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. 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"]

Para a escolha das features a serem usadas, treinei uma árvore de decisão e peguei a pontuação para cada feature, assim, peguei todas as que estavam acima de 10% de

importância o resultado foi = `['total_payments'(0.164), 'bonus'(0.230), 'restricted_stock'(0.120), 'expenses'(0.170)]`, porém ao realizar os testes e tentar ir retirando um por um percebi que o `'restricted_stock',` estava piorando o meu recall, então o retirei, assim como a feature que eu criei. A feature criada é basicamente a soma do escalonamento dos atributos de email. Também realizei o escalonamento para das features para o svm , porém o resultado não foi tão satisfatório, para os demais, então retirei

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

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 final utilizei os seguintes:`

   ```
   {'min_samples_split':range(2,20),'max_depth':range(3,100),'max_features':range(1,5),'max_leaf_nodes':[10,50,100,200,500]}
   ```

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”]

   É dividir os dados de forma a realizar um melhor treinamento e um melhor teste da machine, o problema é quando dividimos o dados sequencialmente o que acaba que o treinamento e o teste podem não abranger situações reais. para a validação da análise fiz uso do *`GridSearchCV`* `que realiza uma validação cruzada enquanto testa diferentes paramentros.`

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 0.41 na precision e 0.31 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.