Detecção de transações digitais fraudulentas

Matheus Túlio Pereira da Cruz, Pedro Gabriel Gengo Lourenço, Vinicius Domingues Ribeiro

11053015, 11020615, 11025415

Introdução

Vivemos em um mundo onde a internet tem se popularizado cada vez mais. Hoje, por exemplo, existem mais de 5 bilhões de *smartphones* no mundo, onde é possível fazer de tudo: conversar com os amigos, marcar reuniões, realizar pagamentos e transferências bancárias, etc. Tudo isso para nos dar conforto e poupar-nos tempo.

Contudo, essa popularização tem atraído muitos criminosos e tem nos deixado vulneráveis caso cometamos algum deslize de segurança. Ao dar algum tipo de brecha, como abrir algum arquivo de fonte desconhecida, facilmente um *cracker* conseguirá utilizar os dados de outras pessoas para cometer crimes, como realizar transações bancárias. Com isso, devido aos seguros para casos assim, as empresas do setor financeiro tem tido perdas financeiras na casa de bilhões.

Segundo a Federação Brasileira de bancos (FEBRABAN), em 2016, Cerca de 9,5 milhões de correntistas fazem mais de 80% de suas transações pela internet ou pelos aplicativos dos bancos. Com o aumento do acesso, crescem também as reclamações no Banco Central (BC). Em 2016, foram 425 queixas referentes ao sigilo e à segurança dos canais eletrônicos, contra 1.688 no ano posterior. O aumento é de 297% de 2016 para 2017.

Sendo assim, o intuito deste trabalho é se utilizar de algoritmos de aprendizado de máquina para prever de fraudes em transações bancárias. Para isso, iremos gerar nossa própria base de dados com informações dos clientes e suas transações. Iremos utilizar a linguagem de programação Python3 com as bibliotecas:

- sklearn;
- numpy;
- pydotplus.

Base de dados

Optamos pela criação da Base de dados que vamos utilizar ao longo do projeto e para cada atributo escolhido foi utilizado uma determinada distribuição de probabilidade, foram elas:

- **Renda:** Foi utilizado uma distribuição normal, pois, segundo dados do IBGE, a média dos brasileiros que receberam algum rendimento, possuem uma renda média de R$2112. Pode-se concluir que a distribuição se comporta na forma de uma Gaussiana, portanto a distribuição normal é a melhor opção. Além disso, excluímos da função, as pessoas com rendimento abaixo de R$700, pois temos a hipótese de que essas pessoas não realizaram compras pela internet.
- **Valor da compra:** Foi utilizado uma distribuição normal, pois, segundo estudos realizados pela Kantar IBOPE Media, o brasileiro gasta por volta de R$661 a cada 3 meses em compras feitas na internet, portanto, de forma similar à renda, percebe-se que a distribuição da função acaba tendo o comportamento de uma Gaussiana, fazendo com que a distribuição normal seja a melhor distribuição que represente o comportamento do atributo.

- **Média de valor gasto nos últimos 3 meses:** Como o atributo é muito similar aos dois últimos atributos citados, podemos inferir que o comportamento da distribuição desse atributo também seja melhor representado por uma distribuição normal.
- **Categoria do produto da compra:** Utilizamos uma distribuição ponderada através de pesquisas feitas, descobrimos as categorias que mais eram compradas e demos pesos maiores na probabilidade de serem escolhidas na hora que fizemos uma escolha randômica.
- **Compra parcelada:** Definimos que, para ter uma fraude, o criminoso não faria uma compra parcelada, portanto, toda compra que é uma fraude, não é uma compra parcelada.
- **Horário da Compra:** Nós temos a hipótese de que o criminoso faria uma compra fraudulenta em períodos em que mais houvesse compras, para que seja algo não muito suspeito. Segundo pesquisa divulgada pela empresa Rakuten, o horário de pico das compras online é no período entre 12h e 14h. Portanto definimos nossa distribuição como normal, onde o pico da gaussiana era nesse espaço de tempo.

Com base nesses campos, nosso grupo avalia que os que serão mais importantes para o aprendizado serão: **Renda, Valor da compra, Média de valor gasto nos últimos 3 meses, Compra parcelada.** Ou seja, serão os campos preditores para o nosso algoritmo classificar a transação. Isso se dá pois a relação entre esses campos é muito forte e deve manter um padrão, assim, uma disparidade entre essas relações pode caracterizar fraude, por exemplo: se o valor de uma transação for cinco vezes maior que a renda ou que a média das transações nos 3 últimos meses, provavelmente é uma transação fraudulenta.

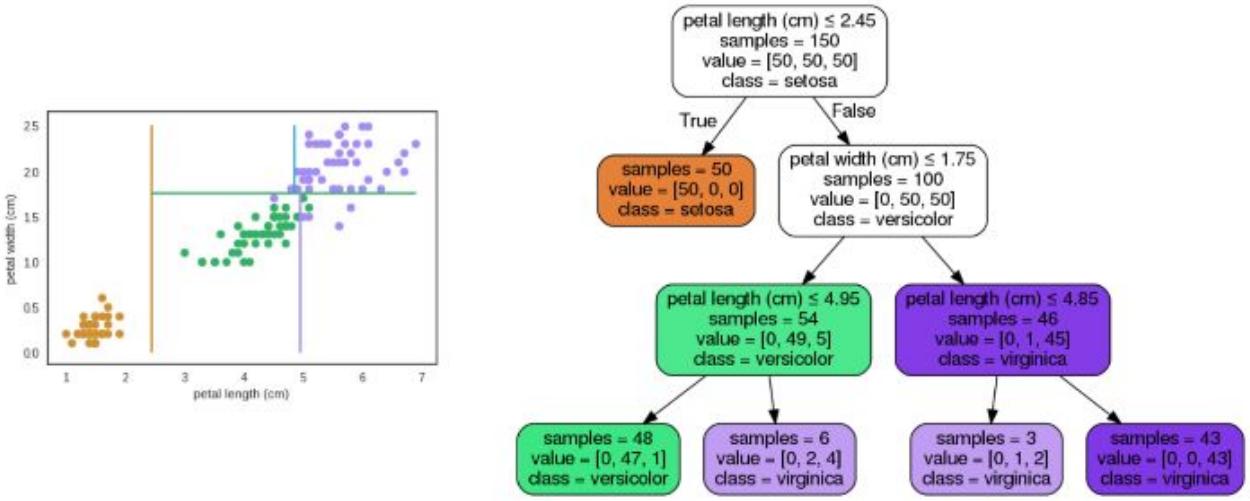
Algoritmo

Dado o problema, temos por objetivo predizer, para outros dados, se uma transação ocorrida é fraudulenta ou não. Nesse sentido, temos um problema onde o que queremos predizer são valores discretos, ou seja, é fraudulenta ou não, 1 ou 0. Sendo assim, devemos fazer uso de um algoritmo de classificação.

Além disso, ainda teremos que comparar o que o modelo aprendeu com nossa hipótese de aprendizado. Portanto, o ideal é que o algoritmo escolhido seja um de alta interpretabilidade, ou seja, que seja possível visualizar e entender quais foram os critérios que ele usou ou as variáveis que ele deu mais peso para a classificação.

Nesse contexto, decidimos que iremos usar árvores de decisão (figura 1), as quais são um método de aprendizado de máquina que estratifica o espaço de preditores em “retângulos”, como pode ser visto na figura 1, tendo em vista sempre a maximização da pureza das sub-regiões geradas. Tal pureza pode ser pode ser medida pela entropia ou pelo índice de Gini, no caso de problemas de classificação.

Figura 1 - À direita, exemplo do particionamento do espaço de preditores. À esquerda, exemplo de árvore de decisão



Disponível em: <https://medium.com/machine-learning-beyond-deep-learning>. Acessado em: 23/03/2019

Resumidamente, a árvore de decisão funciona da seguinte forma: dado o espaço de preditores, para cada um dos atributos calculamos o ganho de informação, que está intimamente relacionado com a pureza, ao usá-lo para dividir os dados de treinamento. Depois, selecionamos o atributo que nos dá o maior ganho. Em seguida, seguimos fazendo isso para as sub-regiões. Ao final, teremos a árvore de decisões com os critérios já definidos nos nós e com as classificações nas folhas. Um ponto de atenção deve ser a escolha de um critério de parada, ou seja, um valor de aumento do ganho de informação que já não representa um ganho muito expressivo, pois assim, tornamos nosso modelo menos suscetível ao *overfitting*.

Considerações Finais

Como considerações finais, é importante comentar a respeito das discussões que tivemos em relação a escolha do problema e a geração da base de dados. Um dos pontos bastante discutidos foi a escolha de um problema real que tivesse casos que se sobrepunham, pois dentre várias ideias, a maioria poderia ser facilmente resolvida com regras simples, uma vez que era bem claro a diferença dos atributos entre as classes propostas. Sendo assim, passamos a pensar em problemas que, para nós humanos, seria difícil encontrar com clareza e facilidade quais seriam os atributos mais importantes e os critérios para predição da classe.

Em relação a distribuição estatística dos dados da base gerada, sentimos certa dificuldade na hora de escolher qual tipo de distribuição usar, talvez pela falta de experiência e pelo costume de sempre usar bases já prontas. Contudo, conseguimos contornar esse problema utilizando como base o documento disponibilizado no Piazza onde havia um “kit iniciante” e um exemplo de geração de base.

Um dos receios que temos no projeto é que os dados gerados sejam sendo melhor modelados por algum outro tipo de classificador, e que o algoritmo que escolhemos não tenha uma performance boa. Entretanto, conseguiremos descobrir mais adiante, quando começarmos a treinar o modelo e analisar os resultados.

Referências

DELOITE, "Pesquisa FEBRABAN de Tecnologia Bancária 2018." Disponível em:<http://www.ciab.org.br/download/researches/research-2018.pdf>. Acessado em: 23/03/2019.

JAMES, Gareth et al. Tree-Based Methods: The Basics of Decision Trees. In: JAMES, Gareth et al. An Introduction to Statistical Learning. New York: Springer, 2013. p. 303-315.

CAMPOS, Raphael. Árvores de Decisão. Disponível em: <https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69>. Acessado em: 23/03/2019.

10% da população concentrava43,3% da renda do país em 2017, diz IBGE. UOL, 2018. Disponível em: <https://economia.uol.com.br/noticias/redacao/2018/04/11/concentracao-renda-ibge.htm>. Acessado em: 24/03/2019.

A cada 3 meses, o brasileiro gasta cerca de R$661 em compras pela internet. Terra, 2019. Disponível em: <https://www.terra.com.br/noticias/dino/a-cada-3-meses-o-brasileiro-gasta-cerca-de-r661-em-compras-pela-internet,70f7a51ef268f3648d98bec463372741o1lsr05w.html>. Acessado em: 24/03/2019.

Brasil tem horário de pico de compras online das 12h às 14, diz pesquisa.E-Commerce News, 2012..Disponível em: <https://ecommercenews.com.br/noticias/pesquisas-noticias/brasil-tem-horario-de-pico-de-compras-online-das-12h-as-14-diz-pesquisa/>. Acessado em: 24/03/2019.