- Melhor compreensão do problema escolhido (Qual o impacto de uma fraude não detectada? Quem sofre com isso? Qual a dor gerada? Quais consequências acontecem após a descoberta? );

Qual o impacto de uma fraude não detectada?

A fraude não detectada tem sozinha o impacto do valor $ da transação, mas quando não detectada indica um caminho para que o fraudador possa realizar inúmeras transações similares.

Quem sofre com isso?

Geralmente a empresa responsável pela transação é quem sofre. Este mercado funciona da seguinte forma: a empresa recebe um percentual em $ das transações que gerencia e as fraudes são arcadas pela mesma, portanto, já é esperado um percentual de fraudes no valor cobrado da pela operação, mas quanto menor o índice de fraudes consumadas melhor para a empresa gerando maior lucro, quando o índice sobre ela pode acabar por empatar ou até mesmo ficar no prejuízo.

5 personas

Cliente – Estabelecimento – (Subadiquirente – Adquirente) – banco – bandeira

Dependendo do nível de risco adotado pelo estabelecimento, aumentando o risco assumido aumenta o número de transações e consequentemente o índice de fraudes. Onde o estabelecimento e a adquirente estabelecem quem fica responsável pelos prejuízos.

Qual a dor gerada?

Caso o índice de fraudes consumadas fica acima do planejado a empresa tem prejuízo – dor para o adquirente.

Transações válidas identificadas como fraude – dor para o estabelecimento e cliente final.

Quais consequências acontecem após a descoberta?

A transação é negada e a operadora do cartão recebe um aviso de possível fraude identificando o cartão, estabelecimento, valor etc.

- Pesquisar o que está sendo feito hoje, no mercado, indústria e acadêmia. (O que a indústria utiliza hoje? É um modelo simples ou complexo?);

Pelo podcast sugerido pela Sandra entendo que a indústria começou com técnicas simples, mas hoje em dia já contam com ferramentas bastante complexas.

Podemos investigar mais detalhes

- Se possível, uma análise de dados, trazendo suas contribuições. (Análise estatísticas descritivas, gráficos e o que achar interessante).

-A base de dados foi coletada em setembro de 2013

-Possui 284.807 transações

-Dentre as transações 492 são identificadas como sendo fraude –

A base de dados é desbalanceada - somente 0,172% das amostras são da classe (1) fraude

-A base de dados contem dados somente numéricos

-As features de V1 a V28 são resultado de uma PCA ( principal component analysis) e não temos a indicação do que representam somente que são as variáveis mais

significativas segundo quem fez a análise com os dados reais. No kaggle alegam que foi disponibilizada desta forma para preservar a identidade/privacidade dos usuários de cartão.

-As features time e amount não foram incluídas na PCA

-A feature time mostra a diferença em segundos da operação em relação a primeira operação da base de dados.

-A feature amount é a quantida em $ transacionada na amostra

-A feature class indica 1 em caso de fraude e 0 em caso de transação validada ( sem fraude)

-No kaggle recomendam a utilização da área abaixo da curva precisão-recall para avaliação da classificação, alegando que a matriz de confusão não tem significado para base de dados desbalanceadas.

Iniciei algumas verificações superficiais na base de dados e me parece bastante integra

Criar uma visualização de dados balanceados com undersampling para avaliar como que os dados de fraude ficam agrupados.

Pesquisar que mais técnicas podemos usar para este caso de classificação de dados desbalanciados