Enron: Maior caso de corrupção e lavagem de dinheiro do Estados Unidos da América

Vilmar dos Santos Alves

Fonte: BBC.

A Enron era uma companhia tradicional de serviços públicos que possuía usinas elétricas, companhias de água e saneamento e unidades de distribuição de gás que se tornou realmente conhecida por atuar com um estilo ousado para o seu setor, baseado nas práticas do mercado acionário.

O achado da equipe da Enron foi perceber que energia, água e mesmo produtos mais obscuros, como espaço em linhas de telecomunicação, poderiam ser negociados como eram commodities. A partir dessa percepção, a companhia passou a atuar como uma espécie de grande corretor do setor de energia, comprando, vendendo e fazendo apostas financeiras muito maiores do que os negócios diretamente realizados pela companhia.

Essas apostas fizeram a Enron se tornar, por um breve período, a maior empresa de energia do mundo, com vendas de US$ 101 bilhões no ano passado, rivalizando com nomes como Shell e Exxon. Sua ousadia também a levou para o mercado europeu, quando este começou seu processo de liberalização.

Mas o que provocou o colapso da Enron?

Por quase uma década, o sistema e a ousadia da Enron foram aplaudidos mundialmente. A empresa parecia ter encontrado a fórmula para fazer muito dinheiro com o negócio de suprir energia.

Ela foi eleita vária vezes como a empresa mais admirada do mundo. Mas a magia não durou muito. As operações de comércio da companhia se baseavam na maior parte das vezes em transações financeiras extremamente complexas, algumas se referindo a negócios que deveriam ocorrer vários anos depois.

Auditar esse tipo de transação é sempre difícil, mas no caso da Enron a situação foi piorada ainda mais por incompetência ou por uma possível ação criminosa de executivos de alto escalão da companhia.

Quando a empresa apresentou o resultado de seu terceiro trimestre em outubro de 2001, revelou um enorme e misterioso buraco em suas contas que derrubou os preços de suas ações. Depois desse anúncio, a comissão responsável pela fiscalização do mercado acionário americano, a SEC, começou a investigar os resultados da empresa.

A Enron então acabou admitindo que havia inflado os seus lucros, o que rebaixou ainda mais o valor de suas ações. A queda afastou a alternativa de venda da companhia como forma de solucionar sua crise financeira, o que a levou para o processo de concordata em 2 de dezembro de 2001.

A rápida transformação da Enron de uma das companhias mais admiradas do mundo em protagonista da maior concordata da história corporativa dos Estados Unidos levantou grandes suspeitas sobre as transações da empresa.

Uma série de investigações realizadas pelo Congresso americano e por órgãos reguladores chegaram ao ponto máximo quando foi anunciado que, além das investigações financeiras, uma investigação criminal seria instalada. Uma vez levantada a possibilidade de que altos executivos da companhia esteavam envolvidos em fraudes.

Com o objetivo de maquiar o balanço da companhia, foi usado um complexo sistema de parcerias financeiras para esconder prejuízos. Além disso, vários executivos da Enron supostamente tiveram grandes lucros vendendo suas ações antes que elas despecassem. Por outro lado, os 20 mil empregados da empresa, porém, perderam bilhões de dólares porque foram impedidos pela direção da companhia de vender suas ações quando elas começaram a cair.

Estudo de caso com machine learning

Neste projeto, utilizamos informações financeiras e de e-mail de 146 executivos para identificar a pessoa de interesse (POI) na fraude da Enron. O POI é referido como alguém envolvido em caso da Enron, foi indiciado por fraude, resolvido com o governo ou testemunhado em troca de imunidade.

```
Shape: (146, 21)

Number of POI in the dataset: 18

Number of non-POI in the dataset: 128
```

Dos 146 executivos, 18 são POI e 128 não POI, conforme tela acima.

Remoção de Outlier

A partir do gráfico de dispersão de salário e bônus e de salário e exercício de opções, podemos ver que há um ponto no conjunto de dados que tem um valor muito maior. Este ponto acaba por ser o “TOTAL”, o que é considerado um outlier uma vez que não é uma pessoa real. Também foi retirado a 'THE TRAVEL AGENCY IN THE PARK' por ser uma agência de viagem. Já o 'LOCKHART EUGENE E' foi retirado por ter muitos dados discrepantes.

Depois de remover os dados, ainda existem vários pontos de dados que têm um valor muito maior do que o restante, conforme se verifica nas figuras 1a, 1b, 2a e 2b. Esses dados representam pessoas reais no caso da Enron. Alguns deles são mesmo POI e, portanto, não serão considerados outliers.

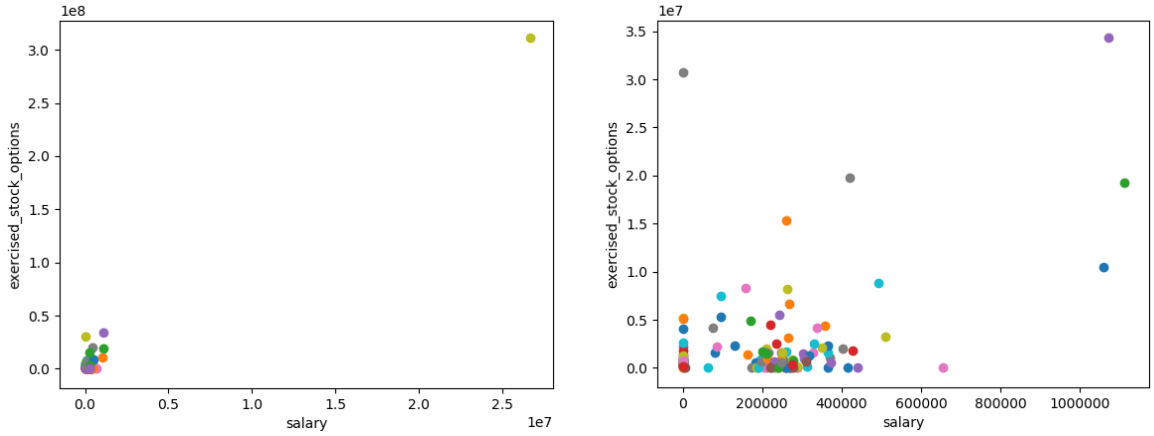
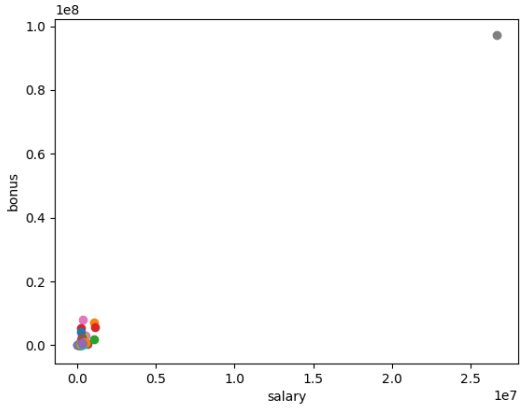


Figura 1a e 1b - Salário versus Opções de ações exercidas antes e depois da remoção de outliers



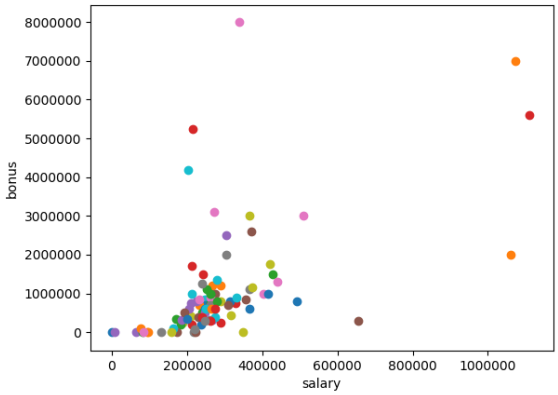


Figura 2a e 2b - Salário versus Bônus

Classificação dos recursos

Os dados podem ser classificados em duas classes, dados financeiros (salário, bônus, valor do estoque, etc.) e dados de e-mail (número total de e para e-mail e número de e para e-mails relacionados ao POI).

Describe:

:	count	unique	top	freq
salary	146	95	NaN	51
to_messages	146	87	NaN	60
deferral_payments	146	40	NaN	107
total_payments	146	126	NaN	21
exercised_stock_options	146	102	NaN	44
bonus	146	42	NaN	64
restricted_stock	146	98	NaN	36
shared_receipt_with_poi	146	84	NaN	60
restricted_stock_deferred	146	19	NaN	128
total_stock_value	146	125	NaN	20
expenses	146	95	NaN	51
loan_advances	146	5	NaN	142
from_messages	146	65	NaN	60
other	146	93	NaN	53
from_this_person_to_poi	146	42	NaN	60
poi	146	2	False	128
director_fees	146	18	NaN	129
deferred_income	146	45	NaN	97
long_term_incentive	146	53	NaN	80
email_address	146	112	NaN	35
from_poi_to_this_person	146	58	NaN	60

Esses recursos serão redimensionados ou recombinados para criar novos recursos para que eles sejam mais eficazes para distinguir os POIs dos não POIs.

Criação de novos atributos (novas features)

Não estamos interessados apenas no valor absoluto da renda, mas também na composição de sua renda. A suposição é que POI e não POI pode ter composição diferente de sua renda. Uma maneira é olhar para a relação entre suas salário e bônus ou ainda relação salário e opção de ações exercidas. O POI pode receber mais bônus em comparação ao salário ou realizar grandes transações de ações.

Os outros tipos de recursos são dados de e-mail. Em vez de olhar para o número total de e-mail de ou para POI, estamos mais interessados na parte do e-mail que envolve o POI.

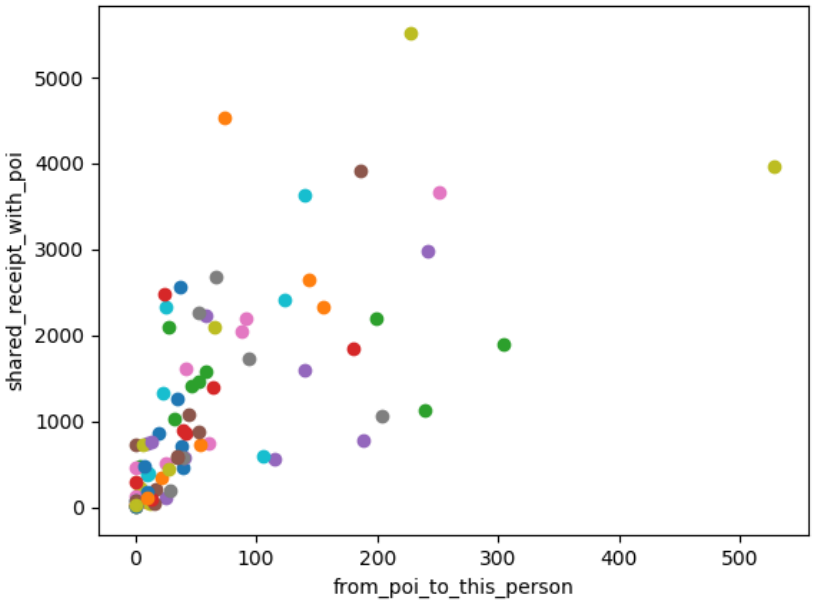


Figura 3 - e-mail enviados por um POI versus recebido de um POI

A suposição é que é um envolvido em fraude são mais propensos a se comunicar mais frequentemente com POI. A relação de e-mail enviados por um POI ou recebido de um POI pode ajudar a explicar a comunicação entre os envolvidos no crime.

Foram implementadas novas features, as quais foram denominadas como fraction_from_poi que consiste na fração de e-mail envidadas por um POI a uma pessoa em relação ao total de mensagens enviadas e fraction_to_poi que consiste na fração de e-mail por uma pessoa para um POI em relação ao total de mensagens enviadas, conforme segue código.

```
my_dataset = data_dict

def compute_fraction(poi_messages, all_messages):
    """ return fraction of messages from/to that person to/from POI"""
    if poi_messages == 'NaN' or all_messages == 'NaN':
        return 0.
    fraction = poi_messages / all_messages
    return fraction

for name in my_dataset:
    data_point = my_dataset[name]
    from_poi_to_this_person = data_point["from_poi_to_this_person"]
    to_messages = data_point["to_messages"]
    fraction_from_poi = compute_fraction(from_poi_to_this_person, to_messages)
    data_point["fraction_from_poi"] = fraction_from_poi
    from_this_person_to_poi = data_point["from_this_person_to_poi"]
    from_messages = data_point["from_messages"]
    fraction_to_poi = compute_fraction(from_this_person_to_poi, from_messages)
    data_point["fraction_to_poi"] = fraction_to_poi
```

Com a implementação das novas features o banco de dados passou a ser constituído pelos seguintes atributos:

Posição no Array	Atributo	Tipo	Features/Labels
0	Bônus	Int	Feature
1	deferral_payments	Int	Feature
2	deferred_income	Int	Feature
3	director_fees	Int	Feature
4	email_address	object	Feature
5	exercised_stock_options	Int	Feature
6	expenses	Int	Feature
7	fraction_from_poi	Float	Feature
8	fraction_to_poi	float	Feature
9	from_messages	Int	Feature
10	from_poi_to_this_person	Int	Feature
11	from_this_person_to_poi	Int	Feature
12	loan_advances	Int	Feature
13	long_term_incentive	Int	Feature
14	Other	Int	Feature
15	poi	bool	Label
16	restricted_stock	Int	Feature
17	restricted_stock_deferred	Int	Feature
18	salary	Int	Feature
19	shared_receipt_with_poi	Int	Feature
20	to_messages	Int	Feature
21	total_payments	Int	Feature
22	total_stock_value	Int	Feature

Ajuste de escala das características

Para ser possível aplicar os algoritmos de validação foi necessário classificar os atributos em features e label e em seguida escalar utilizando-se do fit_transform para tonar os dados adequados a aprendizagem supervisionada. Segue código utilizado:

```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
labelencoder_features = LabelEncoder()

features = data_pd.iloc[:,[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,20,21,22]].values
labels = data_pd.iloc[:, 15].values

features[:,4] = labelencoder_features.fit_transform(features[:,4])

onehotencoder = OneHotEncoder(categorical_features=[4])

labelencoder_labels = LabelEncoder()
labels = labelencoder_labels.fit_transform(labels)

import warnings
warnings.filterwarnings("ignore")
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
features = scaler.fit_transform(features)
```

Seleção de Algoritmo

Inicialmente criamos uma árvore de decisão, a qual consiste em ferramenta de suporte à tomada de decisão que usa um gráfico no formato de árvore e demonstra visualmente as condições e as probabilidades para se chegar a resultados. A representação visual da árvore pertence ao grupo de aprendizado de máquina supervisionado, e funciona tanto para regressão quanto para classificação.

Como parâmetro foi utilizado a entropia, a qual trata-se de uma forma de medir a probabilidade de obter um elemento positivo (ocorrência do evento) a partir de uma seleção aleatória do subconjunto de dados e cujo valores possíveis estão entre 0 e 1.

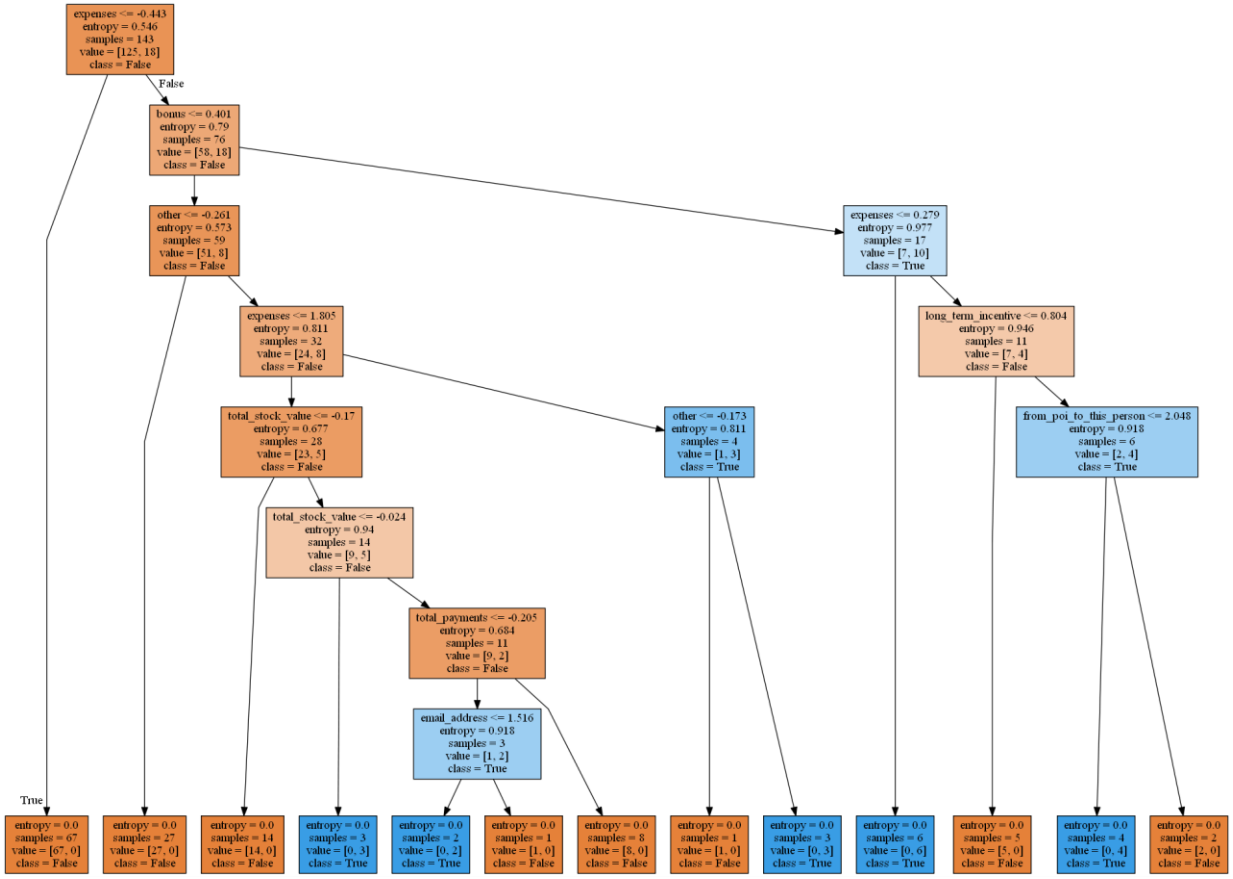


Figura 4 – Árvore de decisão

Em seguida cinco algoritmos foram comparados com seu desempenho, cada um representando um tipo diferente de algoritmo, os quais a seguir descrevemos:

- Arvore de Decisão (Decision Tree): um algoritmo não-paramétrico;
- Classificador de Floresta Aleatório (Random Forest Classifier): é um meta estimador que se ajusta a vários classificadores de árvore de decisão em várias sub-amostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo.
- Naive Bayes é um método paramétrico que assume que as distribuições dos recursos sãonormais. A suposição é que os recursos do POI e do não-IPO terá diferentes distribuição (média e covariância). O algoritmo tem essa vantagem quando o os recursos estão em altas dimensões. No entanto, a suposição subjacente para Naive Bayes é que os recursos são normalmente distribuídos. Então, os recursos financeiros foram redimensionados para transferir para a distribuição normal.

- Naive Bayes (um algoritmo paramétrico), árvore de decisão (um algoritmo não-paramétrico) algoritmo), e floresta aleatória (um método conjunto).
- SVC é um algoritmo de classificação de vetores de suporte, é um conceito na ciência da computação para um conjunto de métodos do aprendizado supervisionado que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão.
- KNeighbors Classifier é um algoritmo de reconhecimento de padrões, é um método não paramétrico usado para classificação e regressão. Em ambos os casos, a entrada consiste nos k exemplos de treinamento mais próximos no espaço de recursos.

Na tabela a seguir estão demonstradas a acurácia, o recall, a matriz de confusão e os parâmetros de cada um dos algoritmos utilizados.

Algoritmo	**Acurácia**	**Recall**	**Matriz de Confusão**	**Parâmetros**
Decision Tree	0,86	0,20	$\begin{bmatrix} 36 & 1 \\ 5 & 1 \end{bmatrix}$	N/A
Random Forest Classifier	0,86	0,14	$\begin{bmatrix} 37 & 0 \\ 6 & 0 \end{bmatrix}$	n_estimators=10, criterion='entropy', random_state=0
Naive Bayes	0,40	0,16	$\begin{bmatrix} 12 & 25 \\ 1 & 5 \end{bmatrix}$	N/A
SVC	0,86	0,43	$\begin{bmatrix} 37 & 0 \\ 6 & 0 \end{bmatrix}$	kernel = 'rbf', random_state = 1, C = 0.2
KNeighbors Classifier	0,86	0,14	$\begin{bmatrix} 37 & 0 \\ 6 & 0 \end{bmatrix}$	n_neighbors=5, metric='minkowski', p = 2

Discussão

Quatro algoritmos tiveram uma acurácia de 0,86, mas somente o SVC teve um recall que atende as especificações de superior a 0,3. Quanto à matriz de confusão, os testes SVC, Random Forest Classifier e KNeighbors Classifier tiveram desempenho iguais, com acerto de todos os não POI, porém considerou erroneamente como não POI os 6 POI. Assim, baseando-se na matriz de confusão, conforme modelo teórico abaixo demonstrada, o algoritmo decision tree foi capaz de acertar 37 dos 37 não POI (97,3%) e 1 de 6 POI (16,7%) do subconjunto de testes.

É não POI e o modelo retornou não POI	É não POI porém o modelo retornou como um POI
É um POI porém o modelo retornou não POI	É um POI e o modelo retornou como um POI

Destaca-se ainda, que apesar da baixa acurácia do Naive Bayes, foi o modelo como melhor detecção de POI, mas por outro lado coloca muitos não POI como um deles. Tanto é que o resultado de identificação de POI foi de 83,3%, porém não POI nesse modelo caiu para 32,4%.

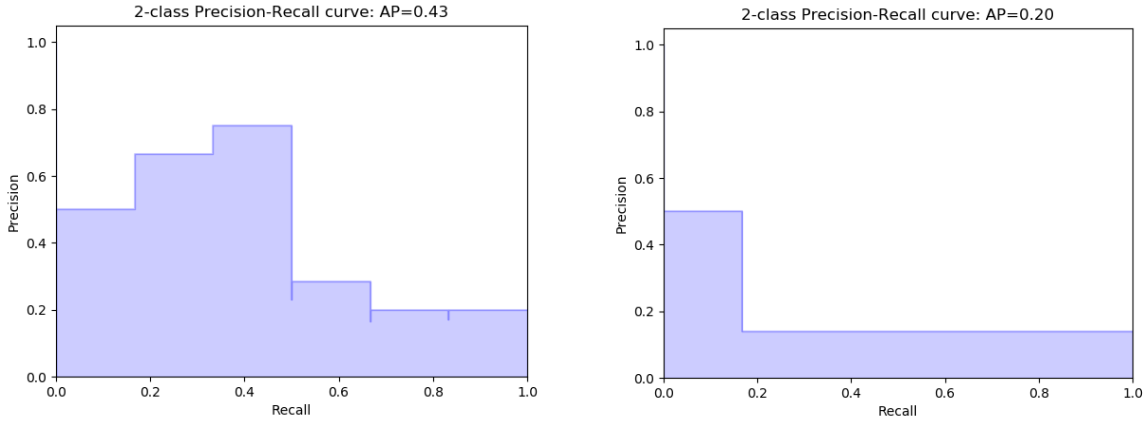


Figura 5a e 5b – Curva de precisão do recall do algoritmo SVC e Decision Tree

Conclusão

Neste projeto, usei técnicas de aprendizado de máquina para identificar o POI na fraude da Enron. Para isso, utilizei as informações financeiras e de e-mail disponíveis e criei outros atributos complementares.

Após a remoção dos outliers e transformação de escala dos dados, de forma a criar melhor uniformidade, apliquei cinco algoritmos (Decision Tree, Random Forest Classifier, Naive Bayes, SVC e KNeighbors Classifier). SVC e Decision Tree teve desempenho geral mais equilibrado.

Entre desafios do estudo, destaca-se que o conjunto de dados é pequeno e desequilibrado, uma vez que há apenas 18 POI em o total de 146 pessoas.

Bibliografia

https://www.wrprates.com/o-que-e-arvore-de-decisao-decision-tree-linguagem-r/

https://scikit-learn.org/stable/

https://www.bbc.com/portuguese/economia/020128_esp_eronqa.shtml