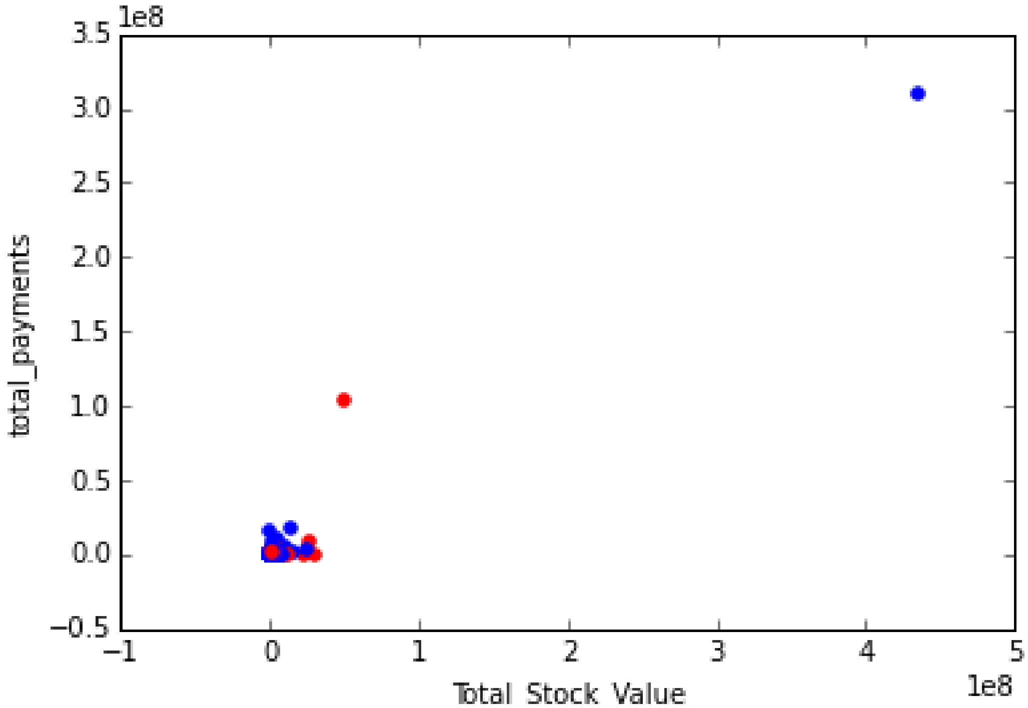
Enron POI Identifier Report

Introduction

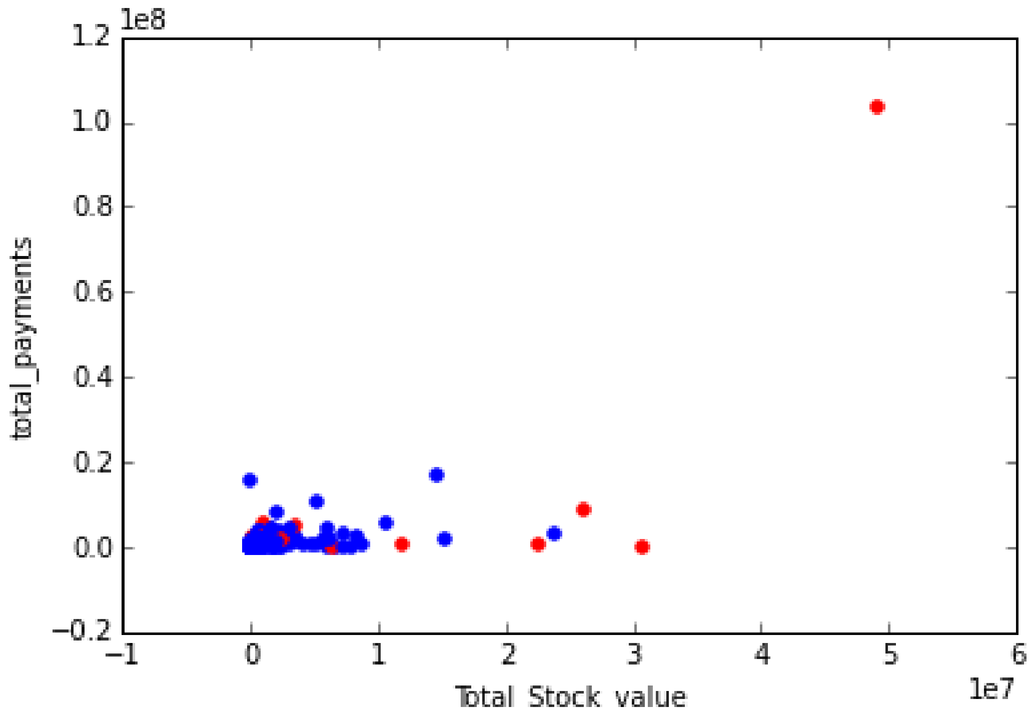
Enron is one of the largest corporate fraud cases in American history and also is one of the well-documented ones. It has a large corpus of corporate e-mail database available for the public for study and research purpose. In this project, the e-mail data and the financial data has been used to identify the Person of Interest (POI). POI is someone who is indicted for fraud, settled with the government or testified in exchange for immunity. This report documents the machine learning techniques used in building the POI identifier

The Enron Data

The Enron Dataset has 146 records. Of these 18 are labeled as POI and the rest are labeled as non-POI. There are 21 features for each Person. Some of the fields like loan_advances, director_fees and restricted_stock_deferred are sparsely populated. loan_advances field has values only for 4 records, director_fees has values only for 17 records and restricted_stock_deferred has values only for 18 records.



On examining the data points plotted for total_stock_value and total_payments, one outlier is distinct. This is the “Total” and has been removed from the data set. The other data point which stood out while examining the PDF for financial data was “THE TRAVEL AGENCY IN THE PARK”. This data point has only data for “Other” and “total_payments” fields. The rest of the data is not available for it. Also, it appears to be an agency and not one of the Enron Executives. After removing these 2 data points we get the graph as shown below



There is still a very distinct outlier standing out for the data point of Ken Lay. The total payment for Ken Lay is around 103 million and the Total Stock Value is around 49 million which stands out as compared to rest of the data points. There are a few more outliers like Joseph Hirko with over 30 million of total stock value, Jeff Skilling with over 26 million of total stock value. But, we would like to retain these data points as they are Enron executives and the fact that so much money was taken out of the company by them and they are identified as the POIs, we want to retain them in the dataset

Feature Processing

After Outlier removal, the next step was to assess if any new feature was to be created from the given features. We created the following 3 new features

1. **fraction_shared_receipt_with_poi:** Created as a fraction of **shared_receipt_with_poi** over **to_messages**
2. **fraction_from_poi:** Created as a fraction of **from_poi_to_this_person** over **to_messages**
3. **fraction_to_poi:** Created as a fraction of **from_this_person_to_poi** over **from_messages**

The premise for creating these features was that POIs among themselves would be having a strong e-mail connection among each other and the fraction will give the percentage of the total e-mails received / sent /shared between the person and the POI and if this fraction is high then it might imply that the person has strong connection with the POI

Once the data was cleaned of outliers, the next step was selecting the best features to use in the classifier. This was an iterative process, where we used the "SelectKBest" algorithm from the sklearn package and the feature importance property of the Decision Tree Classifier to identify the best features having discriminatory power. The K-Best algorithm selects features according to the k highest scores of the scoring function. We are using f_classif score for selecting the best features. We started with 1 feature and started increasing the count 22 i.e. till all the quantitive features (i.e. excluding email_address and including the newly derived features) were incorporated. For each k-value we used the GridSearchCV to test these features across the range of values from 2 to 20 for the parameter min_samples_split for the best f1 score. We then used the best classifier identified by the GridSearchCV to perform the classification using the DecisionTreeClassifier. We also found the feature importance for each feature in the classification for each k-value. The test results are as below:

# of features (k-value)	Feature Name	Feature Importance	min_sample_split (Best value)	precision	recall
1	exercised_stock_options	0.999000999	14	0.39615	0.309
2	exercised_stock_options total_stock_value	0.589210899537 0.409790099464	17	0.36292	0.231
3	exercised_stock_options total_stock_value bonus	0.347725074395 0.268394301665 0.382881622941	2	0.36642	0.3895
4	exercised_stock_options total_stock_value bonus salary	0.310019480728 0.209178343651 0.309449174641 0.17035399998	2	0.32329	0.3325
5	exercised_stock_options total_stock_value bonus salary fraction_to_poi	0.252714059344 0.115566030501 0.240815747507 0.172007974203 0.217897187445	2	0.28922	0.3245

6	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income	0.220052571876 0.0715250058342 0.247013217687 0.056175562737 0.2581066189 0.146128021966	14	0.33693	0.281
7	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive	0.209326543096 0.0721430628984 0.236405811668 0.0441287023119 0.265871760548 0.141789108863 0.029336009615	14	0.33665	0.2705
8	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock	0.209484712441 0.0912000996979 0.229837369818 0.036978312059 0.250168923212 0.124004786418 0.0177610753837 0.0395657199715	14	0.32036	0.2675
9	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi	0.200785033178 0.0784302644175 0.22911479058 0.00858134842999 0.183341865261 0.0987914529196 0.014510220755 0.0512158624559 0.134230161003	14	0.31389	0.269

10	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments	0.204652945266 0.0706978835107 0.204429425105 0.00902137143215 0.179133439112 0.0990254268594 0.00993201934732 0.0442097253219 0.132974781481 0.0449239815655	14	0.2956	0.252
11	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi	0.177251216296 0.0683163285642 0.121355454093 0.0472641567923 0.128934802285 0.0840695153179 0.0398750249528 0.0599446923355 0.0840577796066 0.0756518778151 0.112280150942	2	0.26001	0.2305
12	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances	0.17709633429 0.0637839792378 0.122119773999 0.0475311791158 0.127246327311 0.0855117888243 0.0403072500326 0.0605288574112 0.0853652465276 0.074305621586 0.111095314619 0.0041093260481	2	0.27323	0.247

13	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances expenses	0.203661463888 0.0509214603582 0.141728678545 0.00360434337288 0.188570074419 0.039573974884 0.00368661123119 0.0484943576518 0.0825574646945 0.0171463714768 0.109285544461 0.0 0.109770654019	18	0.32516	0.2785
14	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances expenses from_poi_to_this_person	0.206479726651 0.0491319323065 0.14005158062 0.00423036400532 0.190572684453 0.0388850864137 0.00337394382964 0.0479337951898 0.0817118736234 0.0165801843782 0.108497006814 0.0 0.10899550484 0.00255731587725	18	0.32185	0.2755

15	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances expenses from_poi_to_this_person other	0.151682997287 0.0574068883987 0.0965396803924 0.0231716109831 0.120190944579 0.0414534006492 0.0225826982468 0.0546451422128 0.0653781901732 0.0381152649484 0.0894036944 0.000879307767063 0.108984039686 0.0150754830283 0.113491656249	2	0.30386	0.299
16	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances expenses from_poi_to_this_person other fraction_from_poi	0.145463645413 0.0618783361328 0.0937542438427 0.0228443228929 0.120280523967 0.0401935754911 0.0202038631449 0.0551161139065 0.0638312531623 0.0381645355937 0.0882637792739 0.00104759186392 0.106806372507 0.0158985939644 0.111900860836 0.0133533870091	2	0.3016	0.301

17	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances expenses from_poi_to_this_person other fraction_from_poi from_this_person_to_poi	0.150487783289 0.0535548459285 0.0927138050841 0.0237094821319 0.118909652958 0.0390876636379 0.0191249856797 0.0518327572825 0.0628480544993 0.0365900221716 0.0843752082459 0.00105829544605 0.103685973688 0.0142164326698 0.110136362274 0.0113649709809 0.0253047030349	2	0.29167	0.2835
18	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances expenses from_poi_to_this_person other fraction_from_poi from_this_person_to_poi director_fees	0.147480372214 0.0563577737592 0.0938475708871 0.0228918590313 0.120469317339 0.0391705725156 0.0200747930038 0.0515431494061 0.0624347507693 0.0349544907209 0.0847465930537 0.000662149641741 0.104728982229 0.0134992183957 0.110085121511 0.011572890415 0.0244813941095 0.0	2	0.29333	0.286

19	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances expenses from_poi_to_this_person other fraction_from_poi from_this_person_to_poi director_fees to_messages	0.147829553683 0.0554152656867 0.0931470177062 0.021757744488 0.118765604388 0.0393047508631 0.0176456629784 0.0506760063506 0.0620906011526 0.0354958891983 0.0809293857257 0.000775811490097 0.103487491457 0.0140654942905 0.11246468746 0.0118046114984 0.0246171693405 0.0 0.00872825124366	2	0.29707	0.284
20	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances expenses from_poi_to_this_person other fraction_from_poi from_this_person_to_poi director_fees to_messages deferral_payments	0.147292535342 0.0528572252111 0.0938295958511 0.0216722260345 0.119492497731 0.0386472107923 0.0190909546799 0.0516882464185 0.0649644409158 0.0344360734223 0.0807270338787 0.000897232019681 0.0999670961496 0.0096150115798 0.110480607771 0.013159168598 0.0246235588785 0.0 0.00681811028213 0.00874217344533	2	0.29433	0.283

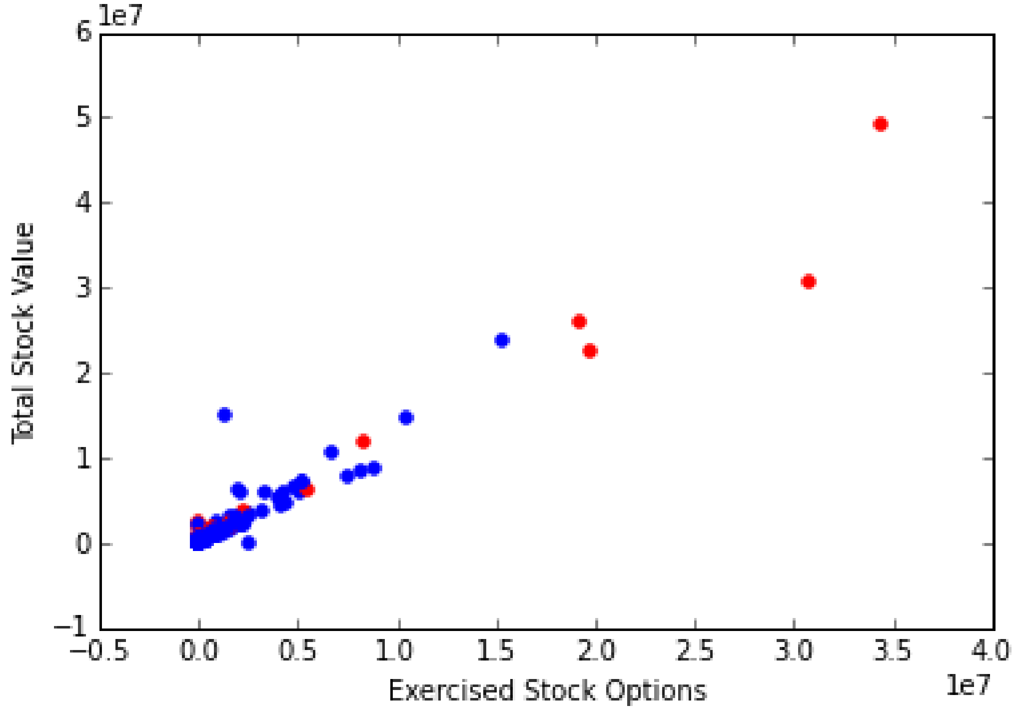
21	exercised_stock_options	0.147741192251	2	0.28843	0.2805
	total_stock_value	0.0533134039233			
	bonus	0.0930473333754			
	salary	0.0208672244808			
	fraction_to_poi	0.119791248544			
	deferred_income	0.0389059076787			
	long_term_incentive	0.0187482010177			
	restricted_stock	0.0514408177403			
	fraction_shared_receipt_with_poi	0.0583448692614			
	total_payments	0.0337958045157			
	shared_receipt_with_poi	0.079903569517			
	loan_advances	0.000387537632436			
	expenses	0.102152284046			
	from_poi_to_this_person	0.00979852020319			
	other	0.108570037757			
	fraction_from_poi	0.0106737474316			
	from_this_person_to_poi	0.0147239463872			
	director_fees	0.0			
	to_messages	0.00826803317495			
	deferral_payments	0.00837543539663			
	from_messages	0.0201518846661			

22	exercised_stock_options total_stock_value bonus salary fraction_to_poi deferred_income long_term_incentive restricted_stock fraction_shared_receipt_with_poi total_payments shared_receipt_with_poi loan_advances expenses from_poi_to_this_person other fraction_from_poi from_this_person_to_poi director_fees to_messages deferral_payments from_messages restricted_stock_deferred	0.148566771885 0.0539877338159 0.0922349968573 0.0216026999308 0.118538189733 0.0373082372421 0.019193109337 0.0503231693659 0.0609117629039 0.033748444456 0.0812439111602 0.000583260277138 0.0999114146649 0.010578347543 0.108029398038 0.0109867443798 0.0147183506078 0.0 0.00671194704949 0.0105093135092 0.0186345948009 0.000678601443908	2	0.29366	0.285

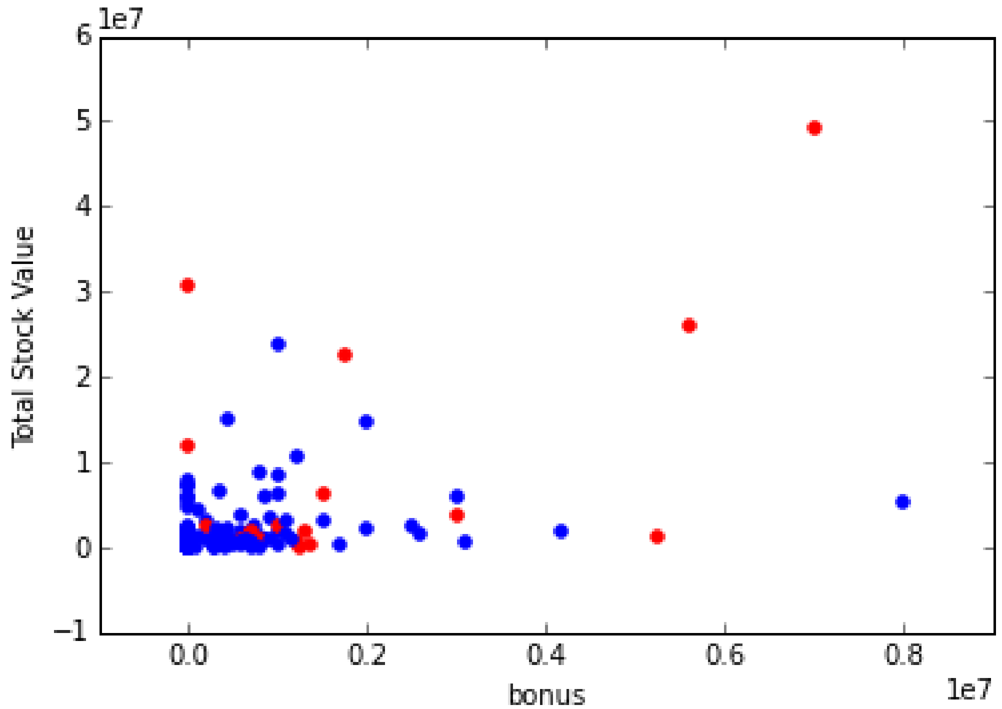
We found that for 3 features we had the best results for the Decision Tree classifier with min_samples_split as 2.

The best 3 features identified are:
bonus
exercised_stock_options
total_stock_value

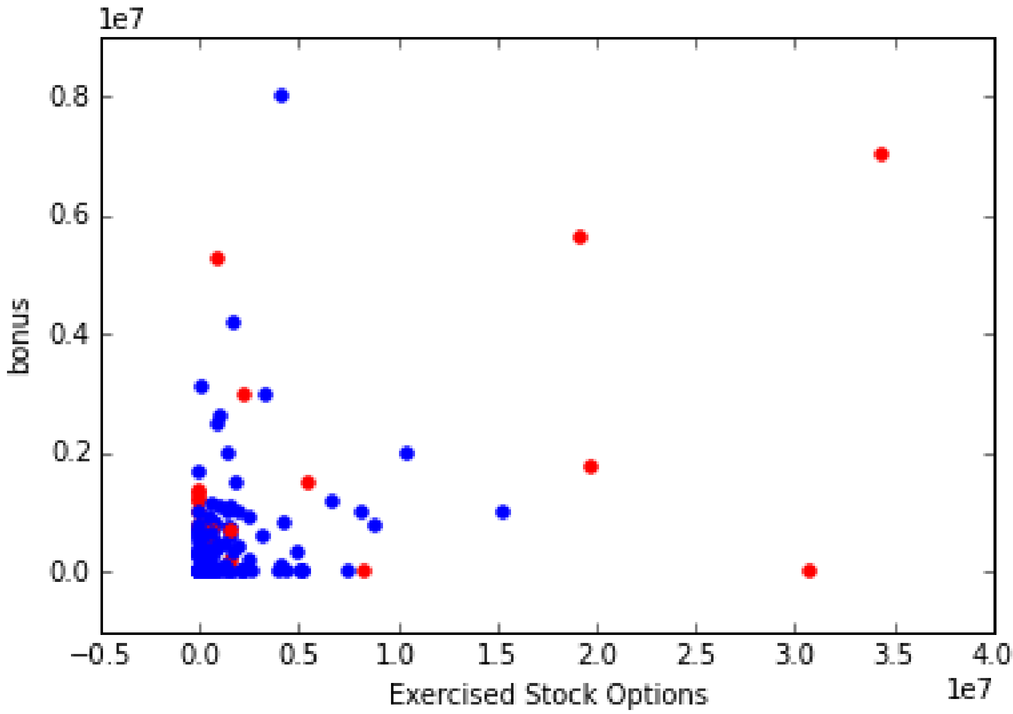
We then plotted the visualizations for these 3 features identified by the K-Best algorithm as having more discriminatory power as compared to other features



1e7
6
5
4
3
2
1
0
−1
Total Stock Value
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Exercised Stock Options
1e7



1e7
6
5
4
3
2
1
0
−1
Total Stock Value
0.0
0.2
0.4
0.6
0.8
bonus
1e7



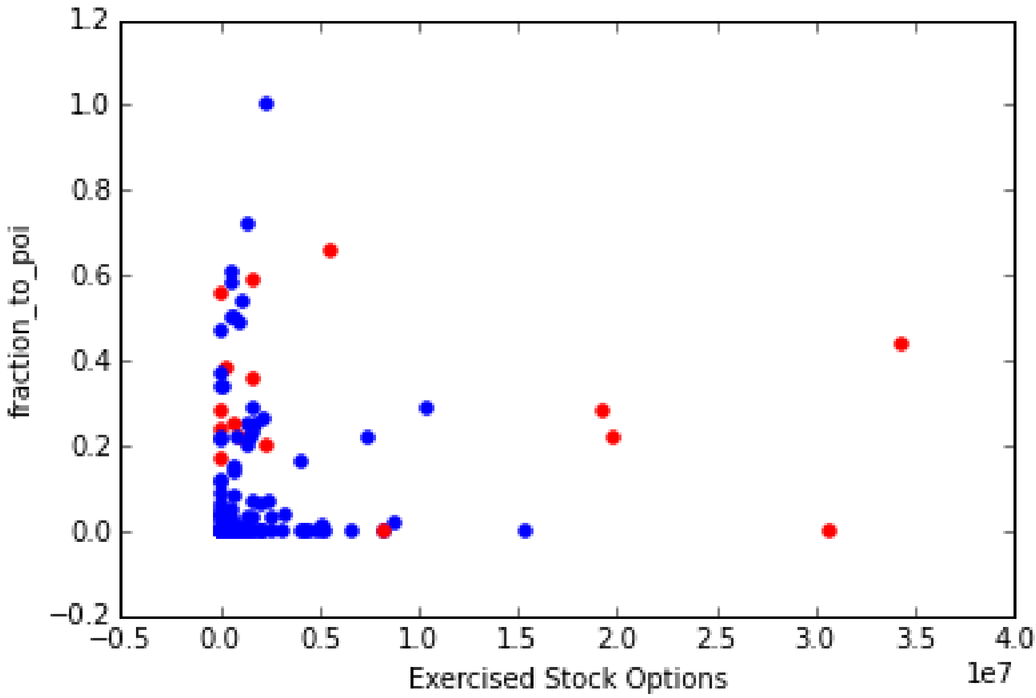
From the visualizations all of these features seemed to have discriminatory power.

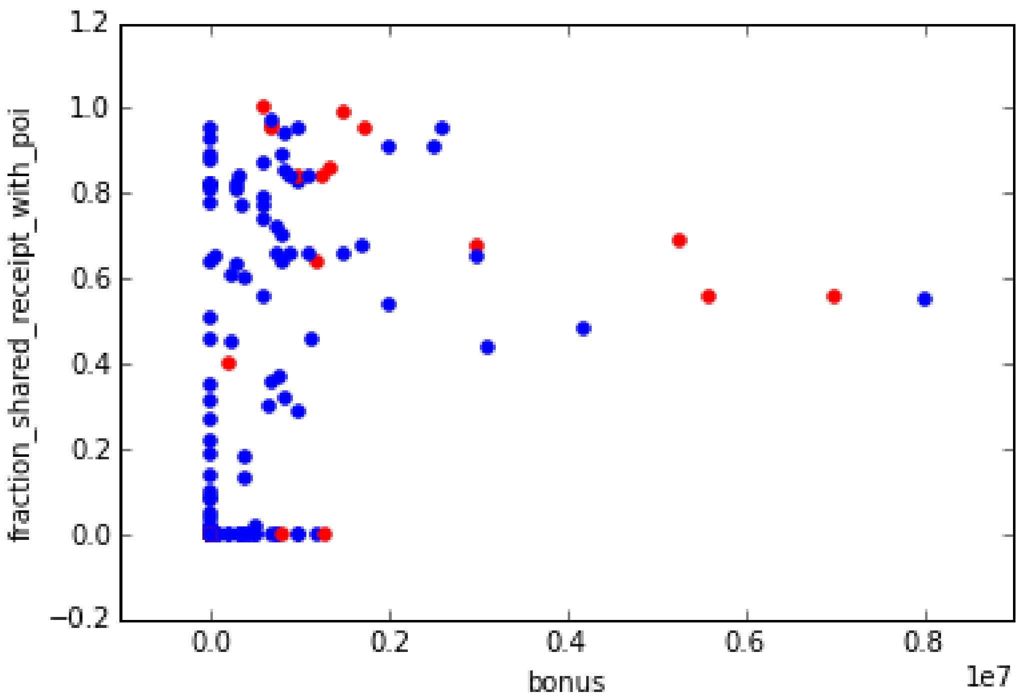
We then checked the precision and recall values for these 3 features using the DecisionTreeClassifier. We used the GridSearchCV to get the best parameter value for min_sample_split from 2 to 20. We got the best min_sample_split as 2 and the precision and recall scores as below:

precision: 0.36183
recall: 0.383

On analyzing further found that of the top 10 features, the following 4 features have maximum importance
exercised_stock_options
fraction_to_poi
bonus
fraction_shared_receipt_with_poi

We plotted the visualizations for these 4 features





These features also seem to have the discriminatory power

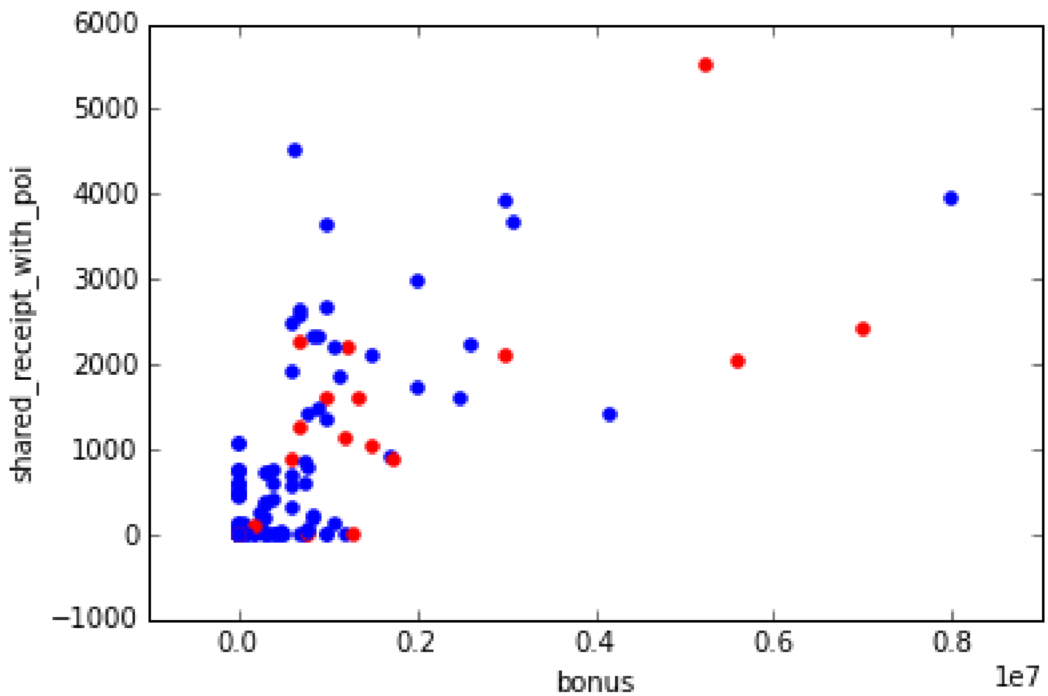
We then checked the precision and recall values for these 4 features using the DecisionTreeClassifier and using the GridSearchCV to get the best parameter value for min_sample_split. We got the best min_sample_split as 2 and the precision and recall scores as below:

precision: 0.37976
recall: 0.394

On analyzing further found that of the top 11 features, the following 4 features have maximum importance
exercised_stock_options
fraction_to_poi
bonus
shared_receipt_with_poi

i.e. fraction_shared_receipt_with_poi was replaced with shared_receipt_with_poi

We plotted the graph of bonus and shared_receipt_with_poi as below



The feature shared_receipt_with_poi also seem to have the discriminatory power

We then checked the precision and recall values for these 4 features using the DecisionTreeClassifier and using the GridSearchCV to get the best parameter value for min_sample_split. We got the best min_sample_split as 2 and the precision and recall scores as below:

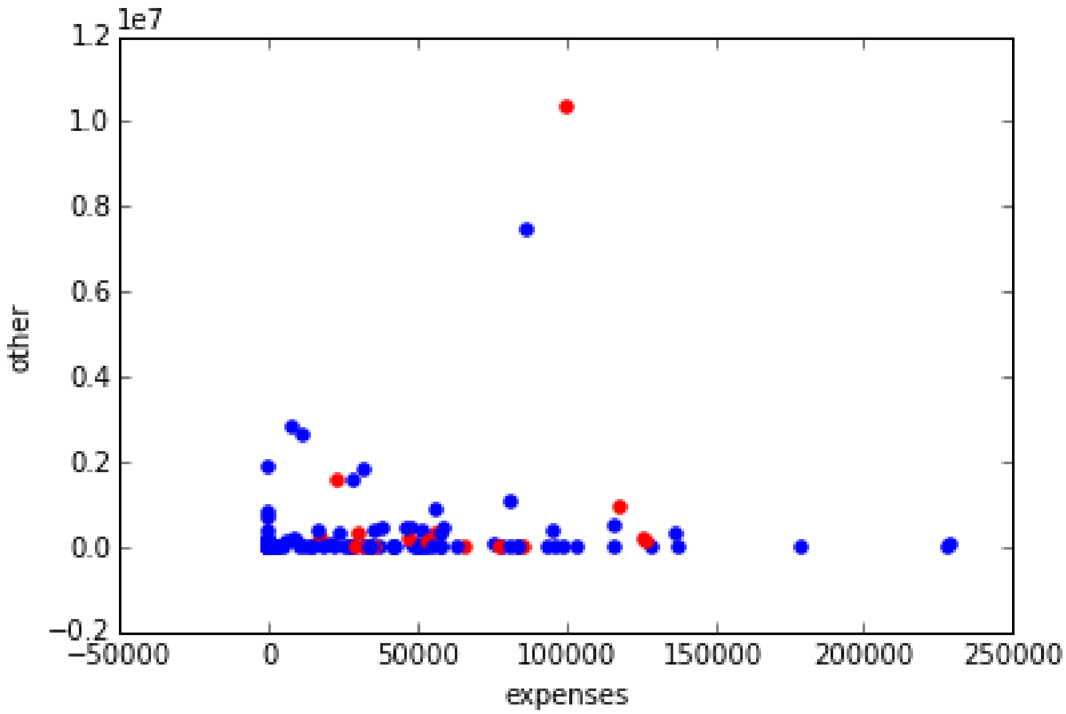
precision: 0.45553
recall: 0.42

These 4 features arrived after replacing fraction_shared_receipt_with_poi with shared_receipt_with_poi seemed to have the more discriminatory power

On analyzing further found that of all the features, the following 6 features have maximum importance
exercised_stock_options
fraction_to_poi
bonus
shared_receipt_with_poi
expenses
other

i.e. the following 2 features “expenses” and “other” got added to the 4 features which we analyzed in the previous step

We plotted the graph of bonus and shared_receipt_with_poi as below



The “expenses” feature seems to have some discriminatory power. For the feature “other” there doesn’t seem to be much discriminatory power as most of the points are at the same level

Based on this we select the following 5 features:
exercised_stock_options
fraction_to_poi

bonus
shared_receipt_with_poi
expenses

We then checked the precision and recall values for these 5 features using the DecisionTreeClassifier and using the GridSearchCV to get the best parameter value for min_sample_split. We got the best min_sample_split as 16 and the precision and recall scores as below:

precision: 0.44678
recall: 0.426

We are getting a slightly better recall score but a slightly less precision score when using these 5 features as compared to the 4 features which we analyzed earlier

We will now use these set of best 4 features (exercised_stock_options, fraction_to_poi, bonus, shared_receipt_with_poi) and best 5 features (exercised_stock_options, fraction_to_poi, bonus, shared_receipt_with_poi, expenses) for algorithm selection and tuning

Algorithm Selection and Tuning

Started with the **Naive Bayes** algorithm.

For the best 4 features this Algorithm had the following values:
precision: 0.38295
recall: 0.292

For the best 5 features this Algorithm had the following values:
precision: 0.37389
recall: 0.3165

Then, I used the **Support Vector Machine, SVC** classifier from sklearn algorithm. For using this classifier, I had to scale the features. I used the MinMaxScaler to scale the features and passed the scaled features to the SVC algorithm. I used Pipelining to achieve this. MinMaxScaler was used as the first stage of the pipeline and SVC algorithm was used as the second stage.

Tuned the SVC for the following combinations of parameters using the GridSearchCV algorithm for the best f1 score

kernel=['linear','rbf']
C=[1,10,100,1000]
gamma=[1.0,2.0,3.0,4.0,5.0,6.0,7.0,8.0,9.0,10.0]

For the best 4 features:
The parameters which was selected by GridSearchCV were as follows:
kernel='rbf'
C=100
gamma=6.0

The precision and recall values are:
precision: 0.48671
recall: 0.412

For the best 5 features:
The parameters which was selected by GridSearchCV were as follows:
kernel='rbf'
C=1000
gamma=3.0

The precision and recall values are:
precision: 0.39437
recall: 0.4275

Then, the **Decision Tree** classifier was selected as the algorithm

The Decision Tree classifier was tuned for the min_samples_split parameter for the following range of values using the GridSearchCV algorithm for the best f1 score

[2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20]

Since, for cross-validation we are using StratifiedSuffleSplit and the results for the Decision Tree algorithm varies slightly for each run. I took average of 5 readings

For the best 4 features:

min_samples_split	**precision**	**recall**
2	0.45297	0.419
	0.46488	0.407
	0.45352	0.422
	0.45916	0.43
	0.45043	0.418
Average	**0.456192**	**0.4192**

For the best 5 features:

min_samples_split	precision	recall
16	0.44334	0.4225
	0.44602	0.4255
	0.44555	0.4255
	0.44351	0.422
	0.44369	0.4235
Average	**0.444422**	**0.4238**

The Naïve Bayes algorithm has low precision and recall scores as compared to the SVC and Decision Tree. Hence, we need not consider it for determination of the best suited algorithm for POI identification.

Following are the best scores we have for precision and recall
For 4 Features:
Algorithm: SVC
kernel='rbf'
C=100
gamma=6.0
precision: 0.48671
recall: 0.412

Algorithm: Decision Tree
min_samples_split = 2
precision: 0.456192
recall: 0.4192

For 5 Features:
Algorithm: SVC
kernel='rbf'
C=1000
gamma=3.0
precision: 0.39437
recall: 0.4275

Algorithm: Decision Tree
min_samples_split = 16
precision: 0.444422
recall: 0.4238

SVC algorithm applied on 4 features has a highest precision 0.48671 but, lower recall 0.412. SVC applied on 5 features has lowest precision 0.39437 but, has highest recall 0.4275. The Decision Tree classifier applied on 5 features has recall almost the same as the SVC (0.4238) but has a better precision of 0.444422. We want high recall as compared to the high precision because high recall signifies identifying a true POI as a POI with a higher probability. Lower precision means erring and misclassifying more number of Non-POIs as a POI. We would want to trade off the erring towards misclassifying more Non-POIs as POIs rather than not identifying a True POI as a POI. This is because once we have the POI flagged then in the further steps of Investigation we can drop off a Non-POI identified as POI if evidences states so. Since, Decision Tree and SVC for 5 features have almost similar and best recall scores and the Decision Tree has better precision than the SVC, we will select the Decision Tree classifier with min_samples_split = 16 applied on the 5 features as the best suited algorithm for POI identification.

From above we see that parameter tuning is very important. For the same dataset, and same algorithm the results of the classifier varies for the different parameters selected. Hence we need to tune our parameters well so that our classifier gives the best possible metrics for the features which we believe explains the trend in our data

Validation and Performance

Validation is a means to measure the performance of the classifier on the features selected for the dataset. The classifier is trained using a set of data points (training set) and is tested on a different set of data points (testing set). We perform predictions on this testing set and compare it to the labels to check for the performance metrics of the classifier. It also serves as a check on over-fitting

The Enron dataset has 146 data points. Of these only 18 are POI and the rest are non-POI. The distribution of the 2 classes is imbalanced with almost 88% being Non-POI and only 12% are POI. Hence, if we employ simple K-fold method to split the data points into training and testing sets the ratio of POI to Non-POI in each of the folds might be very different. This is because it will depend on the order of the data points and how they get distributed in each of the folds. Hence, the metrics for the predictions which we get for our classifier might vary significantly for each of the folds. In order to overcome this problem we have used StratifiedShuffleSplit function to split the dataset into training and testing sets. We have set the folds to 1000 which denotes the number of re-shuffling and splitting iterations and random_state to 42 used for random sampling. This function provides the indices to split data into training and testing sets and we use it to train our classifier and perform the testing

Evaluation Metrics

The distribution of the POIs and Non-POIs in the Enron dataset is imbalanced. Hence, if we use accuracy as a metric to measure the performance of the classifier then even if the classifier classifies all of the data points as Non-POI still the accuracy will be around

88% since around 88% of the data are of Non-POI. Hence, we can't rely on accuracy metric to measure the performance of our classifier.

The metrics precision and recall turns out to be good metrics for such imbalanced datasets. The precision determines the likelihood if a person is identified as POI by the classifier he truly is POI. Given that a person is POI, recall measures the likelihood that our classifier flags it as POI

Hence precision and recall have been used as the metrics to measure the performance of the POI identifier classifier.

Following are the best scores which we have for our classifiers:

The average precision and recall scores for our Decision Tree Classifier is 0.444422 and 0.4238 respectively

Since the best average precision and recall was found when min_samples_split=16 for the Decision tree classifier, this algorithm and parameter has clearly emerged as the best case POI identifier classifier.

Discussion and Conclusions

The precision score of 0.444422 indicates that if our classifier identifies a person as POI, the likelihood that the person is truly POI is 44.44%

The recall score of 0.4238 indicates that given that a person is POI, the likelihood that our classifier will flag the person as POI 42.38%

Given the fact that only 18 out of the 146 data points are POI, these scores are good. But, there seems to be further scope for improvement in the scores. The e-mail data in the Enron starter dataset were financial data and the aggregated count of the e-mail messages. We didn't use the actual e-mail text data in our analysis. By exploring the actual e-mails we could extract Text features. There is a possibility that the text features might provide us with some more patterns in the data which might help in classifying the POIs better