Machine Learning on ENRON dataset

Jader Martins Camboim de Sá

29 de Maio de 2018

1 Introdução

Este projeto tem por objetivo investigar a base de dados da ENRON em busca de "Pessoas de Interesse" que são aqueles que de alguma forma de envolveram com o escandalo de corrupção da empresa. Para isso foram dadas a classe '0' para pessoas inocentes e '1' para as que tiveram algum envolvimento, essas são chamadas de classes alvo, além disso atributos financeiros e outros foram fornecidos sobre cada pessoa através do machine learning conseguimos criar um model que dentro de uma certa metrica consegue prever quais pessoas são inocentes e quais estão envolvidas a base contém 143 registros e 20 atributos, muitos deles vazios, como mostra a contagem a seguir.

Atributo	Qt NaN
bonus	62
deferral_payments	105
deferred_income	95
director_fees	127
exercised_stock_options	42
expenses	49
from_messages	57
from_poi_to_this_person	57
from_this_person_to_poi	57
loan_advances	140
long_term_incentive	78
other	52
poi	0
restricted_stock	34
restricted_stock_deferred	126
salary	49
shared_receipt_with_poi	57
to_messages	57
total_payments	20
total_stock_value	18

Atributo	Qt NaN

Para aplicar o modelo aos dados foi necessário realizar analises em busca de inconsistencia e realizar algumas limpezas, como remoção de outliers, atributos outliers carregavam informações importantes foi preferido substituir seu valor por zero ou pela media, dependendo do contexto deste outlier.

2 Modelagem dos Dados

Muitos atributos não contem informações relevantes sobre o objetivo da predição por isso adiciona-los apenas adicionará ruído ao modelo atrapalhando a predição, como o objetivo é generalizar, utilizei do algoritmo SelectKBest para me escolher os atributos visando que não ficasse enviesado aos meus dados de teste e que cada modelo pudesse trabalhar com uma quantidade diferente de dados com o GridSearchCV.

Também foi necessário fazer um *rescaling* dos dados pois alguns modelos assumem ou se beneficiam de uma distribuição normal, por isso dependendo do atributo foram aplicadas funções raiz, log ou normalização. Também foram criados alguns atributos como 'total_stock_value'/'salary' para capturar algum anormalidade no padrão de investimento dado conhecimento externo de um "poi", outro atributo criado foi o percentual de mensagens enviadas ou recebidas para/de um poi.

3 Metodologia

Alguns algoritmos escolhidos conseguiam prever apenas os "não-poi", tendo em vista que dos 143 registros apenas 18 são "poi", por isso dei preferência aqueles que permitiam balancear o peso das classes, para minha surpresa a regressão logistica (junto ao GradientBoost) teve bons resultados sendo escolhida para o fine tuning final. Na etapa final (tuning) me atentei a fazer um validação razoavel, de forma que não fossem tantos "folds" podendo demorar muito e nem poucos de forma que o modelo pode ter overfitting, consultando diversos posts do forum e atraves de testes empiricos cheguei em um modelo balanceado de validação cruzada. Dado que a regressão logistica tem poucos parametros (comparado a modelos de arvore por exemplo) decidi explorar uma vasta gama de valores no começo porém conforme ia modificando os dados vi que não seria necessário reduzindo meu conjunto ao pequeno atual.

Na validação, que é a analise da capacidade de generalização do algoritmo, coloquei a principio como métrica o f1-score, porém vi que o mesmo como alternativa final não tinha bons resultados e voltei ao default.

4 Conclusão

Obtive os seguintes resultados:

Metrica	Resultado
Accuracy	0.79767
Precision	0.33863
Recall	0.54300
F1	0.41713

Dada uma amostra desbalanceada, maioria de registros de uma unica classe, é importante que usemos outras metricas de acerto, pois se chutassemos que todo valor é da classe mais presente acertariamos na maioria dos casos, porém como nos dados aqui avaliados, é mais importante conhecer os registros mais raros, usando assim as metricas precisão e revocação que nos permite avaliar a relevância da inferencia, dentro do percentual de amostras como foi minha predição.