ENRON DATASET INVESTIGATION

-K.NIKHIL

ABOUT ENRON :-

Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. It was founded in 1985 as the result of a merger between Houston Natural Gas and InterNorth, both relatively small regional companies in the U.S. Before its bankruptcy on December 2, 2001, Enron employed approximately 20,000 staff and was one of the world's major electricity, natural gas, communications and pulp and paper companies, with claimed revenues of nearly $101 billion during 2000.[1] *Fortune* named Enron "America's Most Innovative Company" for six consecutive years.

At the end of 2001, it was revealed that its reported financial condition was sustained by institutionalized, systematic, and creatively planned accounting fraud, known since as the Enron scandal. Enron has since become a well-known example of willful corporate fraud and corruption. The scandal also brought into question the accounting practices and activities of many corporations in the United States and was a factor in the enactment of the Sarbanes–Oxley Act of 2002. The scandal also affected the greater business world by causing the dissolution of the Arthur Andersen accounting firm.[2]

Enron filed for bankruptcy in the Southern District of New York in late 2001 and selected Weil, Gotshal & Manges as its bankruptcy counsel. It ended its bankruptcy during November 2004, pursuant to a court-approved plan of reorganization, after one of the most complex bankruptcy cases in U.S. history. A new board of directors changed the name of Enron to **Enron Creditors Recovery Corp.**, and emphasized reorganizing and liquidating certain operations and assets of the pre-bankruptcy Enron.[3] On September 7, 2006, Enron sold Prisma Energy International Inc., its last remaining business, to Ashmore Energy International Ltd. (now AEI).

--- SOURCE : WIKIPEDIA

1. Goals and dataset :-

The goal of this project is to build a predictive model that can identify persons of interest based on features included in the Enron dataset. We could use this model to find additional suspects who were not indicted during the original investigation,

The dataset contains **146 data points** with **21 features**. Of the 146 records,**18** are labeled as persons of interest.

After visualizing the dataset as a scatter plot, I identified an outlier named total

2. FEATURE SELECTION :-

In this part, we performed feature selection using cross validation and to select . We used the f_classif function as the scoring function for SelectKBest method in scikit-learn, and get

the best k features for k = 1:21,in each case, we built random forest classifiers and adaboost classifiers and tested their performances tusing cross validation.

3 Feature Selection :-

- The accuracy of both classifiers are roughly the same at different number of features used.
- Random Forest classifiers tend to have a higher precision score than AdaBoost classifiers. T
- he recall scores are relatively low for both classifiers no matter how many features used.
- The highest recall scores are obtained for both classifiers when we use the best 3 features bonus, total_stock_value and exercised_stock_options.
- We also obtained the importance of these three features from the random forest model and the scores are 0.33, 0.36 and 0.30.
- The features scores of these three features using SelectKBest are 21.06, 24.47 and 25.10 respectively.

4 TUNING THE ALGORITM :-

- Since we noticed that the recall scores for both classifiers are relatively low, In order to get higher recall score we used GridSearchCV for random forest classifier and adaboost classifier.
- These hyper-parameters determines the complexity of the classifiers
- For random forest classifier, we tuned the values for max_features, min_sample_splits as well as n_estimators.
- The higher values of max_features and n_estimators the higher complexity our classifier will be.
- The optimal values we find with respect to recall score are:
  2 for max_features,
  2 for min_samples_split
  100 for n_estimators
- For adaboost classifier, we tuned the values for n_estimators and learning_rate
- The optimal values for them are
  200 for n_estimators
  0.6 for learning_rate

5) Engineer Features :-

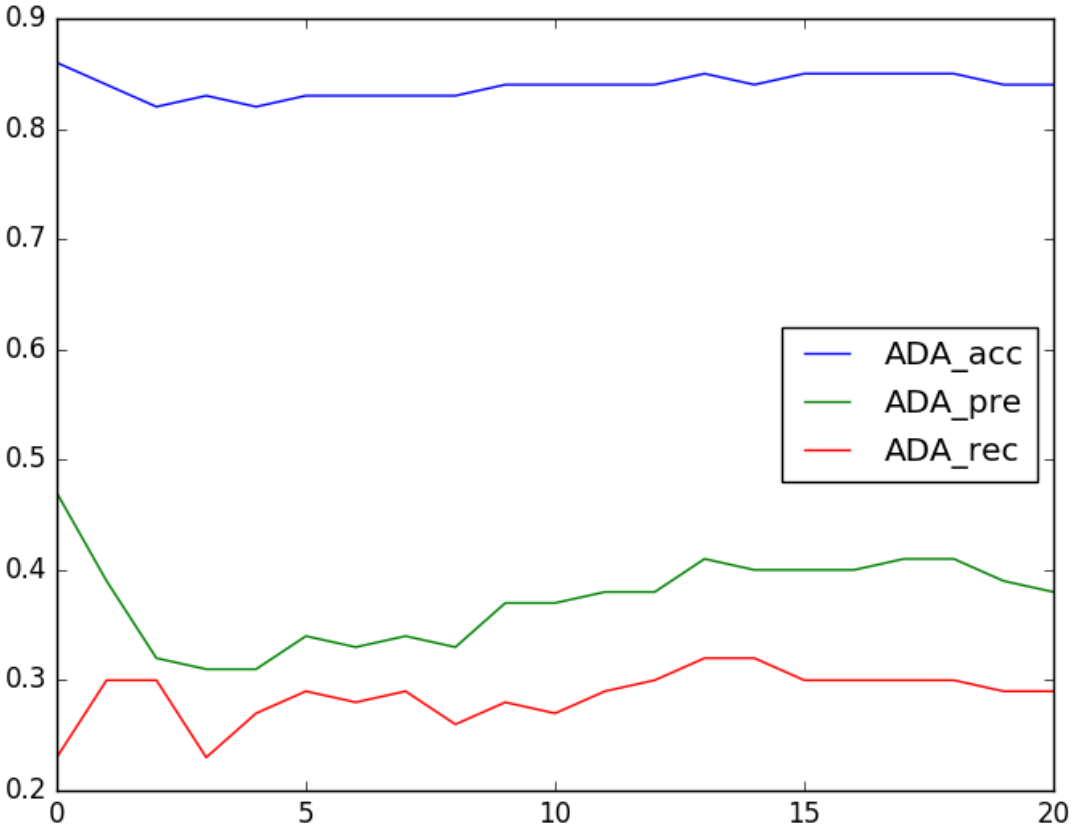
we create two new features here 'to_poi_message_ratio' and 'from_poi_message_ratio'

Using this two features we will be able to know how frequentlyor mails will be sent by him and how frequently he will recivie mails and why i included this two features is

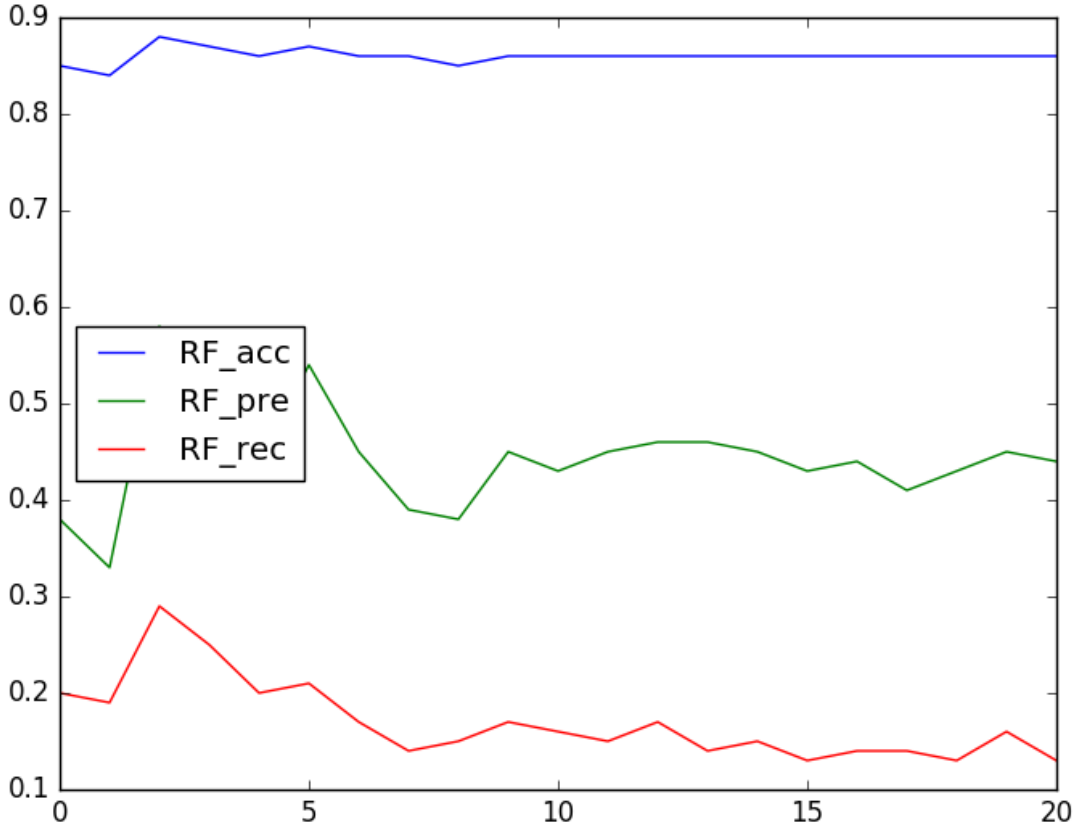
the person who did fraud mostly will have high ratios of either 'to_poi_message_ratio' 'from_poi_message_ratio'

Classification with Engineer Features consideration

Adaboost Classifier performance



RandomForest Classifier performance



1) It appears that using random forest model will tend to have higher precision and lower recall
2) For adaboost model it tend to have higher recall but lower precision,And the accuracy

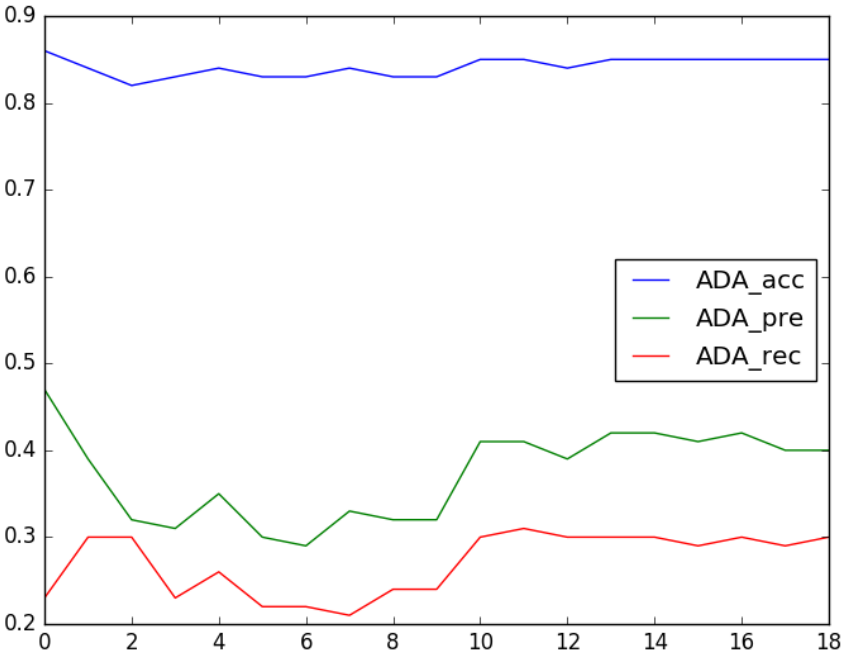
In the two engineered features (‘to_poi_message_ratio' 'from_poi_message_ratio') ‘to_poi_message_ratio’ played a important role

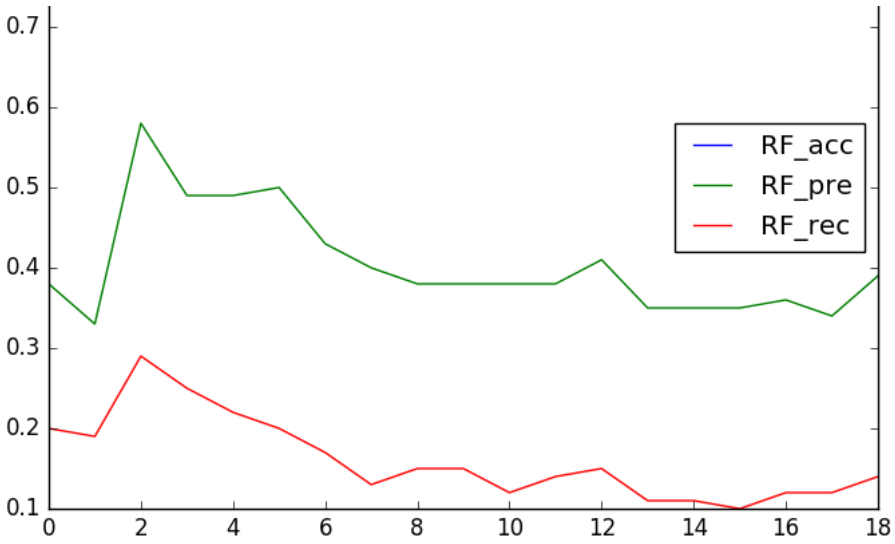
to_poi_message_ratio future importance score 0.11939078
to_poi_message_ratio future importance SelectKBest score is: 16.6417070705

Classification without Engineer Features consideration

Adaboost Classifier performance

Random Forest Classifier performance





From above graphs we are able to say that Random forest recall value and Predision score is effected by Engineer features
From above graphs we are able to say that Adaboost recall value and Precision score is effected by Engineer features

From above graphs Engineer features “to_poi_message_ratio”
Played a crucial role in all Precision Recall and accuracy

VALIDATE AND EVALUATE :-

resampling method – cross validation we use to extract additional information about the performance of our classifiers. We used this method in feature selection part as well as in the algorithm parameter tuning part. Since the number of observations in our dataset is rather small and the proportion of POIs in the dataset is relatively low, we used the StratifiedShuffleSplit method to do cross validation in this project. Below is a summary table of some performance measurements of our two tuned classifiers generated via cross validation.

Classifier	Precision	Recall	Accuracy
Random Forest	0.56201	0.39200	0.85946
Adaboost	0.33204	0.29950	0.79954

Percision = (No of persons classified as pois and are indeed pois)/(no of persons classified as poi’s)

Recall = (No of persons classified as pois and are indeed pois)/(no of persons indeed as poi’s)

5) Final Word :-

We originally thought this project is a bit ill proposed since the task is to predict who is a POI and some of the features given to us, like the number of emails the person sent to POIs, are the information only available when POIs are correctly identified. However the final features selected using cross validation doesn’t contain features of these types. And thus we got some confidence that the classifiers should have some usage in real life.

Missing value Appendix:-

Feature	Value
poi	0
salary	51
deferral_payments	107
total_payments	21
loan_advances	142
bonus	64
restricted_stock_deferred	128
deferred_income	97
total_stock_value	20

expenses	51
exercised_stock_options	44
other	53
long_term_incentive	80
restricted_stock	36
director_fees	129
to_messages	60
from_poi_to_this_person	60
from_messages	60
from_this_person_to_poi	60
shared_receipt_with_poi	60