Modelo de Classificação de Risco para clientes do Banco Caja

Descrição do objetivo

O objetivo da análise é identificar o perfil de clientes com risco de inadimplência, montar uma pontuação de crédito através da análise de dados e avaliar o risco relativo, possibilitando assim, classificar os clientes e futuros clientes do Banco Caja em diferentes categorias de risco com base na sua probabilidade de inadimplência.Esta classificação permitirá ao banco tomar decisões informadas sobre a quem conceder crédito, reduzindo assim o risco de empréstimos não reembolsáveis. Além disso, a integração destas métricas fortalecerá a capacidade do modelo de identificar riscos, contribuindo para a solidez financeira e a eficiência operacional do Banco.

Ferramentas

As limpezas, transformação de informações e criação de novas variáveis foi realizada no BigQuery. A visualização dos dados sob a forma de tabelas dinâmicas e gráficos foi realizada no Looker Studio que também serviu como ferramenta para montagem do dashboard com as principais informações a serem apresentadas na conclusão da análise.

Pré-processamento de dados

O Banco Caja forneceu um conjunto de dados que contém dados sobre empréstimos concedidos a um grupo de clientes do banco. Os dados estão divididos em 4 tabelas, a primeira com dados do usuário/cliente, a segunda com dados do tipo empréstimo, a terceira com o comportamento de pagamento desses empréstimos, e a quarta com a informação dos clientes já identificados como inadimplentes. Os dados foram avaliados quanto a ausência de informações ou presença de informações faltantes. Foram identificados 7199 clientes que não informaram o valor do último salário e 943 que não forneceram informações sobre a quantidade de dependentes. A fim de melhorar a precisão do perfil traçado, esses clientes foram desconsiderados e a base que contava com 36.000 clientes passou a ser composta por

28.848 clientes com todas as informações. Por questões éticas a informação de sexo dos clientes foi desconsiderada nas análises. Dados no formato de string foram padronizados a fim de evitar que a mesma informação fosse contabilizada como diferente dentro de uma mesma coluna.

Durante a avaliação de outliers para a variável taxa de endividamento foram encontrados valores muito altos e discrepantes dos outros. Como forma de reduzir o impacto desses outliers no restante dos dados, foram removidos os clientes que possuíam taxa de endividamento superior a 10.000 contabilizando um total de 47 usuários e culminando no total de 28.801 clientes a serem considerados na construção do modelo de risco. Uma nova variável foi criada para contabilizar a quantidade de empréstimos realizada por cada cliente.

Métodos e técnicas

Foram feitos cálculos de correlação entre as variáveis a fim de entender como estavam relacionadas umas às outras e identificar possíveis redundâncias nos dados. Como haviam 3 variáveis que traziam informações sobre atrasos de 30, 60 e 90 dias respectivamente as 3 estavam se sobrepondo em relação aos demais dados e optou-se por manter apenas a variável que contabilizava os clientes que atrasaram pagamentos por mais de 90 dias.

Foi feita a segmentação de cada uma das variáveis em 4 quartis para avaliar os valores médios e perfil dos clientes dentro de cada um desses quartis. A correlação para todas as combinações de variáveis e quartis foi calculada e encontrou-se uma correlação de 0,83 entre os quartis de total de empréstimos e atrasos com mais de 90 dias.

Avaliando mais a fundo, percebe-se que a decisão de manter apenas a coluna com o maior atraso no pagamento talvez não tenha sido a melhor, porque a grande maioria dos clientes nunca atrasaram mais do que 90 dias. Com isso, mesmo no quartil 4 tenho clientes que nunca tiveram atrasos superiores e essa coluna acaba não sendo uma boa opção para basear meus cálculos. Diante disso, foi realizado o cálculo de risco relativo para todas as demais variáveis do banco, ainda que não se tenha encontrado uma boa correlação entre elas. A partir do risco relativo das variáveis foi calculado o score de risco dos clientes, desconsiderando a variável de atrasos superior a 90 dias.

O score de risco relativo foi calculado considerando a taxa de risco das variáveis idade, número de dependentes, salário e taxa de endividamento. As variáveis using_lines_not_secured_personal_assets (representada como lines na tabela 1) e

more_90_days_overdue (representada como days na tabela 1) foram desconsideradas no cálculo de risco relativo por apresentaram taxas muito altas e discrepantes das demais, o que promoveria um grande deslocamento no score de risco (tabela 1).

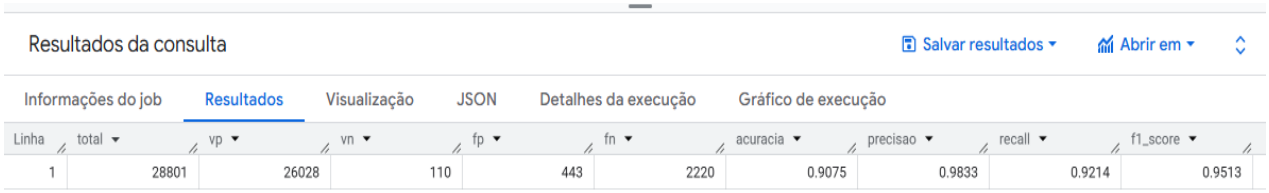
	quartil_base / risco_relativo			
variavel	1	2	3	4
days	0,05	0,05	0,08	47,27
lines	0,03	0,01	0,18	40,66
loan	2,44	0,87	0,59	0,58
age	2,14	1,24	0,76	0,29
salary	2	1,25	0,83	0,29
dependents	1,15	0,45	1,17	1,37
debt	0,78	0,89	0,75	1,71

Tabela 1: taxa de risco por variável e quartil

Além disso, como a esmagadora maioria dos clientes nunca atrasou mais que 90 dias, há em todos os quartis dessa variável clientes que nunca atrasaram. Com isso, considera-se que não seja uma boa variável para usar na composição do score de risco. A variável que contabiliza a quantidade de empréstimos por cliente também foi desconsiderada por apresentar dado ilógico em que o quartil 1, que é composto por clientes com menor quantidade de empréstimos, possui a maior taxa de risco.

O score de risco do cliente foi calculado como score simples em que o seu valor corresponde à soma da taxa de risco de cada uma das quatro variáveis consideradas. A classificação dos clientes em bons e maus pagadores foi feita a partir do valor de score de risco e o valor de corte que divide os dados entre bons e maus pagadores foi estabelecido com base no valor do score de risco em relação a variável default_flag, que informa o histórico de inadimplência ou adimplência dos clientes. Olhando para a coluna default_flag, avaliou-se a distribuição do score de risco para os inadimplentes (default _flag =1). Como a média de score de risco para inadimplente é 5,08 e o desvio padrão é 1,07, considerou-se o valor de 6,1 como ponto de corte, pois considera a média mais o desvio padrão dos dados. Os clientes foram classificados em bons ou maus pagadores com base nesse ponto de corte e essa classificação foi transformada em uma variável binária onde 0 indica bons pagadores e 1 indica maus pagadores e relacionadas na coluna flag_pagador.

Como forma de validar a classificação dos clientes, foram utilizadas como parâmetro as informações contidas na coluna default_flag. A partir da comparação dos valores mostrados nas duas colunas (default_flag e flag_pagador) para cada cliente, foi construída uma matriz de confusão que permitiu calcular o desempenho do modelo em relação a sua capacidade de classificar corretamente os clientes. O desempenho foi medido para acurácia, precisão, recall e F1_Score (tabela 2).



Linha	total	vp	vn	fp	fn	acuracia	precisao	recall	f1_score
1	28801	26028	110	443	2220	0.9075	0.9833	0.9214	0.9513

Tabela 2: cálculo de desempenho para o modelo de risco

Conclusões

Com base na avaliação das taxas de risco para as variáveis podemos concluir que clientes com idade de até 41 anos possuem um risco 2 vezes maior de se tornarem inadimplentes. Clientes com taxa de endividamento a partir de 0,48 possuem um risco 1,71 vezes maior de se tornarem inadimplentes. Clientes que possuem a partir de 2 dependentes têm um risco 1,37 vezes maior de se tornarem inadimplentes, Clientes cujo último salário foi de até R$3.400,00 possuem duas vezes mais risco de se tornarem inadimplentes. Clientes que utilizam a partir de 0,57 do seu limite de crédito, em linhas que não são garantidas por bens pessoais possuem um risco 40x maior de se tornarem inadimplentes. Clientes que atrasaram o pagamento por mais de 90 dias possuem um risco 40 vezes maior de se tornarem inadimplentes do que aqueles que nunca atrasaram.

O modelo de risco desenvolvido classificou de forma correta 92,2% dos clientes adimplentes como bons pagadores e 7,8% dos clientes adimplentes como maus pagadores. Por outro lado, classificou apenas 19,9% dos inadimplentes como maus pagadores, enquanto 80,1% deles ficaram classificados como bons pagadores.

Em número gerais, o banco de dados teve um aumento de 6,3% no número de maus pagadores, indo de 553 clientes com histórico de inadimplência para um total de 2330 maus pagadores. A partir disso, podemos estabelecer uma taxa de inadimplência prevista em 8,09%. A taxa de inadimplência pode ser considerada durante a decisão de concessão de crédito aos atuais ou novos clientes e ajudar na previsão de possíveis prejuízos em relação à inadimplência.

Considerações

Me parece que a divisão por quartis não é a melhor forma de segmentar os dados, pois fornece um ponto de corte que reflete a quantidade de dados ordenados e não considera os valores dos dados em si. Dessa forma, clientes com a mesma característica para uma dada variável (por exemplo, clientes que nunca atrasaram o pagamento) podem ficar em quartis distintos e com isso a classificação não é feita da mesma forma para os dois.

Apesar dos excelentes resultados obtidos nos teste de desempenho do modelo, pode-se perceber que o mesmo parece ser melhor em classificar bons pagadores do que maus pagadores. A grande discrepância entre a quantidade de clientes adimplentes e inadimplentes no banco de dados pode ter contribuído com a tendência do modelo em classificar melhor os clientes com histórico de bons pagadores. Como os bons pagadores estavam representados em maior quantidade, o modelo se tornou mais sensível a esse perfil do que ao de maus pagadores.

Além disso, acredito que ter realizado a normalização das taxas discrepantes para variáveis de atraso superiores a 90 dias e uso de linhas não seguras de crédito e as incluído na composição do score poderia ter promovido uma precisão maior na classificação dos clientes. Outro viés que pode ter contribuído negativamente foi a utilização de um cálculo de score simples. A utilização de uma investigação mais profunda da relevância das variáveis e o estabelecimento de diferentes pesos para o cálculo de uma média ponderada para score de risco poderia ter resultado em um score de risco mais robusto e fiel à realidade dos dados.

Referências

Os recursos utilizados foram os arquivos e vídeos explicativos disponibilizados dentro de cada um dos passos referentes à realização do projeto 3. Em alguns momentos foram consultados documentos explicativos de projetos anteriores a fim de relembrar alguns conceitos e passos realizados anteriormente. Como auxílio, utilizei o assistente de IA Copilot principalmente para a correção de erros nos comandos do BigQuery ou entender qual caminho poderia seguir para realizar uma determinada consulta.