1.Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it. As part of your answer, give some background on the dataset and how it can be used to answer the project question. Were there any outliers in the data when you got it, and how did you handle those? [relevant rubric items: “data exploration”, “outlier investigation”]

R: O objetivo do projeto era investigar a base de dados sobre funcionários da Enron, que continha vários dados financeiros e de emails trocados entre as pessoas, para criar um modelo preditivo que detectasse se um funcionário estava envolvido no esquema de frauda na empresa ou não. Machine learning é útil pois neste caso temos um rótulo que indica se a pessoa está envolvida ou não e vários dados de cada pessoa, neste caso se utiliza o aprendizado supervisionado. Desta forma é possível treinar um algoritmo para prever se a pessoa está envolvida ou não no esquema. O valor TOTAL era um outlier, mostrava a soma de cada variável para o número total de funcionários,este oulier foi retirado. Existem outros ouliers como John Lavorato,Kenneth Lay,Jeffrey Skilling,Mark Fevert,Thimothy Belden e Phillip Allen, com valores muito altos de bonus e salários mas nada foi feito, pois são dados válidos de executivos da empresa.A funcionária 'LOCKHART EUGENE E.' foi retirada, pois possuía 20 colunas nulas.'THE TRAVEL AGENCY IN THE PARK' é pessoa jurídica e foi retirado.

A exploração dos dados mostrou o seguinte:

Total de dados: 146 pessoas e 21 variáveis, Número de POIs: 18 , Número de não POIs: 128

```
Index: 146 entries, ALLEN PHILLIP K to YEAP SOON
Data columns (total 21 columns):
bonus                      82 non-null float64
deferral_payments          39 non-null float64
deferred_income            49 non-null float64
director_fees              17 non-null float64
email_address              111 non-null object
exercised_stock_options    102 non-null float64
expenses                   95 non-null float64
from_messages              86 non-null float64
from_poi_to_this_person    86 non-null float64
from_this_person_to_poi    86 non-null float64
loan_advances              4 non-null float64
long_term_incentive        66 non-null float64
other                      93 non-null float64
poi                        146 non-null bool
restricted_stock           110 non-null float64
restricted_stock_deferred  18 non-null float64
salary                     95 non-null float64
shared_receipt_with_poi    86 non-null float64
to_messages                86 non-null float64
total_payments             125 non-null float64
total_stock_value          126 non-null float64
```

Figura1: Variáveis e número de valores não nulos

Percebe-se que director_fees só tem 17 valores não nulos,restricted_stock_deferred tem 18 e loan_advances apenas 4 não nulos. Estas são as variáveis que possuem mais valores faltando, por este motivo não serão utilizadas.

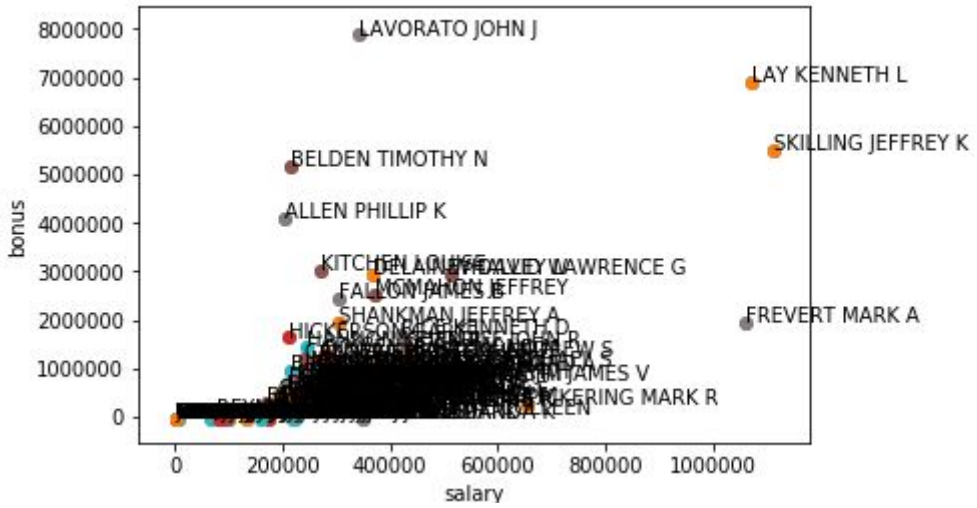

Figura2: Outliers financeiros

2.What features did you end up using in your POI identifier, and what selection process did you use to pick them? Did you have to do any scaling? Why or why not? As part of the assignment, you should attempt to engineer your own feature that does not come ready-made in the dataset -- explain what feature you tried to make, and the rationale behind it. (You do not necessarily have to use it in the final analysis, only engineer and test it.) In your feature selection step, if you used an algorithm like a decision tree, please also give the feature importances of the features that you use, and if you used an automated feature selection function like SelectKBest, please report the feature scores and reasons for your choice of parameter values. [relevant rubric items: “create new features”, “intelligently select features”, “properly scale features”]
R: As variáveis utilizadas são listadas em azul na tabela abaixo, o método de seleção utilizado foi o SelectKBest, testando o número k que resultava no maior score F1.Os valores de escore F1 para diversos números de variáveis são mostrados na tabela 2. O escalamento foi feito mas neste caso não é necessário, já que ele só deve ser feito para os algoritmos SVM e K-means, que não foram utilizados aqui.Sendo assim, servindo apenas como exercício extra. As variáveis novas que foram criadas são fraction_to_poi e fraction_from_poi, estas variáveis mostram a fração de e-mails recebidos ou enviados para POI’s com relação ao número total de e-mails recebidos ou enviados.Desta

forma,um valor alto para estas variáveis indica que a pessoa enviou ou recebeu muitos e-mails de POI's.

Variável	Escore
exercised_stock_options	6,84
total_stock_value	5,47
bonus	5,12
fraction_to_poi	4,64
salary	3,05
total_payments	2,78
long_term_incentive	2,53
shared_receipt_with_poi	2,43
other	1,71
expenses	1,48
from_poi_to_this_person	1,37
from_this_person_to_poi	1,00
fraction_from_poi	0,82
restricted_stock	0,58
to_messages	0,43
deferred_income	0,34
deferral_payments	0,06
from_messages	0,06

Tabela1:As 12 variáveis selecionadas (em azul) e escores

Nº de Variáveis	Escore F1
2	0.3
3	0.4
4	0.46
5	0.47
6	0.44
7	0.39
8	0.46
9	0.44
10	0.48
11	0.44
12	0.50
13	0.41
14	0.45
15	0.43

Tabela2: Escore F1 para k de 2 até 15, com RandomForest

As novas variáveis adicionadas aumentaram significativamente o escore F1, isto é mostrado na tabela abaixo:

12 variáveis	Escore F1 sem novas variáveis	Escore F1 com novas variáveis
	0.36	0.50

Tabela3: Aumento do escore F1 com as novas variáveis, usando RandomForest

3.What algorithm did you end up using? What other one(s) did you try? How did model performance differ between algorithms? [relevant rubric item: “pick an algorithm”]

R: O algoritmos testados foram Naive Bayes e RandomForest. Foi selecionado o RandomForest, pois apresentou o maior escore F1.Abaixo é mostrada a performance para os dois algoritmos.

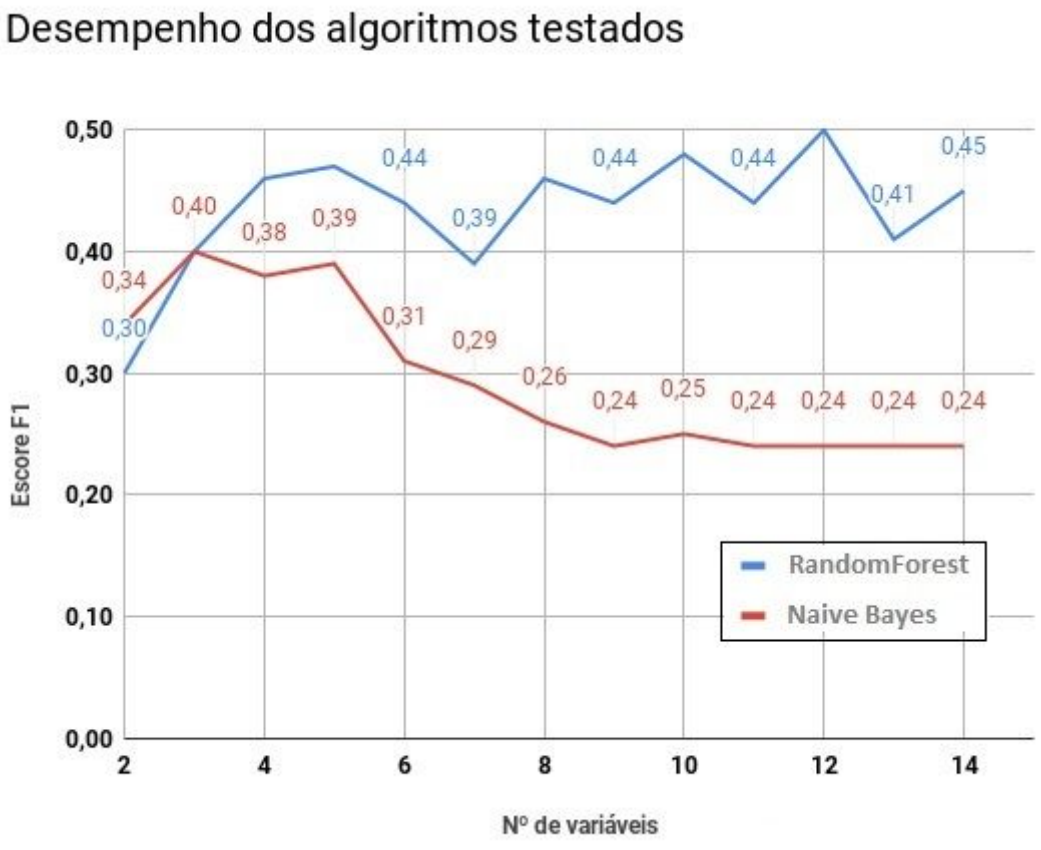

Figura3: Algoritmos testados

4.What does it mean to tune the parameters of an algorithm, and what can happen if you don't do this well? How did you tune the parameters of your particular algorithm? What parameters did you tune? (Some algorithms do not have parameters that you need to tune -- if this is the case for the one you picked, identify and briefly explain how you would have done it for the model that was not your final choice or a different model that does utilize parameter tuning, e.g. a decision tree classifier). [relevant rubric items: "discuss parameter tuning", "tune the algorithm"]

R: É o ajuste de parâmetros de um algoritmo que vai fazer com que ele tenha o seu melhor desempenho possível com as variáveis utilizadas no conjunto de dados. Se não for feito este ajuste não teremos o melhor desempenho do algoritmo. Para ajustar os parâmetros do algoritmo RandomForest foi utilizada a função GridSearchCV do Sklearn. Os parâmetros ajustados foram os seguintes:

```
"max_depth": [3,4,5,6],
"max_features": [2,3,4],
"min_samples_split": [0.1,0.2, 0.3],
"criterion": ["gini", "entropy"],
 "n_estimators":[5,10,15]
```

5.What is validation, and what's a classic mistake you can make if you do it wrong? How did you validate your analysis? [relevant rubric items: "discuss validation", "validation strategy"]

R: Validação é o processo em que se testa se o algoritmo utilizando um conjunto novo de dados de teste para verificar se ele está fazendo oque deveria fazer, de acordo com as métricas de avaliação escolhidas. Se a validação não for feita da maneira adequada o modelo resultante terá um desempenho ruim. Um dos problemas de uma validação feita no mesmo conjunto de dados em que o algoritmo foi testado é o 'overfitting', quando o desempenho com o conjunto de dados utilizados é muito bom mas se torna ruim para dados completamente novos. Uma boa validação possui um erro de treinamento baixo e um erro com o conjunto de dados de teste baixo também. Neste projeto foi utilizada a validação cruzada, com a função 'cross_val_score' do sklearn.A validação cruzada funciona da seguinte forma:

- Um modelo é treinado usando k-1 divisões como dados de treinamento
- O modelo treinado é validado com a parte restante dos dados

A performance final é a média dos valores obtidos em cada iteração, na etapa de validação

6.Give at least 2 evaluation metrics and your average performance for each of them. Explain an interpretation of your metrics that says something human-understandable about your algorithm's performance. [relevant rubric item: "usage of evaluation metrics"]
R: Precisão e Revocação. Valores médios obtidos:

Precision: 0.48

Recall: 0.51

Precisão indica entre todas as previsões para POIs, qual a proporçao dos corretamente previstos.Neste caso 48% das previsões de POIs estão corretas.

Revocação indica entre todos os casos que deveriam ser rotulados como POIs qual a proporção dos corretamente previstos.Neste caso 51% das pessoas que deveriam ser rotuladas como POIs foram rotuladas corretamente.