Enron Submission Free-Response Questions

A critical part of machine learning is making sense of your analysis process and communicating it to others. The questions below will help us understand your decision-making process and allow us to give feedback on your project. Please answer each question; your answers should be about 1-2 paragraphs per question. If you find yourself writing much more than that, take a step back and see if you can simplify your response! When your evaluator looks at your responses, he or she will use a specific list of rubric items to assess your answers. Here is the link to that rubric: [Link] Each question has one or more specific rubric items associated with it, so before you submit an answer, take a look at that part of the rubric. If your response does not meet expectations for all rubric points, you will be asked to revise and resubmit your project. Make sure that your responses are detailed enough that the evaluator will be able to understand the steps you took and your thought processes as you went through the data analysis.

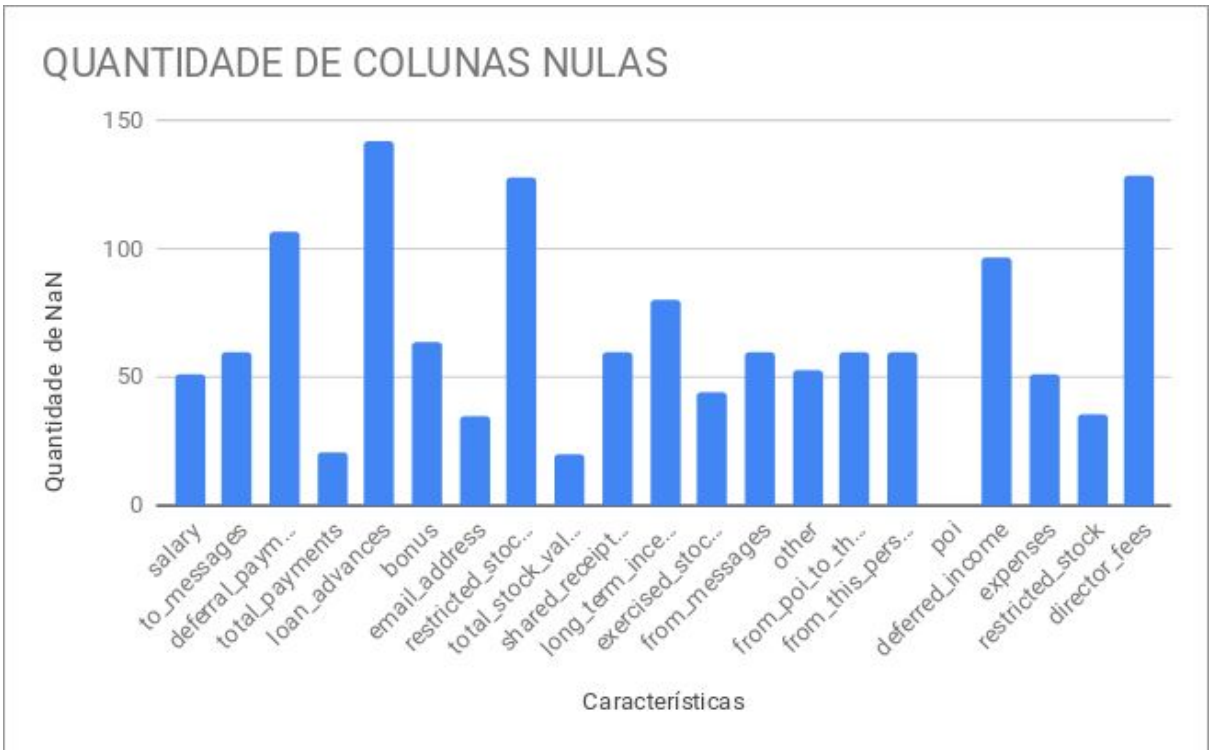
Once you've submitted your responses, your coach will take a look and may ask a few more focused follow-up questions on one or more of your answers.

We can't wait to see what you've put together for this project!

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"]

   Objetivo do projeto é basicamente treinar uma máquina que seja capaz de reconhecer os padrões um POI para que ela consiga identificar estes através de dados como salários, bônus, emails trocados com POIs, emails compartilhados com POIs, etc. O conjunto de dados de dados é bem limitado(146 - 1) possuindo vários campos como NaN, porém com 21 colunas de informação. O número de POIs também não se encontra balanceado no dataset pois existem apenas 18 POI para 128 não POI, o ideal seria 50% para cada. Segue o gráfico de número de NaN do

dataset:



Podemos ver no gráfico que existem muito valores faltantes, o que gera imprecisão nos resultado, portanto é bom evitar o uso de colunas como(deferral_payments, loan_advanceds,restricted_stock_deferred,deferred_income,director_fees).Um valor discrepante era o de TOTAL, que não deveria estar sendo listado como se fosse um funcionário, então, bastou retira-lo. Porém, como muitos campos eram disponibilizados foi possível identificar um padrão nos dados.

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”]

algoritmo\%	100%	50%	25%	10%
GaussianNB	0.24546	0.32894	0.15864	0.29543
DecisionTree	0.19295	0.21150	0.38400	0.46386
SVM	with out positive prediction	with out positive prediction	with out positive prediction	with out positive prediction

Para a escolha das features a serem usadas, treinei uma árvore de decisão e peguei a pontuação para cada feature, assim, peguei todas as melhores 100%, 50%, 25% e 10%. O resultado obtido está na tabela acima. Podemos perceber que o melhor F1 é encontrado na decision tree com 10% dos dados, então esta foi utilizada como final, segue as características utilizadas= `['bonus'(0.230),'expenses'(0.170)]`. A feature criada é basicamente a soma dos atributos `[from_this_person_to_poi,from_poi_to_this_person,shared_receipt_with_poi]` de email, o objetivo era reunir informações relativas ao relacionamento com POIs, porém ao adicionar esta feature o recall diminuía 0.2 . Também realizei o escalonamento para as features do svm utilizando a função preprocessing.MinMaxScaler()

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”]

   O algoritmo final utilizado foi o árvore de decisão, mas testei também o SVM e o GaussianNB, todos deram um resultado próximo, porém a árvore de decisão alcançou o maior recall. Para maior esclarecimento sobre os resultados olhar a tabela da questão anterior, onde apresento os valores de F1 para cada algoritmo.

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”]

   O ajuste dos parâmetros é importante para ajustar a máquina de aprendizado a os dados que serão utilizados. Se isso não for feito pode acontecer um overfit ou outros problemas de aprendizado. Para fazer isso utilizei o *GridSearchCV* que me permitiu testar várias opções para cada parâmetro, para o

```
algoritmo de Decision tree utilizei os seguintes:
{'min_samples_split':range(2,20),'max_depth':range(3,100),'max_fe
atures':range(1,5)}

Para o algoritmo de SVM utilizei:

{'kernel':['rbf'], 'C':[1,
10,100,1000,10000],'gamma':[0.005,0.05,0.1,0.2]}

Para o algoritmo de GaussianNB: {}

A configuração final utilizada no algoritmo de decision tree foi:

DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=5,max_features=None, max_leaf_nodes=None,
min_samples_leaf=1,min_samples_split=5,
min_weight_fraction_leaf=0.0,presort=False, random_state=None,
splitter='best')
```

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"]

   O objetivo da validação é verificar se a máquina treinada realmente aprendeu a identificar os padrões para obter uma resposta, e é preciso verificar isso sem ir diretamente ao problema final, por isso fazemos testes com os próprios dados disponibilizados, partindo deste teste podemos verificar quanto nosso algoritmo está correto. Um grande problema é que se realizamos o teste com os próprios dados que usamos pra treinar, pois sempre vamos obter uma resposta positiva, pois o algoritmo já está ajustado a aqueles dados, então é necessário separar os dados para teste e treino. Isto é dividir os dados de forma a realizar um melhor treinamento e um melhor teste da machine.

   Para a validação da análise fiz uso do *GridSearchCV* `que realiza uma validação cruzada enquanto testa diferentes parâmetros. Ele realiza o teste com dados diferentes do que ele treinou.`

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"]

Eu obtive uma média de 0.49 na precision e 0.43 no recall, o primeiro indica a probabilidade de acerto para o caso de ser POI e o segundo descreve a capacidade do meu algoritmo de acertar como POI quando a pessoa testada é um POI.