1 Primeira Questão

1.1 Questão

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"]

1.2 Resposta

O objetivo desse projeto é criar um mecanismo capaz de prever quem é uma POI (Pessoa de interesse). Ou seja, quem é um possível suspeito Usando a informação financeira da empresa, salários, bônus dentre outras e os e-mails enviados e recebidos pelas pessoas da empresa. Com esses dados se torna viável alimentar um mecanismo de aprendizagem de máquina para que o computador consiga identificar os padrões intrínsecos do fenômeno.

Infelizmente temos somente dados sobre 145 indivíduos, aonde 18 são POI, 127 normais. Foi retirado dos dados a entrada "TOTAL", que era um erro nos dados pois não representava nenhum indíviduo mas a soma de todos, e foi criado ou escalonado 9 features a partir de 10 selecionados do conjunto total. Infelizmente 7 desses 9 features tiveram mais de 50 resultados dos 145 faltando o que diminui a acurácia da análise.

2 Segunda Questão

2.1 Questão

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"]

2.2 Resposta

Escolhi os features, considerando quais poderiam ter peso e sabendo que o PCA iria diminuir o erro de ter escolhido features pouco informativos. Logo escolhi: poi, frac_salary, frac_bonus, frac_total_payments, frac_expenses, frac_total_stock_value, frac_from_poi_to_this_person, frac_from_this_person_to_poi, 'frac_shared_receipt_with_poi e multiply_emails. Os nomes frac só são indicativo de que houve alteração no feature original. Nos features econômicos foi feito e reescalonamento. Foi divido pelo maior valor todas as entradas. Já nos e-mail o reescalonamento foi feito dividindo pela feature to_messages ou from_messages conforme cada caso.

Enquanto o multiply_emails foi criado por mim e é a multiplicação dos atributos de e-mails enviados e recebidos por POI pois considerei que são atributos que poderiam ter comportamento multiplicativo, ou seja, se um for muito baixo cortaria efeito do outro. Infelizmente para o número de componentes escolhida no PCA (que foram 4) Essa variável não apresenta efeito no resultado. Esse teste foi realizado analisando-se o recall e a precisão obtida com e sem essa feature criada. O número de componentes do PCA foi escolhido através de todos os valores de 1 a 8(valor máximo) aonde ficou claro que o melhor resultado (métrica F1) seria para $n_{components} = 4$.

3 Terceira Questão

3.1 Questão

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"]

3.2 Resposta

Terminei utilizando o SVM, mas testei o naive bayes (antes de implementar o PCA) e a árvore de decisão foi amplamente testada (incluindo uso de PCA), ela chegava a cerca de 0.27 de precisão e recall mas nunca passou dos 0.3. Enquanto o SVM conseguiu 0.5 de precisão e recall.

4 Quarta Questão

4.1 Questão

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"]

4.2 Resposta

Boa parte dos algoritmos de aprendizagem de máquina tem parâmetros que vão definir o seu comportamento e se não ajustados para o problema corretamente irão trazer parcos resultados. Inicialmente usei gridSearchCV para afinar os parâmetros. mas não obtive bons resultados para a precisão e recall logo decidi fazer o ajuste manualmente. Os parâmetros que ajustei foram o parâmetro C do SVM e o parâmetro $N_{components}$ do PCA. Com um pouco um pouco de tentativa e erro encontrei o parâmetro $N_{components} = 4$ e $C = 40000$.

Foi testado kernel linear mas os resultados sempre foram pífios assim como $gama = [1 : 10]$ por tanto decidi convencionar usar gama auto que equivale a 0.125 no nosso caso que mostrou bons resultados o parâmetro mais medido foi o C o qual testei vários valores como por exemplo: [1:10,100,1000,10000,20000,30000,40000,50000] E notei que a partir de 40000 os resultados pioravam conforme se aumentava.

5 Quinta Questão

5.1 Questão

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?]

5.2 Resposta

Validação é dizer se seu resultado é válido, aceitável. Um exemplo é se testarmos nossa máquina utilizando-se dos mesmos dados os quais treinamos ela, o resultado pareceria ser excelente porém ao se fazer previsões em outro conjunto

de dados erraríamos muito mais que o esperado. Situação perigosa e indesejável. Nesse caso o aprendizado não foi validado corretamente.

Nesse problema tivemos quase sempre uma acurácia relativamente alta, mas o recall e a precisão eram outra história. Logo se eu tivesse me baseado na acurácia somente teria um máquina incapaz de prever corretamente POIs. Validei meu algoritmo usando em especial a métrica F1 que é a média harmônica do recall e precisão, pois esses eram os valores importantes. F1 final foi de 0.444444. Separei aleatoriamente 30% da amostra para ser de teste e treinei o modelo nos outros 70%

6 Sexta Questão

6.1 Questão

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"]

6.2 Resposta

Medi a acurácia, precisão e o recall. Obtive respectivamente: 0.8605, 0.5 e 0.5. A acurácia me diz o quanto de acerto em suas previsões meu algoritmo consegue ter. Já a precisão diz o quão preciso é ele ao dizer que alguém e uma POI, ou seja qual a probabilidade dele não ter dado um falso positivo e por fim o recall indica qual a probabilidade dele não ter deixado de marcar alguém como POI.

Ou seja com uma precisão baixa estamos identificando de forma incorreta e pode ser que tenhamos acusado alguém injustamente. Logo é na minha opinião a métrica mais importante. Porém alguém poderia argumentar que é o recall a mais importante. Pois com um recall baixo deixaremos possíveis suspeitos escapar. Logo para selecionar previamente pessoas a se investigar a fundo e ter a certeza de que se selecionou todos os culpados tem-se que se ter um alto recall.

Como os resultados de precisão e recall foram ambos baixos não indicaria esse algoritmo para ser usado para encontrar novos POIs.