Metodologia de Análise de dados

Definição do Problema | Análise Eploratória | **Modelagem Estatística** | Algoritmos de Inteligência Artificial | Conclusões

Planejamento

- Objetivos
- Conceitos
- Critérios
- Histórico de dados
- Variáveis

Análise Preliminar

- Medidas resumo
- Análise univariada e bivariada
- Correlações
- Redução de variáveis

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- Randon Forest
- SVM
- Gradient Boosting Tree
- XGBoosting
- LightGBM
- CatBoost

Definição da técnica

- Validação dos resultados com negócios
- Escolha da técnica que melhor se adeque ao uso e estratégias da área de negócio

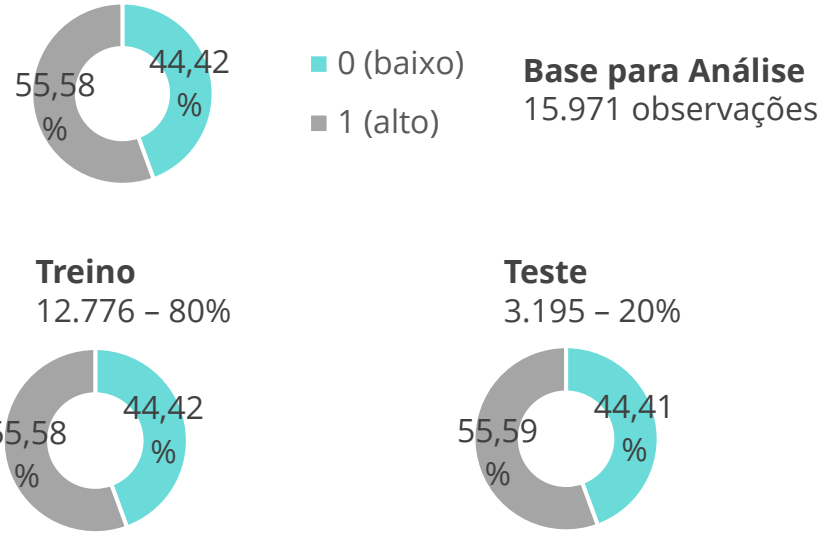
LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

Desempenho dos modelos

MODELAGEM COM ESTATÍSTICA TRADICIONAL | COMPARAÇÃO ENTRE TÉCNICAS

Tratamento das bases de dados para modelagem

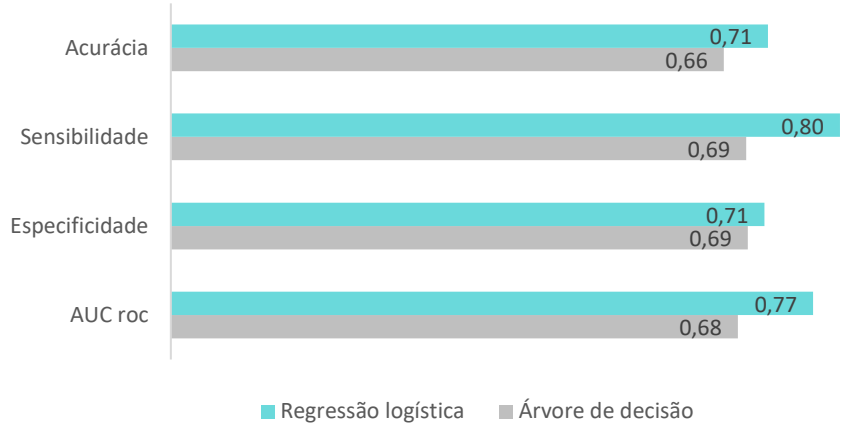
1. Balanceamento da resposta: amostra aleatória
2. 80% aleatório para treino e 20% para teste



3. Regressão Logística

4. Árvore de decisão

- Ambas as técnicas apresentaram **bom acerto** preditivo, com acerto geral de 68%.
- A **Regressão logística** apresentou uma quantidade maior de acertos de todos os casos que de fato eram alto risco (71%), frente a 66% da **Árvore de Decisão**.
- A **Regressão logística** também apresentou **mais estabilidade** quando comparamos os desempenhos das bases de treino e teste.
- A AUC da Regressão logística de 0.77 significa que a classe positiva tem uma chance de 0.77 de receber um score maior do que da classe negativa



Regressão Logística

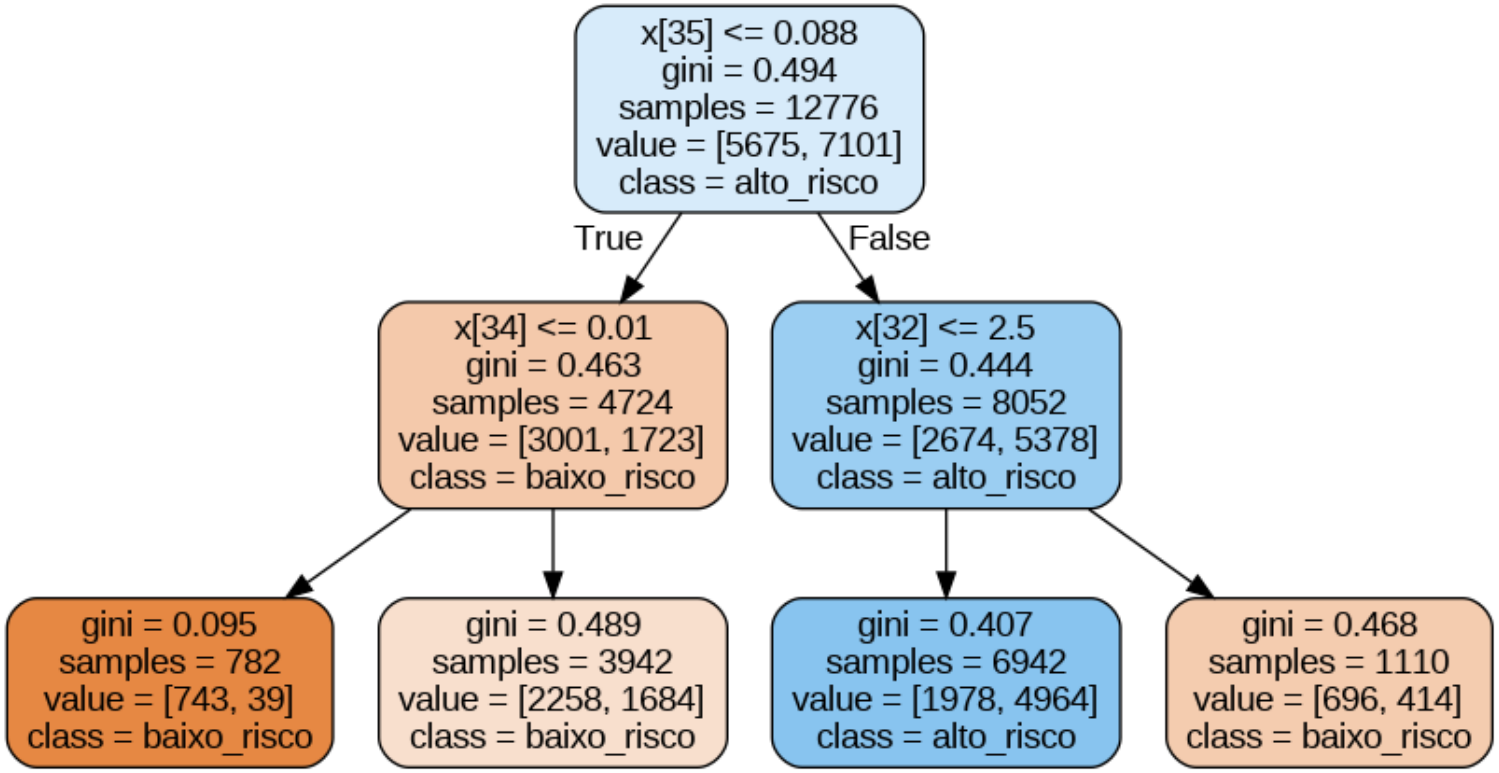
O modelo seleciona as variáveis mais relevantes e estima um peso para cada uma de suas categorias, atribuindo para cada grupo de imóveis de consórcio a probabilidade de ser de alto risco.

$$p = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4}}$$

Variável	Coeficiente (β)	Interpretação em relação ao Risco
Intercepto	-0.61	Quanto maior o intercepto, menor a propensão ao risco
Prazo do grupo em meses	-1.87	Quanto maior o prazo, menor a propensão ao risco.
Valor médio do bem	-2.19	Quanto maior o prazo, menor a propensão ao risco.
Taxa de Administração	2.94	Quanto maior a taxa, **maior** o risco.
Quantidade de cotas ativas em dia	-4.91	Quanto maior a quantidade de cotas ativas em dia, menor a propensão ao risco.
Quantidade de cotas ativas contempladas inadimplentes	-0.95	Quanto maior a quantidade de cotas contempladas inadimplentes, menor a propensão ao risco.
Quantidade de cotas ativas não contempladas inadimplentes	17.23	Quanto maior a quantidade de cotas não contempladas inadimplentes, **maior** a propensão ao risco.
Quantidade de cotas ativas contempladas no mês	2.16	Quanto maior a quantidade de cotas contempladas, **maior** a propensão ao risco.
Quantidade de cotas ativas quitadas	-18.48	Quanto maior a quantidade de cotas quitadas, menor a propensão ao risco.
Quantidade de cotas ativas com crédito pendente de utilização	-2.16	Quanto maior a quantidade de cotas com crédito pendente de utilização, menor a propensão ao risco.
Percentual de inadimplência	1.13	Quanto maior a inadimplência, **maior** o risco.
Receita potencial	2.82	Quanto maior a receita, **maior** o risco.

Árvore de decisão

Classifica as observações pela combinação de características, por meio de uma árvore de classificação, que explique o evento de Risco alto.



- **Intepretação:**
  a variável mais importante é a variável idade, seguida pela quantidade de cotas ativas com crédito pendente de utilização e pela quantidade de cotas ativas contempladas no mês.

- **Grupo de menor risco:**
  Com idade menor ou igual a 8,8% do prazo planejado do grupo e quantidade de cotas ativas com crédito pendente de utilização menor ou igual a 0.01.

- **Grupo de maior risco:**
  Com idade menor ou igual a 8,8% do prazo planejado do grupo e quantidade de cotas ativas contempladas no mês menor ou igual do que 2.5.

Desempenho dos modelos

MODELAGEM COM ESTATÍSTICA TRADICIONAL

Regressão Logística

O desempenho do modelo, tanto nas bases de treino como de teste, apresentou ótimo acerto preditivo, próximo a 72% no percentual geral de classificação correta.

O percentual de acerto do evento de alto risco ficou em torno de 72% e para o evento de baixo risco 71,5%.

As bases de treino e validação apresentaram resultados similares, com um pequena variação no acerto na base de teste, dentro do esperado para este tipo de análise.

Árvore de decisão

O desempenho do modelo, tanto nas bases de treino como de teste apresentou ótimo acerto preditivo, próximo a 67% no percentual geral de classificação correta.

O percentual de acerto do evento de risco alto ficou em torno de 71% e 63% para o evento de risco baixo.

Nota-se que há uma **queda muito expressiva nos indicadores da base de teste**, que deve ser investigada no sentido de tentar utilizar outras técnicas que contornem um possível problema de overfitting no treino.

Base de Treino	Base de Teste
0.7193	0.7142
0.8090	0.8001
0.7204	0.7180
0.7837	0.7681

Indicador
Percentual de classificação Correta
Sensibilidade
Especificidade
AUC Roc

Base de Treino	Base de Teste
0.6779	0.6613
0.6991	0.6881
0.7151	0.6983
0.6961	0.6781

Metodologia de Análise de dados

Definição do Problema → Análise Eploratória → Modelagem Estatística → **Algoritmos de Inteligência Artificial** → Conclusões

Planejamento

- Objetivos
- Conceitos
- Critérios
- Histórico de dados
- Variáveis

Análise Preliminar

- Medidas resumo
- Análise univariada e bivariada
- Correlações
- Redução de variáveis

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- Randon Forest
- SVM
- Gradient Boosting Tree
- XGBoosting
- LightGBM
- CatBoost

Definição da técnica

- Validação dos resultados com negócios
- Escolha da técnica que melhor se adeque ao uso e estratégias da área de negócio

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

Desempenho dos modelos

Modelagem com inteligência artificial

Catboost é uma biblioteca desenvolvida em 2017 pela Yandex, que fornece uma outra implementação do algoritmo *Gradient Boosting*, um método *Ensemble* que combina vários preditores fracos em um forte (preditor fraco significa que sua classificação é apenas um pouco melhor do que adivinhações aleatórias).
A ideia geral é treinar sequencialmente os preditores, cada um tentando corrigir seu antecessor.
Esse algoritmo possui dois grandes objetivos: evitar o *overfitting* e de forma padrão oferecer bons hiperparâmetros.

Modelos	Acuracidade	Precisão	Recall	f1	Roc_auc
CatBoost	**0.8553**	**0.8624**	**0.8803**	**0.8712**	**0.9362**
LightGBM	0.8545	0.8633	0.8773	0.8701	0.9360
XGBoosting	0.8510	0.8601	0.8741	0.8670	0.9334
Random Forest	0.8502	0.8589	0.8742	0.8664	0.9297
Gradient Boosting	0.8461	0.8503	0.8778	0.8638	0.9276
SVM	0.8142	0.8307	0.8364	0.8334	0.8970
Regressão Logística	0.7674	0.7722	0.8248	0.7976	0.8458
Árvore de Decisão	0.7981	0.8217	0.8134	0.8175	0.7962

O melhor modelo foi o **Catboost**, com o **maior Roc_auc** e **melhor acurácia;**

Os modelos tradicionais (Regressão Logística e Árvore de decisão) tiveram uma melhora no desempenho frente ao observado anteriormente (0,77 e 0,68), principalmente pelo inclusão da técnica de **StratifiedKFold** no novo processamento.

Todos os modelos apresentaram um **Roc_auc** bom, pois um classificador com precisão de apenas 0.5 não é melhor do que adivinhar aleatoriamente.

Metodologia de Análise de dados

Planejamento

- Objetivos
- Conceitos
- Critérios
- Histórico de dados
- Variáveis

Análise Preliminar

- Medidas resumo
- Análise univariada e bivariada
- Correlações
- Redução de variáveis

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- Randon Forest
- SVM
- Gradient Boosting Tree
- XGBoosting
- LightGBM
- CatBoost

Definição da técnica

- Validação dos resultados com negócios
- Escolha da técnica que melhor se adeque ao uso e estratégias da área de negócio

Conclusões

- Acurácia: **Catboost (85,5%)** apresentou melhor desempenho em comparação com o **LightGBM (85,4%)**

  Shap Values: revelou as features mais importantes do Catboost.
  - **Quantidade de cotas excluídas**
  - **Numero_da_assembleia_geral_ordinária.**

- Ambos os modelos podem servir como base para predizer grupos de alto risco de consórcio de imóveis.

  Exemplo: alimentação de um **trigger para geração de alertas as áreas de negócios**.

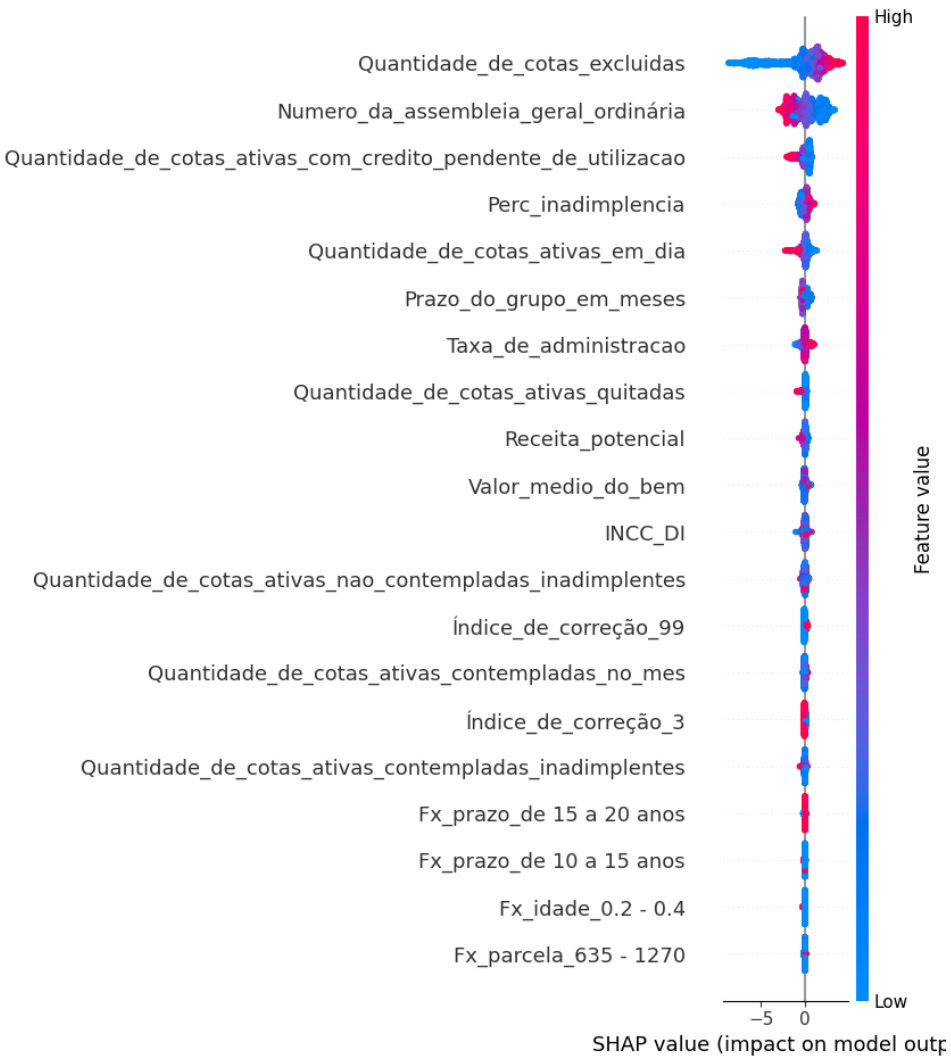
  Squad de vendas: buscar novos consorciados para reposição em grupos específicos com alta taxa de cancelamento;

  Squad de Cancelamentos: fazer ações de retenção com os clientes dos grupos de maior risco:
  - propor alternativas de redução da carta de crédito,
  - renegociação da dívida pelo prazo restante,
  - ou até mesmo a recompra da cota pela Administradora.

Com isso buscar a melhoria contínua e garantir a sustentabilidade/perenidade do negócio.

Shap Values - Feature importance

Modelagem com inteligência artificial



Variáveis + importantes obtidas pelo modelo Catboost

X

Variáveis que pareciam explicar o risco na AED

	Variável	Intepretação em relação ao Risco
	Índice de correção	6, 99, 3
	Faixas de Parcelas	Valores mais baixos
	Faixas de prazos	Prazos mais longos
	Faixas de Idade	Grupos mais novos
1	**Quantidade de cotas excluídas**	Maior quantidade de cancelamentos
	Taxa de Administração	Taxas maiores
2	**Número da assembleia geral ordinária**	Menor número de assembléias
	Var_cotas_excluidas	Variações grandes de cotas excluídas
4	**Perc_Inadimplencia**	Percentuais maiores de inadimplência
	Perc_Churn	Percentuais maiores de churn

Sugestão para Trabalhos Futuros

Realizar a **experimentação de outros modelos;**

Explorar **técnicas de otimização dos hiperparâmetros** dos modelos;

Buscar uma **quantidade maior de dados retroativos** e também adicionar **novas variáveis;**