn Gaither,

Matthew Vaden, Hunter Wilson

In partnership with:

Table of Contents

Introduction

PeopleTec provides emerging technologies, engineering solutions, modeling and simulation, cybersecurity, and mission operations and program support services to the Department of Defense and Civilian Federal sectors. To assist with the company's workflow this project seeks to analyze contract data to find and curate open contracts that will be bid on by PeopleTec.

In order to accomplish this, we analyzed previous contracts to predict which open contracts PeopleTec is likely to win. To begin we determine which contracts PeopleTec is capable of willing to win. As not all contracts are for sectors that PeopleTec intends to win. Therefore, using information given to us by PeopleTec we will pull contracts related to the NAICS codes given to us by the company.

We can gather this data through an open source site called FPDS, which holds all contracts greater than $10,000.00. From here we can combine these files in Python giving us our base dataframe.

Objectives:

- Conduct exploratory data analysis; use derived metrics to verify current approach
- Test models to confirm predictability; present metrics showing reliability
- Provide probability scores attached to open contracts that PeopleTec can win

Assumptions:

- Data provided is limited open source data to maintain company and government confidentiality

Project Deliverables:

- Verify/validate current statistical prediction methods
- Design and develop improved prediction method(s)
- Deliver prediction method(s) via report and code
- Provide PeopleTec, Inc. with a working dashboard that displays open contracts segregated based on the working groups inside the company

Data Description

Below is the data dictionary, which furnishes crucial details to enhance comprehension of its broader context, application, and the interconnections pertinent to its real-world utility.

Field	Type	Description
Action Obligation ($)	Float64	amount obligated or de-obligated by this specific transaction
NAICS	Int64	classification of business establishments by type of economic activity
PSC_encoded	Int32	identifies the products, services, or R&D
PSCType_encoded	Int32	encoded representation of the PSC type
AgencyID_encoded	Int32	encoded representation of Agency Identification

Table 1: Data Dictionary

Feature Selection

For our analysis, once we have gathered our data the next step is to analyze our variables to determine which are useful for analysis and which are not. We must also determine our target variable. In this case that is contracts won by PeopleTec Inc. This variable can be created in a binary class by creating a column for contracts where the business name is PeopleTec. Doing

this we find 235 contracts that are owned by PeopleTec Inc. Many of these contracts, however, are not useful for analysis. This is due to the fact that PeopleTec is a relatively new company beginning operations in 2010. Therefore predicting contracts prior to 2010 will hurt our model as patterns will become murky due to the irrelevant data.

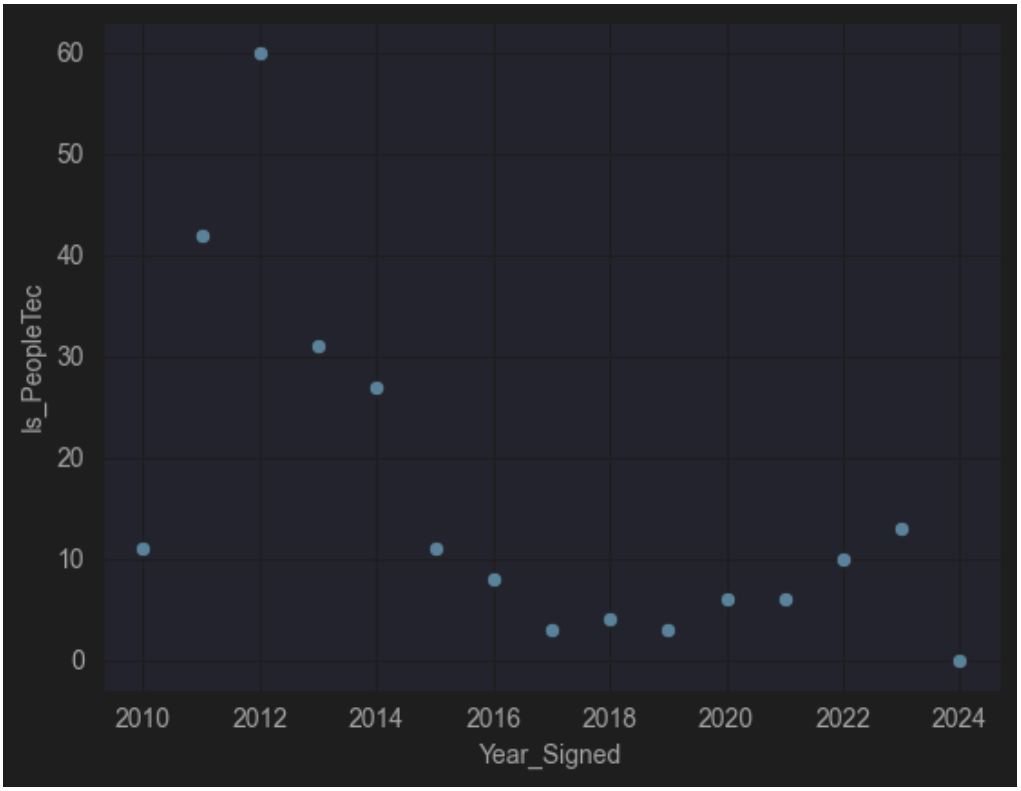


Figure 1: PeopleTec's contracts over time. (Note: Covid Lockdowns from 2019-2021)

After we encode some of our other variables we are left with Action Obligation ($), PSC_encoded, PSCType_encoded, AgencyID_encoded and our target variable Is_PeopleTec. We will scale our variables in order to achieve the best model performance.

Exploratory Data Analysis

Our initial dataset contained the following variables, but through analysis and other methods as we developed our models, we were able to narrow it down to 5 important predictors:

Action Obligation ($)	float64
Additional Reporting Code	object
Additional Reporting Description	object
AgencyID_encoded	int32
Award/IDV Type	object
CAGE Code	object
Contract ID	object
Contracting Agency	object
Contracting Agency ID	object
Contracting Office Name	object
Day_Signed	int32
Entity City	object
Entity State	object
Entity ZIP Code	object
Is_PeopleTec	bool
Legal Business Name	object
Modification Number	object
Month_Signed	int32
NAICS	int64
NAICS Description	object
PSC	object
PSC Description	object
PSC Type	object
PSCType_encoded	int32
PSC_encoded	int32
Reference IDV	object
Solicitation Date	object
Transaction Number	int64
Ultimate Parent Legal Business Name	object
Ultimate Parent Unique Entity ID	object
Unique Entity ID	object
Unnamed: 26	float64
Year_Signed	int32

Figure 2: Initial dataset

Below is a heatmap showing the correlation between variables. To no surprise, PSC_encoded and PSCType_encoded are highly correlated.

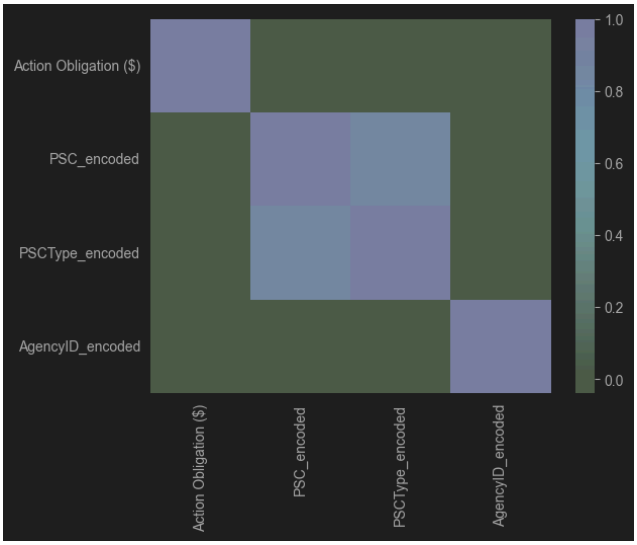


Figure 3: Correlation Heatmap

The below figure shows the quantity and distribution of “PSC_encoded”.

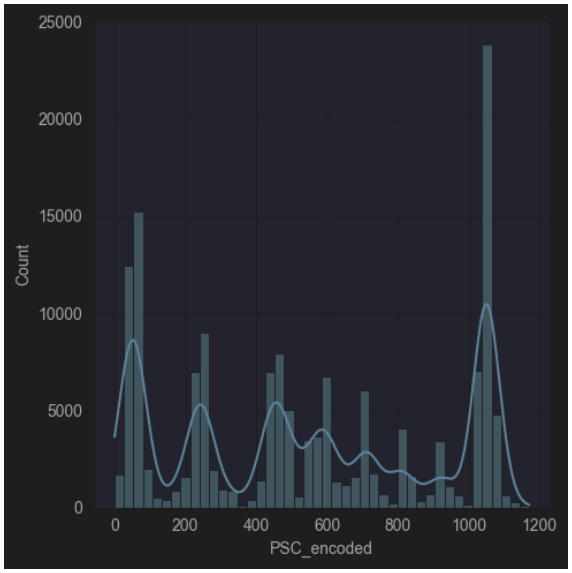


Figure 4: PSC_encoded distribution

Methods and Models

SMOTE:

One issue with our model is the imbalanced shape. The difference in PeopleTec contracts to non-PeopleTec contracts can worsen our models' predictiveness. In order to adjust for this we will use SMOTE. Synthetic Minority Oversampling Technique, or SMOTE for short, is a way of fixing an imbalance available in the imblearn package.The process involves choosing similar examples in the feature space, connecting them with a line, and generating a new sample at a point along that line. Initially, a random instance from the minority class is selected. Subsequently, its k nearest neighbors (k=5) are identified. One of these neighbors is randomly chosen, and a synthetic example is produced at a randomly selected position between the two instances in the feature space.

Model Evaluation:

As our model's goal is to predict the contracts that PeopleTec can win, our aim for the model is not perfect accuracy. Winning a contract is not an exact science, so as a result we should expect our model to predict contracts the company has not won. Our goal is to reduce the total contracts to a reasonable amount that PeopleTec can target which the company can reasonably expect to win. In order to validate our model, however, we need the model to be able to reasonably predict contracts won by PeopleTec while also maintaining a fairly high accuracy score. Therefore, the two statistics that matter to determine the models success is the models accuracy score and recall.

Classification Tree

For our first model we will create a simple classification tree in order to establish a baseline performance standard.

Test accuracy: 0.972958764580253

Train accuracy: 0.9878372820006581

Confusion Matrix:

[[29589 789]

 [34   23]]

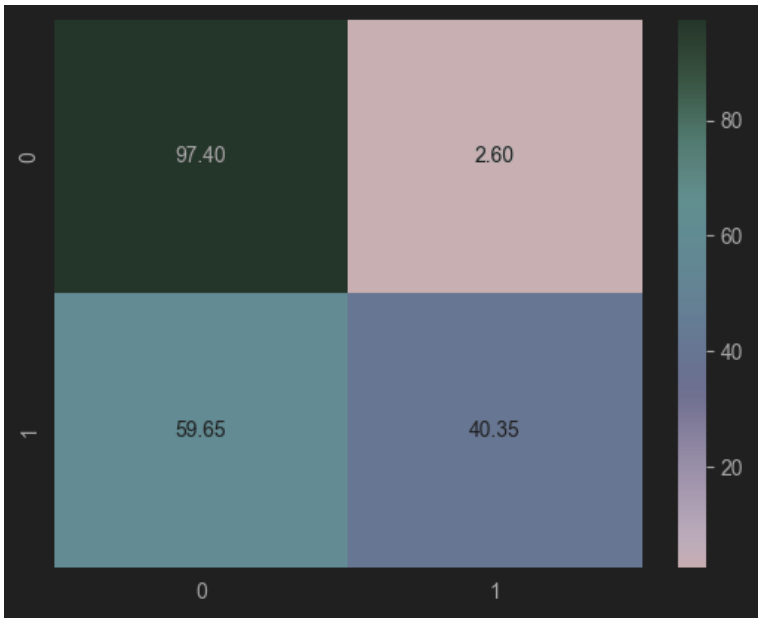


Figure 5: Standardized heatmap

Our initial model provides promising results. It predicts less than 850 of the roughly 30,000 total contracts to be won by PeopleTec and gets 23 correct. This model is able to significantly decrease the total contracts that PeopleTec would need to look at. For future models however we would like to optimize recall in order to miss as few of the company's contracts as possible.

Random Forest Classifier

Our next model will be a random forest classification model. Random forests are an ensemble method, meaning they combine predictions from other models. This will allow the model to improve performance by using previous results to make predictions as to how to better predict our target variable.

Test accuracy: 0.9720716280597995

Train accuracy: 0.9853364593616322

Recall: 0.5964912280701754

Confusion Matrix:

[[29551 827]

 [23   34]]

Our results from the model are significantly better as we the model is able to predict 11 more of the companies contracts while only predicting 30-40 contracts the company did not win.

Gradient Boosting

The next model we will create is using SMOTE and gradient boosting from the XGBoost package in unison.

Our results are

XGBoost SMOTE Test accuracy:

0.9465746673238048

XGBoost SMOTE Confusion Matrix:

[[28768 1610]

 [16   41]]

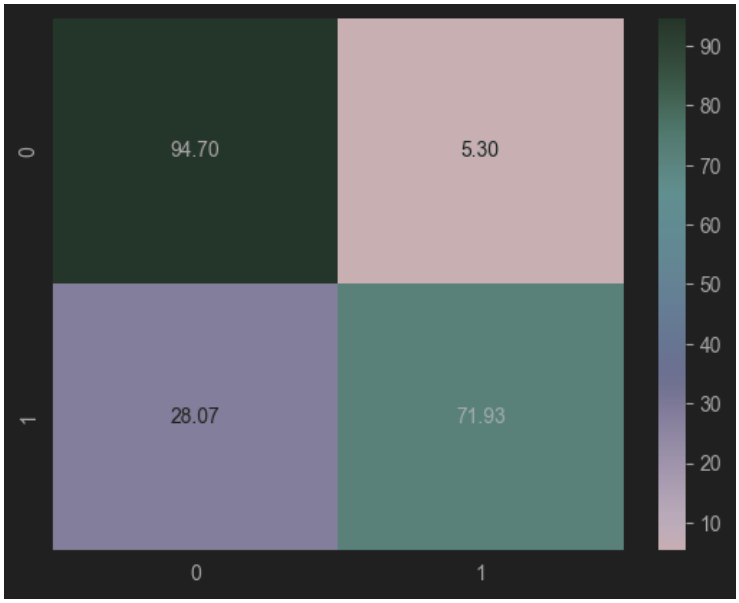


Figure 6: Standardized Heatmap - Gradient Boosting; 72% correctly predicted; 28% incorrect

False Negative Rate (FNR): 0.2807017543859649 .

False Positive Rate (FPR): 0.052998880768977547

Misclassification Rate (MR): 0.05342533267619517

SMOTE Area Under the Curve (AUC): 0.9469759971724689

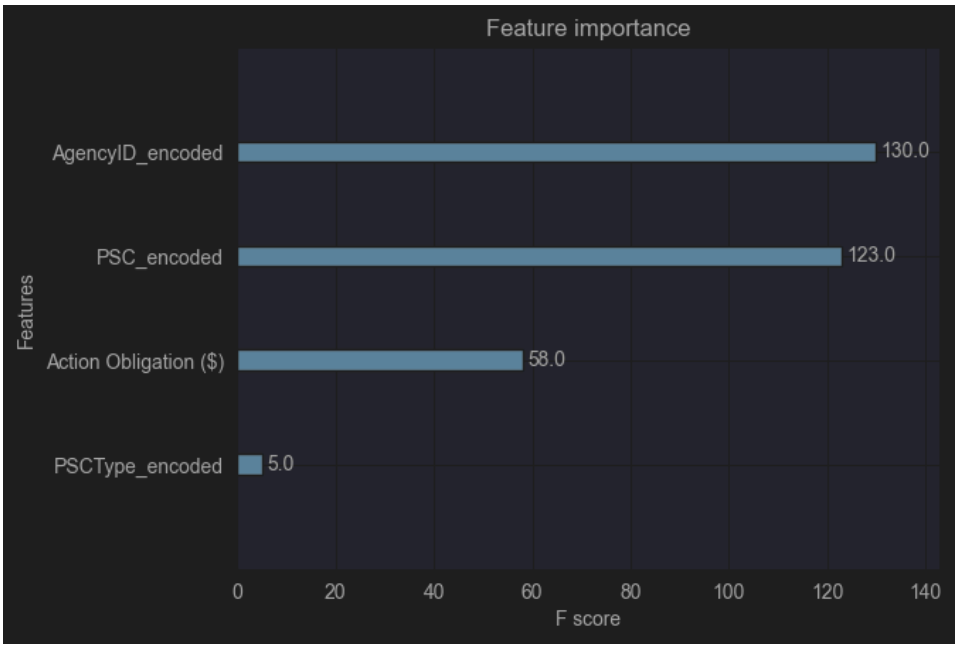


Figure 7:
Feature Importance - Gradient Boosting

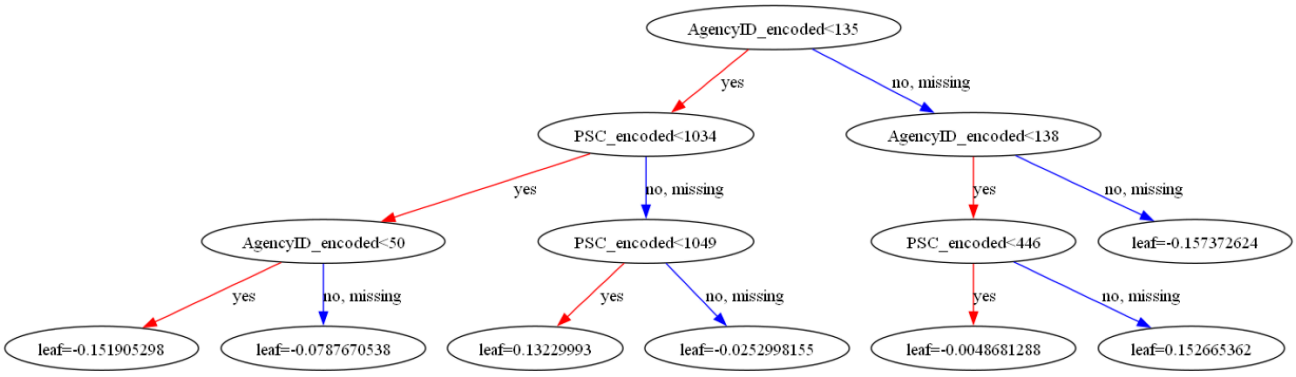


Figure 8: Gradient Boosting Tree

As we can see this model performs well. By using the type of contract, the size of the contract, and who is awarding the contract the model is able to remove 28,768 of the contacts in the test. It predicts 41 of the 57 total contracts correctly with 1,610 other contracts that the model believes PeopleTec should win based on the historic data.This shows the power of fixing the imbalanced class in our data.

Optimized

Our gradient boosted model has significantly better recall than our previous models. As a result, we will attempt to optimize the previous model we created. We can optimize our parameters using gridsearch. The parameters we will try to optimize are max_depth and learning rate. to get these results:

Best model accuracy on test data: 97.01%

Best parameters found: {'learning_rate': 0.3, 'max_depth': 7}

XGBoost SMOTE Test accuracy: 0.9700673566617382

XGBoost SMOTE Confusion Matrix:

[[29494 884]

[27 30]]

False Negative Rate (FNR): 0.47368

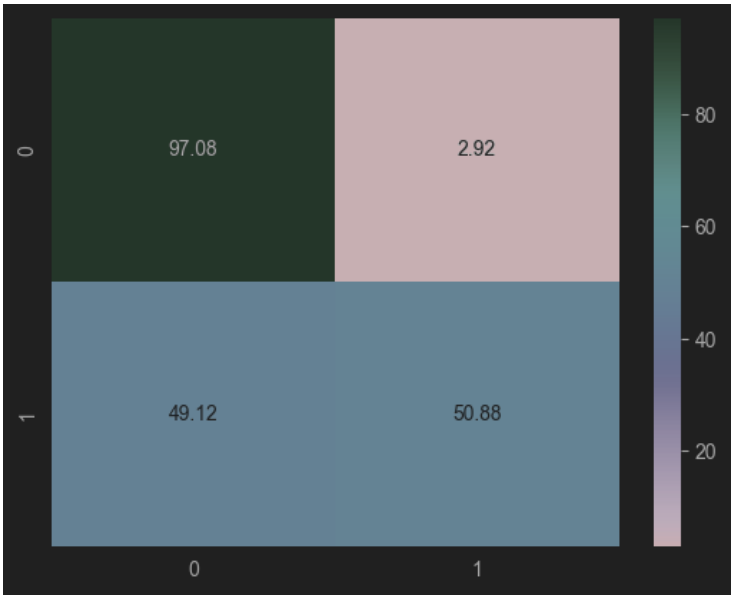


Figure 9: Standardized Heatmap - Optimized

False Positive Rate (FPR): 0.02910

Misclassification Rate (MR): 0.0299326

SMOTE Area Under the Curve (AUC): 0.7486

As we can see from the above statistics we have a much lower misclassification rate. However, we have a trade-off of a decrease in our models ability to accurately predict which contracts PeopleTec wins. We can also see that our feature importance turns out vastly different between the two models with Action Obligation going from third to first place in the optimized model.

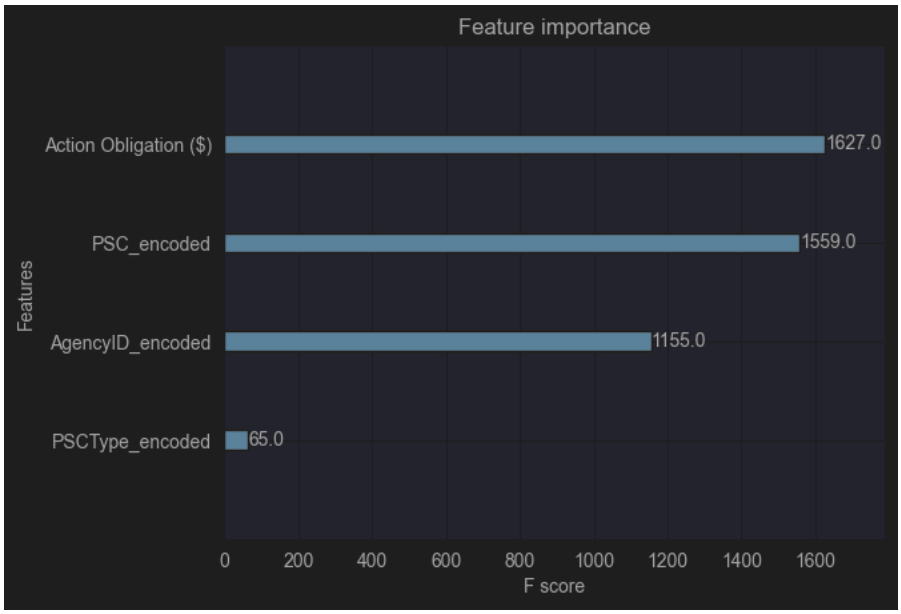


Figure 10: Feature Importance - Optimized

Interpretation

Our best models are the standard XGboost model and Random Forest Classifier. Although other models possess a higher accuracy than the XGBoost the decrease in recall scores hurt the models' reliability. XGBoost, however, is able to accurately predict a large majority of the companies' contracts without a large number of false positives. The Random Forest model however does cut the number of false positives in half, while only missing 7 more of PeopleTec's 57 total contracts. This tradeoff allows both models to be useful predictors, and with further feature selection and optimization, we could see both models have significant increases in performance.

Recommendations

One of our variables is Action Obligation. This variable, however, is not present in open contracts therefore it can't be used for predicting open contracts. However, this variable could be created if a contract description is available. By analyzing contract descriptions against previous contracts, action obligation can be predicted allowing the open contracts to be predicted with better accuracy.

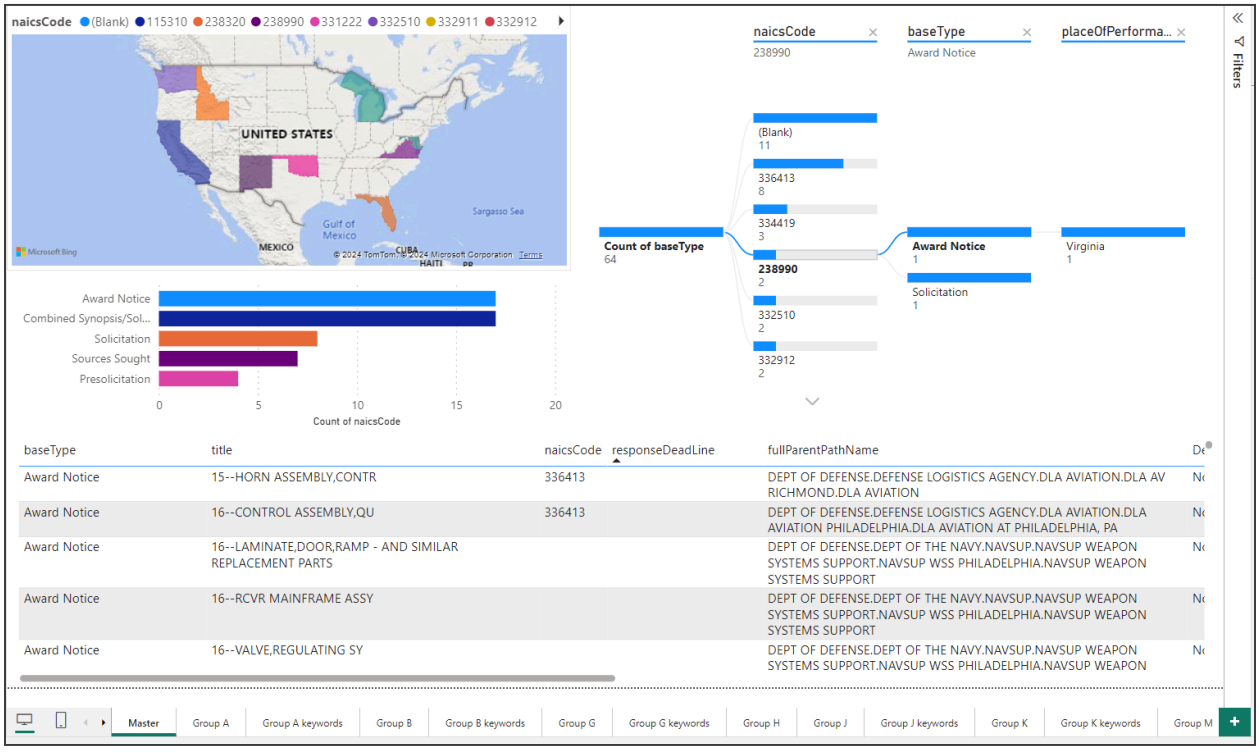
Dashboard

One of the other objectives for this project is to produce a usable dashboard that updates automatically with government contract data as it is published. We built this dashboard in Microsoft Power BI, utilizing API calls to sam.gov in order to obtain data as it is published.

All

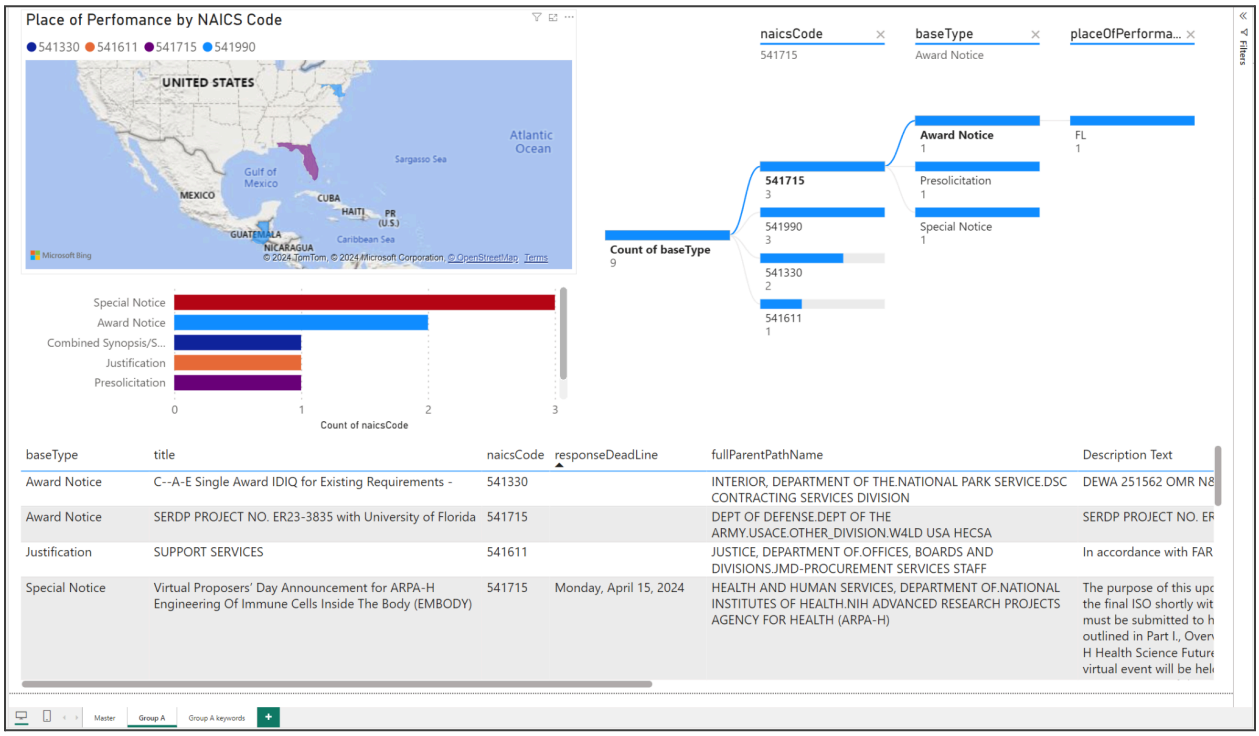
```
let
    // set today's date in the system, then -1 to get yesterday (change to -2 for two days ago, etc)
    today = DateTime.Date(Date.AddDays(DateTime.FixedLocalNow(), -1)),
    // change "today" into correct format to pass to API
    todayfilter = Date.ToText(today, "MM/dd/yyyy"),
    // make call to api.sam.gov, api_key is unique to individual, limit is set to 100 due to limited api key, posted from and to gives just things posted yesterday (could change to cover a certain period of time vice one day)
    Source = Json.Document(Web.Contents("https://api.sam.gov/prod/opportunities/v2/search?api_key=9ntGHL3Eye4JJAHN5FES8E8Os4pymcDpSUOcJX1a&limit=100&postedFrom=" & todayfilter & "&postedTo=" & todayfilter)),
    // "opportunitiesData" is the dataframe we actually care about
    opportunitiesData = Source[opportunitiesData],
    // convert to a table to allow work
    #"Converted to Table" = Table.FromList(opportunitiesData, Splitter.SplitByNothing(), null, null, ExtraValues.Error),
```

The above image shows the method for the API call that was utilized in the main data set using Power Query. This API call set a variable (today) and assigned it to the date yesterday, then passed this along with other necessary elements to the API and took the returned data and cleaned it to a usable format. This resulted in a data table with 49 columns and as many rows as were requested from the API. This data included all actions from the government yesterday, to include solicitations, awards notices, etc. The data was then taken and turned into a visual dashboard in Power BI and resulted in the following dashboard.

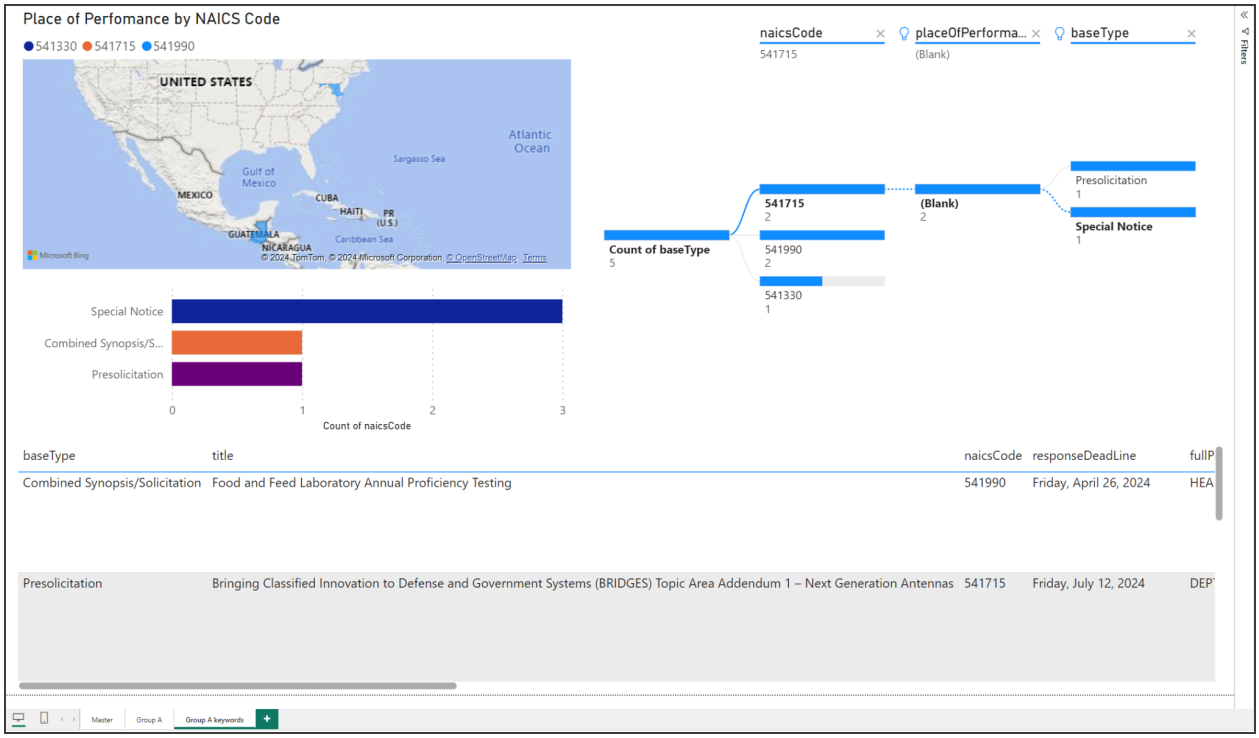


From here, the data was further filtered by NAICS codes that were relevant to various groups within Peopletec. As an example, the dashboard for Group A within PeopleTec was filtered to

only include data from rows that were relevant to NAICS codes 541330, 541611, 541715, and 541990. This gave the following dashboard.



Following this, the data was further filtered to look for certain keywords within each group. This resulted in two dashboards per group, one more broad and one more filtered on specific keywords within the first data set. Filtering the data from Group A gave the following dashboard.



This dashboard is extremely customizable and tailorable for the various PeopleTec groups. Additionally, it can be set to update at predetermined intervals (every day, every time the file is opened, etc.) and will give PeopleTec the ability to save time and effort manually filtering through contract opportunities on a daily basis.

Conclusion

There are opportunities for PeopleTec to design and implement a system of models that will be a support tool in the effort to more efficiently pursue and win contract solicitations. This project provides insight with empirical results that there is a substantiated level of confidence with the limited amount of information we had available. As PeopleTec incorporates proprietary information, this concept of proof will transform into a model of higher fidelity, aiding the company in saving time, resources, and increasing productivity.

Appendix

Scope of Work – Key Artifacts

Deliverable	Date
Statement of Work	2-Feb-2024
Midterm Presentation	15-Mar-2024
Final Poster	15-Apr-2024
Analytics Showcase	22-Apr-2024
Final Presentation	3-May-2024
Final Report	3-May-2024

Master Schedule

The master schedule depicted below is planned out to optimize all of the required tasks. Milestones are shown as the blue squares located in-between tasks. Dates and tasks may vary depending on sponsor feedback.