Enron poi identification

Goal of project:

The goal of project is to find the person of interest POI using finanacial data and email of the enron dataset. Here i'll be using scikit-learn to train my model.

Outlier handling:

In exploratory phase I have sensed that lot of attributes are null so I counted total number of data it was 146, in that few of the attributes having null value more than 100 so that I thought the values are useless so I have removed those attributes then I have made two new atributes namely to_fraction and from_fraction

Features used:

The features that I have selected using SelectK5 are:

- **from_poi_to_this_person**
- **from_this_person_to_poi**
- **from_message**
- **shared_receipt_with_poi**
- **to_messages**

Algorithm Used:

I ended up by using the RandomForestClassifier . The main reason for choosing the algorithm is lable is binary and also it performed very well when compared to remaining Classifier

Parameter Tuning:

min_samples_leaf – number of minimum sample for further split
I tuned the parameter by using grid search cv. It showed high performance when it has 5 min_samples_leaf

Evaluation Metrics:

I evaluated by using scikit-learn's classification_report,
Percison score – 1
Recall score - 1