

RELATÓRIO DE ANÁLISE: SEGMENTAÇÃO DE CLIENTES DE ALTA RENDA

Visão Geral e Objetivo de Negócio

Este projeto teve como objetivo criar uma ferramenta que apoie diretamente a área comercial na identificação de clientes com renda anual acima de US\$ 100 mil, com maior probabilidade de aderir a produtos financeiros de maior valor agregado (fundos exclusivos, previdência e serviços de gestão de patrimônio). Em termos práticos, o modelo permite que as equipes comerciais concentrem esforços nos clientes com maior potencial de aporte, reduzindo disparos genéricos e aumentando a taxa de conversão de clientes de alta renda.

Ao utilizar características de perfil (idade, escolaridade, ocupação, horas trabalhadas, histórico de capital, entre outras), o modelo identifica padrões associados a alta renda e transforma uma base estática de cadastros em uma lista priorizada de clientes, organizada por probabilidade de pertencer ao segmento-alvo.

Tratamento de Dados e Responsabilidade no Uso de Informações

Antes da modelagem, os dados passaram por um processo de limpeza, padronização e criação de variáveis agregadas (por exemplo, faixas etárias, grupos de escolaridade e tipos de vínculo profissional). Essas novas variáveis tornaram o modelo mais aderente à linguagem de negócio (faixas etárias, grupos de profissão) e mais eficiente na diferenciação entre clientes de maior e menor renda.

Ao mesmo tempo, foi tomada uma decisão explícita de não utilizar variáveis sensíveis como raça, sexo e país de origem na etapa preditiva. Essa escolha está alinhada com boas práticas internacionais e com legislações como a LGPD no Brasil e o GDPR na União Europeia, que desestimulam o uso de atributos sensíveis em decisões automatizadas que possam gerar discriminação. Testes internos mostraram que a retirada desses campos praticamente não reduziu o poder de previsão, uma vez que as variáveis econômicas e de emprego já carregam a maior parte das informações relevantes para o objetivo de negócio. Dessa forma, o modelo mantém bom desempenho, respeitando limites éticos e regulatórios.

Comparação dos Modelos Testados

Foram avaliadas diferentes abordagens de modelagem, desde alternativas mais simples e rápidas até modelos mais sofisticados. Entre elas, destacam-se Regressão Logística, KNN, Árvore de Decisão, Random Forest e Gradient Boosting, em versões básica e otimizada.

A principal conclusão é que há um equilíbrio claro entre três dimensões:

- Quantos clientes de alta renda o modelo consegue encontrar (recall);
- Qual a qualidade desses leads (equilíbrio entre acertos e erros – F1 e ROC-AUC);
- Quanto custa, em termos de processamento, treinar e manter cada solução.

Modelos mais simples, como a Regressão Logística, embora rápidos, deixam de identificar uma parcela relevante de clientes de alta renda, o que representa perda de oportunidade comercial. Já soluções mais pesadas, como algumas configurações de SVC, mostraram custo computacional elevado sem ganho proporcional em resultado, o que as torna pouco atrativas para uso em larga escala.

Entre os modelos avaliados, o Gradient Boosting Otimizado apresentou o melhor equilíbrio global: – Maior capacidade de encontrar clientes de alta renda (maior recall); – Melhor combinação entre quantidade e qualidade dos leads (melhor F1-Score); – Maior capacidade de distinguir com clareza clientes de alta e baixa renda (maior ROC-AUC).

Modelos de árvore mais simples, como a Decision Tree, mostraram-se extremamente rápidos e com desempenho competitivo, o que os torna úteis como referência e alternativa leve, mas ainda abaixo do patamar entregue pelo Gradient Boosting Otimizado.

Modelo	Acurácia	Precisão	Recall	F1-Score	ROC-AUC	Tempo (s)
Gradient Boosting (Otimizado)	0,8633	0,757	0,6602	0,7053	0,9212	98,34
Decision Tree (Original)	0,848	0,7088	0,6555	0,6811	0,8984	0,04
Gradient Boosting (Default)	0,8571	0,7704	0,6029	0,6764	0,9124	1,76
Random Forest (Original)	0,8521	0,7972	0,5403	0,6441	0,9161	46,27
KNN (Original)	0,8312	0,6874	0,5837	0,6314	0,8796	45,6
Decision Tree (New)	0,8404	0,7441	0,542	0,6272	0,8874	0,05
KNN (New)	0,8254	0,675	0,5687	0,6173	0,8599	22,37
Logistic Regression (Original)	0,8184	0,7279	0,4257	0,5372	0,8499	3,24
Logistic Regression (New)	0,8027	0,6982	0,3578	0,4731	0,8169	0,53

Escolha do Modelo para o Negócio

Considerando a estratégia comercial da instituição, o maior risco é deixar clientes de alta renda fora do radar. Nessa lógica, é preferível abordar alguns clientes que não se confirmem como alta renda (erro de excesso) do que ignorar potenciais investidores relevantes (erro de omissão). O Gradient Boosting Otimizado é o que melhor atende a essa prioridade: identifica o maior número de clientes de alta renda, mantendo boa qualidade nos leads sugeridos à equipe comercial.

Embora o tempo de treinamento deste modelo seja maior na fase de desenvolvimento, isso não afeta a operação do dia a dia, pois o tempo para avaliar novos clientes é baixo e compatível com uso em ambiente de produção, inclusive em bases amplas. As variáveis derivadas criadas no projeto também se mostraram essenciais para esse resultado, ao traduzir informações brutas em indicadores mais próximos da realidade de negócios.

Conclusão e Próximos Passos

O modelo final selecionado, baseado em Gradient Boosting Otimizado, oferece ao banco uma ferramenta concreta para ampliar a captação de clientes de alta renda, aumentar a eficiência das ações comerciais e fortalecer a estratégia de uso de dados no relacionamento com investidores. Ele permite priorizar esforços, reduzir abordagens pouco relevantes e direcionar produtos adequados ao potencial financeiro de cada cliente.

Recomenda-se avançar com a implantação deste modelo nos sistemas internos da área comercial, com acompanhamento contínuo de desempenho e ajustes periódicos. A coleta de feedback das equipes de atendimento e vendas será fundamental para refinar o uso dos scores gerados, calibrar limiares de corte e, se necessário, ajustar o modelo em futuras versões.

Esse projeto representa um passo concreto na direção de uma tomada de decisão mais orientada por dados, com impacto direto em receita, eficiência operacional e qualidade da experiência do cliente.