Bruno Layus
MVP Análise de Dados PUC Rio

Julho 2025

ÍNDICE

1. INTRODUÇÃO

1. Quais os desafios que este projeto pretende trabalhar?

Este Projeto é o produto principal da Sprint de Análise de Dados e Boas Práticas, segunda etapa do curso de Pós Graduação da PUC Rio sobre Data Science and Data Analytics. A proposta deste projeto é desenvolver uma Análise Exploratória de Dados, (EDA, em inglês), utilizando as ferramentas de programação python, visualização de dados e boas práticas de engenharia de software, disciplinas que foram trabalhadas ao longo desta Sprint.

Neste projeto escolhi fazer uma análise de detecção de fraudes em compras com cartão de crédito em um sistema de e-commerce. Estes dados compreendem um período de pouco mais de 1 mês entre março e abril de 2020 e relata transações de compra de produtos divididos em categorias, data e valor da compra e colunas com dados de score dos clientes, que foram provavelmente adquiridas de bureaus de crédito, junto com algumas outras informações.

Prevenção à fraude é um tema muito importante no universo de Machine Learning, pois algoritmos bem projetados e implementados são capazes de monitorar os milhões de transações diárias que acontecem no varejo físico e digital e identificar anomalias e transações fora do padrão, disparando alertas para barrar transações suspeitas ou liberando aquelas transações legítimas.

Em 2024, segundo dados da Serasa, foram feitas mais de 1.8 M de tentativas de fraudes no setor bancário e de cartões de crédito no Brasil, gerando um prejuízo estimado de R$ 3.5 bi de acordo com a ClearSale. O Brasil ocupa o segundo lugar no ranking dos países que mais sofrem com crimes de fraude, com um risco estimado em 14,25, atrás apenas da China, com um risco de 14,93. Ou seja, de cada 100 mil transações registradas, há um risco de que 14 a 15 dessas operações sejam tentativas de fraude. Agora, em 2025, o foco destas fraudes está sendo dividido entre fraudes a cartões de crédito e fraudes ao Pix. No dia anterior ao dia em que escrevo este texto, um ataque hacker conseguiu invadir o sistema do PIX e desviou um valor próximo a 1 bi do sistema bancário.

Estima-se que cerca de metade da população adulta do Brasil tenha sofrido alguma tentativa de fraude nos últimos cinco anos, enquanto em outros países com menor índice este número gira em torno de 10%. É importante distinguir a tentativa de fraude da efetivação dela. Em termos percentuais, apenas 2,5 % das tentativas de fraude são bem sucedidas e pouco a pouco observa-se uma redução nas fraudes, graças ao letramento digital da população, o que a torna menos suscetível a golpes e a mecanismos mais eficientes de controle e detecção de fraudes.

Olhando para a escala do país, há uma grande concentração das fraudes nos estados de São Paulo, mais de 40%, Rio de Janeiro, com 16% do total de tentativas de fraude, seguido por Minas Gerais e Paraná, com 8% e 6%, respectivamente.

2. ORIGEM E DICIONÁRIO DOS DADOS

1. Origem dos dados

Estes dados foram disponibilizados para um case prático do PED, da professora Renata Biaggi, preparatório para entrevistas de dados, programa do qual participei em 2023 e 2024. A Renata foi cientista de dados no Mercado livre e provavelmente conseguiu obter com eles esta pequena amostra de dados para estudo de fraudes em compras.

Este dataset contém 150 mil transações em um período de tempo compreendido entre março e abril de 2020, em pleno início da pandemia de Covid-19. Das 150 mil transações, 7.5 mil são fraudes, portanto 5% dos casos. Este número redondo de transações e de fraudes e a proporção perfeita (95, 5%), provavelmente indica que este dataset foi preparado e manipulado para esta finalidade de estudo, o que não diminui o sentido didático deste conjunto de dados.

Este conjunto de dados já apresenta uma coluna score_fraude_modelo que é uma probabilidade da compra ser fraudada dada por um modelo legado, já existente, porém que se mostrava pouco eficiente em discriminar corretamente as compras legítimas das compras fraudadas. O objetivo deste estudo é apresentar esta probabilidade do modelo legado é superar este benchmarking, esta linha base.

2. Dicionário de dados

Feature	Descrição	Tipo de Dado	Nulos
produto	Nome do Produto	Object	0
categoria_produto	Categoria do Produto	Object	0
data_compra	Data e hora da compra	Datetime	0
valor_compra	Valor em Reais da compra	Float	0
pais	País de origem da transação	Object	194
entrega_doc_1	Documentação tipo 1	Int / Boolean	0
entrega_doc_2	Documentação tipo 2	Int / Boolean	108857
entrega_doc_3	Documentação tipo 3	Int / Boolean	0
score_1 a 10	Colunas Score (bureau de crédito)	Int e Float	vários
score_fraude_modelo	Score Modelo Legado	Int	0
fraude	Label pós compra	Int / Boolean	0

3. HIPÓTESES E PERGUNTAS DE NEGÓCIO

1. Quais as perguntas que este projeto pretende responder?

- Qual o perfil das compras fraudadas? Elas ocorrem preferencialmente em quais horários, dias da semana, com quais valores e com quais tipos de produtos.

- Quais métricas podemos elaborar e utilizar para monitorar os prejuízos com fraudes?

- Quais as principais características deste dataset que mais se relacionam às compras legítimas e fraudadas?

- A partir destas informações, é possível esboçar um perfil do fraudador?

- Quais seriam as estratégias e quais modelos ideais de Machine Learning que poderiam contribuir para a diminuição de fraudes e como implantá-las?

- Qual seria o melhor modelo de Machine Learning para lidar com este problema de classificação? Seriam modelos do tipo ensemble?

- Estratégias de balanceamento dos dados do tipo Smote (oversampling, undersampling) seriam positivas para a modelagem?

4. TRATAMENTO INICIAL DOS DADOS

1. Tipos de dados

A primeira etapa após abrir o dataset é verificar o número de linhas e colunas e as principais informações dos dados, como data types (tipos de dados) e a existência de nulos. Uma primeira checagem indica que os dados estão relativamente bons para uso: os tipos não estão tão absurdos, aparentemente possuem poucos nulos e os nomes de colunas estão devidamente grafados com letras em caixa baixa e underscores entre palavras. Também constata-se que o tamanho do arquivo é de aproximadamente 23 MB de memória, o que o torna facilmente manipulável no Google Colab.

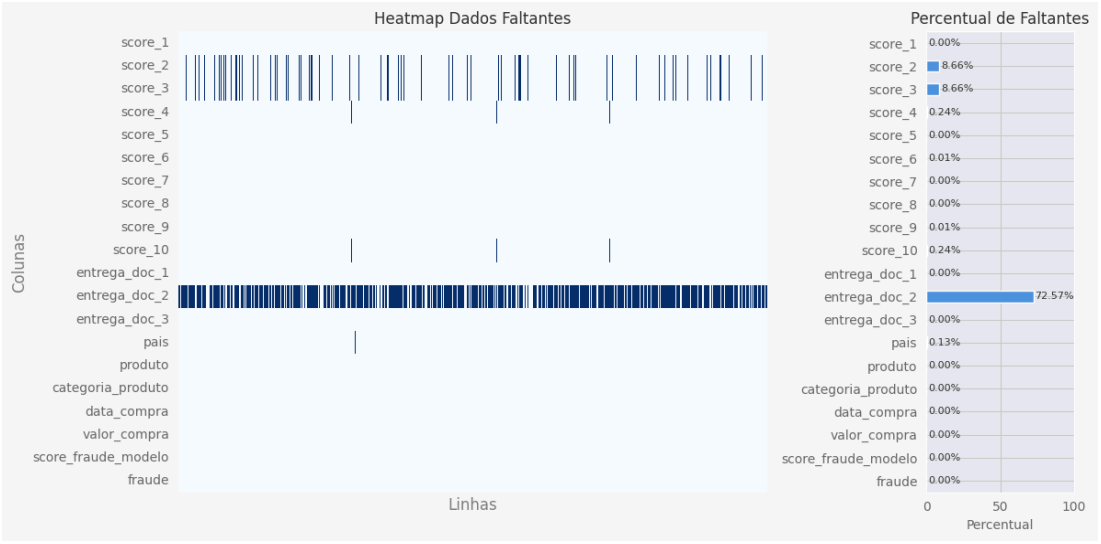
2. Nulos

Ainda através do método df.info(), é possível ver a contagem de nulos e o valor total de linhas de dados presentes no dataset. Sistemas de Machine Learning (ML), na maioria dos casos, não são capazes de lidar com dados faltantes (nulos), por isso a etapa de verificação e tratamento dos valores nulos e posterior imputação / preenchimento ou remoção de dados é fundamental para fazer um pré processamento eficiente, preparando o dataset para futuras etapas de modelagem de ML.

3. Ordenação das colunas

Ao abrir o dataset, optei por reordenar o index de colunas para deixá-las organizadas em blocos. O primeiro bloco de colunas são as colunas com scores. O segundo bloco são as colunas de entrega dos documentos. O terceiro bloco são as informações da compra e, por fim, as informações de fraude.

O gráfico abaixo mostra de uma forma visual e organizada onde estão distribuídos os dados faltantes ao longo de todo dataset:



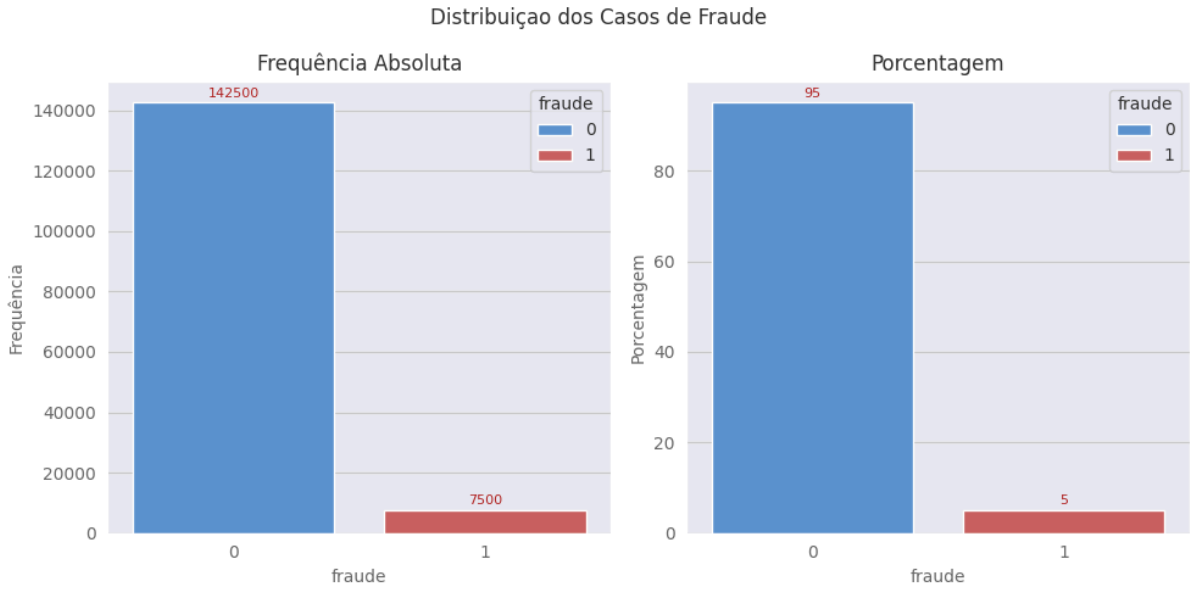
5. ANÁLISE DAS VARIÁVEIS

1. Distribuição da Variável Alvo: Fraude

A primeira avaliação deste conjunto de dados, que busca compreender a dinâmica das transações fraudadas e compará-las às transações legítimas, é justamente olhar para a quantidade absoluta e percentual de fraudes presentes neste dataset.

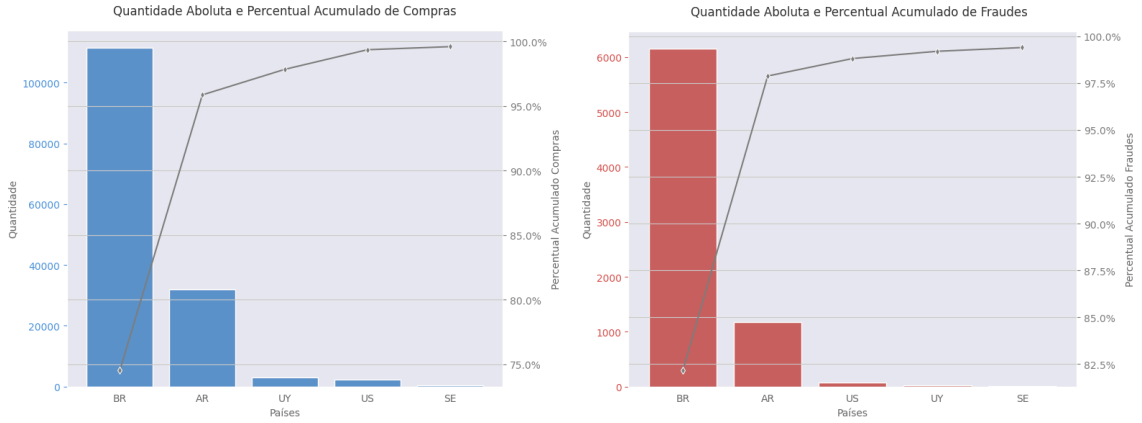
Esta amostra conta com 150 mil linhas de dados, das quais 7500 foram marcadas como fraudes, perfazendo 5% das transações. Apesar do evidente desbalanceamento, ainda assim, 5% é um percentual bastante elevado de fraudes.

Este problema é um clássico problema de classificação em Machine Learning, onde o objetivo é prever se uma transação é (1) ou não (0) fraudada. Este tipo de classificação é feita com base na probabilidade de pertencer à classe minoritária, onde o modelo de ML prevê um score para cada transação e acima de um threshold define se aquele caso é classificado em positivo ou negativo.



2. Análise da Variável País

A próxima variável de interesse é a coluna País. Sabemos que o Brasil é um dos campeões em fraudes, será que com estes dados iremos confirmar esta afirmação? Quantos países aparecem com mais frequência neste dataset e onde acontecem mais fraudes? Será que o país onde ocorre a transação é determinante para definirmos o risco de fraude?



No gráfico acima podemos ver, em azul, a quantidade absoluta de compras e, na linha cinza, a curva acumulada para cada país. Esta visualização é bem interessante pois podemos ver ao mesmo tempo quanto cada país representa em compras tanto em termos absolutos quanto em termos percentuais. Podemos ver no gráfico da esquerda que o Brasil concentra mais de 100 mil transações, o que equivale a aproximadamente 75% do total. Argentina é o segundo país com mais compras, com aproximadamente 30 mil transações, equivalente a mais de 20% do total. Os dois países, Brasil e Argentina, juntos somam mais de 95% de todas as transações. Ao restante dos países sobra uma participação bem pequena neste conjunto de dados.

No gráfico da direita, em vermelho vemos o total e o acumulado das transações fraudadas. De novo, o Brasil é o país que mais possui compras fraudadas, com mais de 6 mil transações, o que equivale a mais de 82%. Argentina também é o segundo país com mais fraudes, aproximadamente 1 mil transações, o que equivale a mais de 15% das fraudes. Os dois países juntos respondem por mais de 97.5% das transações fraudadas.

3. Avaliação da Cardinalidade das Colunas

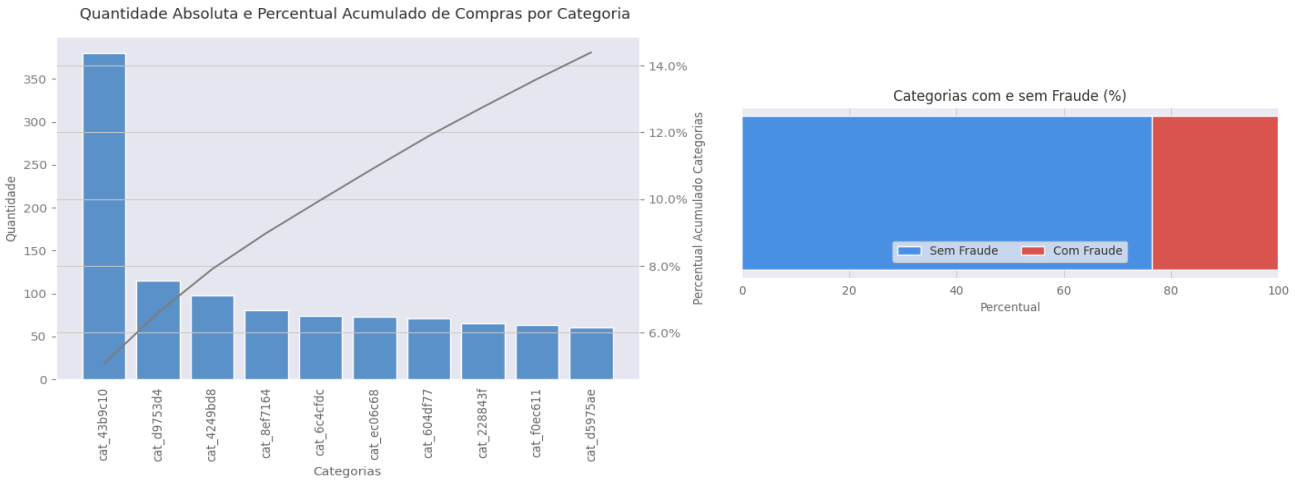
Analisar a cardinalidade, ou seja, quantos valores diferentes cada coluna apresenta, é uma etapa importante da análise de dados, pois é possível, através da cardinalidade, fazer algumas inferências a respeito das variáveis e quanto elas podem ou não contribuir para uma futura modelagem de ML.

Por exemplo, a coluna Score_1, aparentemente é uma coluna numérica, mas ao olharmos a cardinalidade dela, percebe-se que só há 4 valores distintos. Logo podemos afirmar que, na verdade, este tipo de dado é mais categórico do que numérico. Talvez as colunas Score_4 e Score_7, que possuem 51 e 59 valores distintos, respectivamente, também possam ser consideradas como categorias, apesar de serem colunas com valores numéricos discretos. Porém, como não temos mais nenhuma informação destas 2 últimas colunas score, talvez seja melhor mantê-las como estão.

Também vemos uma cardinalidade muito alta na coluna produto, com mais de 127 mil valores diferentes, o que reflete a grande diversidade de itens do varejo digital. A coluna

categoria_produto apresenta aproximadamente 8300 diferentes valores, mas será que há alguma relação entre fraudes e categorias? E relação entre fraude e tipo de produto?

4. Análise de Categorias e Produtos



Apesar de não sabermos o que estes códigos das categorias representam, podemos perceber que há algumas categorias com maior concentração de fraudes do que outras. Neste gráfico acima, podemos ver como as 10 categorias com mais fraudes concentram aproximadamente 15% do total e que a primeira categoria mais fraudada é visivelmente mais atingida por fraudes do que as outras. Vamos olhar esta categoria em detalhe e ver qual tipo de produtos ela agrega.

Também podemos constatar, através do gráfico da direita, que quase 80% das categorias não apresentam nenhuma ocorrência de fraude, enquanto aproximadamente 20% das categorias concentram todas as transações fraudadas deste conjunto de dados.

Ao filtrar os dados pela categoria mais fraudada e agrupar por produto, podemos ver claramente que há uma predileção por comprar aparelhos celulares em transações fraudadas. Isso parece bastante curioso e informativo para traçar o perfil do fraudador.

5. Análise do Valor da Compra

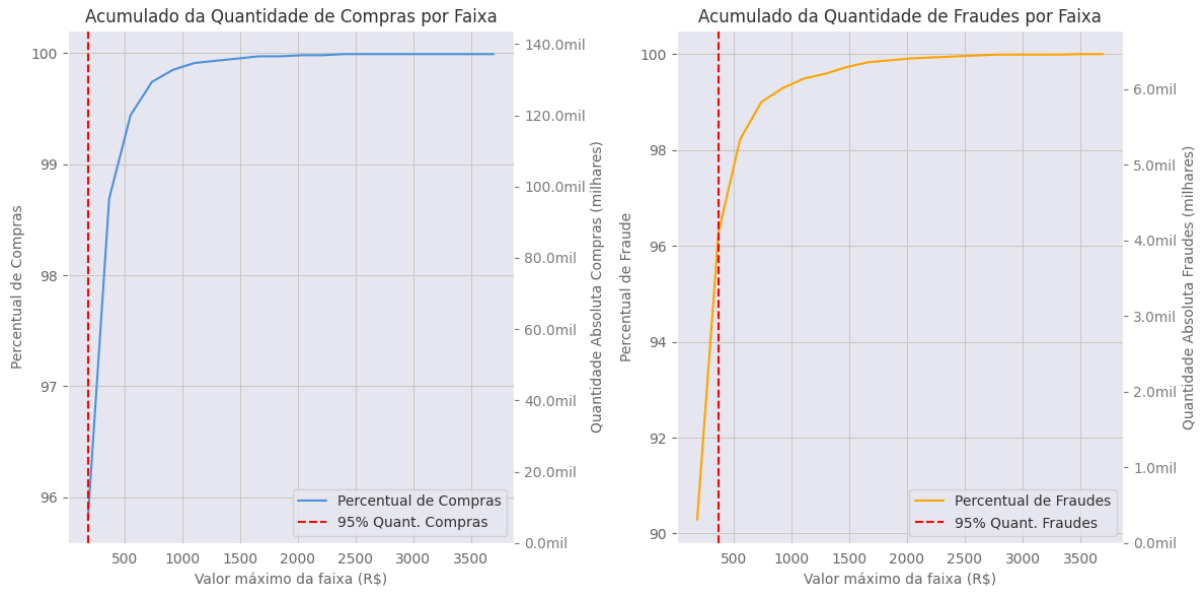
Na coluna valor da compra, muitas informações podem ser extraídas, já que este é uma variável de vital importância para todo comércio, seja ele físico ou digital. A companhia de cartões de crédito obtém um ganho de 10% do valor de cada compra legítima, porém perde o valor integral da transação se a compra for aprovada e se revelar uma fraude. É do valor das compras que a empresa obtém sua receita e os vendedores também, portanto este é o grande propósito de toda a operação.

Em primeiro lugar, podemos olhar para algumas métricas que resumem a situação das compras, tanto legítimas, quanto fraudadas. Depois veremos alguns gráficos mostrando as

faixas de valores mais compradas e mais fraudadas, por fim veremos em quais faixas de preço a companhia tem lucro ou prejuízo com as compras e fraudes.

Métrica	Legítimas	Fraudes
Soma do valor de compras	R$ 5.981.199,00	R$ 547.271,12
Valor médio das compras	R$ 41,97	R$ 72,97
Valor máximo das compras	R$ 3.696,35	R$ 3.424,81
Valor mínimo das compras	R$ 0,02	R$ 0,21
Faturamento / Prejuízo	R$ 598.119,90 (10% da transação)	R$ - 547.271,12

Da tabela acima, constatamos que, apesar da movimentação total no período ser de aproximadamente 6 MI e a companhia faturar 10% deste valor (aprox. 598 K), o prejuízo causado por fraudes foi de 547 K. Descontado o prejuízo causado por fraudes, o lucro da operação foi de apenas algo em torno de 50 K, menos de 1% do total movimentado. Esses valores oferecem uma perspectiva bem preocupante do tamanho do problema causado por fraudes.



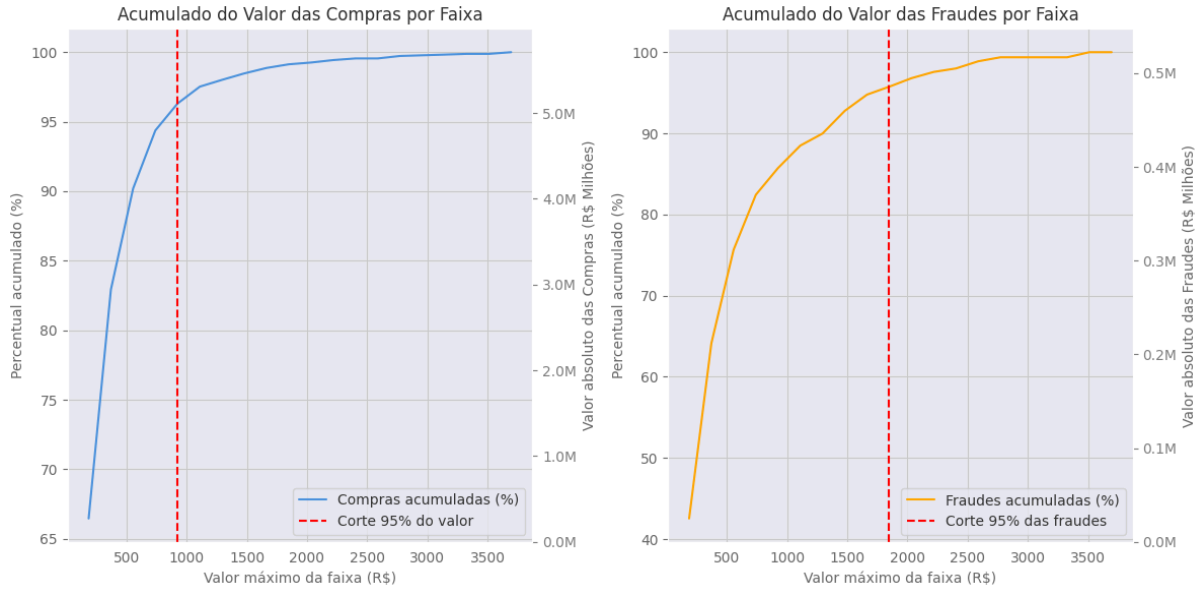
Este par de gráficos mostra a quantidade absoluta e percentual das compras legítimas e fraudadas em relação às faixas de valor em que ocorrem. Do gráfico à esquerda, reparar como a maioria das compras (a linha vermelha tracejada indica o corte em 95% das compras) acontecem na faixa de valores abaixo de R$ 200, 99% das compras acontecem abaixo de R$ 500. Apenas 1% de todas as transações são acima desta faixa de R$ 500, o que resulta em um ticket médio muito baixo.

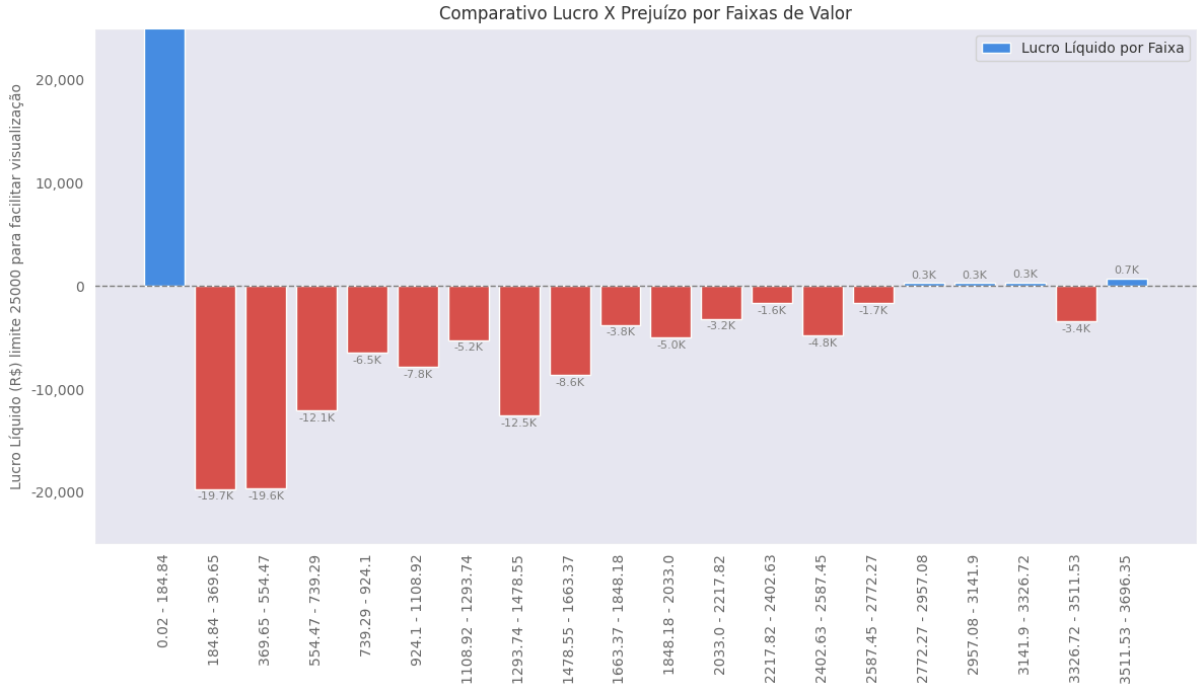
Pelo gráfico à direita, por outro lado, podemos ver como o valor de corte dos 95% das fraudes está em quase R$ 500 e 2% das fraudes ocorrem em valores acima destes R$ 500, o que indica um “ticket médio”das fraudes mais alto do que o das compras legítimas. Vamos

ver no próximo par de gráficos a mesma relação, só que desta vez pensando no valor total das transações.

Do gráfico à esquerda, podemos ver como dos quase R$ 6 MI em transações, os 95%, aproximadamente R$ 5 MI acumulados foram de compras feitas com valores abaixo de R$ 900, onde está a maior fatia do mercado e do faturamento.

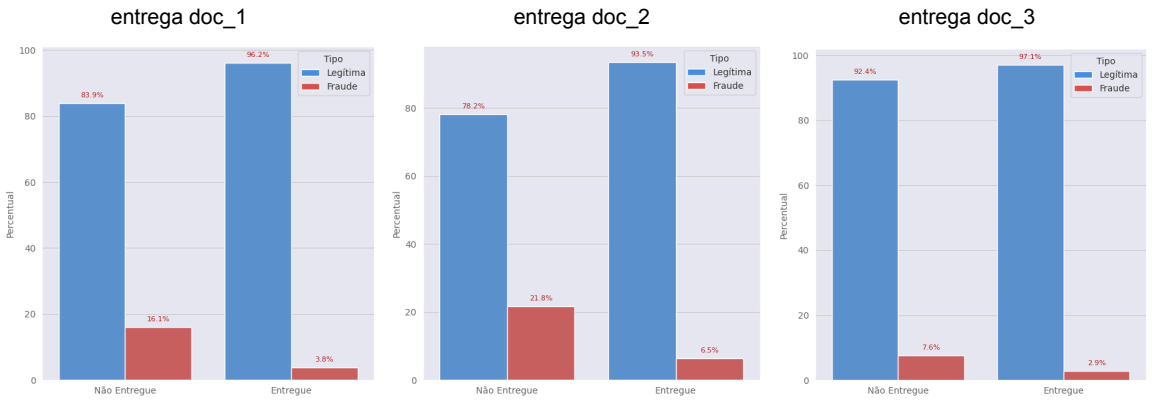
Do gráfico à direita, podemos ver como os 95% acumulados no valor total de fraudes acontece na faixa até R$ 2K. As fraudes se distribuem por uma faixa mais ampla de valores e também acontecem com mais frequência em faixas de valores mais altos.





Este gráfico acima elucida as faixas de preço onde a companhia obteve receita e as faixas de preço onde as perdas por fraude superaram as receitas.
Impressionante ver como em todas as faixas de preço, com exceção da primeira faixa, as transações fraudadas superaram as receitas. No cômputo geral, este período foi muito ruim para a empresa que ficou com uma fatia muito pequena do valor das transações, trabalhando com uma lucratividade muito baixa. Se tivéssemos os dados das despesas da empresa para manter a operação funcionando (aluguel, salários, impostos, etc), com certeza iríamos confirmar que a empresa estava, neste período, operando no prejuízo.

6. Análise da Coluna Entrega de Documento

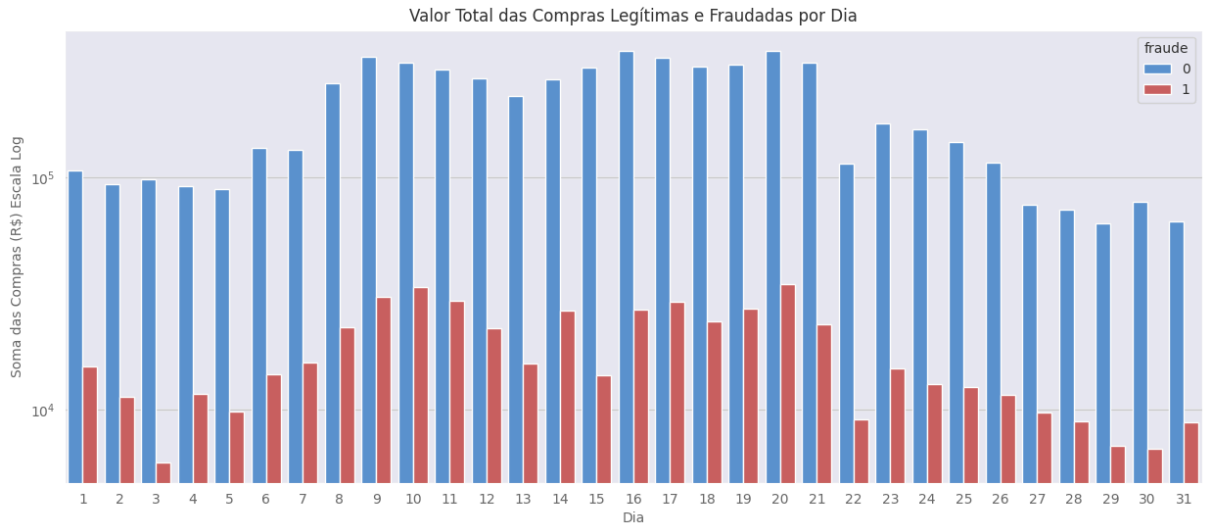


A entrega ou a falta de entrega dos documentos está associada ao cometimento de fraudes. Como podemos ver nas colunas da esquerda, quando os documentos não foram entregues há uma maior quantidade de fraudes cometidas e quando os documentos são entregues, a associação com fraudes diminui. Sobretudo o gráfico do meio, que mostra a entrega do documento 2, possui um percentual maior de transações fraudadas, com cerca de 20% delas.

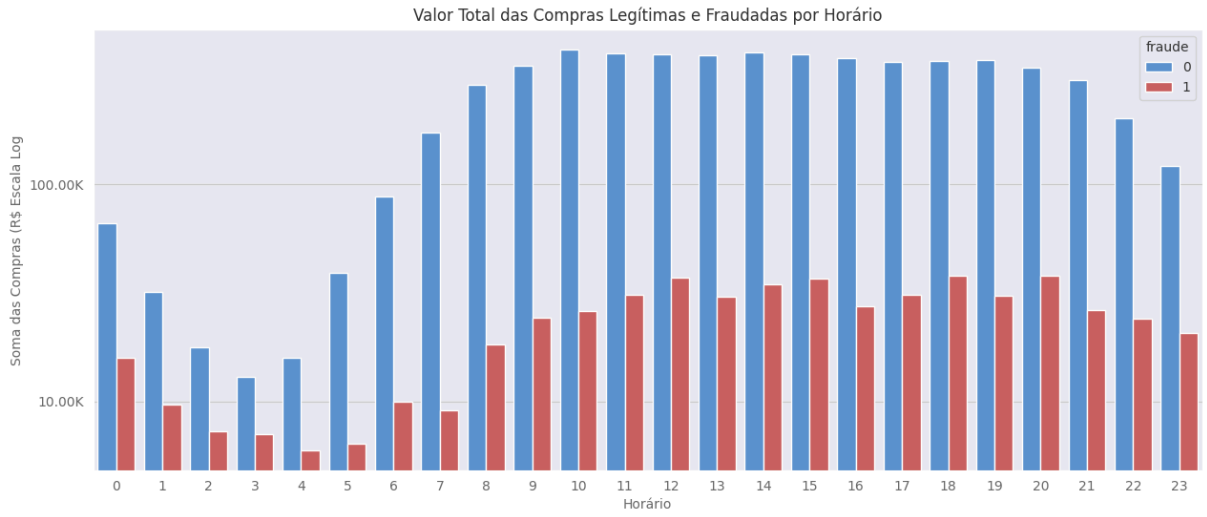
7. Análise da Data da Compra

A intenção de analisar esta coluna é entender se há alguma associação significativa entre o período do mês, dia ou horário do dia em que mais fraudes acontecem. Esta variável pode ajudar a construir o perfil do fraudador, porém é difícil definir com precisão se este perfil traçado pela análise irá se manter ao longo do tempo. O fraudador, como criminoso que é, está sempre buscando formas de disfarçar sua atividade ilícita e mudando suas estratégias de atuação para tentar passar despercebido. Desta forma, se nos basearmos muito em um perfil comportamental relacionado à variável temporal, por exemplo, bloqueando transações em determinado dia ou hora, rapidamente os fraudadores poderão perceber estas “regras”e alterar seu modo de atuação.

Na hora de preparar o modelo de classificação, a variável data não deverá compor o modelo, mas para fins de análise ela pode ser muito útil e trazer informações ricas para compor a interpretação deste tema complexo. Abaixo, vamos ver a distribuição por dia:



Este gráfico evidencia a distribuição das compras legítimas e fraudadas ao longo dos dias do mês, com o objetivo de identificar padrões e tendências que possam estar relacionados a maiores índices de fraudes. Primeiramente, salta aos olhos uma maior concentração de transações na segunda e terceira semanas do mês, com uma queda nos dias entre semanas, talvez final de semana, formando um padrão em “M”. As transações fraudadas acompanham esta tendência, mostrando uma óbvia correlação positiva entre um maior número de transações e um maior número de fraudes. Utilizei o teste estatístico de médias Chi-Quadrado para confirmar esta hipótese de que há uma relação estatisticamente significativa entre o dia do mês e a quantidade de fraudes. O valor chi2 foi de 176.07 e o p-value foi 1.33 e-22, um p-value muito baixo que não rejeita a hipótese nula, ou seja, há uma relação estatística entre dia do mês e fraudes.

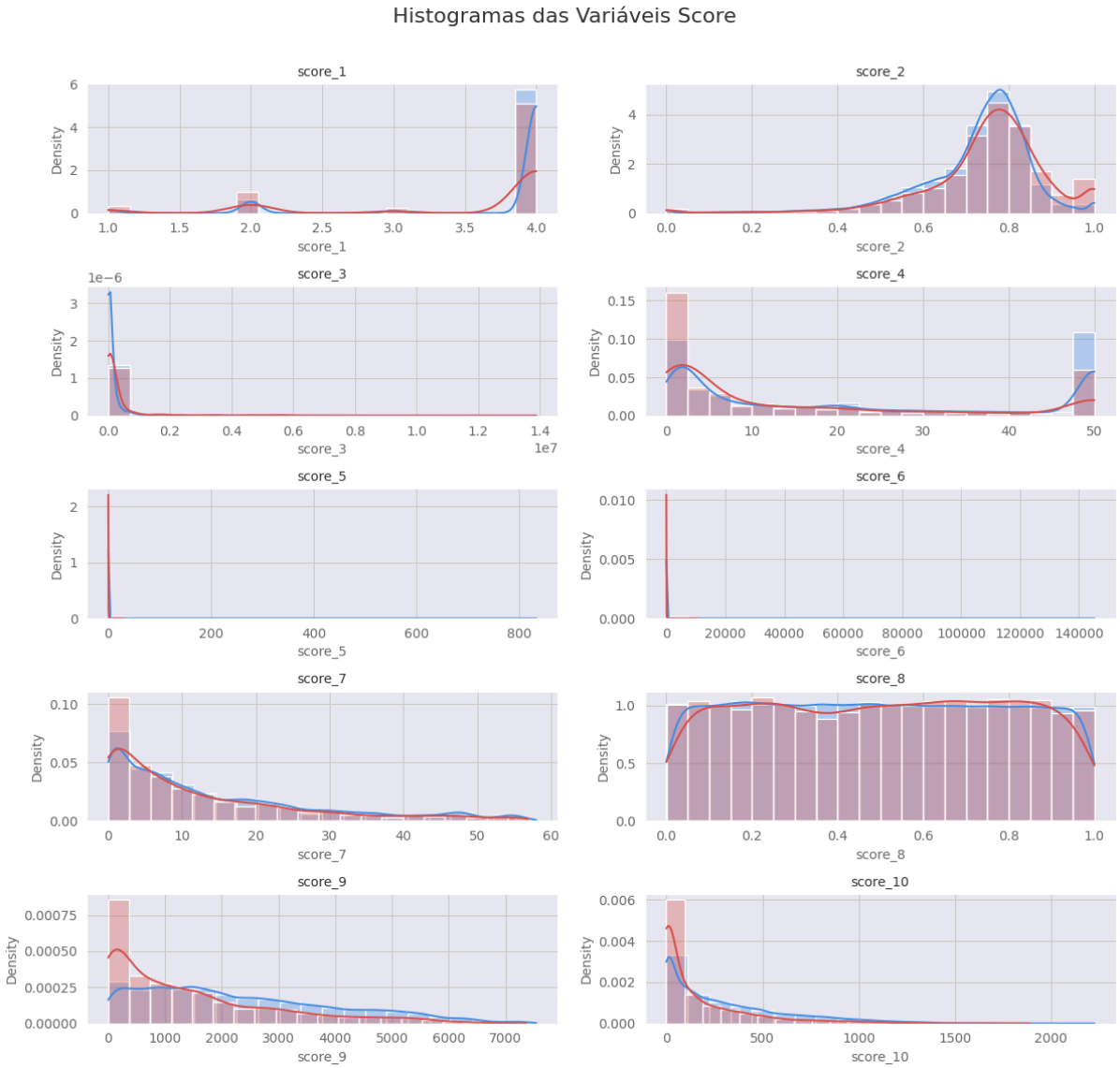


Um segundo momento de análise pretende verificar se há relação associativa entre o horário do dia e o cometimento de fraudes. Este gráfico apresenta uma tendência de aumento na quantidade de transações durante o período comercial e uma diminuição durante as madrugadas. As fraudes acompanham o movimento das transações de forma diretamente proporcional ao longo do dia e da noite, mas durante as madrugadas, das 0

horas até as 4 ou 5 horas, há bem poucas transações e a proporção de fraudes aumenta em relação ao total das transações. Também realizei um teste de hipótese Chi-quadrado para confirmar esta intuição. O resultado do teste chi2 = 841.65 e o p-value = 5.3e-163, confirmam que não é possível aceitar a hipótese nula (de que não há relação entre hora do dia e fraudes), portanto existe sim uma associação estatisticamente significativa entre o horário e o cometimento de fraudes.

8. Análise das Colunas Score_1 a Score_10

As colunas Score são dados dos quais não temos nenhuma informação sobre o significado de cada score, portanto só é possível avaliar as distribuições e tentar extrair alguma associação entre fraudes e cada um deles, bem como averiguar o tipo de variável (se categórica, numérica, discreta ou contínua) e procurar por possíveis semelhanças e diferenças nas distribuições entre as transações legítimas e fraudadas.



A coluna Score_1 aparentemente é uma coluna categórica, pois só tem 4 valores distintos, mas há uma concentração maior no valor 4, tanto de transações legítimas, quanto fraudadas. O valor 2 também apresenta uma concentração alta de dados, principalmente de fraudes. A coluna Score_2 apresenta uma distribuição mais próxima de uma normal, mas com uma cauda à esquerda, onde tanto as transações legítimas quanto as fraudes se distribuem, de maneira semelhante.

A coluna Score_3 apresenta alta cardinalidade, com mais de 135K valores, distribuídos em um intervalo que varia de 0 a 1.4MI, mas apresenta uma distribuição assimétrica positiva, com uma cauda muito longa e a maioria dos dados em faixas de valores baixos. A coluna Score_4 varia entre 0 e 50 e aparenta uma distribuição bimodal, com a maioria dos valores concentrada na parte baixa e na parte alta do intervalo, com o restante dos dados distribuídos mais ou menos uniformemente ao longo do eixo x.

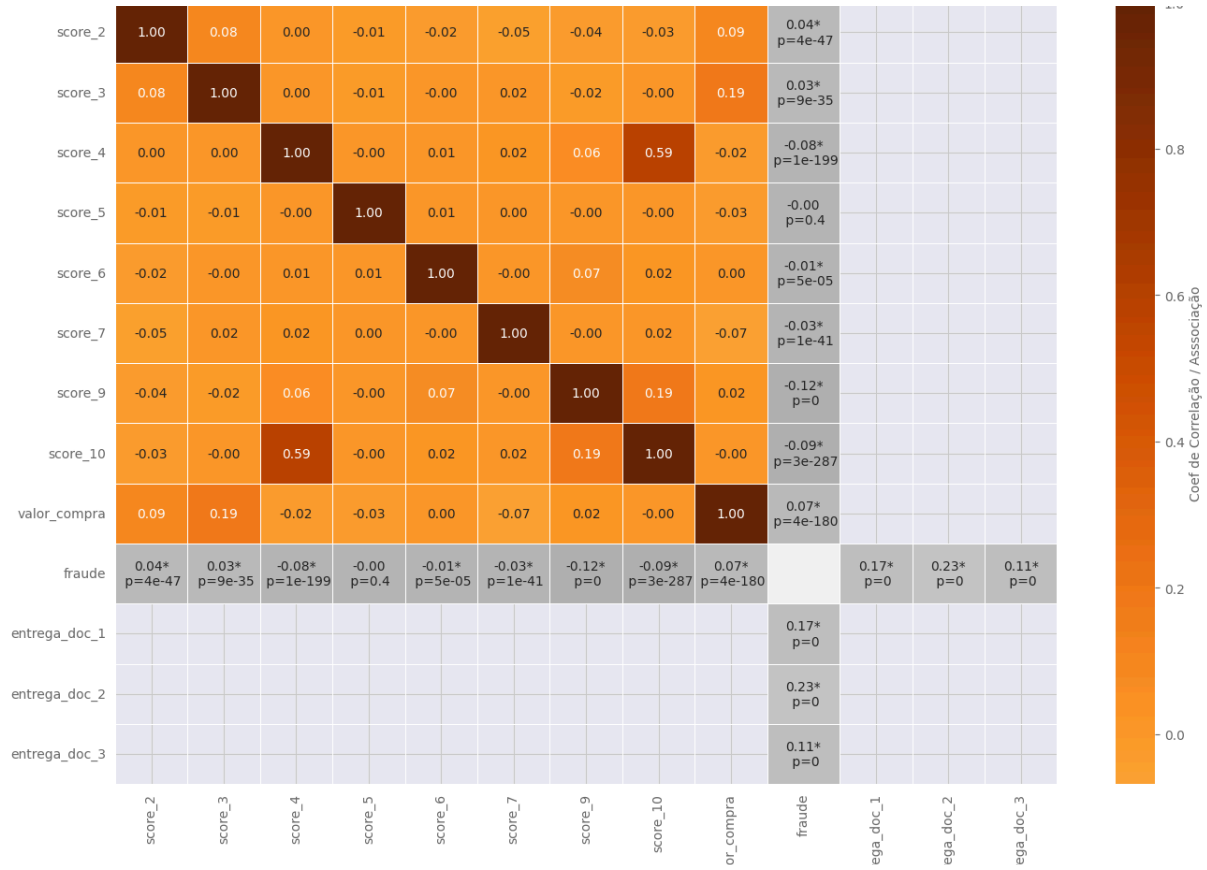
As colunas Score_5 e Score_6 apresentam uma distribuição altamente enviesada para a esquerda, com a quase totalidade dos dados na faixa próxima a zero. As colunas Score_7, Score_9 e Score_10 possuem distribuições assimétricas positivas (enviesada à esquerda) com a maioria dos valores concentrados na faixa baixa de valores, poré com um decaimento ao longo do intervalo que às vezes assume um comportamento exponencial e às vezes assume um padrão linear.

Já a coluna Score_8 varia enre 0 e 1 mas possui 150 mil diferentes valores, em uma distribuição quase uniformemente distribuída. Parece que é uma coluna que já foi normalizada e que possui um único valor para cada usuário ou transação. Desta forma ela não acrescena informação de tendência ou padrão, funciona quase como um ID da transação.

9. Correlações

Este dataset possui colunas com informações categóricas e outras com informações numéricas. As correlações entre variáveis numéricas podem ser avaliadas usando a correlação de Pearson, uma forma de comparação entre médias de variáveis que indica a força e a direção da associação entre duas variáveis. Porém as variáveis categóricas precisam de outra metodologia de avaliação, dependendo do tipo de categoria a que pertencem. Categorias ordinais podem ser avaliadas com a correlação de Spearman, baseada em rankeamento (portanto ordem de valor), que indica a força e direção da associação entre variáveis. A correlação Ponto Bi-Serial é uma forma de calcular a associação entre uma variável binária (no nosso caso fraude= 1, não fraude = 0) e variáveis contínuas. Já para avaliar a correlação entre duas variáveis binárias, podemos usar o teste chi-quadrado e confirmar se há ou não uma associação estatisticamente significativa entre duas variáveis.

Neste projeto utilizei um mix entre correlação de spearman, Bi-Serial e Chi-Quadrado, como podemos ver na figura abaixo:



A correlação mais forte encontrada neste conjunto de dados é entre a coluna Score_4 e a coluna Score_10, com um valor positivo de 0.59. Apesar de ser a maior correlação identificada, não é suficientemente alta para precisarmos escolher apenas uma das variáveis para o modelo (não recomendado utilizar variáveis autocorrelacionadas em ML). O restante das correlações não é significativa, apenas indicam levemente associações entre variáveis.

A associação entre as variáveis entrega dos documentos 1, 2 e 3 com a variável fraude foi calculada usando teste chi-quadrado para medir a força da associação (phi) e o p-value que confirma ou refuta a hipótese de associação. Através dos valores calculados e do heatmap, fica confirmado que há sim uma associação estatisticamente significativa, ainda que a força não seja tanta, entre a entrega ou não dos documentos e o cometimento de fraudes.

6. PRÉ PROCESSAMENTO DE DADOS

1. Divisão (Split) dos Dados em Treino e Teste

Esta etapa de separação de dados para treino e teste é fundamental para os sistemas de Machine Learning supervisionados possam “aprender” de forma generalizável e evitar overfitting, ou seja, quando o modelo vê os mesmos exemplos duas vezes e memoriza as respostas ao invés de generalizar para novos dados.

O conjunto de dados de teste simula dados nunca antes vistos pelo modelo, similar ao que acontece na realidade. A divisão entre treino e teste permite comparar o desempenho no treino com o desempenho no teste. Quando há um desempenho muito bom no treino e muito ruim no teste, isso pode ser um sinal de pouca capacidade discriminativa do modelo. A divisão entre treino e teste também é importante para usar técnicas mais avançadas de validação, como validação cruzada, grid search ou Bayesian search Cross validation. Comparar várias amostras (bootstrap) com cross validation permite uma maior confiabilidade estatística das métricas de erro usadas para avaliação.

2. Definir Funções de Pré-Processamento

Nesta etapa, o objetivo é definir funções que possam ser aplicadas de modo generalizado tanto no dataset de treino como no dataset de teste. Ambos os datasets precisam passar exatamente pelo mesmo processo de tratamento e precisam ter os mesmos tipos de dados, ordem, nomes de colunas, escala, etc.

O uso de funções é uma boa prática pois permite modularizar o projeto e até mesmo utilizar determinadas funções em diferentes projetos. Funções permitem reprodutibilidade e consistência na forma como os dados são transformados, evita duplicação e torna o código mais limpo, organizado e fácil de fazer alterações e manutenções.

As funções de pré-processamento em um pipeline de dados do sklearn exigem que as transformações sejam encapsuladas em objetos do tipo fit e transform, garantindo que apenas os dados de treino sejam usados para fazer o fit e os dados transformados do conjunto de testes não tenham contato com os dados do conjunto de treino.

3. Imputers

Imputers são funções que preenchem os dados nulos ou missing values utilizando critérios para cada tipo de coluna. Constituem uma etapa muito importante pois a maioria dos algoritmos de ML são incapazes de lidar com dados nulos diretamente. A biblioteca sklearn possui a classe SimpleImputer, que pode imputar dados usando estratégias definidas pelo usuário.

No contexto deste projeto, utilizei imputers para preencher dados faltantes usando a estratégia de imputar valor constante para as categorias país e entrega_doc_2 e utilizei a estratégia de imputar a mediana nas colunas score que possuíam dados faltantes. Neste último caso, a opção pela mediana foi decidida após avaliar as distribuições das colunas score e constatar que nenhuma delas seguia uma distribuição normal.

4. Encoders

Encoders são funções que transformam colunas categóricas em colunas com valores numéricos. É uma etapa importante porque a maioria dos modelos de ML não pode lidar diretamente com textos e categorias não ordinais podem ser compreendidas como ordinais. Por exemplo, neste projeto há uma coluna entrega_doc_1 que possui apenas 4 valores numéricos (1, 2, 3 e 4), porém não temos informação se esta ordenação representa um valor da categoria, ou seja, se 4 é maior que 1, ou se 1 é melhor do que 4. Por isso a opção para esta coluna foi encodá-la com One Hot Encoding, que irá apenas indicar a qual categoria pertence, sem gerar uma ordem de valores.

O One Hot Encoding (ou encoding binário) funciona atribuindo uma nova coluna para cada categoria e marcando todas as linhas que não pertenciam àquela categoria com o valor 0 e marcando apenas as linhas que eram da categoria original com o valor 1. Ele aumenta a dimensionalidade do dataset, mas é rápido e eficaz de ser aplicado.

Outro encoder utilizado neste projeto foi o WOE, Weight of Element, para encodar a coluna categoria_produto, que é uma coluna categórica com altíssima cardinalidade mais de 8 mil valores diferentes. O WoE funciona calculando para cada categoria a razão entre a quantidade de transações legítimas e quantidade de transações fraudadas e consegue captar bem a relação entre as categorias que possuem mais e as que possuem menos fraudes.

Outra forma de encoder que este projeto apresenta, derivado do One Hot Encoder, é criar uma coluna para assinalar a falta de dados nas colunas de entrega de documentos e scores. Neste caso, a premissa adotada foi de que a falta de dados pode ser compreendida como informação com potencial relevância para o modelo e que, portanto, deve ser uma informação mantida.

5. Normalização / Escala dos Dados

O propósito de normalizar os dados é diminuir o intervalo de valores presentes no conjunto de dados, trazendo os valores para uma escala comum. Modelos de ML baseados em distância e em gradiente são muito sensíveis à escala dos dados. Sem normalização, as variáveis com escalas maiores dominam o modelo e podem distorcer o aprendizado. Há algumas estratégias de normalização, mas neste projeto optei por usar a normalização por Min Max, que mantém a distribuição original e apenas achata os dados de cada variável para um intervalo entre 0 e 1. Este scaler foi aplicado a todas as variáveis numéricas.

6. Pipeline

O pipeline é um dos conceitos mais importantes desta etapa de pré-processamento, pois ela ajuda a automatizar e organizar os fluxos do projeto de machine learning. É uma função que encadeia e encapsula várias etapas de pré-processamento e modelagem em um único objeto. Neste projeto, para finalizar a etapa de pré-processamento, defini uma função pipeline para encadear todas as etapas anteriores em uma única. Isso facilita a aplicação e teste de diferentes modelos e as etapas de validação cruzada, mantendo a integridade dos conjuntos de treino e teste e a separação correta entre os dados.

7. RESPOSTAS ÀS PERGUNTAS E HIPÓTESES INICIAIS

Este projeto de Análise Exploratória de Dados, respondeu parte das questões levantadas inicialmente, porém este tema de prevenção à fraudes é suficientemente complexo para restarem muitas dúvidas.

Sobre o perfil das transações fraudadas, acredito que foi possível identificar algumas características importantes. Transações fraudulentas ocorrem com maior intensidade no Brasil do que em outros países. Fraudes também podem ser caracterizadas por preferencialmente acontecerem na compra de produtos da categoria “43b9c10” que engloba certos produtos, como aparelhos de telefonia móvel. Também constata-se que as fraudes ocorrem, em média, em faixas de valor mais altas do que as compras normais. Também vimos que as fraudes ocorrem em maior intensidade durante as madrugadas, mas não só nestes períodos.
Fraudes acompanham a quantidade de transações tanto nos dias do mês como nos horários do dia, numa tentativa clara de se misturar entre as compras normais, uma forma de dissimular a atividade ilícita. Também foi possível identificar uma associação entre a não entrega dos documentos, principalmente do tipo 1 e 2, este último concentrando 21% das compras fraudadas entre aqueles que não entregaram a documentação.

O mercado de cartões de crédito e de transações digitais possui seus próprios indicadores de monitoramento das transações saudáveis e fraudadas. Uma delas é a taxa percentual de fraudes, que neste estudo é de 5%. Mas neste caso em particular, esta taxa de 5% parece um pouco artificial, o que pode ser explicado pela amostra escolhida, que é uma fatia temporal de pouco mais de um mês no início da pandemia, quando todo o comércio do mundo (e o crime de fraude) se voltou exclusivamente para o meio digital, que pouco a pouco foi desenvolvendo novas formas de prevenção.
Outra métrica usada é a taxa de fraudes por cada 100 mil transações, métrica que neste caso fica ainda mais inflada, já que temos 150 mil dados, das quais 7500 são fraudes. Outra taxa comumente utilizada é a “basis points , BPS”, que é calculada como o valor total das fraudes dividido pelo valor total das transações aprovadas multiplicado por 10000. Esta métrica é mais comum no mercado financeiro e cada basis point representa 0.01%, um centésimo de um ponto percentual. Neste dataset, o basis point é de 914.99, um valor completamente desproporcional, provavelmente causado pela amostra escolhida.

As features que mais se relacionam a fraudes neste dataset são as entregas de documento e a categoria do produto. Como vimos extensamente, a falta de entrega dos documentos, sobretudo o documento tipo 2 é um indício bem forte de possibilidade de fraude. Também vimos como há uma preferência na aquisição de aparelhos celulares em compras fraudadas.

É bem difícil traçar um perfil do fraudador com base apenas nas informações deste dataset, pois estes dados não trazem informações a respeito dos clientes, e sim das transações. Caso haja um outro dataset com os ids e informações dos clientes, poderíamos cruzá-lo com estas transações e então avaliar separadamente o perfil dos clientes envolvidos em fraudes.

Não há uma estratégia de prevenção à fraudes sozinha que seja suficiente para diminuir as transações falsas. É necessário implantar um conjunto de medidas e um sistema completo de detecção e prevenção que passa por mais segurança no lado do consumidor com senhas e sistemas de dupla ou tripla verificação, alertas em tempo real para que o cliente reconheça aquela compra e também sistemas de machine learning mais precisos para identificar pequenas anomalias e comportamentos fora do normal entre as milhões de transações diárias.

O objetivo desta análise é subsidiar uma próxima etapa onde será de fato desenvolvido um sistema de machine learning que tenha um desempenho melhor do que o sistema usado nesta empresa dentro deste período de março a abril de 2020 e que foi responsável por um percentual alto de fraudes que levaram a uma grande perda de faturamento da companhia.

8. AUTOAVALIAÇÃO

Neste projeto fiz uma análise criteriosa das principais variáveis envolvidas na classificação de transações fraudadas e legítimas com o objetivo de compreender a dinâmica das operações de compra e identificar semelhanças e diferenças marcantes entre os milhares de dados deste dataset.

Algumas observações e possíveis melhorias para uma próxima etapa:

1 - Segmentar os dados por tipo de documentação entregue e olhar separadamente para características que diferenciam e assemelham os grupos. Desta forma poderia gerar novas colunas com estas características, que poderiam ser valiosas para uma modelagem posterior.

2 - Separar as compras por mês e analisar cada período separadamente para identificar sazonalidades. Fazer feature engineering usando a coluna data para gerar novas colunas com semana do mês, dia da semana, final de semana e feriados. Compras acontecem mais em quais dias da semana? E as fraudes, são mais frequentes no início ou no final da semana? Fazer o mesmo por período do dia: manhã, tarde, noite e madrugada. Talvez essas novas features possam enriquecer o modelo de classificação.

3 - Algumas colunas do grupo Scores poderiam ser usadas para agrupar e analisar os conjuntos de fraude / não fraude em separado. Por exemplo a coluna Score_1 que possui apenas 4 valores, que podem ser considerados como categorias. Tentar descobrir o que cada “categoria” desta poderia acrescentar de informação ao modelo.

4 - Possível uso de transformações algébricas para tentar aproximar de uma normal as distribuições de algumas colunas score que possuem distribuições muito enviesadas.

5 - Aplicar algum modelo não supervisionado de clusterização (K-Means, por exemplo) e usar os clusters gerados como mais uma feature informativa para o modelo.

6 - Utilizar a informação de probabilidade de fraude do modelo legado como mais uma informação que pode agregar valor aos modelos testados.

9. PRÓXIMOS PASSOS

Pretendo utilizar este mesmo dataset na próxima sprint de Machine Learning e aprimorar o conjunto através do uso de feature engineering, conforme idéias levantadas na seção anterior, com o objetivo de ganhar mais informação que possa ser útil para o modelo de classificação.

Pretendo investigar as estratégias de balanceamento de dados como Smote, Undersampling e Oversampling e experimentar se elas realmente são favoráveis em modelos de classificação baseados em árvore e em gradiente.

Pretendo também utilizar MFlow para gerir e avaliar o fluxo de experimentação dos modelos e das etapas de fine-tuning, chegando no final a um modelo com melhor capacidade de generalização. Alguns dos modelos que pretendo aplicar incluem Random Forest, Lgbm e Xgboost. Possivelmente usar modelos ensemble, eventualmente Redes Neurais, enfim, explorar as diferentes possibilidades e ver quais se adequam melhor a este tipo de problema.

Ainda pretendo salvar o modelo treinado em formato .pkl e montar um sistema de consulta em uma interface web, possivelmente Streamlit ligado a uma API, onde possa inserir dados e receber uma saída de acordo com o modelo. Provavelmente irei precisar de algum máquina virtual rodando um container docker que encapsule todas as dependências e versões para que o modelo possa funcionar online.

Assim, além do modelo de Machine Learning em si, pretendo chegar a uma tentativa de deploy simulando uma situação de uso real deste modelo, extrapolando a parte mais teórica e inserindo um aspecto mais prático da utilização de machine learning para responder problemas reais de negócio.

10. CONCLUSÃO

Este trabalho de análise de dados e boas práticas desta sprint intermediária (minha segunda de três sprints), foi muito interessante para acrescentar e aprimorar habilidades, tanto de programação python, como em teoria das distribuições e práticas de visualização de dados.

Das disciplinas contidas nesta sprint, a de visualização de dados foi a que mais me despertou interesse, gostei muito dos exemplos e do material usado pela professora Simone. Ela ofereceu vasta bibliografia interessante sobre data viz e também um referencial teórico consistente.

A disciplina de Análise e pré processamento também foi muito produtiva, a professora Tatiana Escovedo é muito carismática e tem uma didática incrível. Como este tema de Machine Learning e Pré-Processamento é um tema no qual já estou imerso há alguns anos, senti que poderia haver um pouco mais de complexidade nos exemplos (dataset Íris) de tratamento e transformações dos dados, porém compreendo que o curso precisa abranger um amplo espectro de experiências e conhecimentos. Mesmo assim ouvi muitos relatos de colegas com dificuldade para compreender conceitos básicos de ciência de dados.

A disciplina de Engenharia de Software, do professor Marcos Kalinowski apresentou um referencial teórico fundamental para quem trabalha na prática com equipes multidisciplinares de engenheiros de software, engenheiros de dados e cientistas de dados, como as metodologias ágeis e Solid, boas práticas de DevOps, CI / CD, etc.

De modo geral o balanço desta sprint foi bem positivo e espero que este trabalho corresponda ao esperado pelos professores. Agradeço especialmente aos colegas professores Hugo Villamizar e Antônio Pedro pelas ótimas orientações e discussões, tanto em aulas online, quanto através do canal do Discord. Agradeço profundamente pela paciência e dedicação com a qual conduziram estas sessões e conversas.