Highlighted fields appear in Open990.com's ~100-variable snack-sized dataset (topic: "Contractor Compensation") for organizations that e-filed a Form 990 tax return for 2016. More information is available in this dataset's Variable Guide. Attribute dataset to Open990.com.

Form **990**

Department of the Treasury
Internal Revenue Service

Return of Organization Exempt From Income Tax

Under section 501(c), 527, or 4947(a)(1) of the Internal Revenue Code (except private foundations)

▶ Do not enter social security numbers on this form as it may be made public.

▶ Go to *www.irs.gov/Form990* for instructions and the latest information.

OMB No. 1545-0047

2017

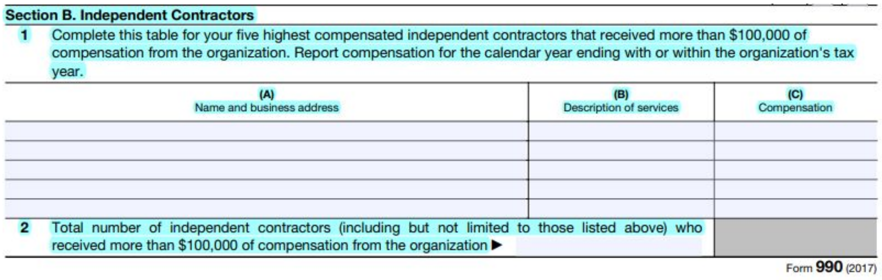
Open to Public Inspection

Classification with Not-For-Profit Tax Returns

Can machine learning be used to identify not-for-profit orgs that might be at higher risk for contractor payment fraud?

Dataset Overview

- Dataset obtained from Open990(https://www.open990.org/catalog/), a site hosted by a company that specializes in not-for-profit tax return research (Applied Nonprofit Research, LLC)
- Prior to 2016 this data was extremely difficult to obtain and work with
- Contractor payments in these tax filings were particularly interesting to me since billing schemes are one of the most common types of fraud among not-for-profits.
- Dataset only includes returns for FY2016.
    - Not-for-profits must list the number of contractors who received in excess of $100K
    - They must also list information for the top 5 contractors paid in excess of $100K
        - Contact information
        - Description of services rendered
        - Amount paid



Section B. Independent Contractors

1 Complete this table for your five highest compensated independent contractors that received more than $100,000 of compensation from the organization. Report compensation for the calendar year ending with or within the organization's tax year.

(A) Name and business address	(B) Description of services	(C) Compensation

2 Total number of independent contractors (including but not limited to those listed above) who received more than $100,000 of compensation from the organization ▶

Form **990** (2017)

Data Preprocessing

```
# formation year has a number of NaN values
# Convert these to None to be able to encode properly

df.loc[df.formation_yr.isna(), 'formation_yr'] = 'None'
```

- Narrowed dataset to only not-for-profits with at least one contractor that was paid in excess of $100K - this left me with 30,185 not-for-profits.
  - I wanted to focus specifically on contractor payments
  - The organizations that pay contractors in excess of $100K in a give year are much larger in scale than most not-for-profits
- Properly encoding data, from object to bool format
- Many columns had NaN for zero values, or the absence of a value
- Organized 501(c) type into single column
- Reconciling "contractor_100K_ct" "with amt_paid_contractor_1"
  - 54 organizations had payments over $100K, but stated that they their "contractor_100K_ct" was zero.
- 3,047 not-for-profits with over $100K in spending to a not-for-profit had zero employees or volunteers. For the analysis of my project, these seemed to be non-standard not-for-profit organizations, so I made the decision to drop them, leaving me with approximately 27K rows.

```
df.grp_return.value_counts()
```

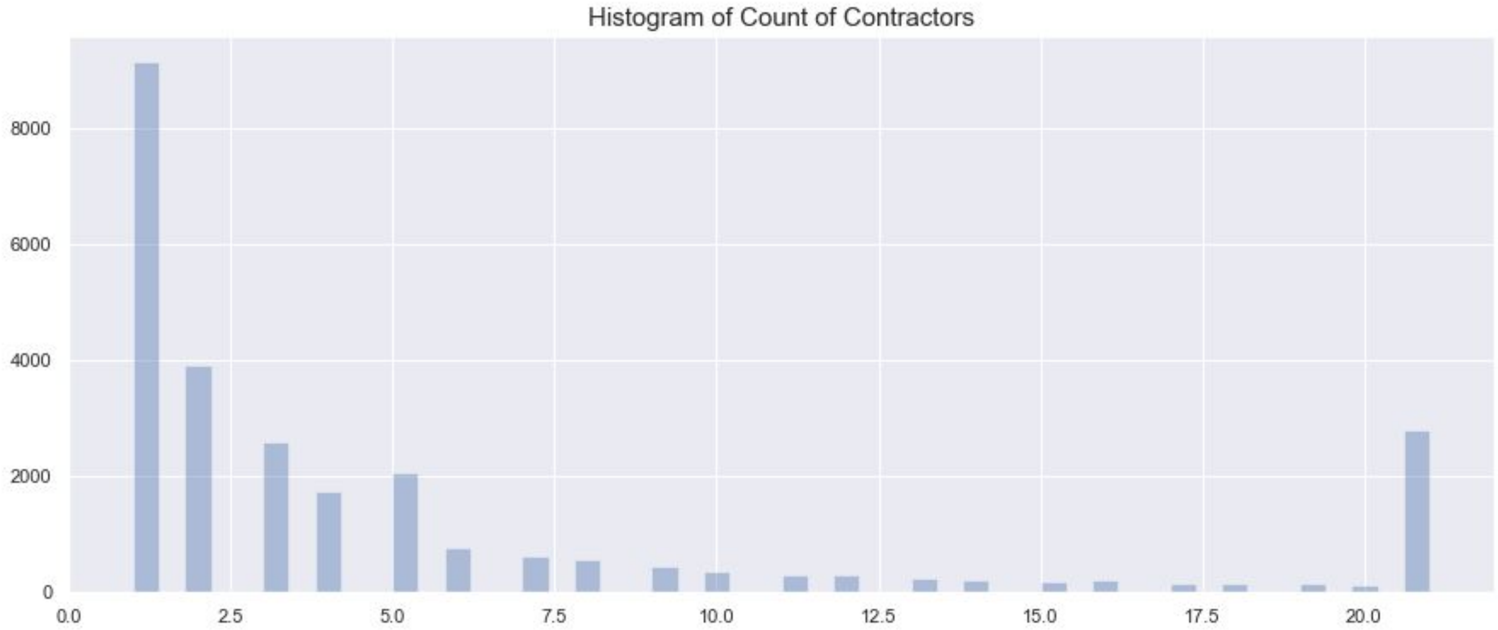
```
no      26777
yes       104
Name: grp_return, dtype: int64
```

```
# what other 501c types are there?
df['exempt_status_501c_txt'].value_counts()
```

```
3_     21229
6_      2163
12_      655
4_       623
5_       577
9_       544
14_      512
7_       377
8_        57
13_       56
2_        28
25_       15
10_       14
19_       11
29_       11
27_        3
26_        2
23_        2
18_        2
```

EDA - Contractor Count Distribution

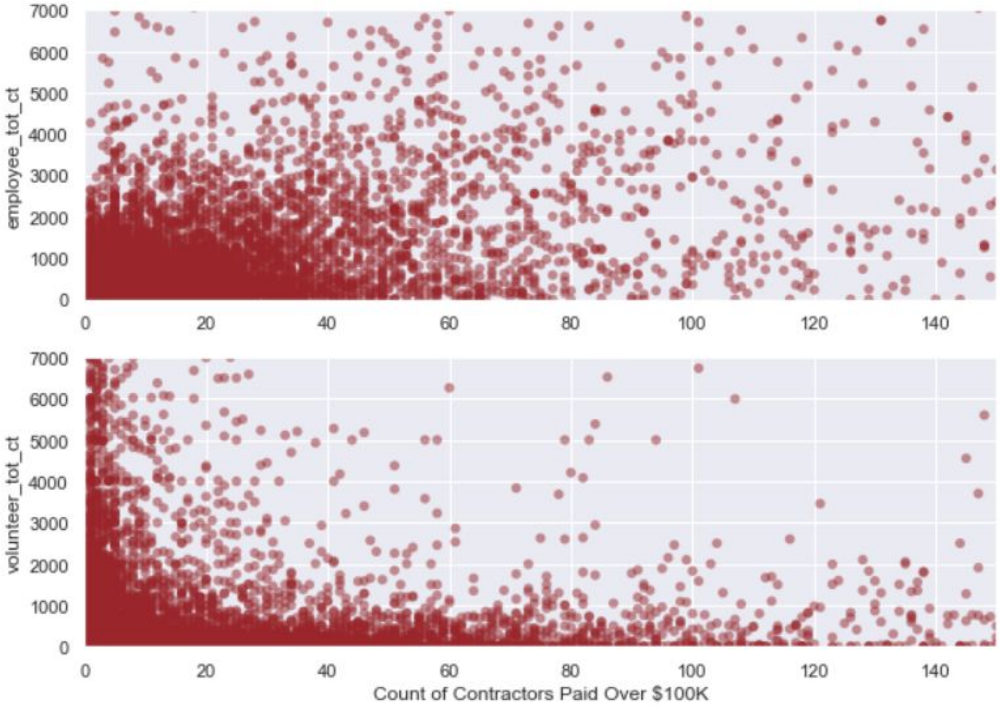
Histogram of contractor_100K_ct, with top 10% winsorized. There are some large outliers to keep in mind.



EDA - Employees & Volunteers

Scatterplot showing relationship of the count of contractors to the count of volunteers and employees.

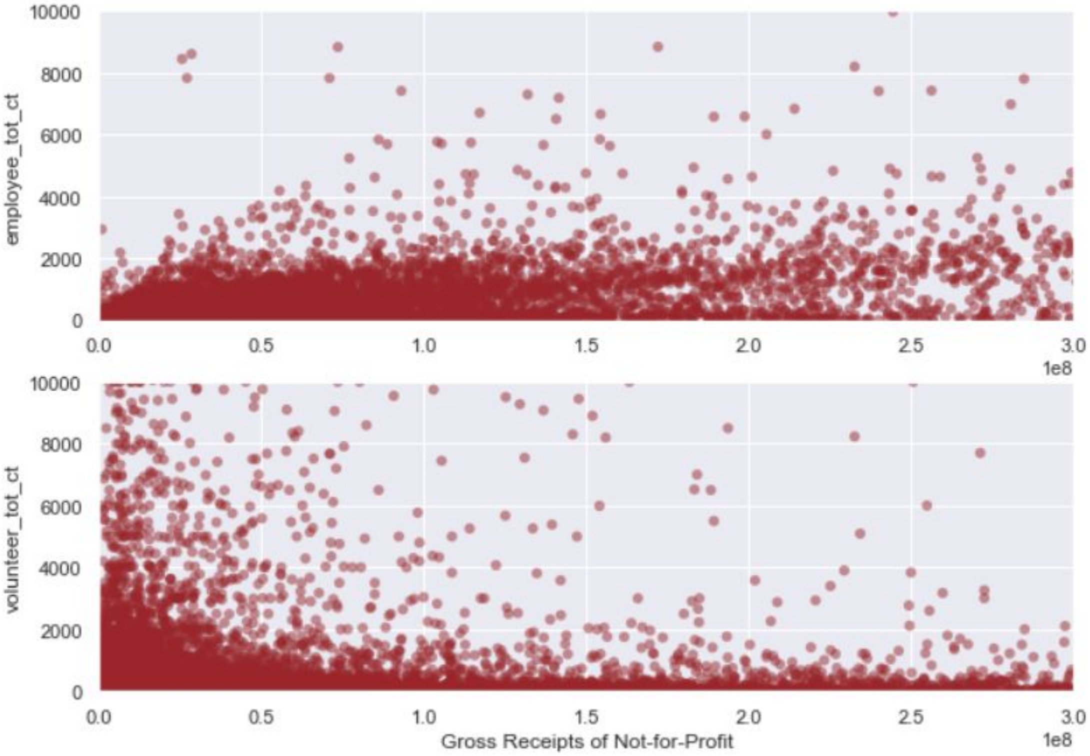
- Employees increase as contractors increase up to about 80 contractors.
- Volunteer count is inversely related to contractor count. However the count of volunteers remains more consistent beyond 80 contractors



EDA - Gross Receipts

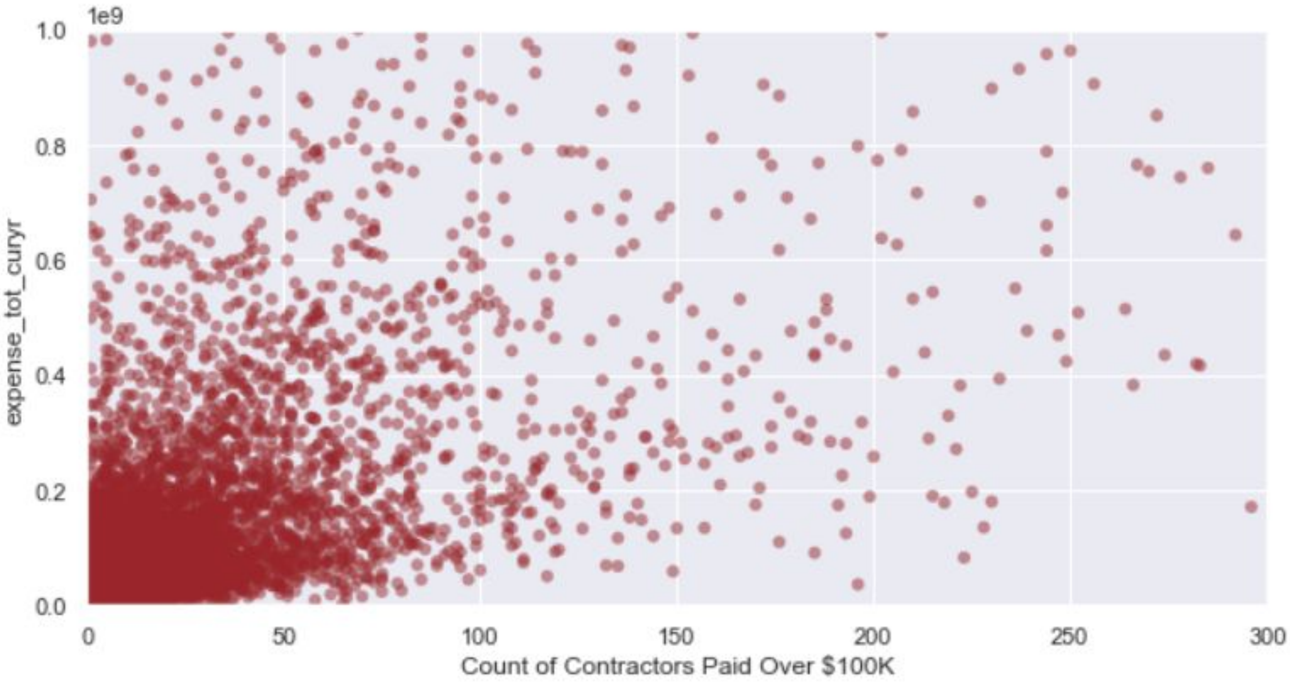
Scatterplot showing relationship of gross receipts or the organization to the count of employees and the count of volunteers.

- Employees increase as gross receipts increase up to about 2,000 employees.
- Volunteer count is inversely related to gross receipts. Organizations with smaller gross receipts seem to be more active at volunteer recruitment



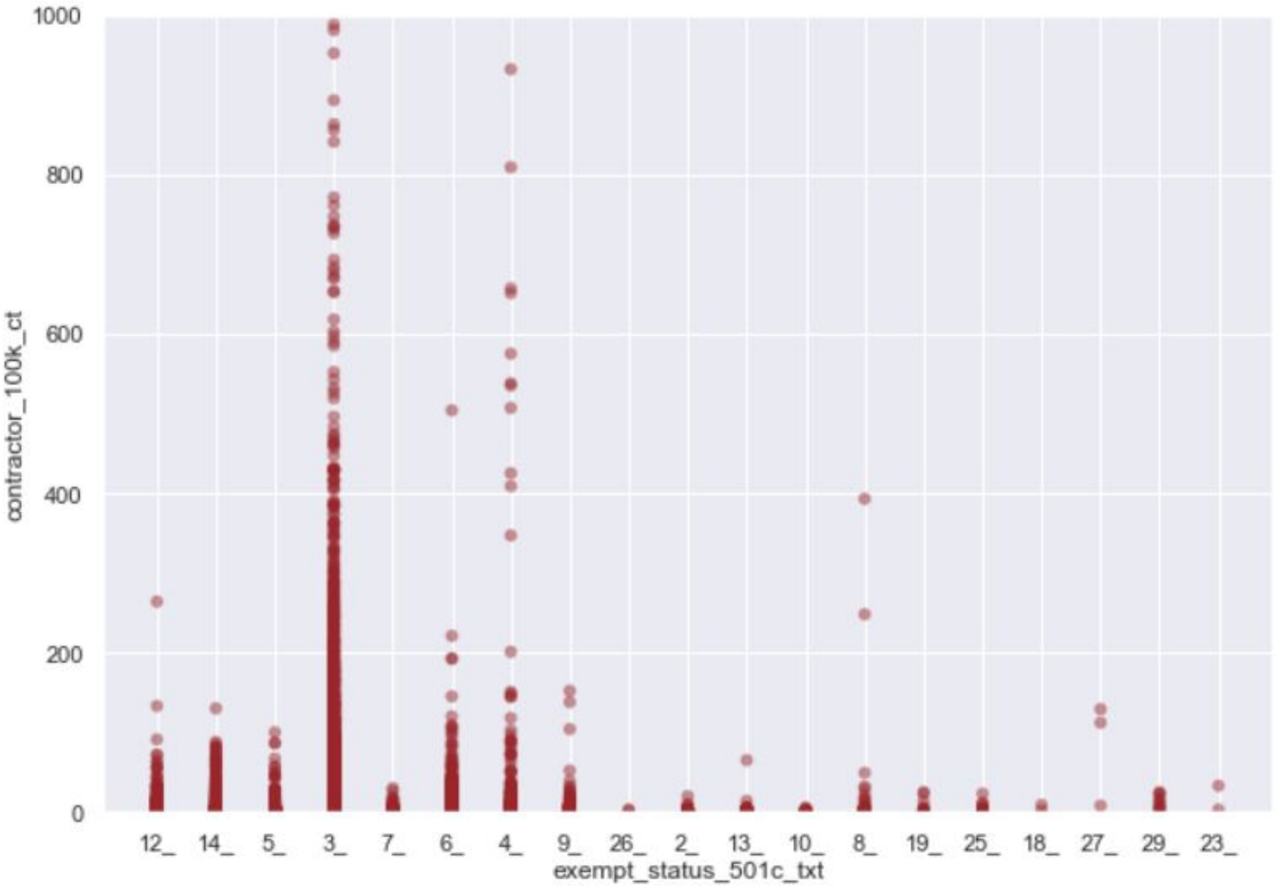
EDA - Expenses

Scatterplot showing relationship of the count of contractors to the current year expenses. There is a strong positive relationship between these two variables.



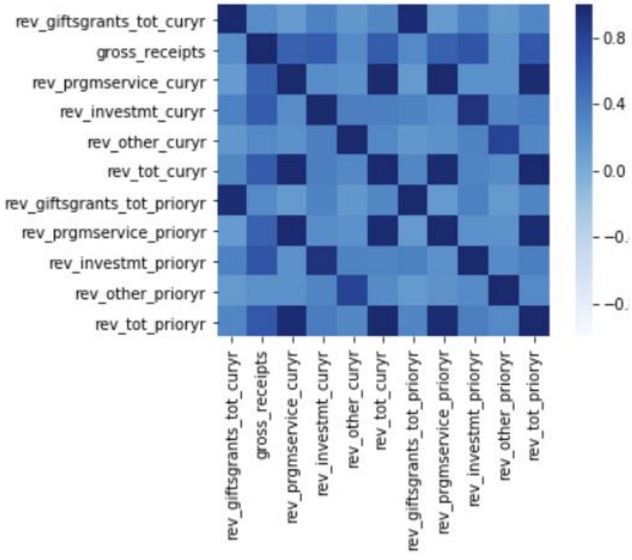
EDA - Count of contractor by 501(c) type

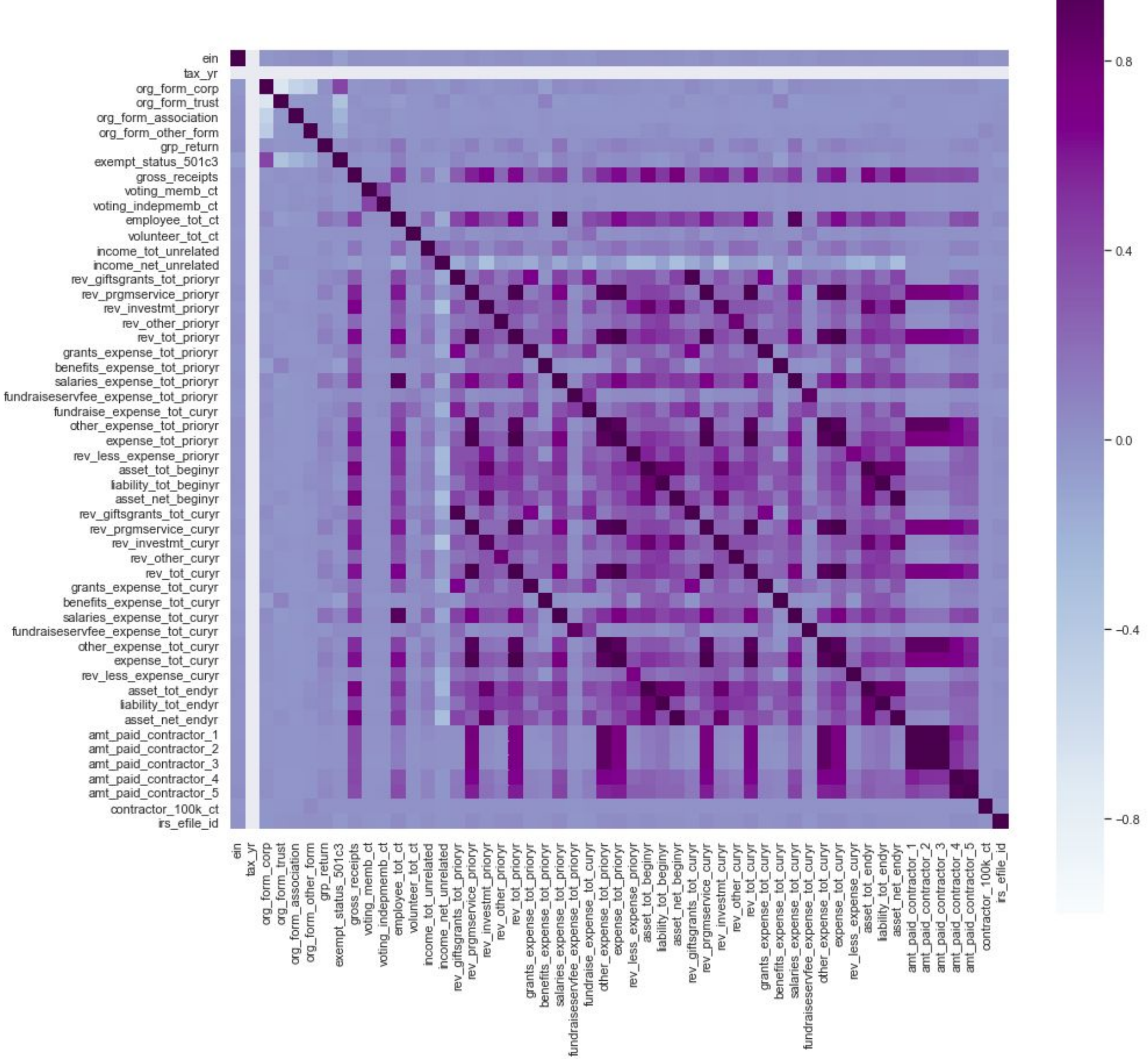
- Most not-for-profits with contractor payments over $100K are registered as 501(c)3 organizations.
- According to IRS regulations 501(c)3 organizations are organized under 8 categories of purpose (religious, charitable, scientific, literary, or educational purposes, for testing for public safety, to foster national or international amateur sports competition, for the prevention of cruelty to children, women, or animals.)



EDA - Correlation heatmap

Financial data shows the most correlation



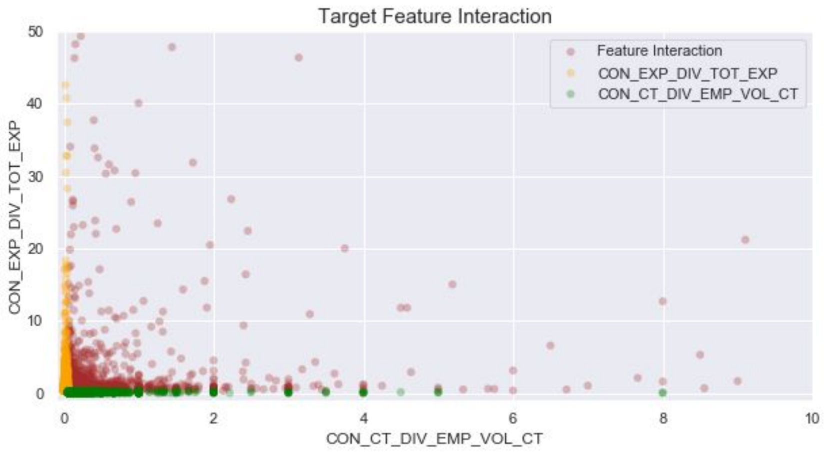


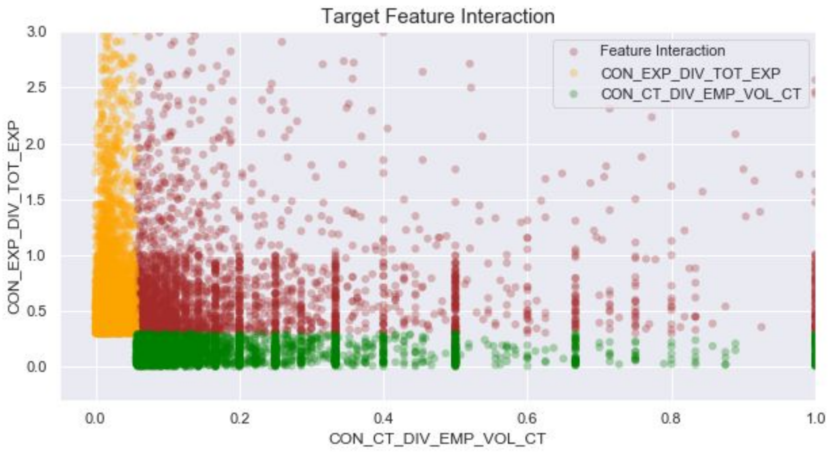
Target Variable Creation

Two features were designed to identify not-for-profit organizations that might be at higher risk of contractor payment fraud.

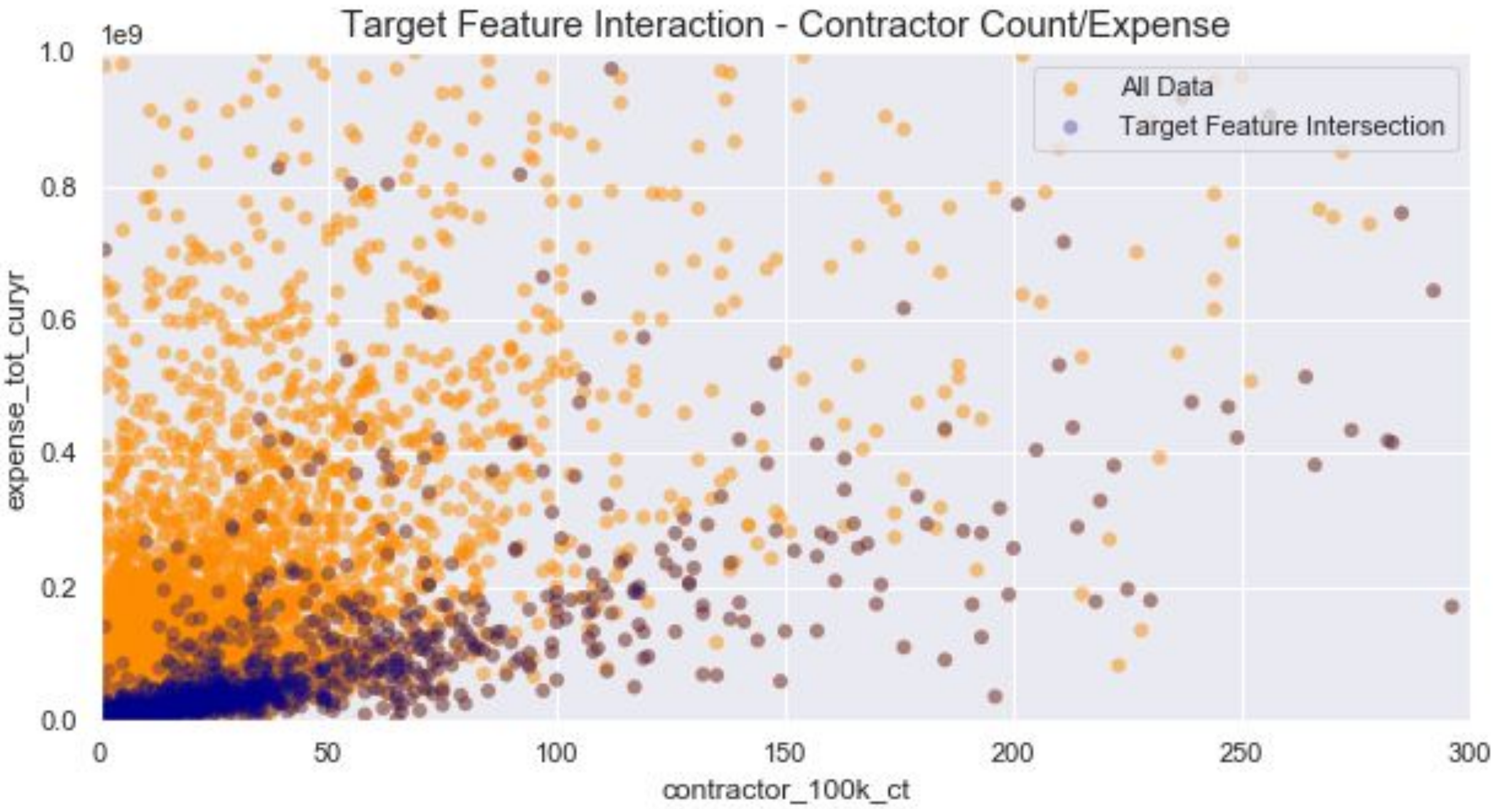
- Contractor count / count of employees + count of volunteers - the higher this ratio, the less oversight of the contractor payment process, creating more opportunity for fraud.
- Mean contractor expense / total current year expenses - organizations that spend a high percent of their money on contractors

For each of these features I selected the top quartile. The intersection of these two quartiles represents 10% of my dataset

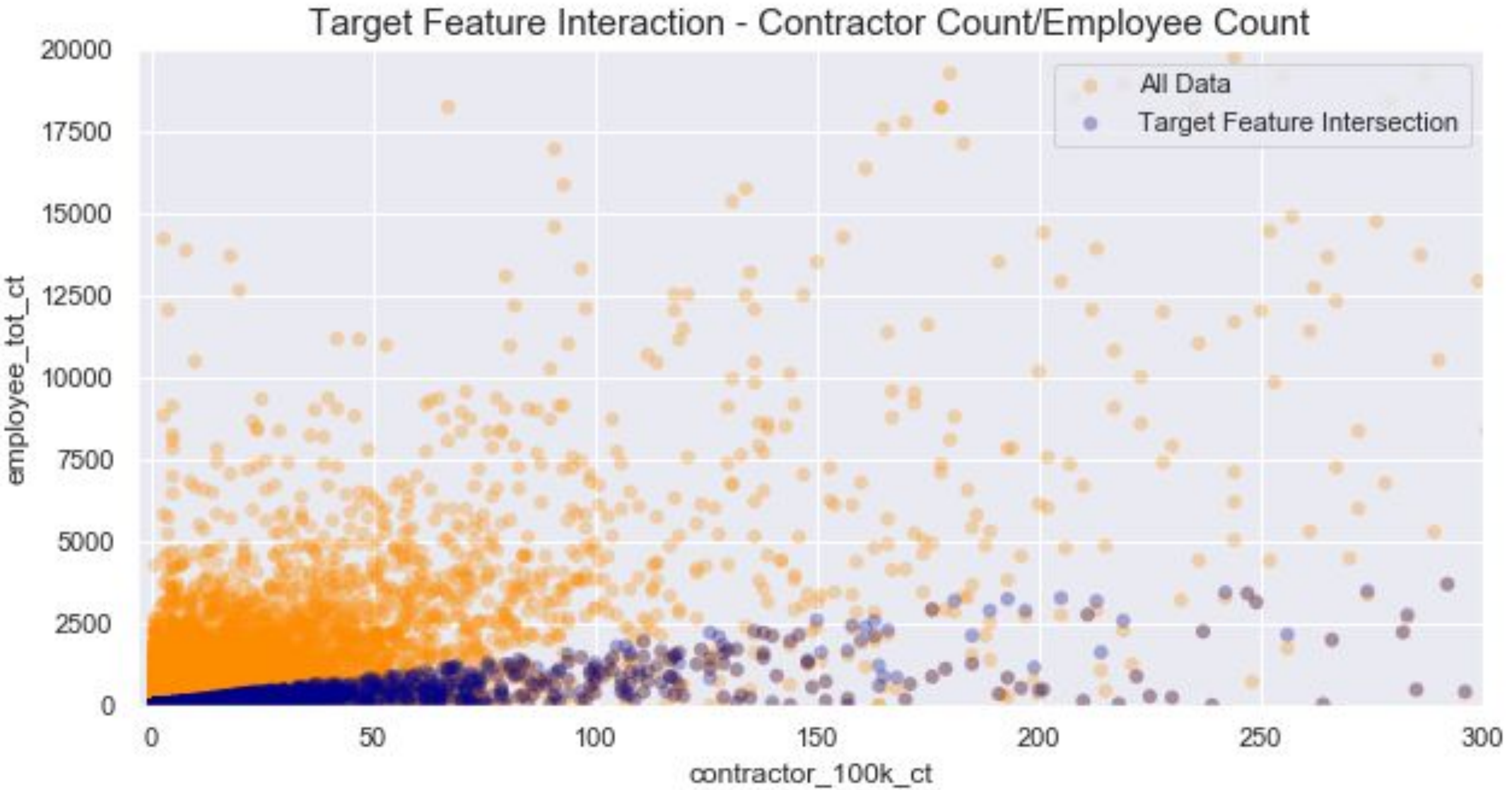




Target Feature Interaction



Target Feature Interaction



Modeling - Feature Iteration #1

- Removed features that intuitively had no bearing on my target variable - EIN of the org, contact information of the org, ect.
- Used PCA to consolidate all revenue and expense features

<u>Gradient Boost - Test Set Confusion Matrix:</u>

	Predicted: No	Predicted: Yes
Actual: No	4710	118
Actual: Yes	287	262

Test Results		
	Type I Errors (%)	**Type II Errors (%)**
Random Forest Classifier	.52	9.3
Gradient Boost Classifier	1.36	8.54
KNN Classifier	.223	9.84
Support Vector Classifier	1.71	9.43

Modeling - Feature Iteration #2

Removed outliers after experimenting with two independent trimming iterations:

- 1% highest outliers - improved performance in all models except SVC
- 10% highest outliers - this hurt all the models

Results: My best model, Gradient Boost, saw slight improvement with 1% outliers removed. SVC did extremely well with Type II errors, but the cost of much higher Type I errors was too high.. As a result, I used features with trimmed outliers on feature adjustments in future feature iterations.

Test Results		
	Type I Errors (%)	**Type II Errors (%)**
Random Forest Classifier	0.45	9.24
Gradient Boost Classifier	1.32	8.52
KNN Classifier	.223	9.8
Support Vector Classifier	38.89	2.94

Modeling - Feature Iteration #3

- After examining feature_importances, dropped more features that weren't contributing to the models
  - 'org_form_corp', 'org_form_trust', 'org_form_association', 'org_form_other_form'
- Also, removed dummy features that weren't contributing to the model, specifically exempt_status_501c_txt designations
- Total number of remaining features = 15

Results - Random forest improved slightly, with all the other models doing poorly. SVC did extremely well with Type II errors, but the cost of much higher Type I errors was not worth it. As a result, I did not use these feature adjustments in future feature iterations.

Test Results		
	Type I Errors (%)	**Type II Errors (%)**
Random Forest Classifier	0.52	9.21
Gradient Boost Classifier	1.1	8.67
KNN Classifier	.223	9.82
Support Vector Classifier	51.3	2.72

Modeling - Feature Iteration #4

- Used Jenks natural breaks classification method(jenkspy python library) to find natural breaks in several continuous features that were contributing the most information to my models.
  - gross_receipts
  - assets_tot_beginyr
  - liability_tot_beginyr

Results - This iteration hurt performance for all models except Gradient Boost Classifier, which still didn't performing as well as it did in feature iteration 2.

Test Results		
	Type I Errors (%)	**Type II Errors (%)**
Random Forest Classifier	.465	9.34
Gradient Boost Classifier	1.54	8.56
KNN Classifier	.223	9.84
Support Vector Classifier	1.71	9.43

Modeling - Feature Iteration #5

Upsampled from minority class to see if this has an impact on any of the models. Rather than only making up 10% of the dataset, my target features now make up 50% of the dataset.

Results:

All the models performed worse in this iteration.

Test Results		
	Type I Errors (%)	**Type II Errors (%)**
Random Forest Classifier	8.69	17.16
Gradient Boost Classifier	10.05	11.94
KNN Classifier	11.98	18.3
Support Vector Classifier	27.9	11.05

Model Summary

Gradient Boost Classifier was the most successful model. Using a combination of 38 features, this model was able to classify not-for-profits at a higher risk for fraud:

- 8.54% Type II error rate (the rate at which the model incorrectly categorized something as negative, when it should have been categorized as positive)
- 1.3% Type I error rate (the rate at which the model incorrectly categorized something as positive, when it should have been categorized as negative)
- This model was also more robust than the other models in that it performed well with fewer preprocessing and feature engineering steps.

```
# fit model based on grid search parameters
params = {'n_estimators': 950,
          'max_depth': 2,
          'subsample': .8,
          'learning_rate': .1,
          'loss': 'deviance'}

clf = ensemble.GradientBoostingClassifier(**params)
clf.fit(X_train, y_train)
```

Gradient Boost - Training Set Confusion Matrix:

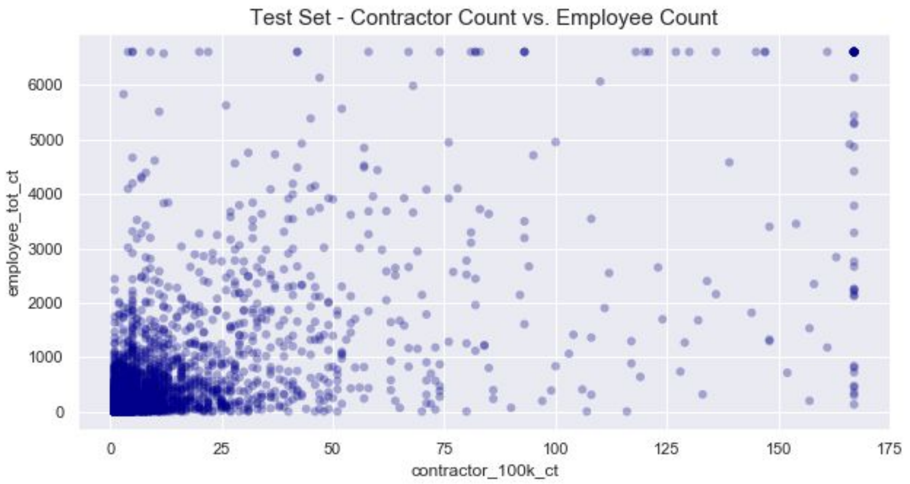
	Predicted: No	Predicted: Yes
Actual: No	19067	77
Actual: Yes	1810	550

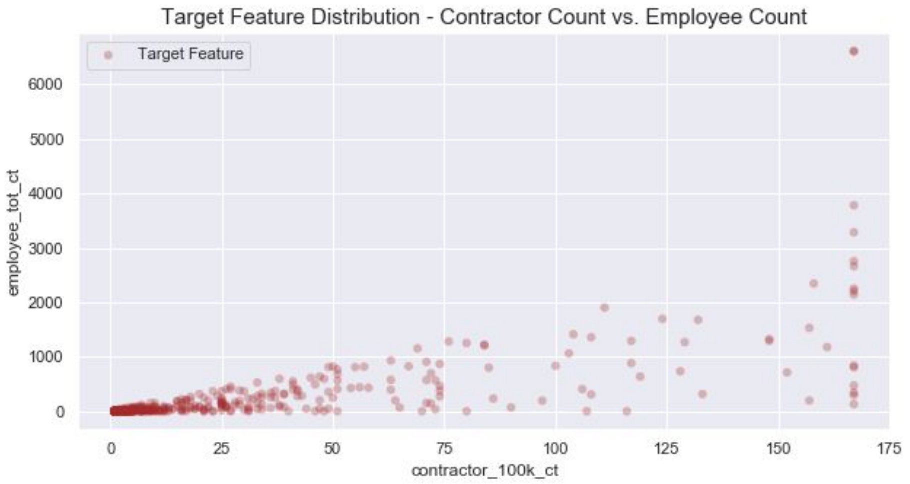
Gradient Boost - Test Set Confusion Matrix:

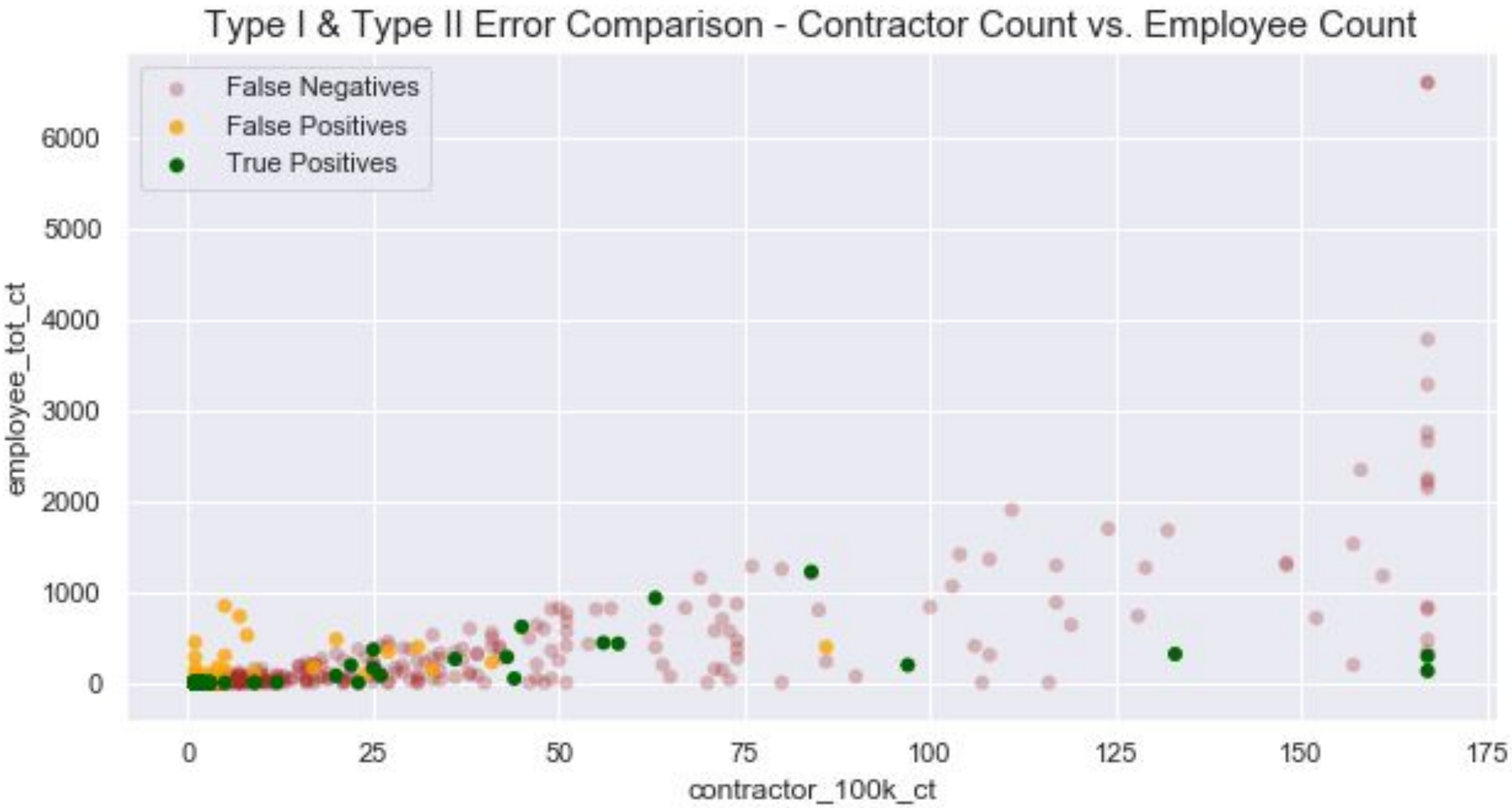
	Predicted: No	Predicted: Yes
Actual: No	4757	71
Actual: Yes	458	91

Model Summary - Contractor Count & Employee Count

Gradient Boost Classifier had issues with false negatives uniformly, however it struggled with false positives for not-for-profit organizations with few contractors and employees (next slide).



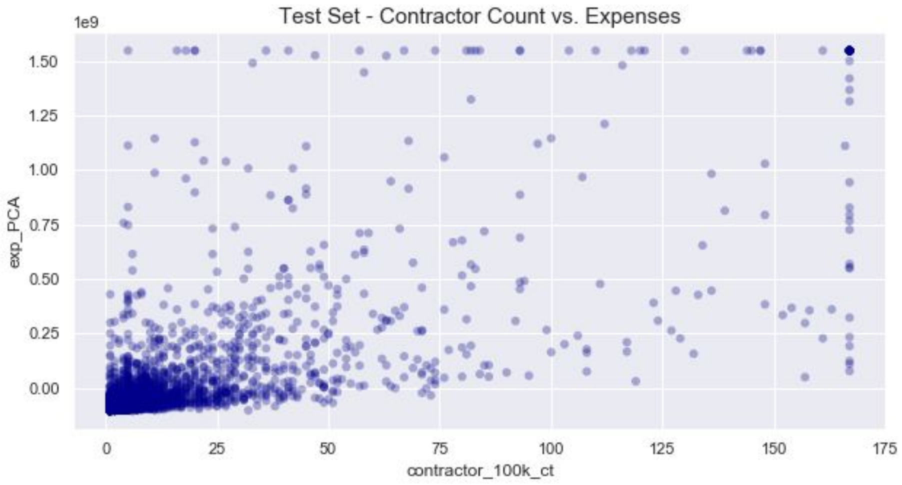


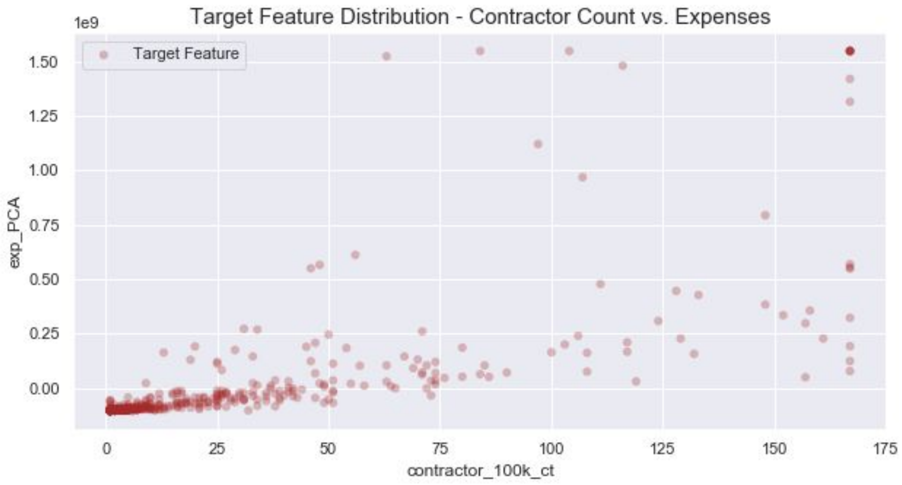


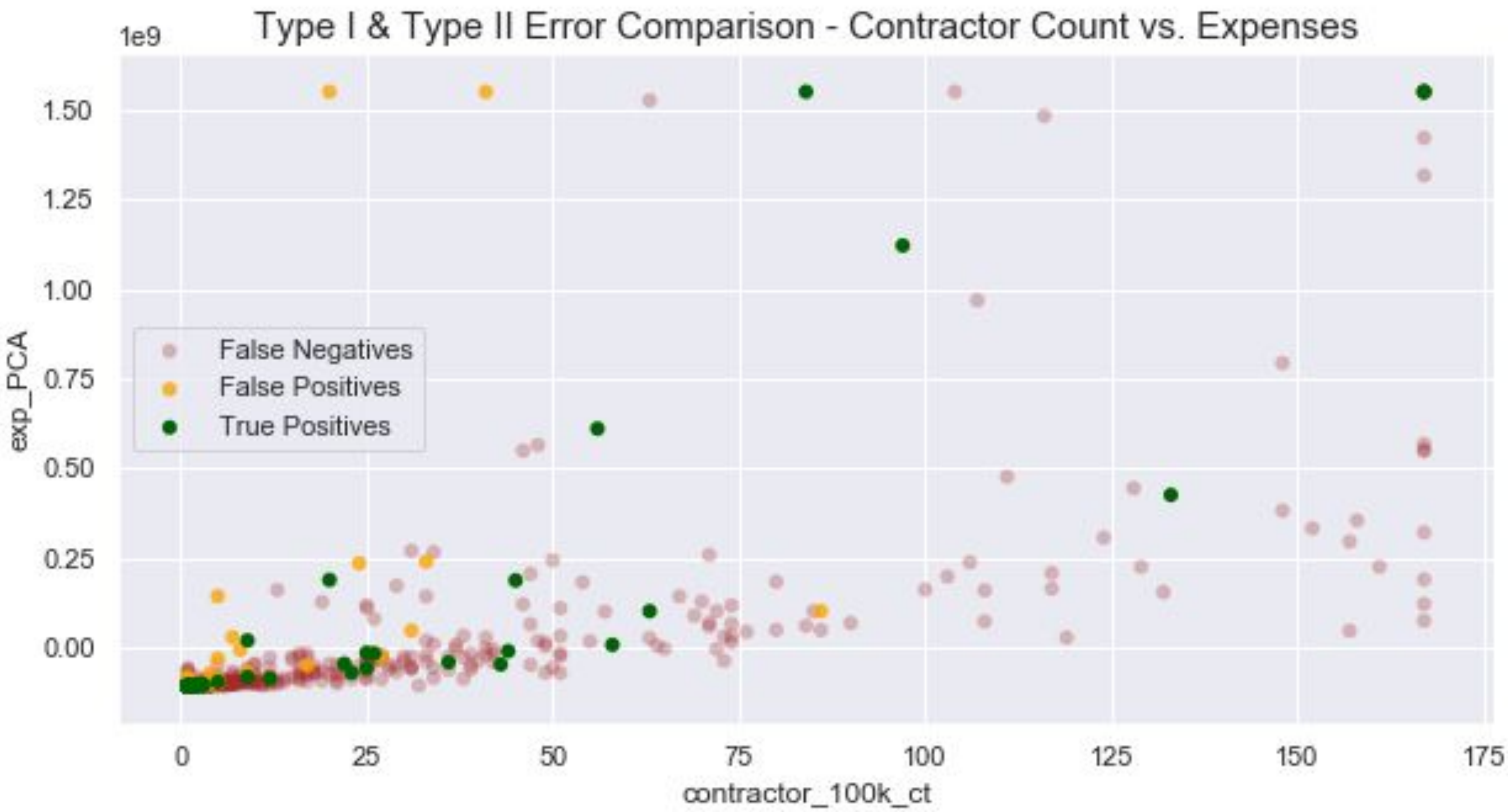
Type I & Type II Error Comparison - Contractor Count vs. Employee Count
False Negatives
False Positives
True Positives
employee_tot_ct
0
1000
2000
3000
4000
5000
6000
contractor_100k_ct
0
25
50
75
100
125
150
175

Model Summary - Contractor Count & Expenses

Looking at the relationship between contractor_100K_ct and expenses, the model had issues with false negatives uniformly. Again it struggled with false positives for not-for-profit organizations with few contractors and employees (next slide).







Type I & Type II Error Comparison - Contractor Count vs. Expenses
1e9
1.50
1.25
1.00
0.75
0.50
0.25
0.00
exp_PCA
False Negatives
False Positives
True Positives
0
25
50
75
100
125
150
175
contractor_100k_ct

Future Considerations

- Efforts I made to categorize 501(c)(3) organizations based on mission statement weren't effective at providing information that allowed my model to perform better. However there may be other ways to extract information from the mission statement of the organizations.
- Contractor payment information includes summary of services rendered by the contractor. It would be interesting to categorize the types of services rendered and how those services compare to the organization based on the size of the organization and any other factors.

Example Mission Statements:

TO PROVIDE FINANCIAL SUPPORT TO OTHER CHARITABLE ORGANIZATIONS WHICH PROMOTE SOCIAL, EDUCATIONAL AND OTHER CHARITABLE SERVICES IN THE UNITED STATES AND ISRAEL. IT ALSO PROVIDES SOCIAL SERVICES TO POOR AND DISADVANTAGED INDIVIDUALS IN THE IRANIAN AMERICAN JEWISH COMMUNITY.

ALBANY COMMUNITY CHARTER SCHOOL PREPARES STUDENTS FOR A LIFETIME OF OPPORTUNITY BY HELPING THEM MASTER PRIMARY RIGOROUS, STANDARDS-BASED CURRICULUM FOCUSED ON LITERACY AND OTHER FOUNDATIONAL KNOWLEDGE.

WALDO COUNTY GENERAL HOSPITAL'S MISSION IS TO BE THE BEST - BETTER, EMPATHY, SERVICE AND TEAMWORK. OUR GOAL IS TO ENSURE QUALITY, ACCESSIBLE AND AFFORDABLE HEALTH CARE SERVICES AND TO IMPROVE THE HEALTH AND WELL-BEING OF OUR COMMUNITY. PLEASE SEE ATTACHED COMMUNITY BENEFITS REPORT.

Questions?