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"]

O objetivo do projeto é, utilizando dados públicos, tentar encontrar um padrão que permita encontrar os chamados Funcionários de Interesse (POI). Os POI são funcionários que de alguma maneira foram condenados ou estiveram envolvidos com a justiça. Com base nos dados financeiros e de emails disponíveis, todos os funcionários foram classificados como POI ou não-POI.

A base de dados possui um total de 146 pessoas, destas 18 são classificadas como POI. Para cada pessoa listada é disponibilizado informações sobre atributos financeiros e dados sobre os emails. Para esse projeto foi utilizado 11 atributos financeiros e 5 atributos relacionados aos emails (melhor detalhado na resposta seguinte).

Foram encontrados dois problemas iniciais para a construção do algorítmo. Primeiro, a lista de POI é relativamente pequena para a realização do treinamento. Segundo, poderiam haver pessoas que cometeram fraudes e seus dados teriam características de POI mas não foram envolvidas com a justiça. Esse segundo caso poderia dificultar o treinamento do algorítmo ou apresentar falsos positivos. Também foi encontrado bastante heterogeneidade entre os dados financeiros e de email. Os outliers encontrados foram as linhas de "TOTAL" e "'THE TRAVEL AGENCY IN THE PARK" que não representavam nenhum indivíduo, alguns valores negativos (provavelmente extornos financeiros) e "LOCKHART EUGENE E" por não possuir nenhuma informação.

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”]

As features utilizadas foram:

financial_features: 'salary', 'bonus', 'deferral_payments', 'deferred_income', 'exercised_stock_options', 'expenses', 'long_term_incentive', 'other', 'restricted_stock', 'total_payments', 'total_stock_value' e 'bonus_ratio'.

email_features: 'from_messages', 'from_poi_to_this_person', 'from_this_person_to_poi', 'shared_receipt_with_poi', 'to_messages', 'from_this_person_to_poi_ratio' e 'from_poi_to_this_person_ratio'.

Além das features dadas foram criadas três novas features basedas na razão de duas features existentes. 'from_this_person_to_poi_ratio' e 'from_poi_to_this_person_ratio' é a razão entre as mensagens trocadas com POI e o total de mensagens, e 'bonus_ratio' é a razão do bonus em relação ao valor do salário. Suspeitamos que, além dos valores absolutos, os valores relativos de algumas variáveis também podem ajudar a classificar um pessoa como POI.

Em seguida, cada feature foi avaliada manualmente quanto a sua contribuição na classificação entre POI/não-POI. Por exemplo, não haviam dados de 'director_fee' para não-POI, portanto essa feature não foi utilizada no classificador. O mesmo ocorreu com 'restricted_stock_deferred' e 'loan_advances', features com poucos dados disponíveis. Em seguida foi utilizado um histograma para verificar qual o formato da distribuição de cada feature, e para as que possuiam uma "cauda longa" foi realizada uma transformação logarítmica. Todas as features foram normalizadas entre 0 e 1.

Por fim, na etapa de avaliação dos classificadores, foi testado os algorítmos k-best para escolher os melhores atributos e o PCA para decompor e diminuir a dimensionalidade dos atributos. Foram testadas diversos arranjos de algorítmos e parâmetros para se escolher a melhor opção.

Ao rodarmos o script com e sem as novas features chegamos aos seguintes valores:

Com novas features – f score = 0,47

Sem as novas features – f score = 0,43

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”]

Foram testados os algorítmos Gaussian Naive Bayes, Support Vector Machine, Decision Tree e Random Forest. Para cada algoritmo foi testado duas estratégias de redução do número de atributos. A primeira foi usando o algoritmo k-best que seleciona os k-principais atributos que influenciam na métrica selecionada, no caso f-score. Foram testados valores de k de 1 a 19. A segunda estratégia foi utilizar o algoritmo PCA que reduz a dimensionalidade dos atributos para n compondentes. O valor de n também variou de 1 a 19.

Todas as estratégias foram testadas 5 vezes e foi selecionado pipeline (algoritmo de redução de atributos mais algoritimo de classificação) com melhor f-score.

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"]

Para cada algorítmo escolhido há uma série de parâmetros, relacionados com sua lógica de funcionamento, que podem ser ajustados. A escolha dos valores desses parâmetros podem implicar em um melhor ou pior resultado para cada situação, portanto afinar esses parâmetros é uma etapa importante do projeto. É necessário tomar cuidado com o sobreafinamento, onde os resultados ficará muito bem ajustados aos dados de treinamento, mas pode gerar um modelo pouco realista para predição de novos dados. Já o subafinamento pode gerar um modelo que gerará resulados aquém da capacidade do algoritmo.

Desta forma, após a escolha dos algorítmos que seriam testados foram identificados os principais parâmetros que poderiam influenciar a performance. Esses parâmetros foram afinados utilizando o GridSearchCV permitindo um afinaamento de forma automática.

Os parâmetros ajustados foram:

- Support Vector Machine – kernel, C e degree
- Decision Tree – criterion, splitter e min_sample_split
- Random Forest – n_estimators, criterion e min_sample_split

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"]

Validação são técnicas utilizadas para separar uma parte dos dados, que não será utilizado para treinamento, para verificar se a performance do algorítmo desenvolvido está adequada. Isso evita que a avaliação seja feita com os mesmos dados em que o algorítmo foi treinado, o que pode ocasionar um sobreajuste. Para esse projeto foi utilizado Stratified K-fold para separar os dados de treino e de teste .

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"]

Foram utilizados os seguintes indicadores para medir a performance dos algoritmos: Accuracy, Precision, Reccal e f Score.

A Accuracy mede a quantidade de previsões corretas em relação ao total de medições. Como nesse caso o número de POI é sempre muito baixo, mesmo se o modelo nunca prever um POI o valor da Acurácia será relativamente alto, portanto passamos para os outros indicadores.

O Precision é a razão entre os verdadeiros positivos e todos os positivos detectados (verdadeiros ou falsos). Já o Reccal é a razão entre os verdadeiros positivos e todos os positivos existentes (verdadeiros positivos e falsos negativos). Podemos notar que um alto Precision indica que todos os positivos detectados são de fato positivos (podendo haver

outros não detectados) e um alto Reccal indica que detectamos todos os positivos (podendo haver detectado alguns falsos positivos. O f Score é uma relação entre Precision e Reccal.

A decisão do algorítmo a ser utilizado foi com base no f Score e obedecendo o limite mínimo de 0,3 tanto para o Precision quanto para o Reccal.