

1. Perfil de transações

Esta primeira parte do relatório apresenta uma análise descritiva das estatísticas de transações. As informações foram extraídas do *dataset* enviado, que conta com 11.004 transações na Aba 1. Vê-se a seguir as Estatísticas Descritivas do Valor das Transações (Aba 1).

A média do valor das transações indica que a maioria delas são de valor relativamente baixo. No entanto, o desvio padrão vultoso demonstra uma grande dispersão nos valores das transações, com algumas destas situando-se muito acima ou abaixo da média. O valor mínimo de R\$ 1,00 sugere a possibilidade de transações fraudulentas (teste do cartão) ou com erros de digitação (menos provável, pois o próximo valor ainda seria 10 reais, baixo). A mediana de R\$ 99,00 indica que metade das transações tem valor inferior a R\$ 99,00, enquanto a outra metade tem valor superior a tal quantia financeira. O terceiro quartil de R\$ 154,00 sugere que 75% das transações têm valor inferior a R\$ 154,00. Por final, o valor máximo de R\$ 2.920,00 revela a existência de transações de alto valor no conjunto de dados, destoando bastante dos outros valores reportados.

Na Figura 1, a seguir temos a distribuição do valor das transações. Este gráfico representa a distribuição geral dos valores das transações. A maioria das transações concentra-se em valores baixos, particularmente na faixa de R\$0 a R\$100, com uma redução na frequência para valores maiores.

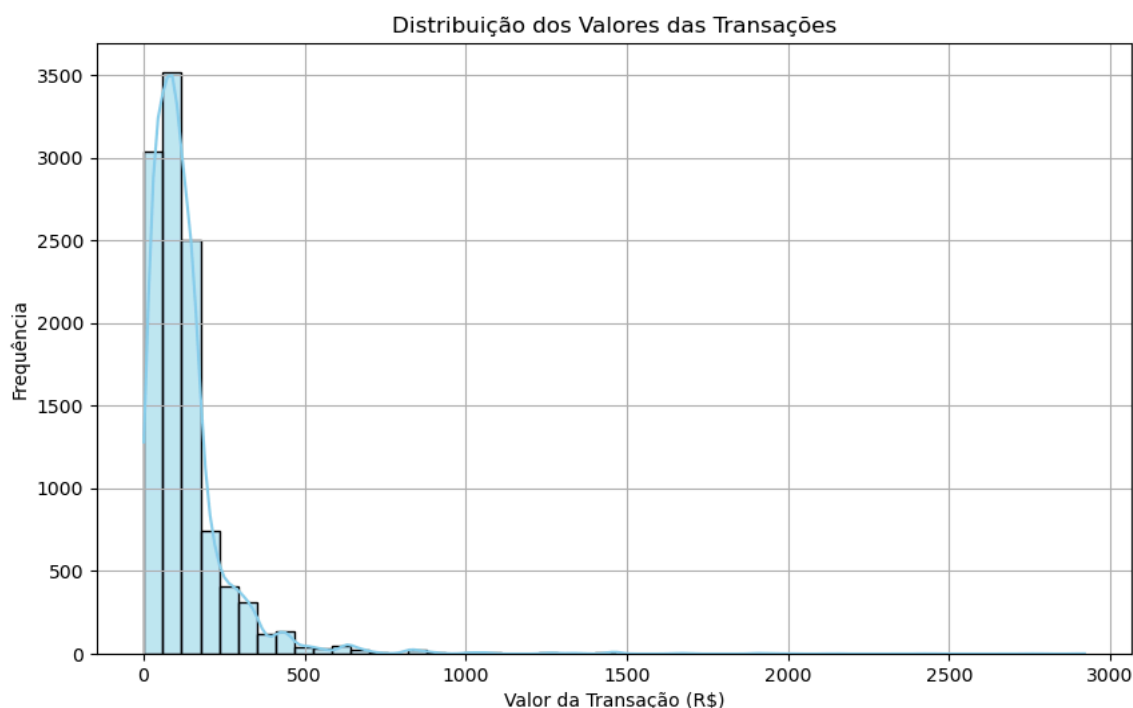


Figura 1 - Distribuição dos valores das transações.

Fazendo uma análise mais detalhada, vemos que a moda da distribuição está novamente na faixa de R\$0 a R\$100, sugerindo uma “predileção” de transações de baixo valor; a assimetria positiva ($\text{skewness} > 0$) indicam que a maioria das transações é de valor baixo, mas existem transações esporádicas de valores significativamente altos.

Sobre a Frequência de Transações por Hora do Dia, na Figura 2 temos um gráfico de barras que mostra a frequência das transações ao longo do dia. Observa-se um aumento constante no número de transações a partir das 8:00, atingindo um pico entre 17:00 e 19:00, seguido por uma diminuição gradual.

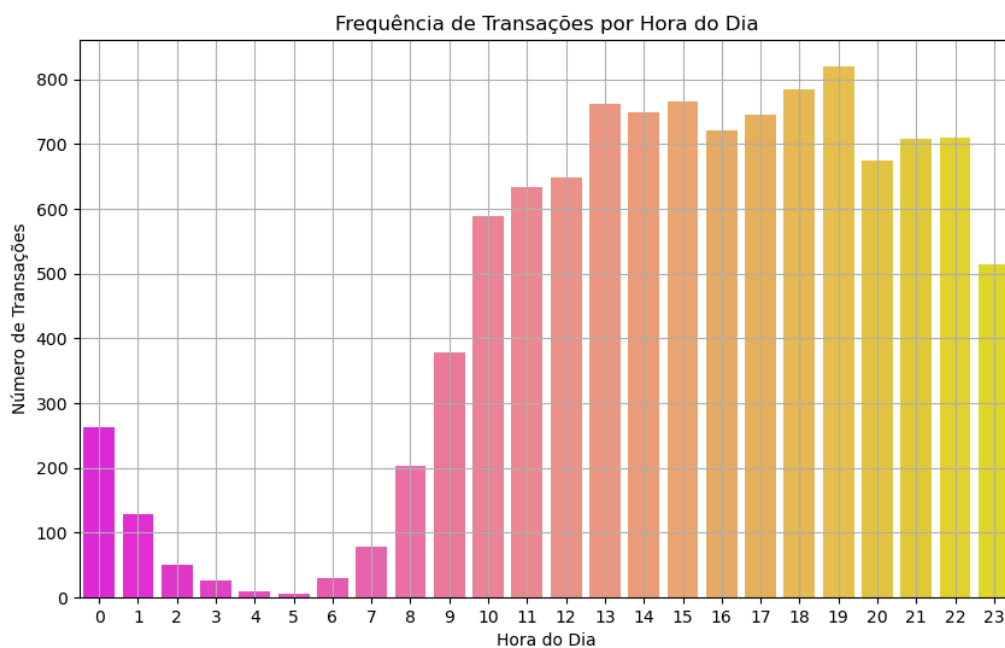


Figura 2 - Frequência das Transações por Hora do Dia.

A maior atividade de transações ocorre no final da tarde e início da noite, com um pico notável no período do *rush* (18/19 horas). O padrão de aumento ao longo do dia até o pico, seguido de um declínio, sugere uma correlação com os horários comerciais e hábitos de consumo, não especificamente com evento de fraude, mas, pode ser estudado pelo fraudador a fim de encontrar o horário mais oportuno para fazê-lo.

Na **Figura 3**, temos um histograma que ilustra a distribuição dos valores das transações realizadas entre 19:00 e 23:00 horas. A distribuição exibe uma concentração alta de transações em valores baixos, particularmente entre R\$0 e R\$100. A análise mostra uma diminuição progressiva na frequência à medida que o valor das transações aumenta. Existe, visivelmente uma assimetria positiva com formação de cauda.

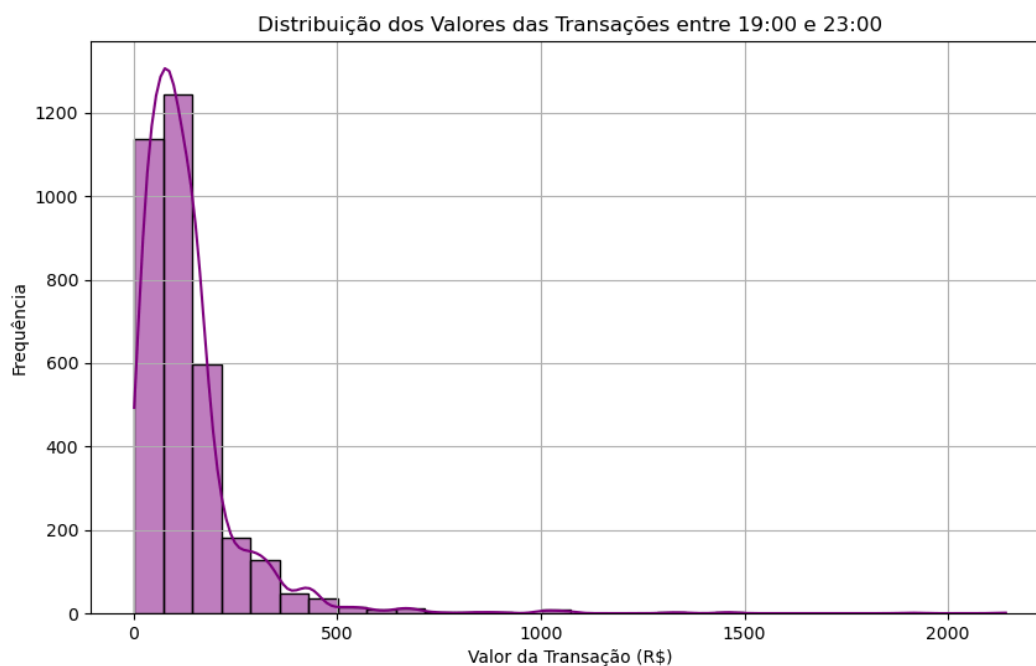


Figura 3 - Dsitribuição dos Valores das Transações entre 19 e 23 horas.

Na Figura 4, temos a incidência de *chargebacks*. Este gráfico de pizza demonstra a porcentagem de transações que resultaram em chargebacks. Apenas 5.2% das transações culminaram em estornos, enquanto 94.8% das transações foram concluídas sem problemas. A baixa incidência de chargebacks (5.2%) é um indicativo positivo da integridade das transações e da confiança no sistema de pagamentos. Apesar da baixa taxa, é de suma importância manter um monitoramento rigoroso e implementar melhorias contínuas para reduzir ainda mais essa incidência. Para isso, é interessante conter mais variáveis na base de dados, a fim de detalhar e melhor discriminar essas operações.

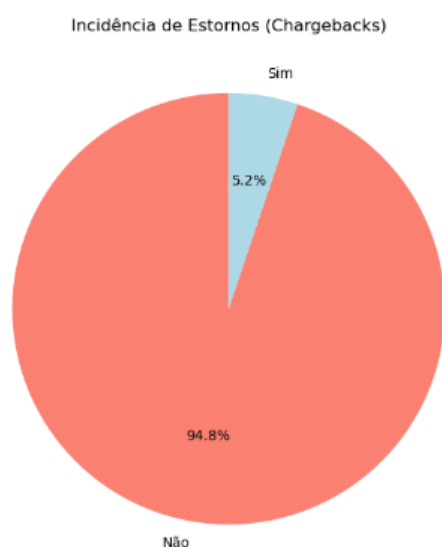
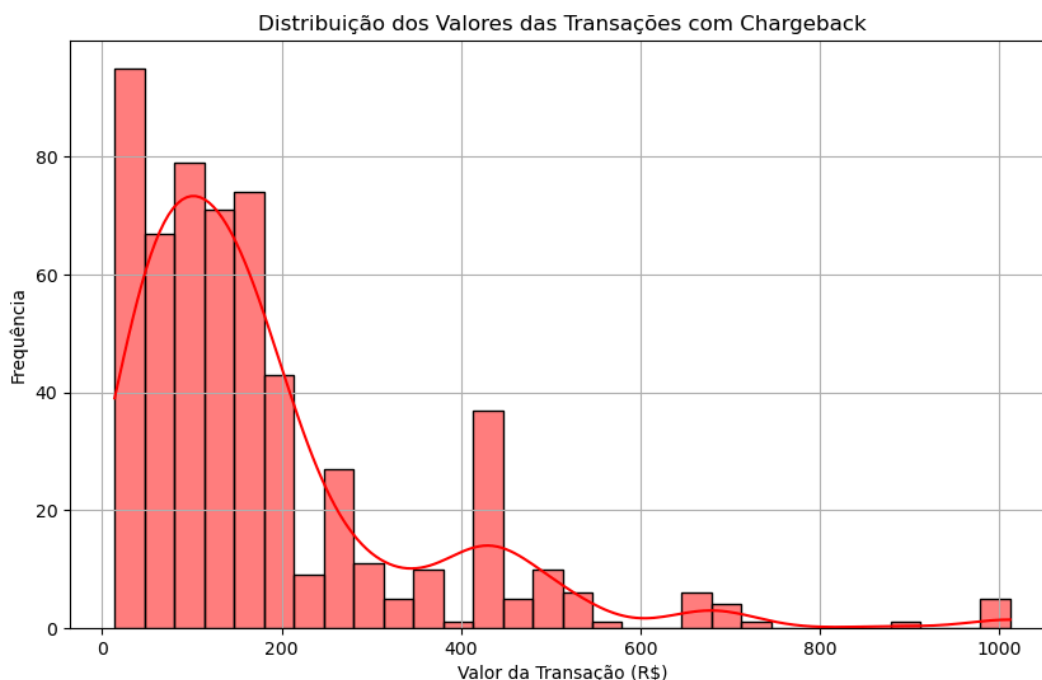


Figura 4 - Gráfico de incidência de estornos.

No conjunto de dados analisado, fora percebido um total de 568 transações com *chargeback*, representando 5,16% do total. A média da hora em que os *chargebacks* ocorrem é 15:05, indicando que a maioria dos *chargebacks*, pelo menos para este cenário, acontece no período da tarde. O valor médio das transações com *chargeback* é de R\$ 182,86, significativamente maior do que a média geral das transações. O desvio padrão da hora dos *chargebacks* é 5:42, mostrando uma grande variabilidade no horário em que ocorrem. Da mesma maneira, o desvio padrão do valor das transações com *chargeback* é de R\$ 165,53, indicando grande variabilidade nos valores.

Chargebacks podem ocorrer a qualquer hora do dia, com o horário mínimo registrado sendo 0:00, e para valores relativamente baixos, com o valor mínimo sendo R\$ 13,99. O primeiro quartil da hora é 13:00, mostrando que 25% dos *chargebacks* ocorrem antes desse horário, enquanto o primeiro quartil do valor é R\$ 68,25, indicando que 25% das transações com *chargeback* têm valor inferior a esse montante. A mediana da hora é 16:00, significando que metade dos *chargebacks* ocorre antes desse horário, e a mediana do valor é R\$ 132,00, com metade das transações de *chargeback* abaixo desse valor. O terceiro quartil da hora é 19:00, mostrando que 75% dos *chargebacks* ocorrem antes desse horário, e o terceiro quartil do valor é R\$ 209,25, indicando que 75% das transações com *chargeback* são menores que esse valor. Finalmente, os *chargebacks* podem ocorrer até às 23:00, com o maior valor de *chargeback* registrado sendo de R\$ 1.012,00.

Na Figura 5, temos a Distribuição dos Valores das Transações com *Chargeback*.



Este gráfico de distribuição foca nos valores das transações que resultaram em *chargebacks*. Os dados mostram uma alta incidência de *chargebacks* em transações com valores até R\$100, com uma diminuição gradual à medida que o valor das transações aumenta. Transações de baixo valor são mais suscetíveis a *chargebacks*, conforme evidenciado pela moda na faixa de R\$0 a R\$100. A distribuição é assimétrica positiva, similar ao padrão geral, com uma cauda longa à direita. Esta informação é crucial para a implementação de mecanismos de prevenção de fraudes, direcionando esforços para monitorar transações de valores mais baixos com mais rigor.

2. Classificando as operações

Nesta parte do *case* nos propomos a criar um método de identificar se uma transação futura retornará *chargeback* e aplique-o nas transações da Aba 2, que representam as transações do mesmo lojista depois de um mês. Os dados foram importados da planilha fornecida, utilizando a biblioteca *pandas* para carregamento e manipulação, devido à sua eficiência e facilidade de uso em tarefas de utilizar-se de dados tabulares. Durante o pré-processamento, as colunas foram convertidas para tipos numéricos apropriados, valores ausentes foram removidos, e dados duplicados foram eliminados para assegurar a qualidade e consistência dos dados. A variável dependente 'CBK' (*chargeback*) foi dicotomizada/binarizada para representar a ocorrência dos eventos, facilitando a tarefa de classificação binária.

Para garantir a validade da avaliação dos modelos, os dados foram divididos em conjuntos de treino e teste (*split* de dados) com uma proporção de 70/30 (%), para mitigar *overfitting* e assegurar uma avaliação imparcial. O desbalanceamento das classes, um problema comum em muitas aplicações práticas (exemplo também de aferição de risco clínico, risco de crédito), foi abordado por meio do Random Under Sampler (RUS), equalizando o número de exemplos em cada classe no conjunto de treino e garantindo que os modelos não fossem tendenciosos para a classe majoritária.

Diversos algoritmos de aprendizado supervisionado foram selecionados para avaliação, a saber, Regressão Logística, Floresta Aleatória (*Random FOrrest* - *RF*), MLP (Perceptron Multicamadas), XGBoost, LightGBM e CatBoost. Para a avaliação dos modelos de classificação na predição de *chargebacks*, foram selecionadas diversas técnicas baseadas em suas características e vantagens específicas.

A Regressão Logística foi escolhida por sua simplicidade e interpretabilidade, facilitando a compreensão dos resultados e das relações entre as variáveis independentes e a variável dependente. O *Random FOrrest*, por ser um método de *ensemble*, foi incluída devido à sua robustez contra *overfitting* e à capacidade de lidar com dados de alta dimensionalidade, embora não seja nosso caso no *case*, mas com toda certeza é a realidade do dia a dia [1]. Métodos de *ensemble*, como a Floresta Aleatória, combinam as previsões de múltiplos modelos base, geralmente árvores de decisão, para melhorar a precisão e a estabilidade das previsões.

O XGBoost, um outro método de *ensemble* [2], foi escolhido por sua eficiência computacional e alta performance. XGBoost é uma implementação otimizada de *boosting*, que ajusta iterativamente modelos fracos para corrigir os erros dos modelos anteriores, resultando em um modelo robusto e preciso. O LightGBM [3] foi incluído devido à sua eficiência e escalabilidade para grandes conjuntos de dados, permitindo um treinamento rápido e utilização de memória eficiente. LightGBM também utiliza técnicas de *ensemble*, especificamente o *gradient boosting*, para construir modelos preditivos poderosos ao combinar várias árvores de decisão ajustadas para corrigir erros residuais iterativamente.

Ainda entrou no *hull* de escolhas o CatBoost [4], foi selecionado por seu manejo eficiente de variáveis categóricas e mitigação de *overfitting*. CatBoost, outro método de *ensemble* baseado em *boosting*, é particularmente eficiente na incorporação de variáveis categóricas diretamente no modelo, evitando a necessidade de pré-processamento extenso.

O Perceptron Multicamadas (MLP), foi escolhido por sua capacidade de capturar relações não lineares complexas, sendo especialmente útil em problemas onde as relações entre as variáveis não são lineares [5].

Para a avaliação do desempenho dos modelos de classificação na predição de *chargebacks*, utilizamos diversas métricas para obter uma visão abrangente da eficácia de cada modelo. A métrica ROC AUC (Área sob a Curva ROC) foi escolhida por ser especialmente útil em conjuntos de dados desbalanceados, como é o caso dos *chargebacks*, fornecendo uma medida da capacidade do modelo em distinguir entre as classes positiva e negativa ao longo de diferentes limiares de decisão (maior acomodação visual de incertezas). Outra métrica explorada aqui a fim de explicar os resultados é o *recall*.

O *recall*, que mede a proporção de transações de *chargeback* que foram corretamente identificadas pelo modelo. Faz-se importante para minimizar os falsos negativos, garantindo que o maior número possível de transações fraudulentas seja detectado. Embora se espere baixa precisão (métrica) neste contexto, usamos ela em combinação com *recall* para calcular o *F1-score*. O *F1-score*, que é a média harmônica da precisão e do *recall*, que fornece um equilíbrio entre essas duas métricas, sendo particularmente útil quando há uma necessidade de equilibrar a importância dos falsos positivos e falsos negativos.

Iniciando a avaliação, temos na Figura 6 o gráfico de Curva ROC, que compara o desempenho de diferentes modelos de classificação na predição de *chargebacks*, destacando suas taxas de verdadeiros positivos contra falsos positivos em diversos limiares. O LightGBM apresenta o melhor desempenho com uma AUC de 0.7177, seguido pelo XGBoost com AUC de 0.6981 e pela Floresta Aleatória com AUC de 0.6824, indicando uma superior capacidade de distinguir entre transações com e sem *chargebacks*. Modelos como a Regressão Logística, MLP e CatBoost também são avaliados, mas com AUCs menores, refletindo menor eficácia comparativa. Em geral, vê-se que algoritmos de árvore entregam melhor resultado em dados tabulares.

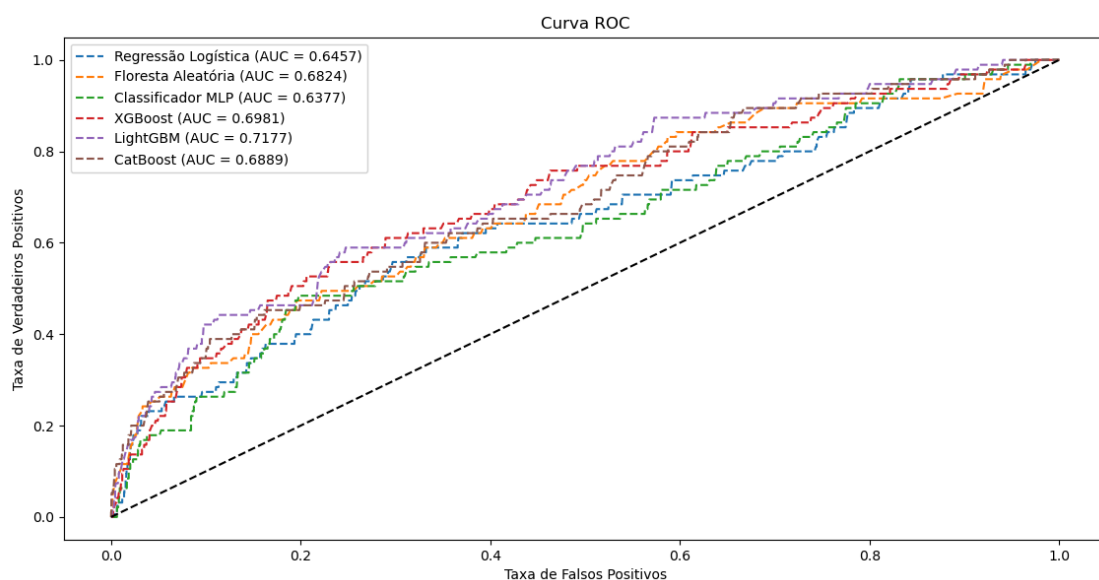


Figura 6 - Curva ROC.

As matrizes de confusão para cada classificador podem ser evidenciadas a seguir. Para determinar qual classificador erra menos na identificação de *chargebacks*, analisamos o valor de FN (False Negative), que representa as transações de *chargeback* incorretamente identificadas como não-chargeback. Quanto menor o valor de FN, melhor o classificador na identificação correta de chargebacks. Temos, portanto, os seguintes valores, CatBoost: 42 FN, LightGBM: 37 FN, Regressão Logística: 46 FN, Classificador MLP: 65 FN, Floresta Aleatória: 38 FN e XGBoost: 36 FN.

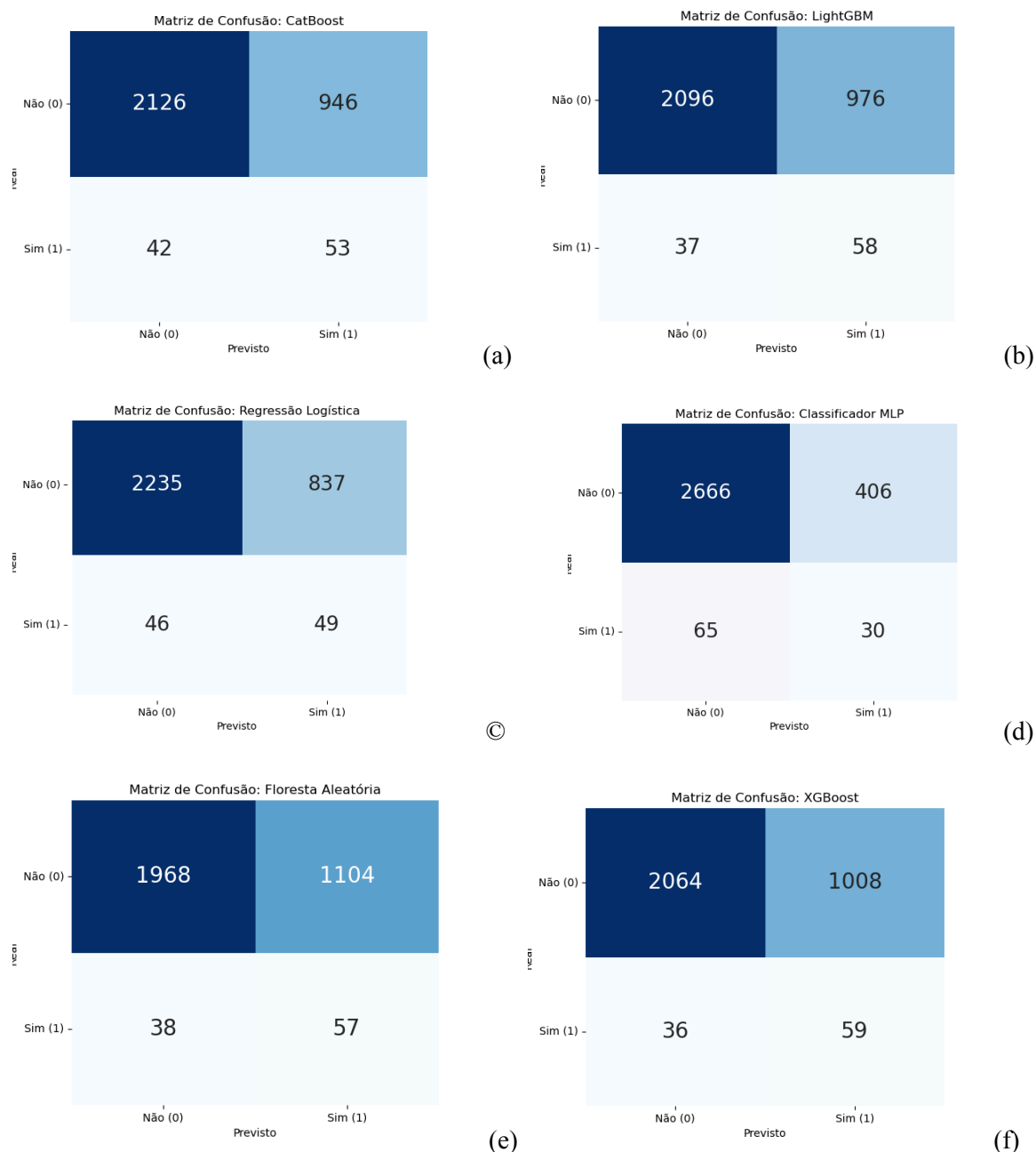


Figura 7 - Matriz de confusão dos classificadores.

Ter um bom recall na classe que aponta para *chargeback* significa que o modelo é eficaz em identificar a maioria das transações que realmente resultam em estorno. Em outras palavras, um alto valor de recall indica que o modelo consegue detectar quase todas as instâncias de chargeback presentes nos dados. Na Tabela 1, a seguir, temos as métricas para a situação sem estorno (No) e para

o caso que queremos mapear, com estorno (YeS). Proucupando-se em especial com a classe onde há *chargeback*, vemos que o *recall* obtido pelos três últimos algoritmos foi o melhor, com a vitória do Catboost, mas que perde em F1-score para XGBoost e LightGBM.

Entende-se portanto que na média das métricas, os algoritmos de *ensemble* ficam bem ranqueados e que podemos escolher com poucos prejuízos quaisquer um destes para usar como classificador. Escolhemos portanto, para dar prosseguimento as simulações o XGBoost, por entender que possui *recall* na faixa dos campeões e um bom F1-score também.

Tabela 1 - Métricas de avaliação.

Modelo	ROC AUC Score	Accuracy	Precision (No)	Recall (No)	F1-Score (No)	Precision (Yes)	Recall (Yes)	F1-Score (Yes)
Regressão Logística	0.6457	0.7212	0.9798	0.7275	0.835	0.0553	0.5158	0.0999
Floresta Aleatória	0.6824	0.6277	0.9812	0.6283	0.7660	0.0483	0.5579	0.0896
Classificador MLP	0.6377	0.5917	0.9785	0.5921	0.7378	0.4200	0.5789	0.0784
XGBoost	0.6981	0.6704	0.9829	0.6719	0.7981	0.0553	0.6105	0.1015
LightGBM	0.7177	0.6801	0.9827	0.6823	0.8054	0.0561	0.6105	0.1027
CatBoost	0.6889	0.6881	0.9806	0.6921	0.8115	0.0531	0.6211	0.0969

Continuando o processo, agora iremos prever e preencher a coluna de CBK na Aba 2. Os dados utilizados para o treinamento e avaliação do modelo foram obtidos da Aba 1 da planilha Excel. Após a limpeza e preparação dos dados, as características ('Valor' e 'Hora') e o alvo ('CBK') foram separados. Em seguida, os dados foram divididos em conjuntos de treino e teste utilizando uma proporção de 70/30.

Para lidar com o desbalanceamento das classes, foi aplicado o RUS novamente nos dados de treino, equilibrando as classes ao reduzir a quantidade de exemplos da classe majoritária. Após o balanceamento, um modelo XGBoost foi treinado com os dados de treino balanceados. O desempenho do modelo foi avaliado no conjunto de teste utilizando métricas como AUC-ROC, matriz de confusão e novamente as métricas oriundas desta ultima.

Uma vez treinado e avaliado, este foi utilizado para prever a coluna 'CBK' na Aba 2 da planilha Excel. Os dados da Aba 2 foram pré-processados de forma semelhante aos dados da Aba 1, com a conversão dos valores para tipos numéricos e a extração da hora. As previsões foram realizadas utilizando as características ('Valor' e 'Hora') presentes na Aba 2. As previsões do modelo foram então mapeadas para os valores categóricos 'Sim' e 'Não', representando a ocorrência ou não de *chargebacks*.

Para este problema de preenchimento da coluna de CBK, em particular temos as taxas de *recall* (yes) e 0.62 em comparação com 0.68 de *recall* (no), o que dada natureza do problema é factível.

3. Proposição de regras de negócio e seu impacto

Baseado na análise exploratória de dados realizada inicialmente, foram gerados insights valiosos que permitiram a criação de regras de negócio eficazes para melhorar a classificação das operações em *chargeback* ou não. Esses insights foram fundamentais para entender melhor os padrões e tendências presentes nos dados, permitindo a formulação de estratégias específicas para a detecção e prevenção de *chargebacks*, resultando em um sistema de classificação mais robusto e preciso (isso inicia nos dados, logo, foi realizada composição de novas variáveis que descrevem essas regras). Baseado nisto, foram criadas variáveis novas, estas, incorporadas a Aba 1 para testar o quanto elas impactam na classificação. O quadro com as variáveis e seu significado encontra-se a seguir na Tabela 2. Pontua-se que a partir daqui, está sendo usado as variáveis dispostas inicialmente na base juntamente com as desenvolvidas pela lógica do negócio.

Tabela 2 - Dicionário de Variáveis.

Nome da Variável	Descrição Completa
NUM_TRS_CRT	Número total de transações realizadas por cada cartão.
REP_TRS_VAL	Indica se uma transação é duplicada com base no valor, dentro do mesmo conjunto de dados.
TRS_19_23	Flag que indica se uma transação foi realizada entre as 19h e 23h.
TRS_FDS	Flag que indica se uma transação ocorreu durante um fim de semana (sábado ou domingo).
ALT_VAL_TRS	Flag que indica se o valor da transação é superior à média de valores de todas as transações.
STD_VAL_CRT	Desvio padrão dos valores de transações para cada cartão.
TX_ESTR_CRT	Taxa de transações que resultaram em estorno (chargeback) por cartão.
TP_ENT_TRS	Tempo em minutos entre transações consecutivas do mesmo cartão.

VAR_VAL_CP	Variabilidade (desvio padrão) dos valores das últimas cinco transações para cada cartão.
TRS_FRD	Flag que indica se a transação ocorreu em um feriado nacional brasileiro.
TRS_NTN	Flag que indica se a transação foi realizada entre a meia-noite e as 6h da manhã.
FRQ_TRS_DS	Frequência de transações por dia da semana para cada cartão.
TND_VAL	Tendência do valor das transações ao longo do tempo para cada cartão, calculada como a média cumulativa.

Os resultados de classificação obtidos com essas inclusões foi o seguinte, com os métodos de *ensemble* novamente indicando os melhores resultados de *recall* para classe de *chargeback* == 1.

Tabela 3 - Métricas de avaliação do modelo com as novas variáveis.

Modelo	ROC AUC Score	Accuracy	Precision (No)	Recall (No)	F1-Score (No)	Precision (Yes)	Recall (Yes)	F1-Score (Yes)
Regressão Logística	0.6515	0.7623	0.9651	0.7796	0.8625	0.7400	0.3846	0.1242
Floresta Aleatória	0.7882	0.6772	0.9845	0.6730	0.7995	0.0973	0.7692	0.1727
Classificador MLP	0.6308	0.4631	0.9666	0.4542	618	0.0523	0.6573	0.0969
XGBoost	0.7868	0.6974	0.9846	0.6944	0.8144	0.1025	0.7622	0.1808
LightGBM	0.7870	0.7060	0.9848	0.7034	0.8206	0.1053	0.7622	0.1851
CatBoost	0.7778	0.7222	0.9834	0.7217	0.8324	0.1078	0.7343	0.1880

Perceba que se focarmos na métrica de *recall* antes da incorporação das regras de negócio e após a incorporação, percebe-se num geral um aumento das métricas de avaliação do modelo, o que corrobora para a situação onde haja melhor identificação e portanto menores percas por *chargeback*.

Tabela 4 - Métricas de *recall* para *chargeback* = 1 antes e depois da incorporação das regras de negócio.

Modelo	Recall (Yes) - ANTES	Recall (Yes) - DEPOIS
Regressão Logística	0.5158	0.3846
Floresta Aleatória	0.5789	0.7692
Classificador MLP	0.2737	0.6573
XGBoost	0.6211	0.7622
LightGBM	0.6105	0.7622
CatBoost	0.5579	0.7343

Levando em consideração que para as métricas do modelo foi possível observar ganho, calcula-se aqui o impacto em termos financeiros associando os valores de cada transação com *chargeback* como perda em algum sentido, estimando esses valores para os três principais classificadores, antes e depois da incorporação da regra de negócio. Percebe-se, em especial pelo XGBoost que a perda foi a menor, seguido por LightGBM e Catboost. Em termos percentuais temos uma margem com o XGBoost de economia de 25,01% (*redução de chargebacks não identificados*). Pontua-se que essa economia diz respeito apenas as regras de negócio, sem precisar fazer *fine tuning* ou calibração do modelo.

Tabela 5- Métricas de valor associado a *chargeback* (em R\$).

Modelo	Antes	Depois (Regra)	Redução (em %)
Floresta Aleatória	219670.43	173319.8	21.10
XGBoost	206807.48	155079.84	25.01
LightGBM	209071.14	168075.73	19.61

Reiterando os resultados supracitados, podemos ver uma comparação entre a situação antes e depois das novas regras de negócio que deram origem a variáveis importantes para o modelo.

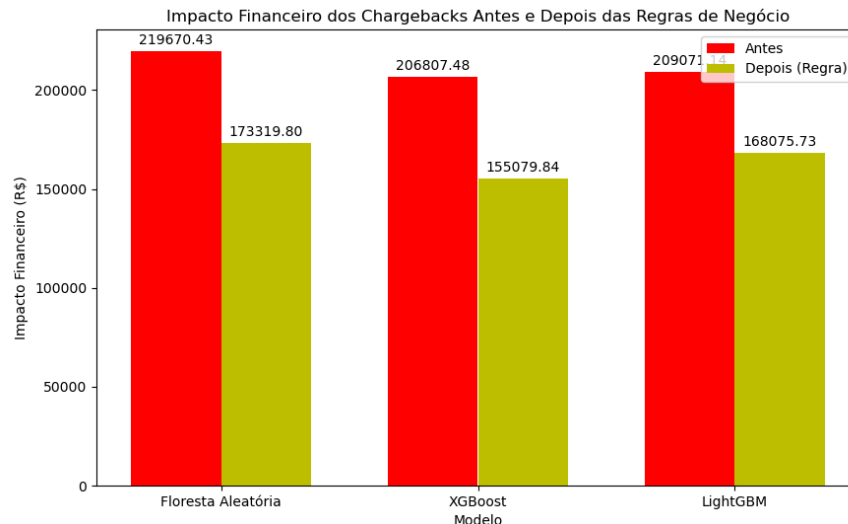


Figura 7 - Comparativo de perdas (em R\$) por *chargeback*.

- Referências

- [1] BREIMAN, L. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5-32, 2001.
- [2] CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. p. 785-794, 2016.
- [3] KE, G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: *Advances in Neural Information Processing Systems* 30. p. 3146-3154, 2017.
- [4] DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.
- [5] HAYKIN, S. *Redes Neurais: Princípios e Prática*. 2. ed. Porto Alegre: Bookman, 2001.
- [6] HE, H.; GARCIA, E. A. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, v. 21, n. 9, p. 1263-1284, 2009.
- [7] KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. p. 1137-1145, 1995.
- [8] GUPTA, Palak et al. Unbalanced credit card fraud detection data: a machine learning-oriented comparative study of balancing techniques. *Procedia Computer Science*, v. 218, p. 2575-2584, 2023.
- [9] NGUYEN, Thanh Thi et al. Deep learning methods for credit card fraud detection. *arXiv preprint arXiv:2012.03754*, 2020.

[10] ALARFAJ, Fawaz Khaled et al. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access, v. 10, p. 39700-39715, 2022.