1. O PROBLEMA:

O problema é a detecção de fraudes em transações financeiras. Em muitos setores financeiros, como bancos, cartões de crédito e serviços de pagamento, seriam de grande valia identificar atividades fraudulentas para evitar prejuízos e proteger os clientes.

Machine learning pode ser aplicado utilizando os dados financeiros e os separando por grupos, onde assim seria possível detectar fraude, pois devido à quantidade baixa de dados rotulado disponível, pode-se recorrer ao aprendizado não supervisionado.

Nessa abordagem, algoritmos como K-Means, Isolation Forest e DBSCAN podem ser aplicados para agrupar os dados e identificar grupos incomuns que podem representar transações fraudulentas.

Os usuários que se importam seriam: Instituições Financeiras, Consumidores dessas instituições, Agências Reguladoras e Empresas de Segurança Cibernética:

2. FORMATO CRISP-DM:

FASE 1: ENTENDIMENTO DO NEGÓCIO

Objetivo: Identificar fraudes em transações financeiras para proteger os clientes e minimizar perdas financeiras.

Contexto: Uma instituição financeira deseja melhorar sua capacidade de detecção de fraudes em suas operações. A ocorrência de atividades fraudulentas pode resultar em prejuízos financeiros e impactar negativamente a confiança dos clientes.

FASE 2: ENTENDIMENTO DOS DADOS

Fontes de Dados: Dados históricos de transações financeiras (valores, data, hora, tipo de transação, etc.).

Descrição dos Dados: Os dados contêm informações sobre várias transações financeiras realizadas pelos clientes. Não há rótulos explicitamente indicando quais transações são fraudulentas ou legítimas.

Objetivo dos Dados: Identificar padrões e características nas transações que possam ajudar a distinguir transações fraudulentas das legítimas.

FASE 3: PREPARAÇÃO DOS DADOS

Limpeza de Dados: Tratar valores ausentes e eliminar dados irrelevantes ou duplicados.

Transformação de Dados: Normalizar as variáveis para que elas estejam na mesma escala e prontas para a modelagem.

FASE 4: MODELAGEM

Técnica de Modelagem: Aprendizado Não Supervisionado (Isolation Forest ou K-Means) para detecção de anomalias.

Descrição do Modelo: Aplicar um algoritmo de aprendizado não supervisionado (Isolation Forest ou K-Means) para identificar grupos incomuns (anomalias) nos dados. Calcular os valores de anomalia para cada transação com base nas características do grupo ao qual ela pertence.

Avaliação do Modelo: Definir um limiar de anomalia para determinar quais transações são potenciais fraudes. Avaliar a capacidade do modelo em detectar fraudes através de métricas como precisão, recall e F1-score.

FASE 5: AVALIAÇÃO

Interpretação dos Resultados: Analisar as transações identificadas como potenciais fraudes e verificar a precisão das detecções.

Revisão do Processo: Ajustar parâmetros do modelo e técnicas de detecção de fraudes para melhorar o desempenho.

FASE 6: IMPLANTAÇÃO

Implementação do Modelo: Implantar o modelo de detecção de fraudes em um ambiente de produção para monitorar as transações em tempo real, utilizando dashboards.

Monitoramento: Monitorar o desempenho contínuo do modelo e realizar ajustes conforme necessário.

3. CONSIDERAÇÕES

O grande desafio seria lidar com o desbalanceamento dos dados, já que as transações fraudulentas geralmente representam uma pequena proporção em relação às transações legítimas.

O tratamento adequado desse desbalanceamento é crucial para evitar viés no modelo e garantir que ele possa detectar efetivamente as fraudes. Técnicas como oversampling (aumento da quantidade de amostras da classe minoritária) ou undersampling (redução da quantidade de amostras da classe majoritária) podem ser empregadas, assim como o uso de técnicas de aprendizado sem desbalanceamento, como aprendizado por transferência.

Outra decisão importante é a definição do limiar de anomalia, que determinará quais transações serão consideradas potenciais fraudes. Ajustar esse limiar é essencial para controlar a sensibilidade do modelo em relação aos falsos positivos e falsos negativos. Um limiar mais alto resultará em menos falsos positivos, mas pode aumentar o risco de falsos negativos, enquanto um limiar mais baixo terá o efeito oposto.