Vinicius Ferreira Santos

Data Science para Negócios

03/06/2018

Identificando Pessoas de Interesse da Enron

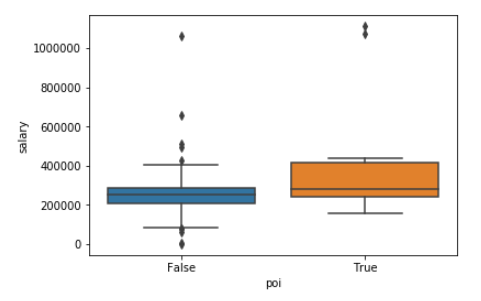
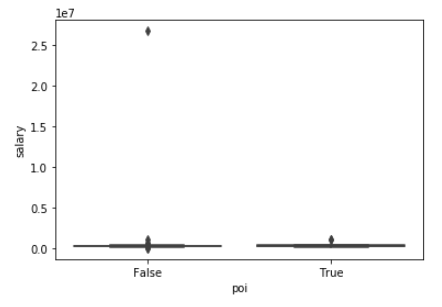
1. Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it. As part of your answer, give some background on the dataset and how it can be used to answer the project question. Were there any outliers in the data when you got it, and how did you handle those?

O objetivo desse projeto é identificar as pessoas de interesse(POI) no caso de fraude da Enron. O projeto usa técnicas de Machine Learning para criar um modelo preditivo, que utiliza dados financeiros e de e-mails de funcionários investigados, o que significa que eles foram indiciados, fecharam acordos com o governo, ou testemunharam em troca de imunidade no processo.

Na exploração dos dados, encontrei 146 funcionários, dos quais 18 eram POI. Cada registro do dataset possui 22 features, com os detalhes abaixo:

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Tipo | Total non-Null | Total Null |
| nome | Object | 146 | 0 |
| bonus | Float64 | 82 | 64 |
| deferral\_payments | Float64 | 39 | 107 |
| deferred\_income | Float64 | 49 | 97 |
| director\_fees | Float64 | 17 | 129 |
| email\_address | Object | 111 | 35 |
| exercised\_stock\_options | Float64 | 102 | 44 |
| expenses | Float64 | 95 | 51 |
| from\_messages | Float64 | 86 | 60 |
| from\_poi\_to\_this\_person | Float64 | 86 | 60 |
| from\_this\_person\_to\_poi | Float64 | 86 | 60 |
| loan\_advances | Float64 | 4 | 142 |
| long\_term\_incentive | Float64 | 66 | 80 |
| other | Float64 | 93 | 53 |
| poi | Bool | 146 | 0 |
| restricted\_stock | Float64 | 110 | 36 |
| restricted\_stock\_deferred | Float64 | 18 | 128 |
| salary | Float64 | 95 | 51 |
| shared\_receipt\_with\_poi | Float64 | 86 | 60 |
| to\_messages | Float64 | 86 | 60 |
| total\_payments | Float64 | 125 | 21 |
| total\_stock\_value | Float64 | 126 | 20 |

Ao analisar os dados, foram encontrados alguns ruídos interessantes que foram prontamente removidos, entre eles, está o funcionário *LOCKHART EUGENE E* que não possui dados em nenhuma feature de e-mail e financeira. Outro ruído encontrado foi um funcionário com nome de empresa *THE TRAVEL AGENCY IN THE PARK.* Por último, ao observar graficamente salário dos funcionários, me deparei com um outlier aberrante de $ 26.704.229 para um funcionário com nome *TOTAL,* que mais parecia ser o totalizador dos salários, causado por algum erro na criação do dataset.



Salários com Outlier Salários sem Outlier

1. 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.

Para selecionar as features do modelo final, inicialmente foi realizado um processo de engenharia para aumentar a probabilidade de identificar um POI. Como resultado da engenharia, surgiram três novas features:

• **fraction\_to\_poi**: que é a fração de e-mails enviados para um POI(from\_poi\_to\_this\_person/to\_messages).  
• **fraction\_from\_poi**: que é a fração de e-mails que foram recebidos de um POI(from\_this\_person\_to\_poi/from\_messages)  
• **income**: que é o rendimento de um funcionário(salary+bônus).

Após a engenharia de features, foi utilizado o SelectKBest para selecionar as 10 melhores levando em consideração a sua variância(ANOVA). O SelectKBest definiu as 10 melhores features abaixo:

|  |  |
| --- | --- |
| Feature | Score |
| bonus | 30.7287 |
| income | 30.5270 |
| salary | 15.8587 |
| fraction\_to\_poi | 15.8380 |
| shared\_receipt\_with\_poi | 10.7225 |
| total\_stock\_value | 10.6338 |
| exercised\_stock\_options | 9.6800 |
| total\_payments | 8.9591 |
| deferred\_income | 8.7922 |

Foi utilizado MinMaxScaler para definir o menor valor para 0 e o maior como 1 para todas as features. Foi necessário escalar as features, devido a quantidade de outliers que poderiam prejudicar o modelo sendo escolhidos como preditores principais. Outro motivo importante para o escalonamento, é que as features com valores ausentes foram preenchidos com 0 e o MinMaxScaler define como 0 os menores valores.

1. 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”]
2. 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”]
3. 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”]
4. 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”]