Modelo preditivo de estudar em um Curso de Graduação

LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

14/10/2021

MBA Analytics e Inteligência Artificial – ADMAI7

Nome do Aluno:
Gustavo Sanches Oliveira

Coordenadores:
Prof.ª Dr.ª Alessandra de Álvila Montini
Prof. Dr. Adolpho Walter Pimazoni Canton

Agenda

1. Objetivo do Trabalho
2. Contextualização do Problema
3. Base de Dados
4. Análise Exploratória de Dados
5. Modelagem com Estatística Tradicional
6. Conclusões (Preliminares)

Metodologia de análise de dados

Definição do problema

- Objetivos
- Contextualização
- Histórico de dados
- Variáveis

Análise preliminar

- Medidas resumo
- Análise de frequências
- Gráficos
- Análise de *outliers*
- Análise de *missings*

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- SVM
- Random Forest
- Boosting

Definição da técnica

- Escolha da técnica que melhor se adeque ao uso e estratégias da área de negócios

Metodologia de análise de dados

Planejamento | Análise Exploratória | Modelagem Estatística Tradicional | Algoritmos de Inteligência Artificial | Conclusões

Definição do problema

- Objetivos
- Contextualização
- Histórico de dados
- Variáveis

Análise preliminar

- Medidas resumo
- Análise de frequências
- Gráficos
- Análise de *outliers*
- Análise de *missings*

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- SVM
- Random Forest
- Boosting

Definição da técnica

- Escolha da técnica que melhor se adeque ao uso e estratégias da área de negócios

1. Objetivo do Trabalho

O objetivo do trabalho é **predizer a probabilidade de uma pessoa estudar em um curso de ensino superior de graduação.**

A predição será realizada utilizando **a pesquisa de orçamentos familiares (POF) de 2017-2018, modelos estatísticos** e **algoritmos de Inteligência Artificial**, que selecionarão as **características mais relevantes** que explicam o evento em questão.

Desta forma, o Ministério da Educação poderá **traçar estratégias de incentivo** para diferentes públicos ou **validar se as ações em prática** estão bem direcionadas.

2. Contextualização do Problema

O IBGE realiza de tempos em tempos a **Pesquisa de Orçamentos Familiares - POF**. O objetivo da pesquisa é avaliar as estruturas de consumo, de gastos, entre outras a partir da análise dos orçamentos domésticos. A pesquisa é realizada por amostragem e tem como unidade de investigação o domicílio. Todos os residentes do domicílio são entrevistados.

Um dos pontos avaliados pela pesquisa é se **os residentes do domicílio frequentam a escola e qual o tipo do curso,** desta forma é possível identificar quem frequenta um curso de graduação.

Hoje o sonho de muitas pessoas é poder cursar uma faculdade e obter um diploma mas muitas pessoas não conseguem ingressar em um curso de graduação.

Dado este cenário o Ministério da Educação solicitou a **construção de um modelo preditivo com o objetivo de identificar o perfil das pessoas que frequentam um curso de graduação** para criar iniciativas de incentivos e validar se as iniciativas em execução estão sendo bem direcionadas.

3.i. Base original

Visão da base

- Residente do domicílio

Filtros de inclusão

- Maiores de 17 anos (IDADE >= 17) – Menores que 17 anos não deveriam estar em um curso de graduação
- Possuem ensino médio completo ou superior incompleto (INSTRUCAO in (5,6)) – Pessoas que não tem o ensino médio completo não podem entrar no ensino superior, para a análise também não posso considerar pessoas que já tem uma graduação completa.
- Possuem 12 anos ou mais de estudo – 12 anos é a quantidade de anos de estudo de quem se formou no ensino médio

Variável Resposta

- Será criada uma variável chamada FREQ_GRADUACAO que será 1 para aqueles que vão na escola e que o tipo do curso é “Superior – Graduação”, os demais serão 0.

  Código: VAI_NA_ESCOLA == 1 & TIPO_CURSO == 8 then FREQ_GRADUACAO == 1
  Else FREQ_GRADUACAO == 0

3.ii. Filtros

Base Original
178.431 entrevistados

Base de partida

Base original da POF com 178.431 entrevistados e a base possui 57 variáveis

Filtrando maiores de 17 anos
134.054 entrevistados

Base filtrada possui 134.054 entrevistados, sendo que 5% deles frequentam um curso de ensino superior

Filtrando instrução
34.050 entrevistados

Base filtrada possui 34.050 entrevistados, sendo que 9% deles frequentam um curso de ensino superior

Filtrando pessoas com mais de 11 anos de estudo
27.543 entrevistados

Base filtrada possui 27.543 entrevistados, sendo que 11% deles frequentam um curso de ensino superior

Base Modelagem
4.712 entrevistados

A base foi balanceada para treino e validação dos modelos, resultando em 4.712 entrevistados

3.iii. Todas as variáveis da Base

A base de dados apresenta **57 variáveis**

Variáveis POF

- Estrato_POF
- Cod_UPA
- Num_dom
- Num_uc
- Cod_informante
- Peso
- Peso_final

Variáveis do Entrevistado

- UF
- Tipo_situacao_reg
- Grau_parentesco
- Morador_presente
- Dia_nasc
- Mes_nasc
- Ano_nasc
- Idade
- Sexo
- Cor_raca
- Tem_plano_saúde
- Trabalhou_ult_12m
- Gastos_sem_renda
- Renda_total
- Composicao
- PC_rend_disp
- Pc_renda_monet
- Pc_renda_nao_monet
- Pc_deducao

Variáveis Bancárias

- Qtd_cartaocred
- Qtd_contacorr
- Qtd_chequeesp
- Uso_chequeesp_90d
- Qtd_contapoup

Variáveis Escolares

- Sabe_ler_escrever
- Vezes_escola_1semana
- Toma_café_manha_escola
- Cafe_manha_escola
- Toma_lanche_escola
- Lanche_escola
- Toma_almoco_escola
- Almoco_escola
- Toma_jantar_escola
- Jantar_escola
- Tipo_escola
- Tipo_curso
- Duracao_curso
- Tipo_período_curso
- Ano_serie_curso
- Possui_curso_superior
- Ja_freq._escola
- Curso_mais_elevado_ant
- Duracao_curso_ant
- Tipo_período_curso_ant
- Conc_1periodo_curso_ant
- Utl_período_conc_curso_ant
- Curc_curso_ant
- Anos_estudo
- Instrucao
- Vai_na_escola

Metodologia de análise de dados

Planejamento | Análise Exploratória | Modelagem Estatística Tradicional | Algoritmos de Inteligência Artificial | Conclusões

Definição do problema

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

Análise preliminar

- Medidas resumo
- Análise de frequências
- Gráficos
- Análise de *outliers*
- Análise de *missings*

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- SVM
- Random Forest
- Boosting

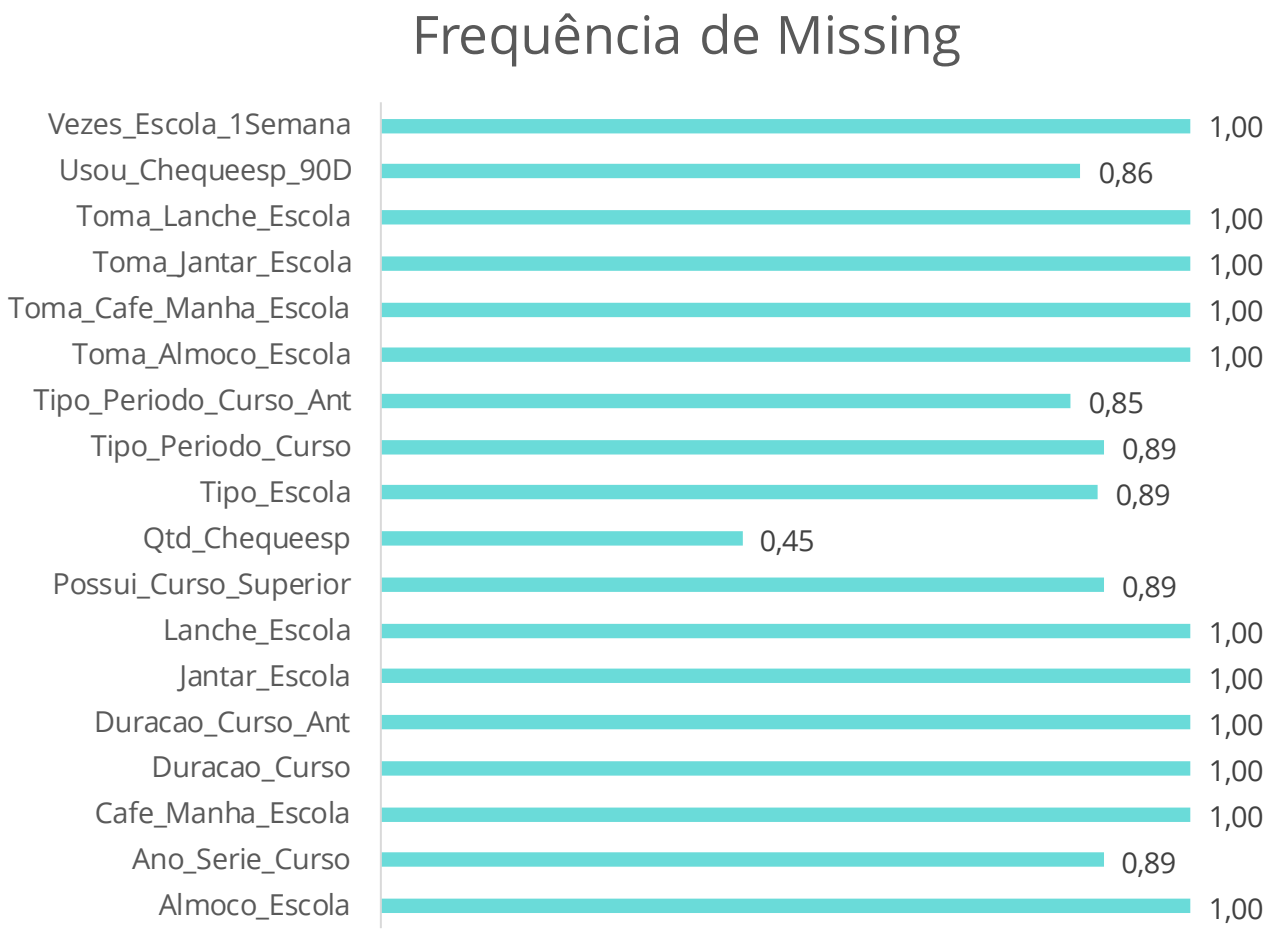
Definição da técnica

- Escolha da técnica que melhor se adeque ao uso e estratégias da área de negócios

4.i. Análise de Missing

ANÁLISE EXPLORATÓRIA

18 variáveis explicativas foram removidas por ter alta quantidade de missing.



4.ii. Correlação Linear

ANÁLISE EXPLORATÓRIA

Variável Pc_Renda_Monet foi removida por ter alta correlação com a variável **Pc_Renda_Disp**

	Idade	Qtd_ Cartaocred	Qtd_ Contacorr	Qtd_ Contapoup	Anos_ Estudo	Renda_ Total	Pc_Renda_ Disp	Pc_Renda_ Monet	Pc_Renda_ Nao_Monet	Pc_ Deducao
Idade	1,00	0,15	0,15	0,02	-0,08	0,10	0,21	0,20	0,13	0,15
Qtd_Cartaocred		1,00	0,36	0,19	0,10	0,14	0,19	0,19	0,08	0,16
Qtd_Contacorr			1,00	0,12	0,16	0,17	0,23	0,24	0,09	0,20
Qtd_Contapoup				1,00	0,03	0,03	0,07	0,08	0,03	0,05
Anos_Estudo					1,00	0,16	0,17	0,17	0,08	0,16
Renda_Total						1,00	0,60	0,61	0,19	0,39
Pc_Renda_Disp							1,00	0,95	0,44	0,46
Pc_Renda_Monet								1,00	0,18	0,59
Pc_Renda_Nao_Monet									1,00	0,12
Pc_Deducao										1,00

- Variáveis Pc_Renda_Disp e Pc_Renda_Monet possuem alta correlação
- Podemos utilizar apenas uma das informações, uma vez que a correlação entre elas é de 0.95 (altíssima correlação).
- Usaremos a **Pc_Renda_Disp** para efeitos de análise.

4.iii. Outras motivos de remoção

ANÁLISE EXPLORATÓRIA

20 variáveis explicativas foram removidas por outros motivos

Detalhes da remoção

Motivos:

- Variáveis só vem preenchidas, ou quase, quando a variável resposta é 0
- Variáveis utilizadas na construção da variável resposta
- Variáveis removidas porque não entendi o significado
- Variáveis removidas porque já temos a variável IDADE
- Variáveis removidas porque é de identificação dos entrevistados
- Variável “Tipo_Situacao_Reg” removida porque é um resumo da variável “Estrato_Pof”
- Variável “Instrucao” removida porque tem relação com a variável “Anos_Estudo”
- Variável “Grau_Parentesco” removida porque não faz diferença qual o grau de parentesco do entrevistado com a pessoa de referência da unidade de consumo
- Variável “Morador_Presente” removida porque não fará diferença saber se o morador estava presente na hora da entrevista
- Variável “Sabe_Ler_Escrever” removida porque todos na base sabem ler e escrever

4.iv. Principais Variáveis

Após as remoções, a base de dados final apresenta **17 variáveis explicativas** e a **variável resposta**.

Variáveis do Entrevistado

- UF
- Estrato_POF
- Idade
- Sexo
- Cor_raca
- Tem_plano_saude
- Trabalhou_ult_12m
- Gastos_sem_renda
- Renda_total
- Composicao
- PC_rend_disp
- Pc_renda_nao_monet
- Pc_deducao

Variáveis Bancárias

- Qtd_cartaocred
- Qtd_contacorr
- Qtd_contapoup

Variável Escolar

- Anos_estudo

Variável Resposta

Freq_Graduacao:
1 = Sim
0 = Não

4.v. Raio-X da base

ANÁLISE EXPLORATÓRIA | UNIVARIADA

Variáveis do Entrevistado

- UF
- Estrato_POF
- Idade
- Sexo
- Cor_raca
- Tem_plano_saude
- Trabalhou_ult_12m
- Gastos_sem_renda
- Renda_total
- Composicao
- PC_rend_disp
- Pc_renda_nao_monet
- Pc_deducao

Persona

- A maior parte dos entrevistados são representados por **mulheres (53%)**
- É um base de **pessoas de meia idade,** sendo a **metade** dos clientes **abaixo de 37 anos**
- A maior parte deles está concentrada na **região Sudeste** sendo que **SP (9%)** é o estado mas representativo seguido de **MG (7%)** e **RJ (7%)**
- São pessoas que moram em **regiões urbanas** de cidades do interior
- Grande parte **não possui plano de saúde** e **trabalharam nos últimos 12 meses**
- Maior parte das residências dos entrevistados é composta por **mais de um adulto (81%)**
- São pessoas que tem despesas mesmo sem rendimento
- A maior parte dos entrevistados é de **classe média** (Renda_total, que é a renda familiar, entre R$2.674 e R$9.897) – Fonte: http://g1.globo.com/economia/seu-dinheiro/noticia/2013/08/veja-diferencas-entre-conceitos-que-definem-classes-sociais-no-brasil.html

Detalhes das análises

4.v. Raio-X da base

ANÁLISE EXPLORATÓRIA | UNIVARIADA

Variáveis Bancárias

- Qtd_cartaocred
- Qtd_contacorr
- Qtd_contapoup

Variáveis Bancárias

Todas as informações bancárias possuem **distribuições assimétricas acentuadas à direita,** com a presença de **outliers superiores.**
75% dos entrevistados possuem 1 cartão de crédito, 50% possuem 1 conta corrente e 75% possuem uma conta poupança

Detalhes das análises

4.v. Raio-X da base

ANÁLISE EXPLORATÓRIA | UNIVARIADA

Variável Escolar

- Anos_estudo

Variável Escolar

Cerca de **metade dos entrevistados possuem 12 anos de estudo** o que é condizente com o tempo médio necessários para se formar no ensino médio.

Detalhes das análises

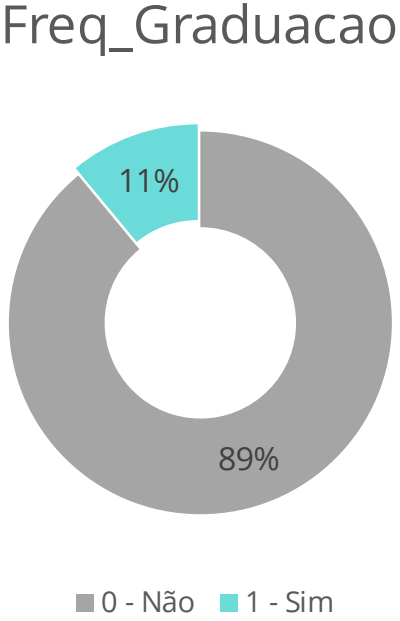
4.v. Raio-X da base

ANÁLISE EXPLORATÓRIA | UNIVARIADA

Dos 27.543 entrevistados da base, 11% frequentam um curso de ensino superior

Variável Resposta

Freq_Graduacao
1 = Sim
0 = Não



O percentual real na base de dados é de 4% mas após os filtros de elegibilidade o percentual foi para 11%.

4.vi. Análise das covariáveis x Target

Principais variáveis que explicam o target:

Variável	Interpretação em relação ao target
UF	MT e AC
Idade	Pessoas mais jovens
Trabalhou_Ult_12M	Não trabalharam nos últimos 12 meses
Gastos_Sem_Renda	Não tiveram gastos sem renda
Composição	Mais de um adulto sem criança
Anos_Estudo	Mais de 12 anos de estudo

Detalhes das análises

Metodologia de análise de dados

Definição do problema

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

Análise preliminar

- Medidas resumo
- Análise de frequências
- Gráficos
- Análise de *outliers*
- Análise de *missings*

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

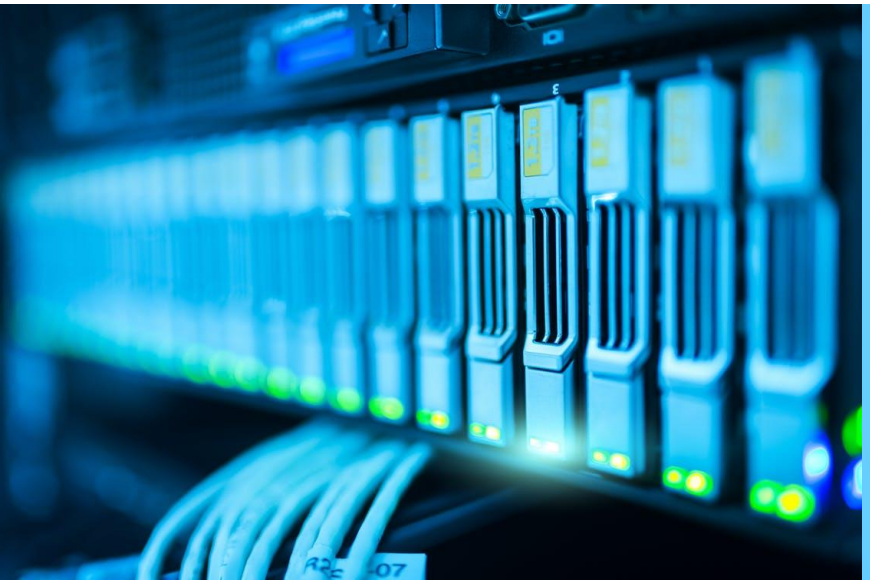
Avaliação das técnicas

- SVM
- Random Forest
- Boosting

Definição da técnica

- Escolha da técnica que melhor se adeque ao uso e estratégias da área de negócios

5.i. Modelagem Estatística Tradicional



Tratamento das base de dados para modelagem

1. 80% aleatório para treino e 20% para teste
   - Treino: 22.034 clientes
   - Teste: 5.509 clientes

2. Balanceamento da resposta na base de treino: amostra aleatória do grupo 0 e total de 1
   - 2356 pessoas que frequentam ensino superior (target =1)
   - 2356 pessoas que não frequentam (target =0)

5.ii. Regressão Logística

O modelo seleciona as variáveis mais relevantes e estima um peso para cada uma de suas categorias, atribuindo para cada cliente a probabilidade de se tornar inadimplente.

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

Variável	Categoria	Coeficiente (β)	Interpretação em relação ao target
Intercepto		-0,07	-
Idade	Valores inteiros de 0 a 111	-1,35	Idades maiores possuem menor propensão
Tem_Plano_Saude	1 - Sim 0 - Não	0,13	Quem tem plano de saúde tem maior propensão
Anos_Estudo	Valores inteiros de 0 a 16. A categoria 16 engloba pessoas com 16 anos ou mais de estudo.	0,90	Quem tem mais anos de estudo possui maior propensão
Pc_Renda_Nao_Monet	Valores em R$	0,24	Quem tem maior renda não monetária possui maior propensão
Pc_Deducao	Valores em R$	0,09	Quem tem maior dedução possui maior propensão

Detalhes das demais variáveis que não entraram no modelo

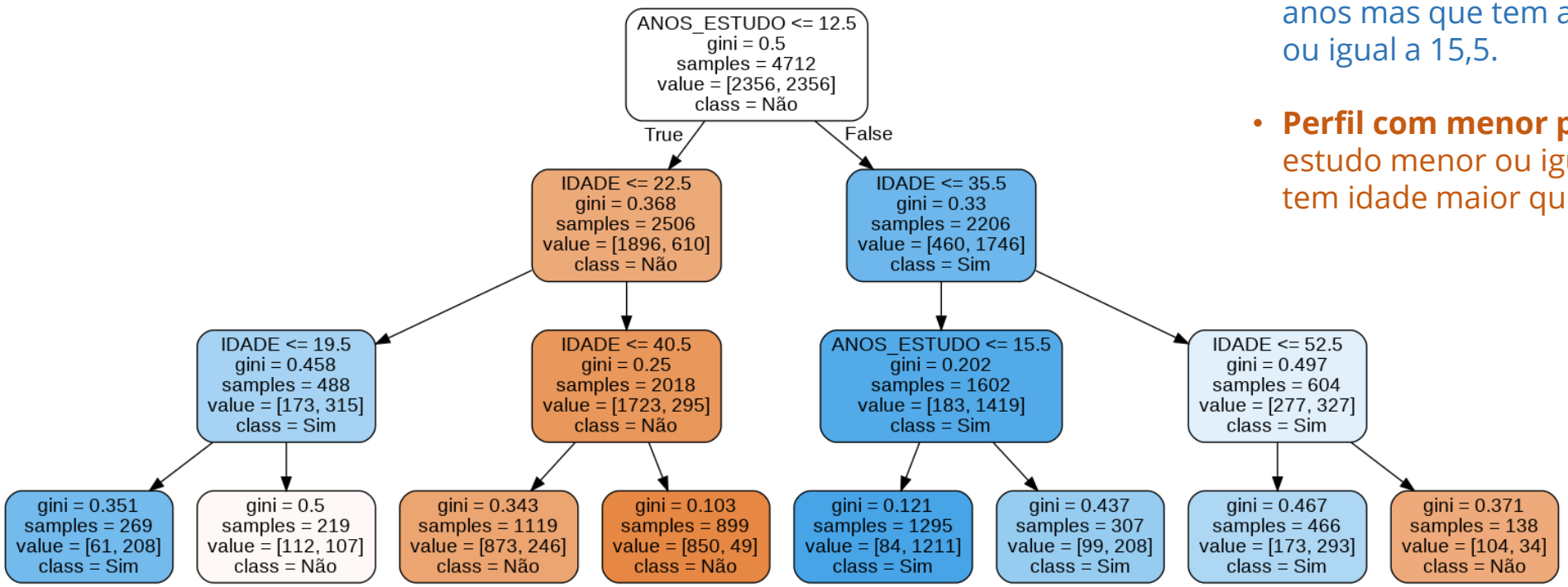
Variáveis removidas na Regressão Logística

Variáveis removidas na ordem abaixo

	Motivo
Composicao	p-value NAN
UF	p-value NAN
Cor_Raca	p-value > 0.05
Estrato_POF	p-value > 0.05
PC_Renda_Disp	p-value > 0.05
Renda_Total	p-value > 0.05
Sexo	p-value > 0.05
Qtd_Cartaocred	p-value > 0.05
Gastos_Sem_Renda	Multicolinearidade
Trabalhou_Ult_12m	p-value > 0.05
Qtd_Contapoup	Multicolinearidade
Qtd_Contacorr	p-value > 0.05

5.iii. Árvore de decisão

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

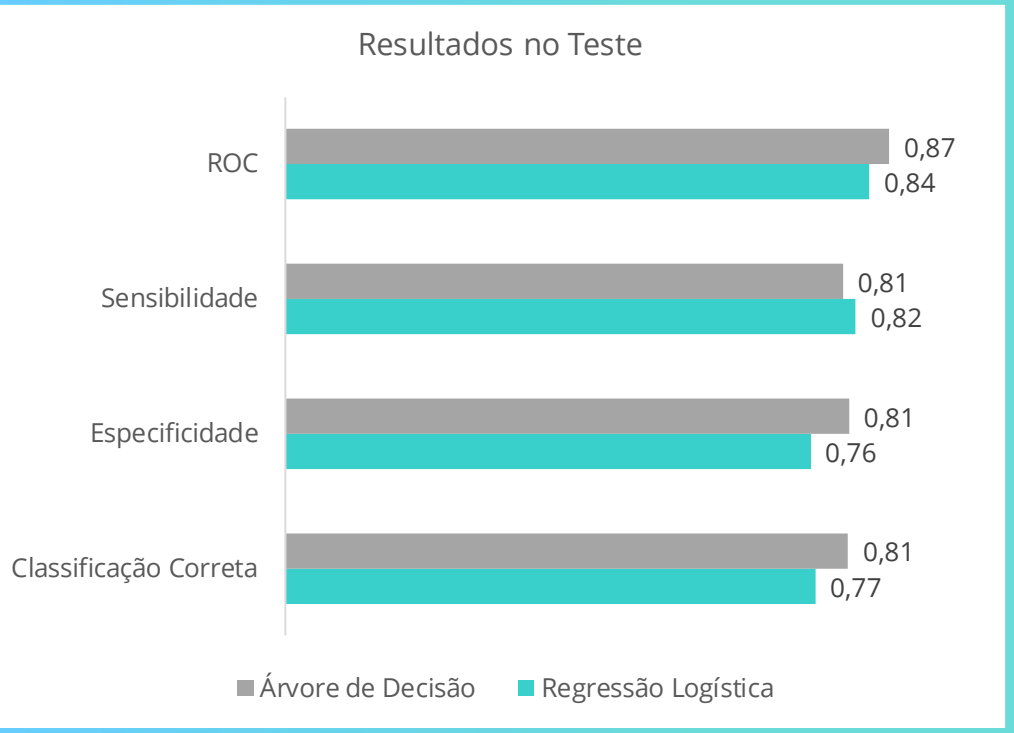


Intepretação:

- A variável mais importante é a variável ANOS_ESTUDO, seguida pela IDADE
- **Perfil com maior probabilidade:** Mais de 12,5 anos de estudo, menores que 35,5 anos mas que tem anos de estudo menor ou igual a 15,5.
- **Perfil com menor probabilidade:** Anos de estudo menor ou igual a 12,5 anos e que tem idade maior que 40,5 anos.

5.iv. Desempenho dos modelos

MODELAGEM COM ESTATÍSTICA TRADICIONAL | COMPARAÇÃO ENTRE TÉCNICAS



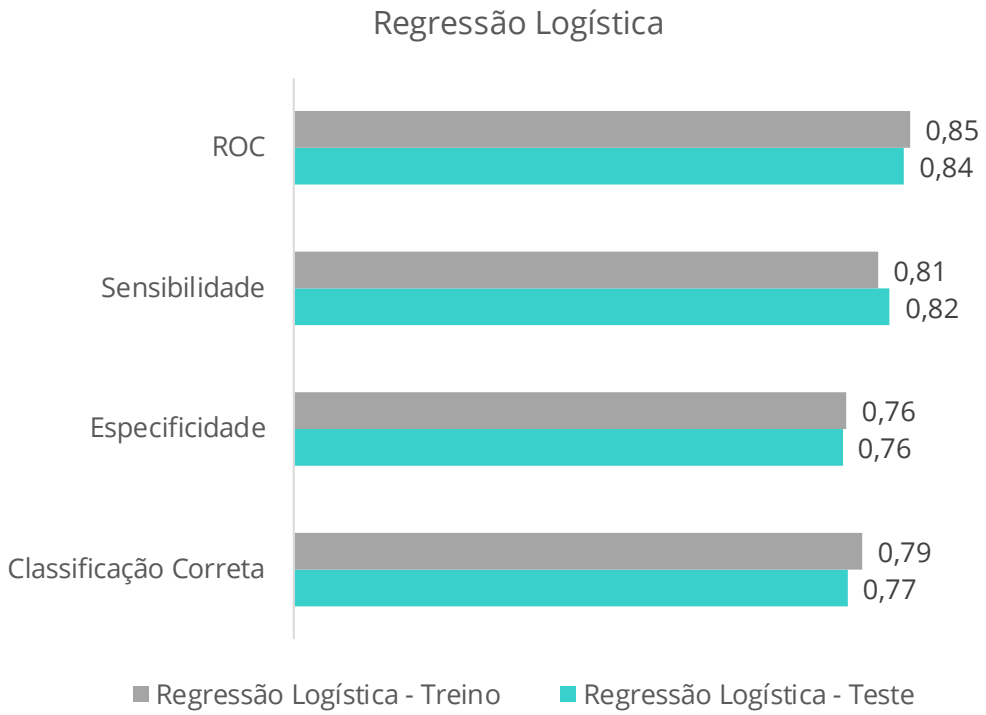
- Ambas as técnicas apresentaram **ótimo acerto** preditivo.
- A **Árvore de Decisão (AD)** teve maior acerto geral que a **Regressão Logística (RL),** 83% x 77%, e ROC, 87% x 84%, mas a **RL** acertou mais as pessoas que realmente frequentam uma graduação (Sensibilidade 82% x 81%).
- Dentre estes dois modelos, o melhor é a **AD** porque ela tem um maior acerto geral, e como o propósito do modelo é gerar insights tanto para quem tem alta probabilidade de cursar uma graduação quanto quem tem baixa, é muito interessante eu acertar o maior número possível.

Detalhes do desempenho: Treino x Teste

Regressão Logística

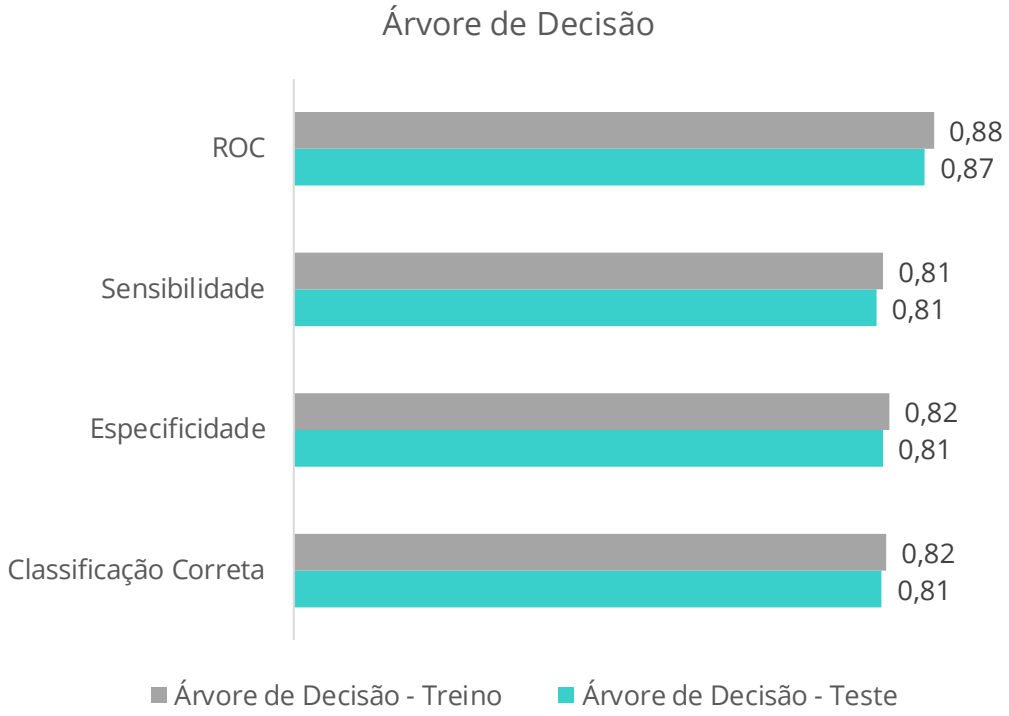
MODELAGEM COM ESTATÍSTICA TRADICIONAL | DESEMPENHO DO MODELO

O desempenho do modelo, tanto nas bases de treino como de teste, apresentou ótimo acerto preditivo, próximo a 78% no percentual geral de classificação correta. O percentual de acerto do evento de frequentar ensino superior ficou em torno de 81%. O percentual de acerto do evento de não frequentar o ensino superior ficou em torno de 76%. A ROC também ficou estável perto de 84%.



Árvore de Decisão

O desempenho do modelo, tanto nas bases de treino como de teste, apresentou ótimo acerto preditivo, próximo a 82% no percentual geral de classificação correta. O percentual de acerto do evento de frequentar ensino superior ficou em torno de 81%. O percentual de acerto do evento de não frequentar o ensino superior ficou em torno de 82%. A ROC também ficou estável perto de 88%.



Metodologia de análise de dados

Planejamento | Análise Exploratória | Modelagem Estatística Tradicional | **Algoritmos de Inteligência Artificial** | Conclusões

Definição do problema

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

Análise preliminar

- Medidas resumo
- Análise de frequências
- Gráficos
- Análise de *outliers*
- Análise de *missings*

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

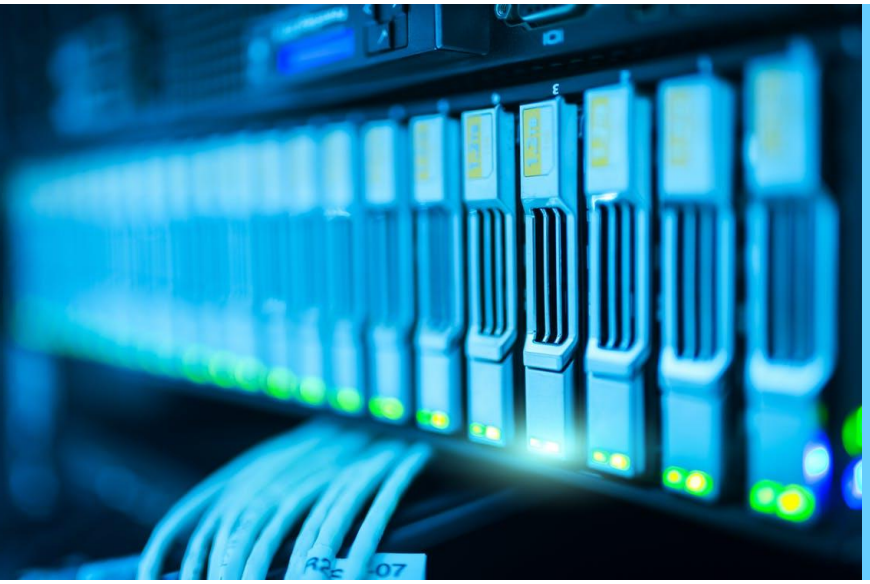
- SVM
- Random Forest
- Boosting

Definição da técnica

- Escolha da técnica que melhor se adeque ao uso e estratégias da área de negócios

6.i. Modelagem com Machine Learning

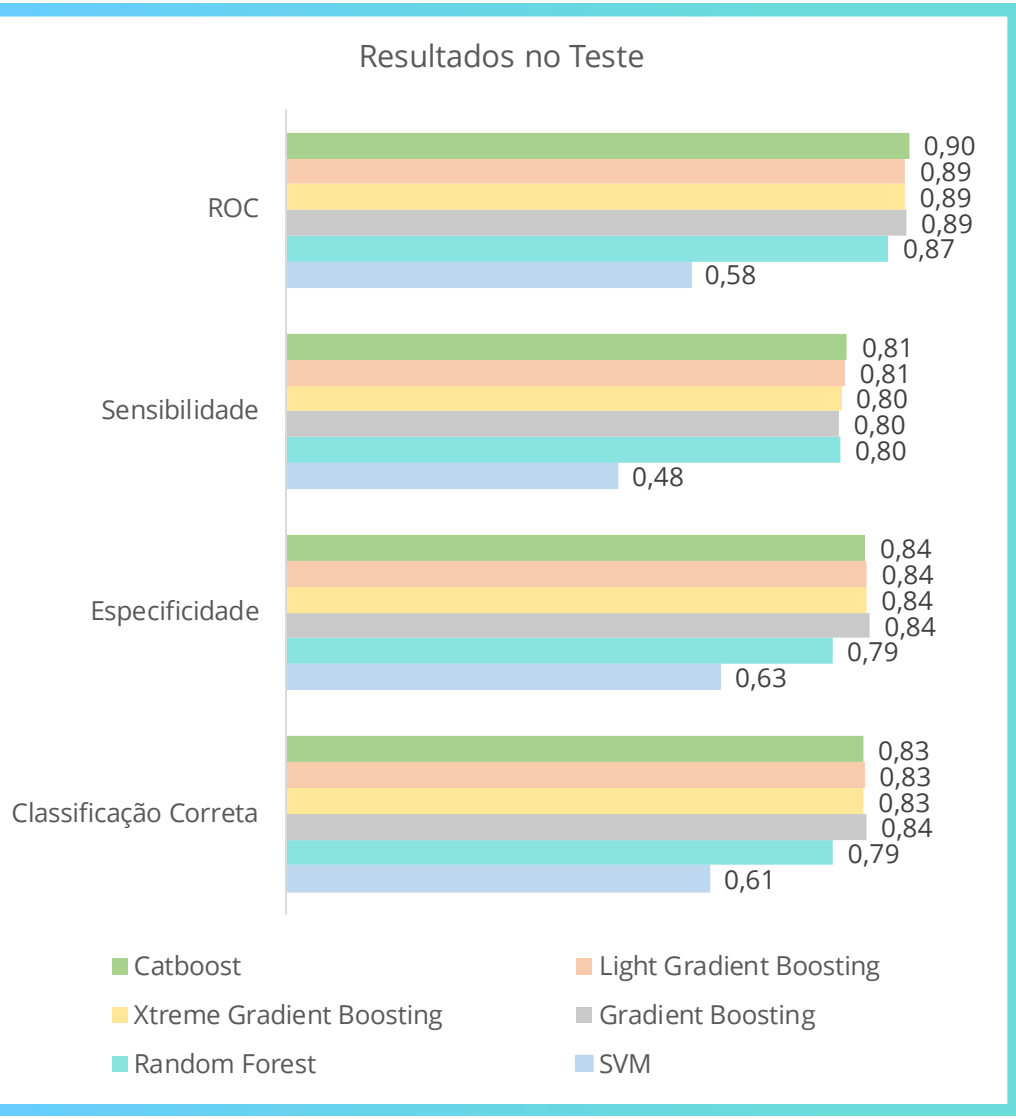
LISTA DE MODELOS



Foram implementados os modelos abaixo:

- SVM
- Random Forest
- Gradient Boosting
- Xtreme Gradient Boosting
- Light Gradient Boosting
- Catboost

6.ii. Desempenho dos modelos

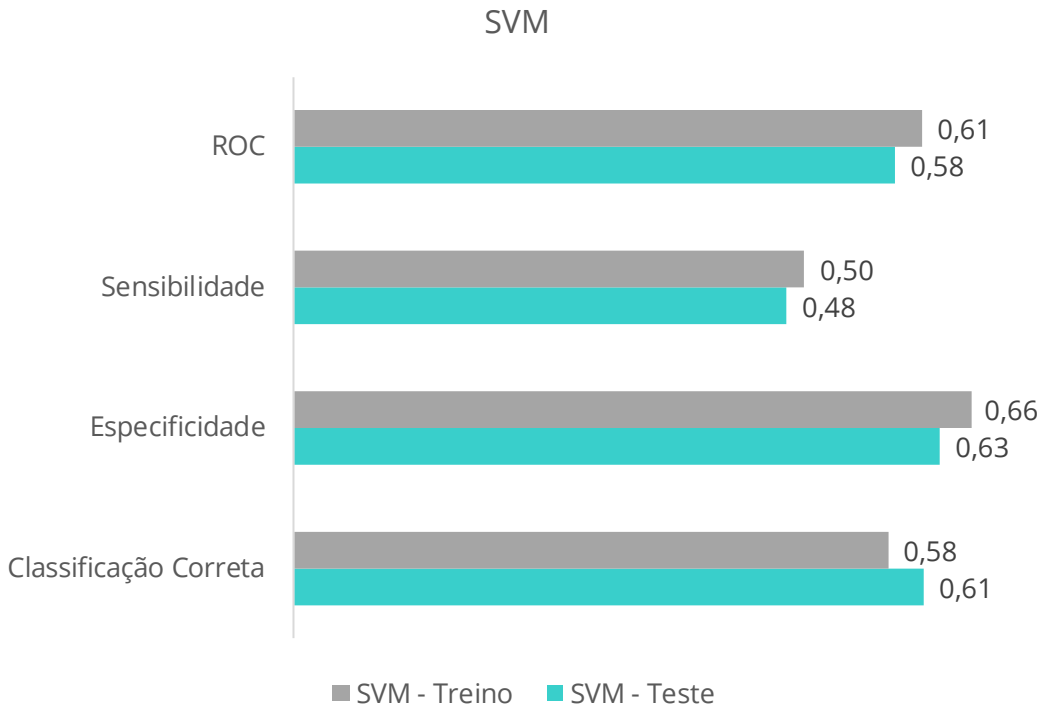


- Com exceção do SVM, todas as demais técnicas apresentaram **ótimo acerto** preditivo e métricas parecidas.
- O **Catboost** apresentou **melhor desempenho** na sensibilidade e ROC, além de ser um dos melhores na outras métricas.
- Comparando com a **Árvore de Decisão** (melhor modelo dos modelos baselines), o **Catboost** conseguiu aumentar o nível de acerto geral e melhorou a ROC, tornando o **Catboost** o melhor entre os dois.

Detalhes do desempenho: Treino x Teste

SVM

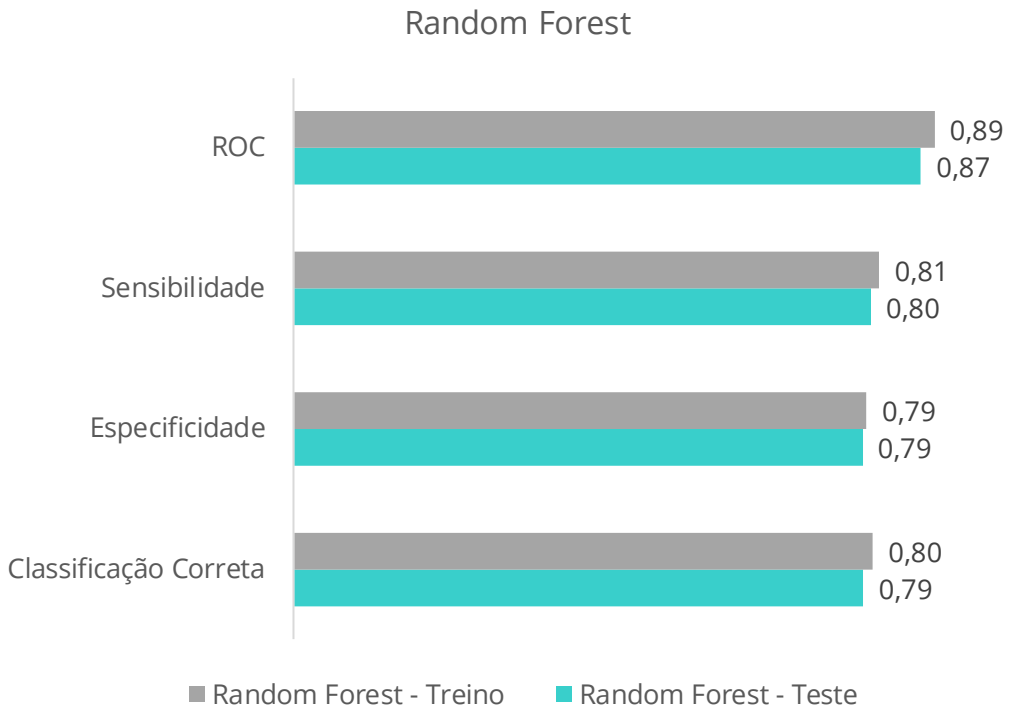
O desempenho do modelo, tanto nas bases de treino como de teste, apresentou bom acerto preditivo, próximo a 59% no percentual geral de classificação correta. O percentual de acerto do evento de frequentar ensino superior ficou em torno de 49%. O percentual de acerto do evento de não frequentar o ensino superior ficou em torno de 64%. A ROC também ficou estável perto de 60%.



Random Forest

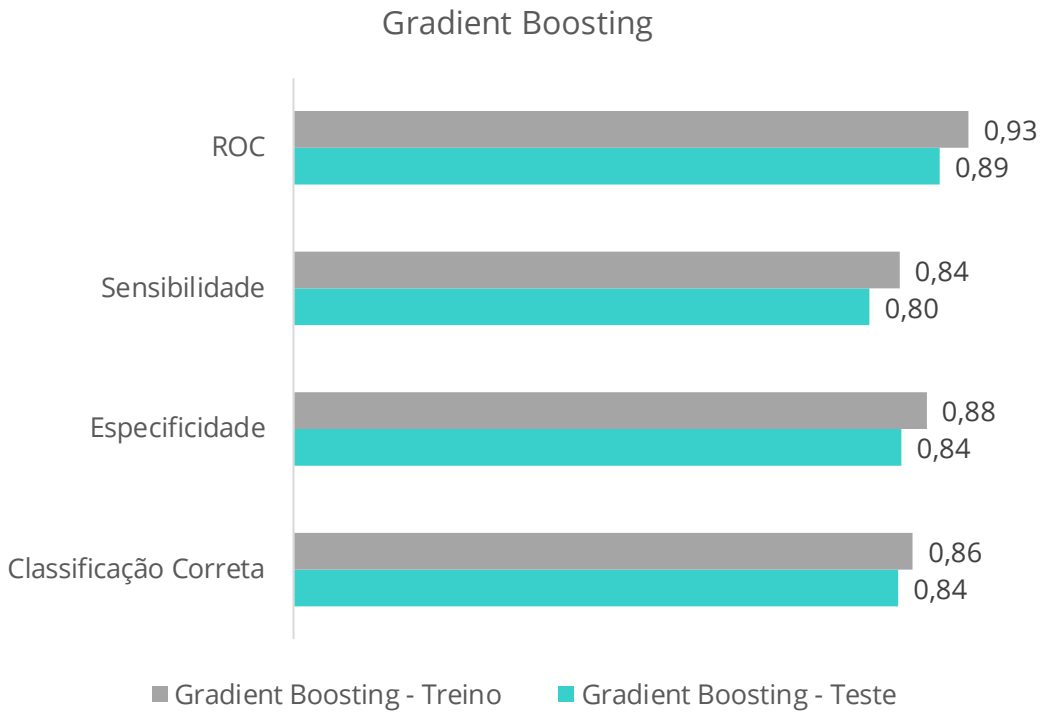
MODELAGEM COM MACHINE LEARNING | DESEMPENHO DO MODELO

O desempenho do modelo, tanto nas bases de treino como de teste, apresentou ótimo acerto preditivo, próximo a 80% no percentual geral de classificação correta. O percentual de acerto do evento de frequentar ensino superior ficou em torno de 81%. O percentual de acerto do evento de não frequentar o ensino superior ficou em torno de 79%. A ROC também ficou estável perto de 88%.



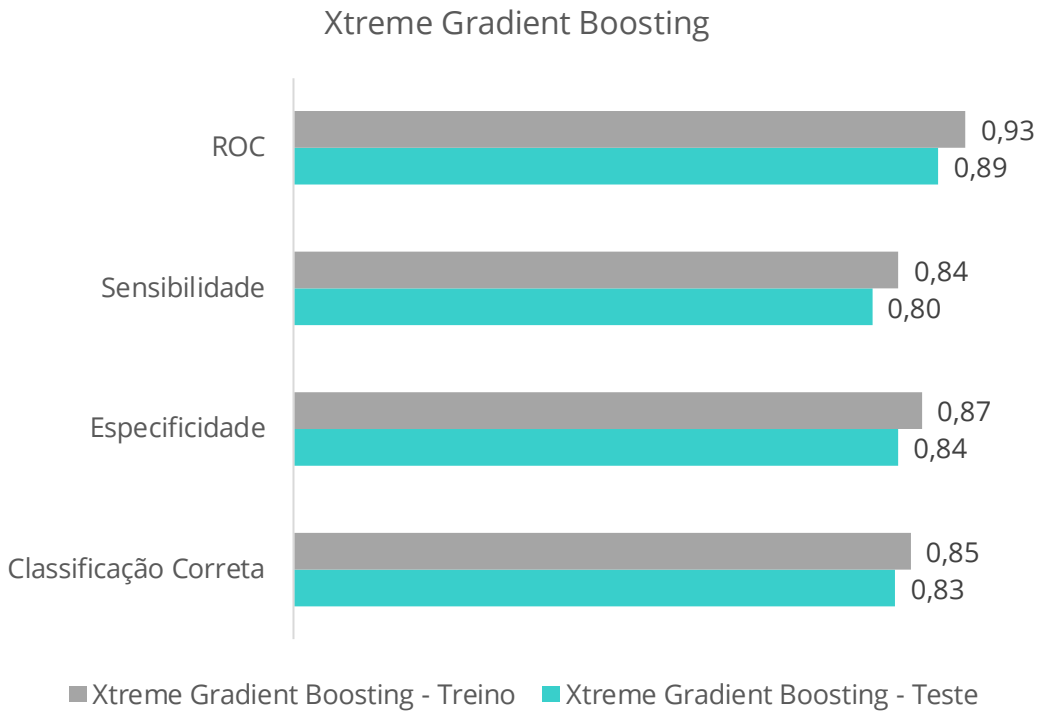
Gradient Boosting

O desempenho do modelo, tanto nas bases de treino como de teste, apresentou ótimo acerto preditivo, próximo a 85% no percentual geral de classificação correta. O percentual de acerto do evento de frequentar ensino superior ficou em torno de 82%. O percentual de acerto do evento de não frequentar o ensino superior ficou em torno de 86%. A ROC também ficou estável perto de 91%.



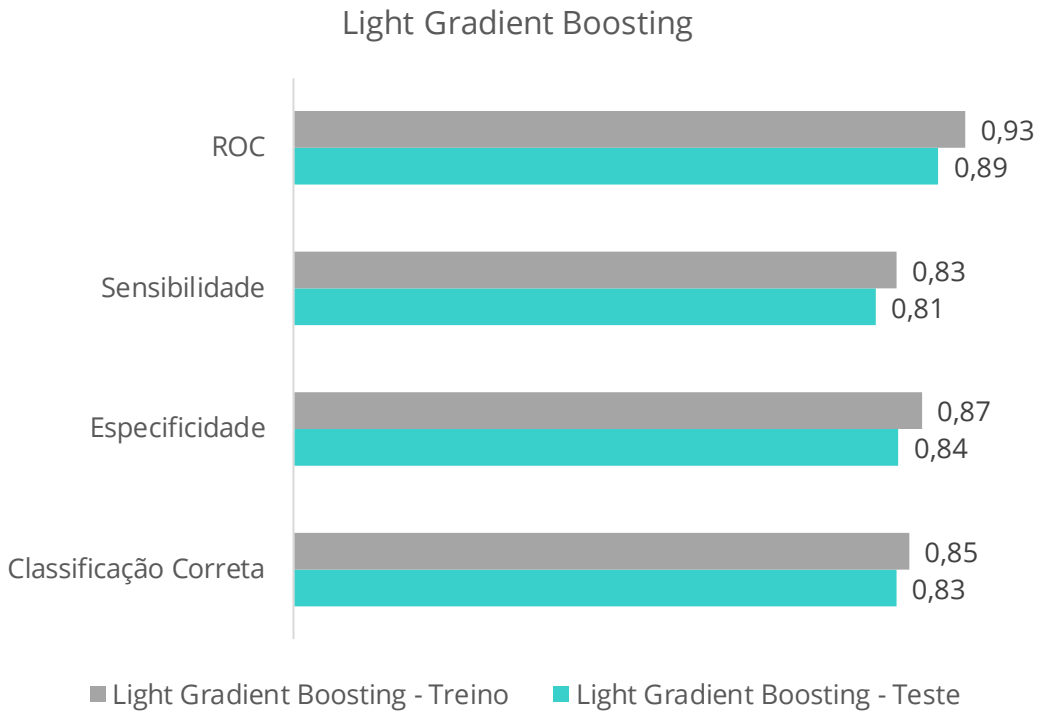
Xtreme Gradient Boosting

O desempenho do modelo, tanto nas bases de treino como de teste, apresentou ótimo acerto preditivo, próximo a 84% no percentual geral de classificação correta. O percentual de acerto do evento de frequentar ensino superior ficou em torno de 82%. O percentual de acerto do evento de não frequentar o ensino superior ficou em torno de 85%. A ROC também ficou estável perto de 91%.



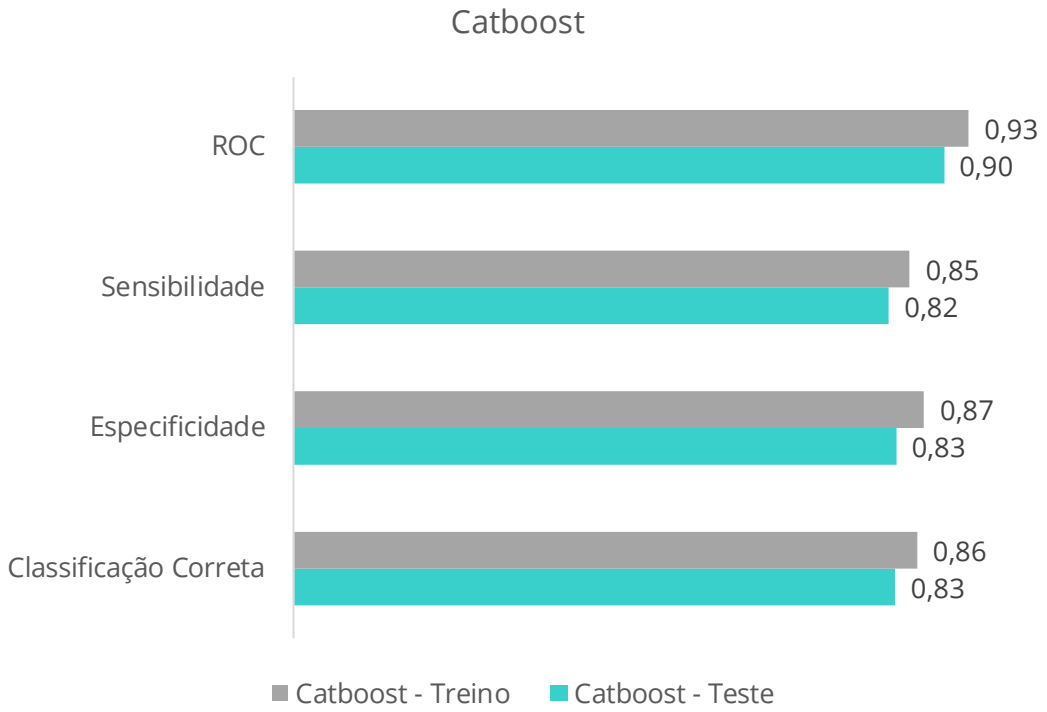
Light Gradient Boosting

O desempenho do modelo, tanto nas bases de treino como de teste, apresentou ótimo acerto preditivo, próximo a 84% no percentual geral de classificação correta. O percentual de acerto do evento de frequentar ensino superior ficou em torno de 82%. O percentual de acerto do evento de não frequentar o ensino superior ficou em torno de 85%. A ROC também ficou estável perto de 91%.



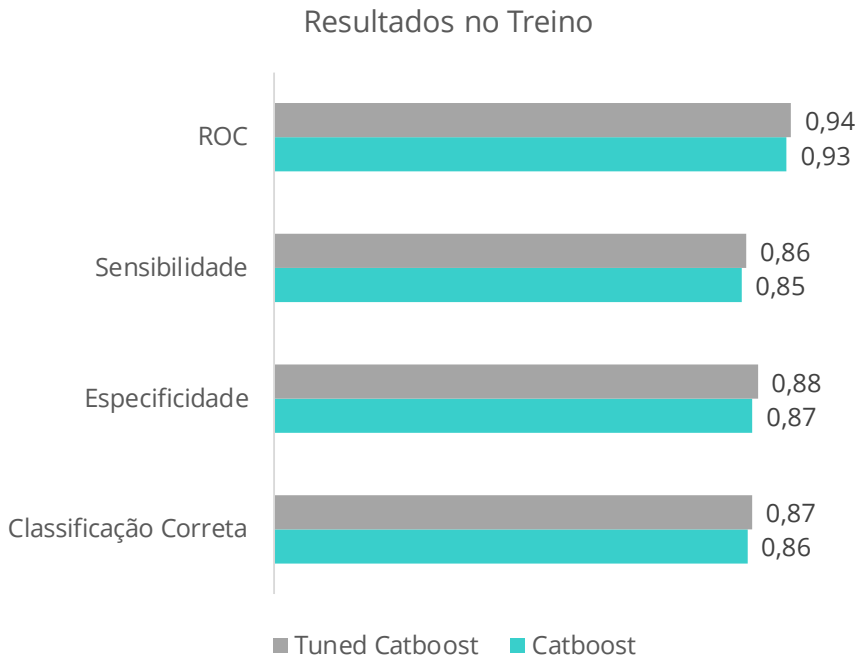
Catboost

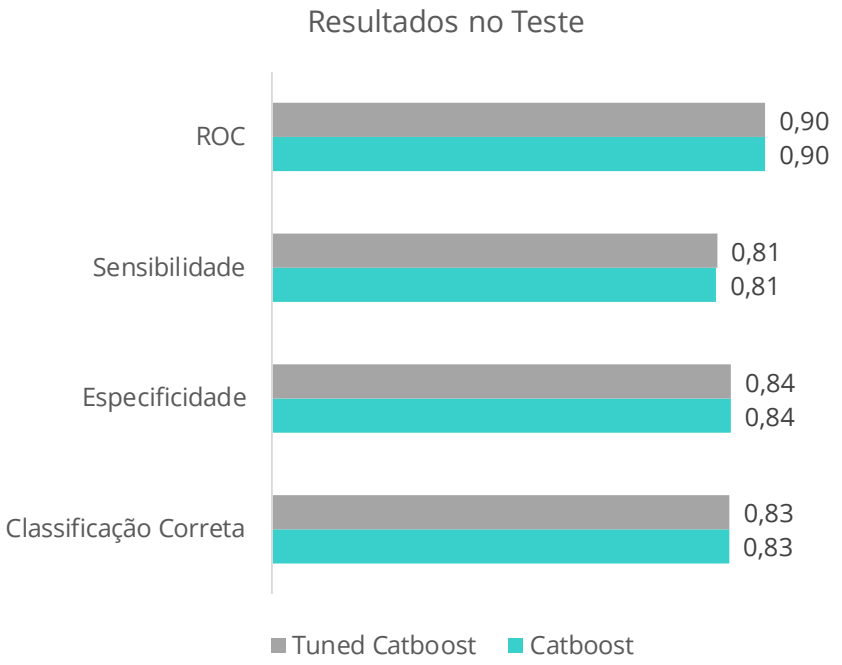
O desempenho do modelo, tanto nas bases de treino como de teste, apresentou ótimo acerto preditivo, próximo a 85% no percentual geral de classificação correta. O percentual de acerto do evento de frequentar ensino superior ficou em torno de 83%. O percentual de acerto do evento de não frequentar o ensino superior ficou em torno de 85%. A ROC também ficou estável perto de 92%.



6.iii. Otimização de Hiperparâmetros

Aplicando a otimização de hiperparâmetros no modelo Catboost fez com que o modelo no treino aumentasse as métricas em 0.01pp, entretanto no teste não houve mudanças significativas.





Metodologia de análise de dados

Planejamento → Análise Exploratória → Modelagem Estatística Tradicional → Algoritmos de Inteligência Artificial → Conclusões

Definição do problema

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

Análise preliminar

- Medidas resumo
- Análise de frequências
- Gráficos
- Análise de *outliers*
- Análise de *missings*

Avaliação das técnicas

- Regressão Logística
- Árvore de Decisão

Avaliação das técnicas

- Random Forest
- Boosting
- SVM

Definição da técnica

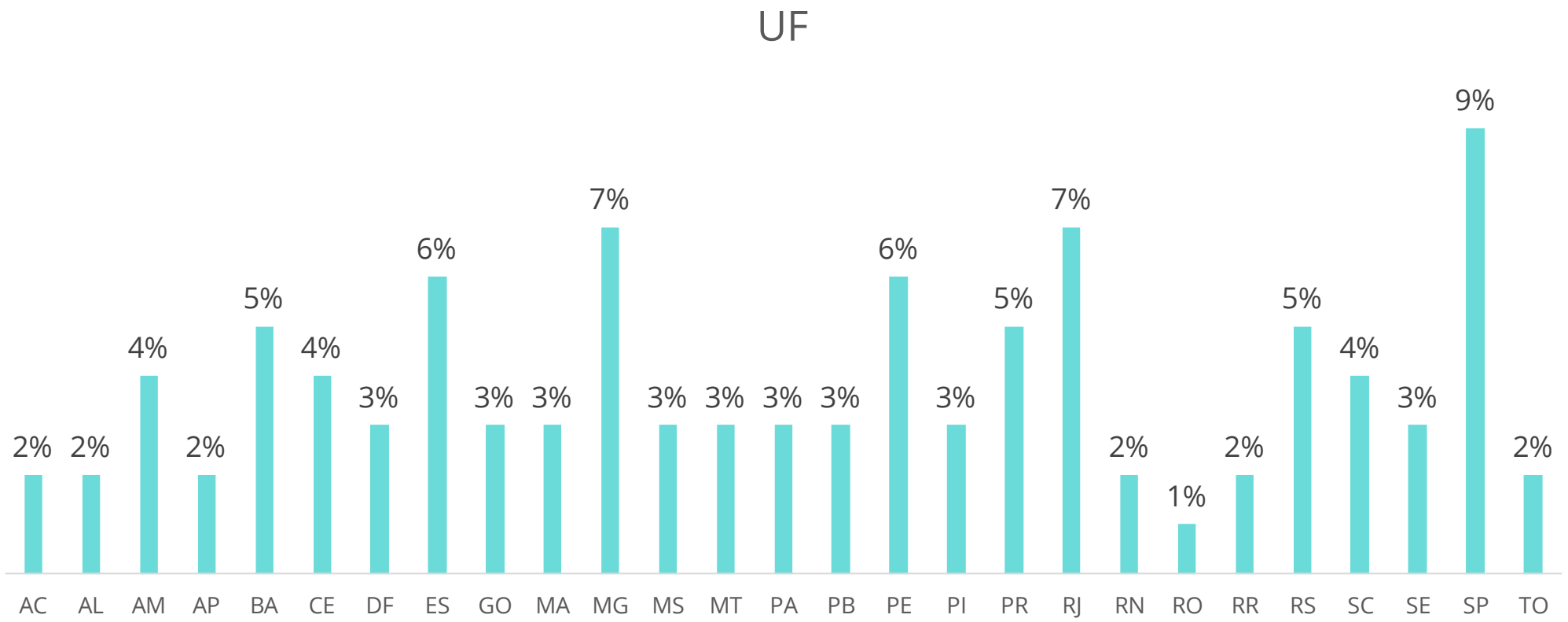
- Escolha da técnica que melhor se adeque ao uso e estratégias da área de negócios

7. Conclusões

- O **Catboost (CB)** apresentou **melhor desempenho** em comparação com os outros modelos, com acerto geral de **83%** e sensibilidade de **81%** (excelente desempenho).
- As métricas de avaliação do **CB** ficaram estáveis quando comparamos treino e teste;
- Como a maioria dos modelos apresentaram ótimos desempenhos, o que levou a decisão do **CB** como melhor modelo foi a sensibilidade e a ROC.
- As variáveis que mais discriminaram no modelo foram:
  - **Tempo de Estudo**: Pessoas com mais tempo de estudo tem mais probabilidade, geralmente essas pessoas começaram um curso de graduação no passado mas não terminaram.
  - **Idade**: Pessoas mais novas possuem mais probabilidade de cursar uma graduação

Detalhes das análises

Variáveis do Entrevistado

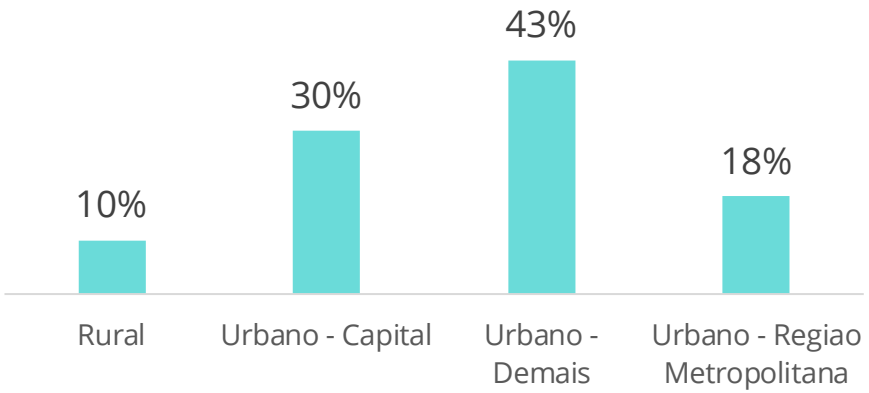


- SP é os estado que possui o maior número de entrevistados, seguido por MG e RJ, todos estados do Sudeste.
- A base é bem distribuída, a maioria das UFs tem frequência próxima de 3%

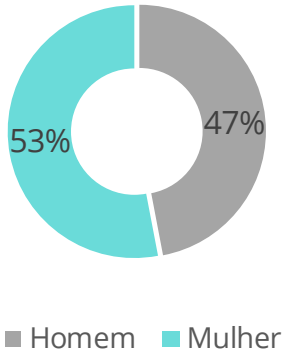
Detalhes das análises

Variáveis do Entrevistado

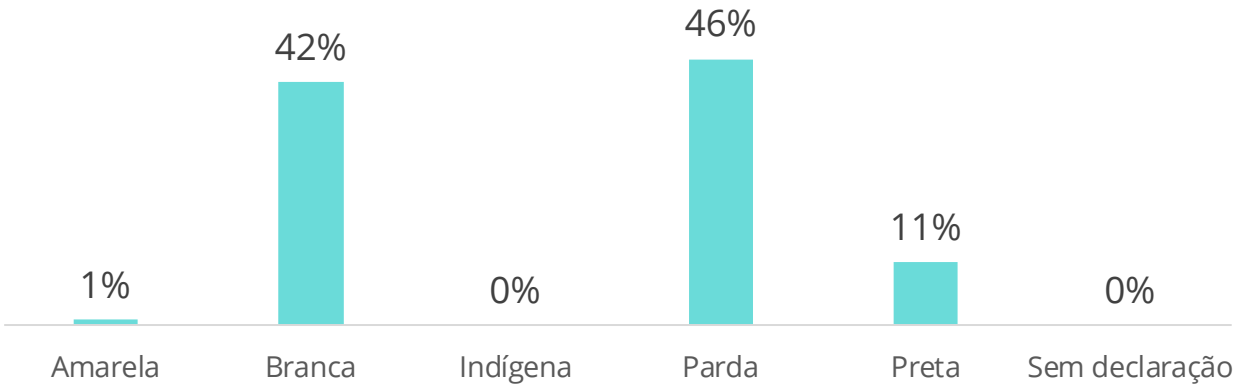
Estrato_POF



Sexo



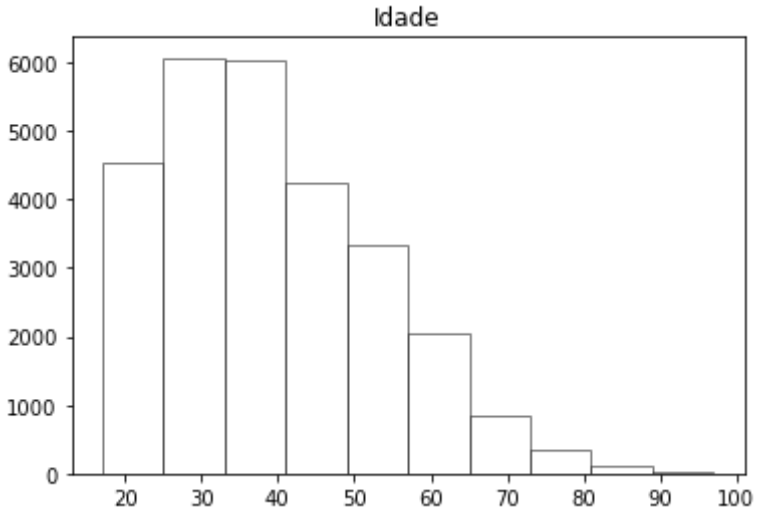
Cor_Raca

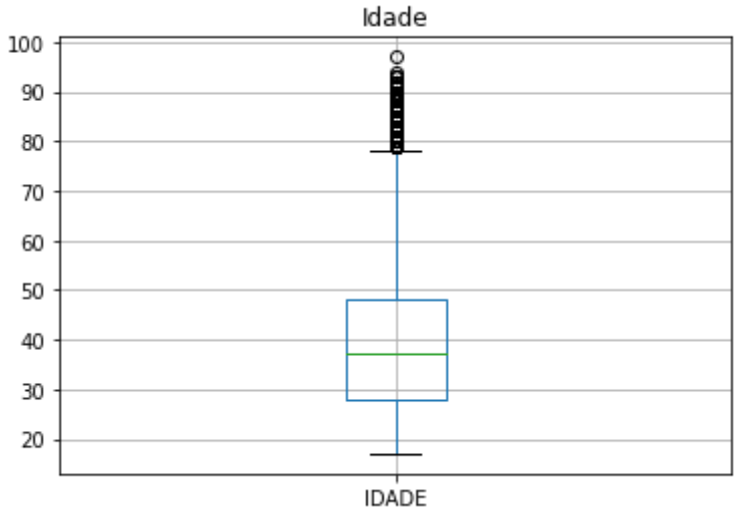


- A maior parte dos entrevistados moram em regiões urbanas mais concentrados em cidades do interior (43%)
- Mulheres são a maioria da base de dados (53%)
- Entrevistados são na maioria de cor Parda ou Branca (88% juntos)

Detalhes das análises

Variáveis do Entrevistado





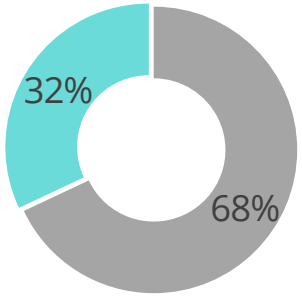
Medida	Valor
Mínimo	17
1º Quartil	28
Mediana	**37**
Média	39
3º Quartil	**48**
Máximo	97

A base de dados é de clientes de meia idade, sendo metade dos clientes com menos de 37 anos e 75% dos clientes abaixo de 48.
Verificamos uma distribuição assimétrica à direita indicando poucos clientes com idades maiores.
A idade mínima é de 17 anos que está condizente com o filtro feito na base de trabalhar apenas com maiores de 17 anos.

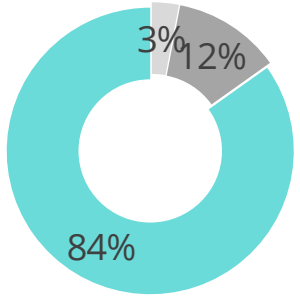
Detalhes das análises

Variáveis do Entrevistado

Tem_Plano_Saude



Trabalhou_Ult_12M

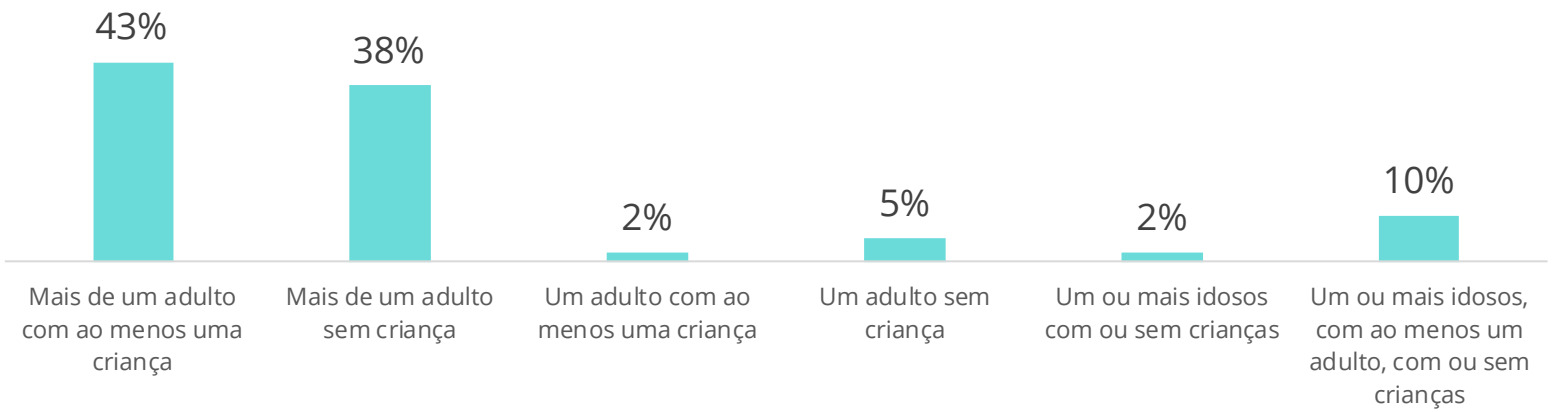


- Entrevistados sem plano de saúde são maioria na base (68%)
- Maioria trabalhou nos últimos 12 meses (84%)

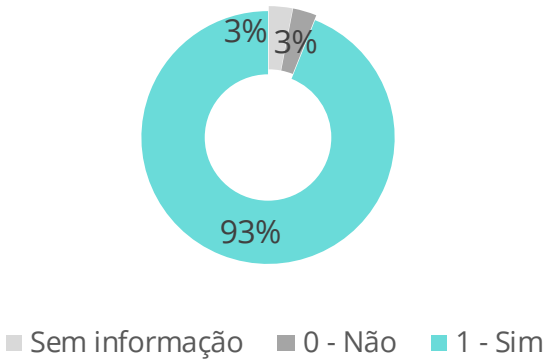
Detalhes das análises

Variáveis do Entrevistado

Composição



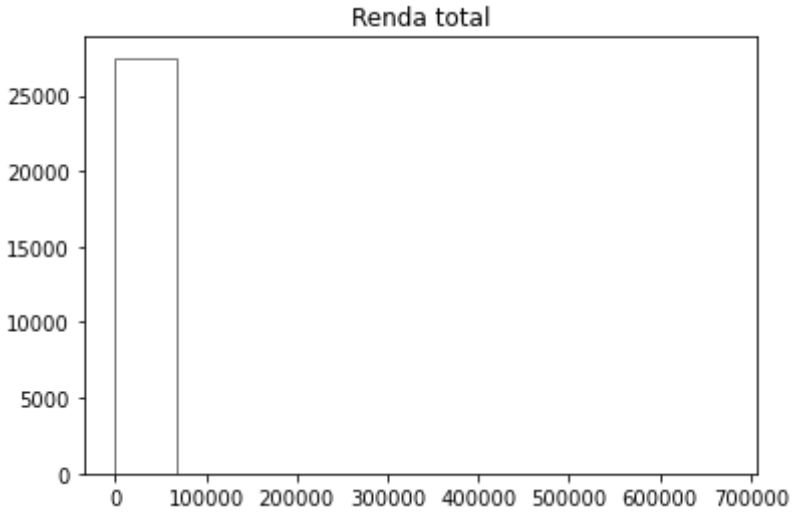
Gastos_Sem_Renda

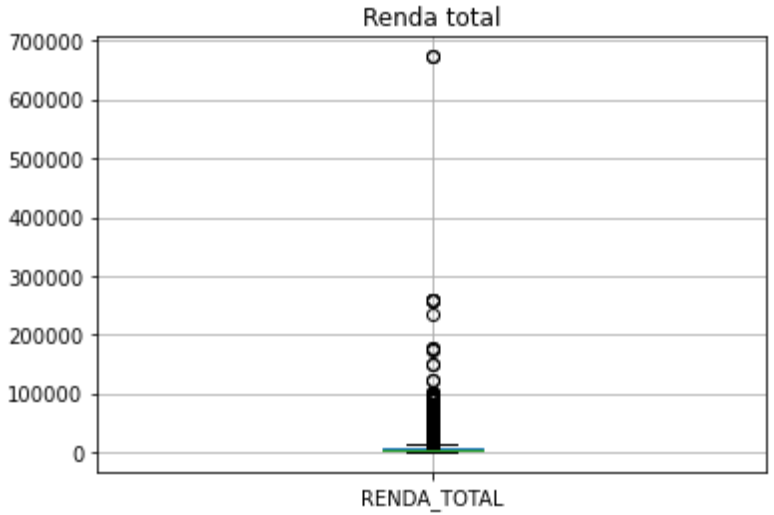


- A residência dos entrevistados geralmente é composta por mais de um adulto (81%)
- Maioria tem o hábito de ter gastos/despesas sem um rendimento próprio (93%)

Detalhes das análises

Variáveis do Entrevistado



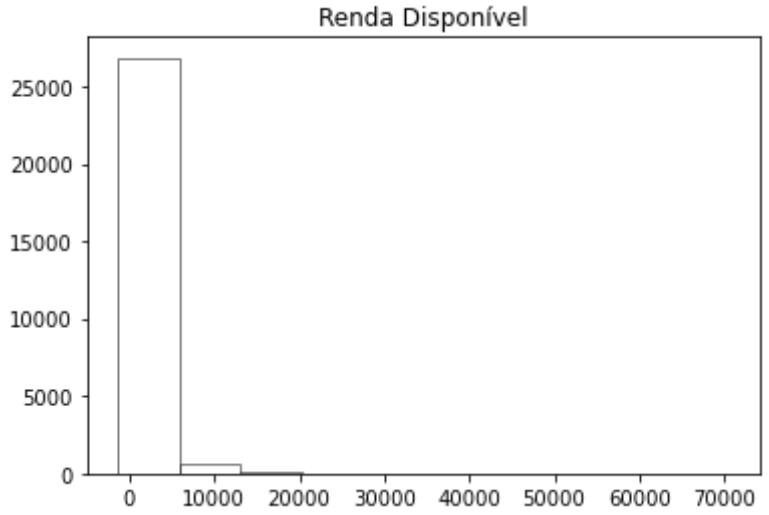


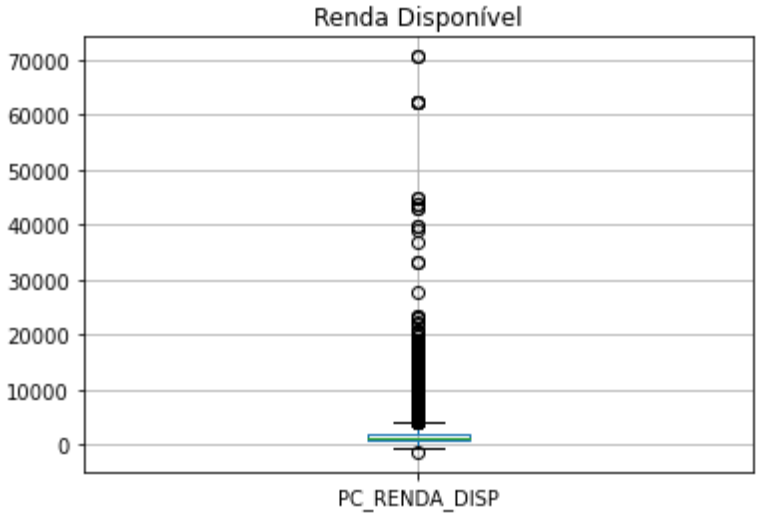
Medida	Valor
Mínimo	8,4
1º Quartil	2.465,2
Mediana	4.117,9
Média	**5.830,7**
3º Quartil	**6.779,7**
Máximo	672.891,0

A renda total da residência possui distribuição assimétrica acentuada á direita, com a presença de vários outliers superiores, sendo que 75% dos entrevistados possuem renda total menor do que R$6.779,7. A média fica entre a mediana e o 3º quartil o que demonstra que a quantidade de outliers é muito pequena.

Detalhes das análises

Variáveis do Entrevistado



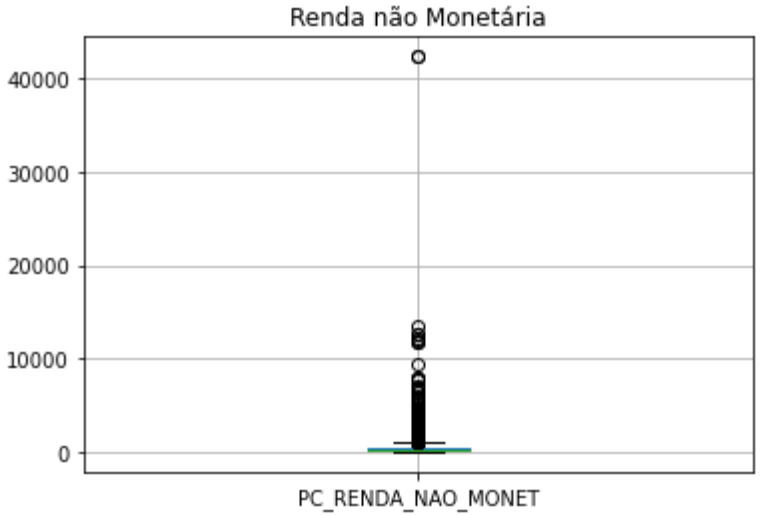


Medida	Valor
Mínimo	-1.343,5
1º Quartil	808,6
Mediana	1.300,4
Média	**1.744,3**
3º Quartil	**2.061,0**
Máximo	70.587,3

A renda disponível per capita possui distribuição assimétrica acentuada á direita, com a presença de vários outliers superiores, sendo que 75% dos entrevistados possuem renda total menor do que R$2061. A média fica entre a mediana e o 3º quartil o que demonstra que a quantidade de outliers é muito pequena. Existem entrevistados com rendas negativas, isso acontece quando a renda não é suficiente para arcar com impostos diretos, contribuições sociais, e outras deduções compulsórias ou quase compulsórias.

Detalhes das análises

Variáveis do Entrevistado

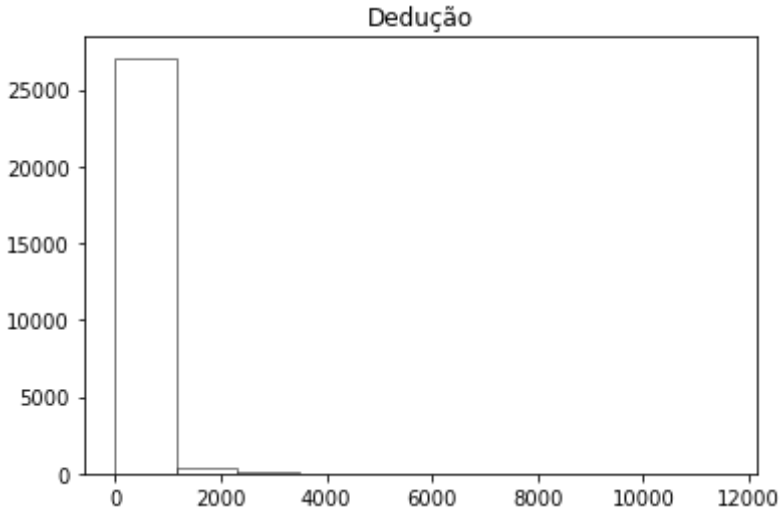


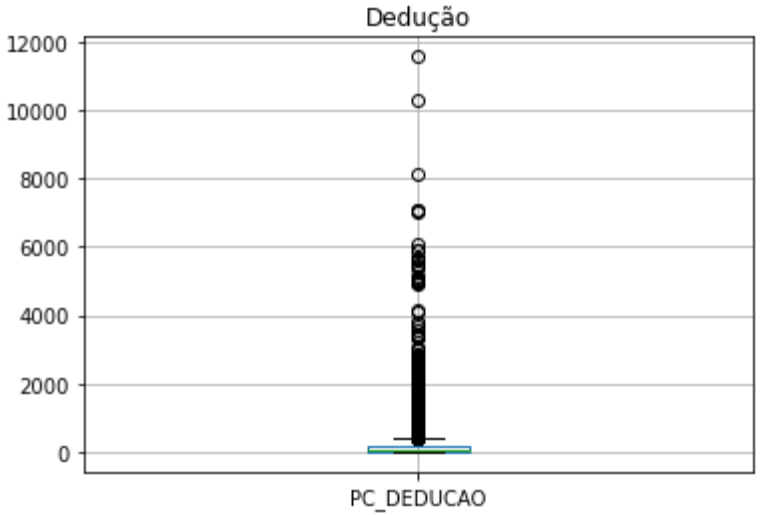
Medida	Valor
Mínimo	0,0
1º Quartil	146,3
Mediana	271,9
Média	**383,5**
3º Quartil	**475,2**
Máximo	42.379,4

A renda não monetária per capita possui distribuição assimétrica acentuada á direita, com a presença de vários outliers superiores, sendo que 75% dos entrevistados possuem renda total menor do que R$475,2. A média fica entre a mediana e o 3º quartil o que demonstra que a quantidade de outliers é muito pequena. Existem entrevistados que ninguém na sua residência possui renda não monetária.

Detalhes das análises

Variáveis do Entrevistado



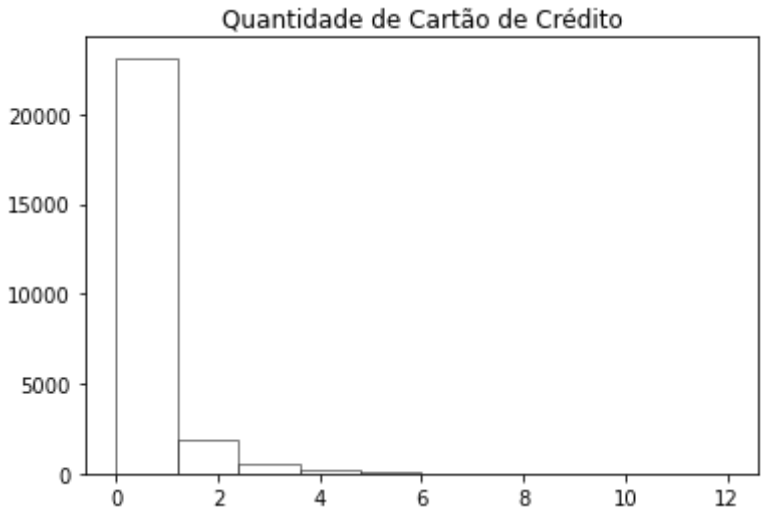


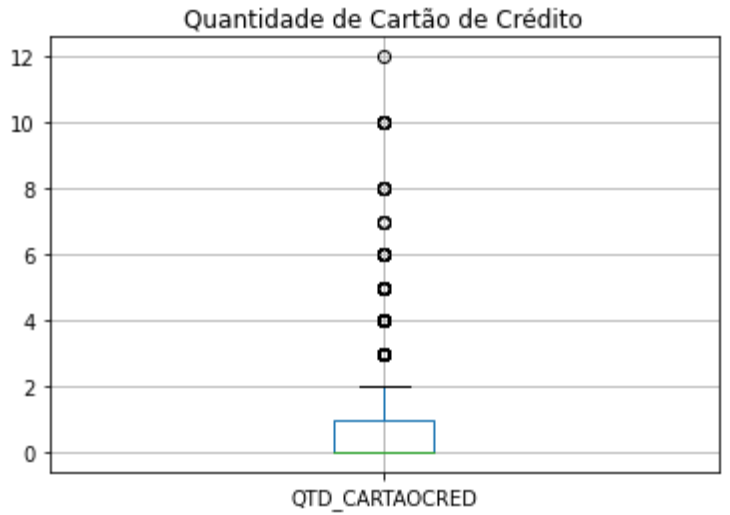
Medida	Valor
Mínimo	0,0
1º Quartil	25,5
Mediana	72,4
Média	**165,4**
3º Quartil	**171,0**
Máximo	11.563,8

A dedução per capita possui distribuição assimétrica acentuada á direita, com a presença de vários outliers superiores, sendo que 75% dos entrevistados possuem renda total menor do que R$171. A média fica entre a mediana e o 3º quartil o que demonstra que a quantidade de outliers é muito pequena. Existem entrevistados que ninguém na sua residência possui dedução.

Detalhes das análises

Variáveis Bancárias



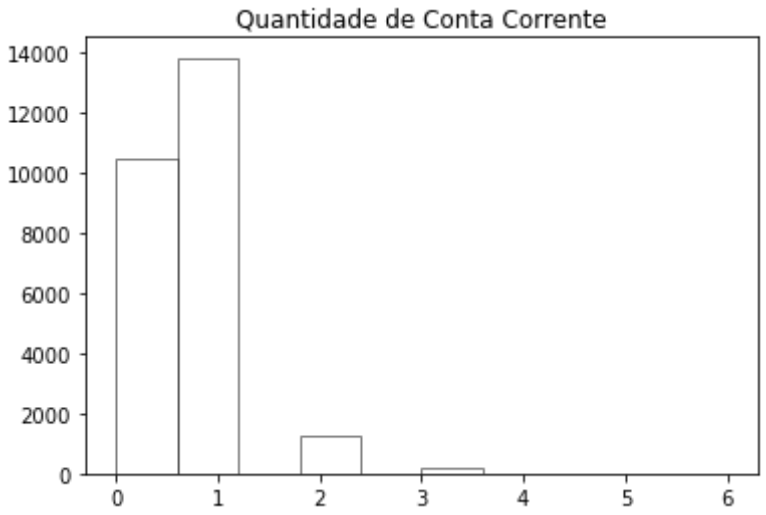


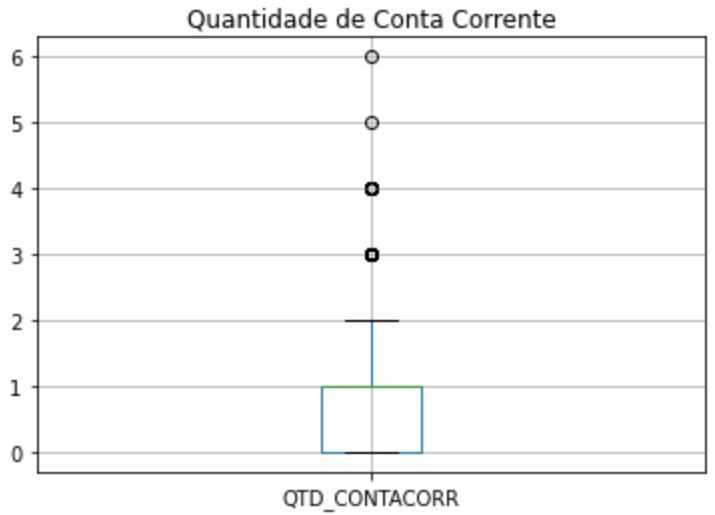
Medida	Valor
Mínimo	0
1º Quartil	0
Mediana	0
Média	**1**
3º Quartil	**1**
Máximo	12

A quantidade de cartão de crédito possui distribuição assimétrica acentuada á direita, com a presença de outliers superiores, sendo que 50% dos entrevistados não possuem cartão de crédito e até 75% possui apenas 1 cartão.

Detalhes das análises

Variáveis Bancárias



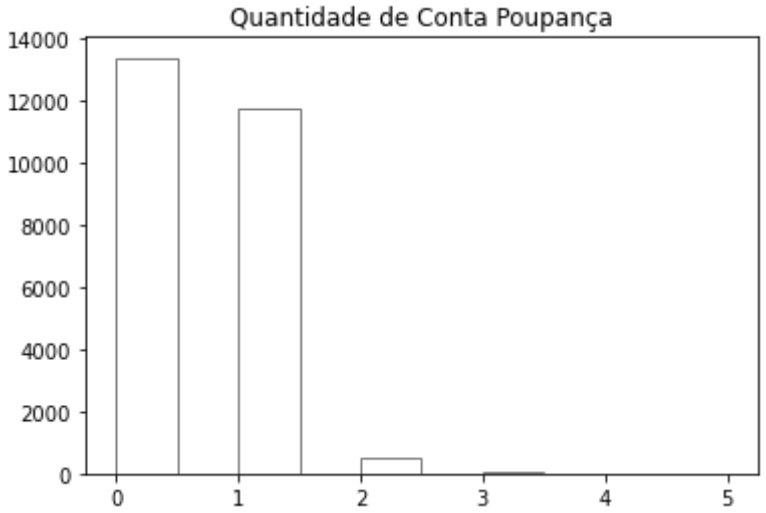


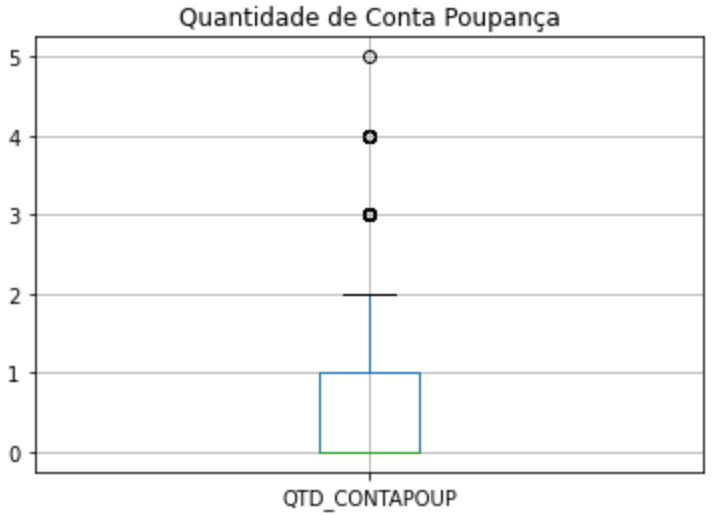
Medida	Valor
Mínimo	0
1º Quartil	0
Mediana	1
Média	**1**
3º Quartil	**1**
Máximo	6

A quantidade de conta corrente possui distribuição assimétrica acentuada á direita, com a presença de outliers superiores, sendo que 50% dos entrevistados possuem 1 conta corrente.

Detalhes das análises

Variáveis Bancárias



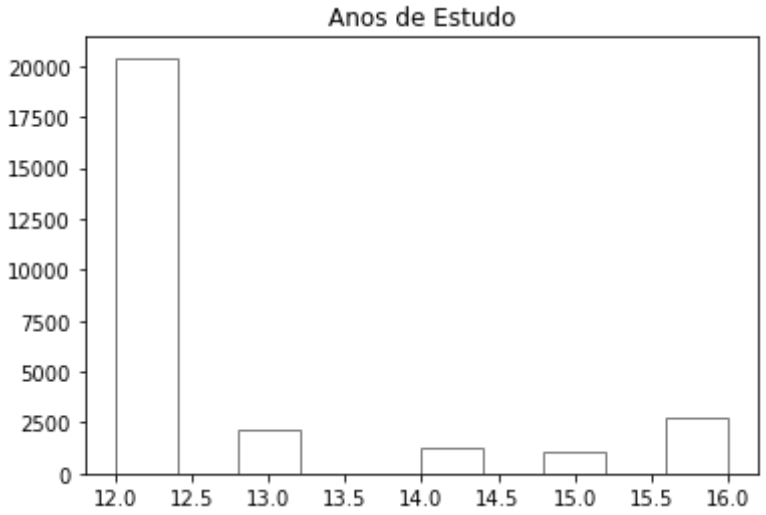


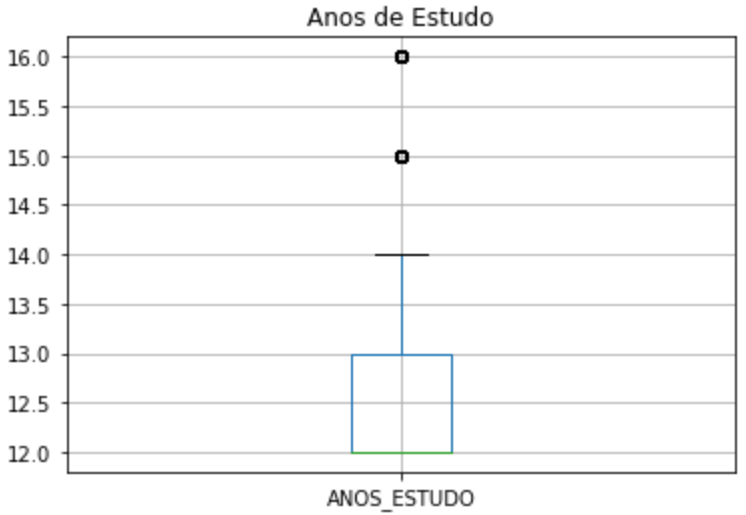
Medida	Valor
Mínimo	0
1º Quartil	0
Mediana	0
Média	**1**
3º Quartil	**1**
Máximo	5

A quantidade de conta poupança possui distribuição assimétrica acentuada á direita, com a presença de outliers superiores, sendo que 75% dos entrevistados possuem 1 conta poupança.

Detalhes das análises

Variável Escolar



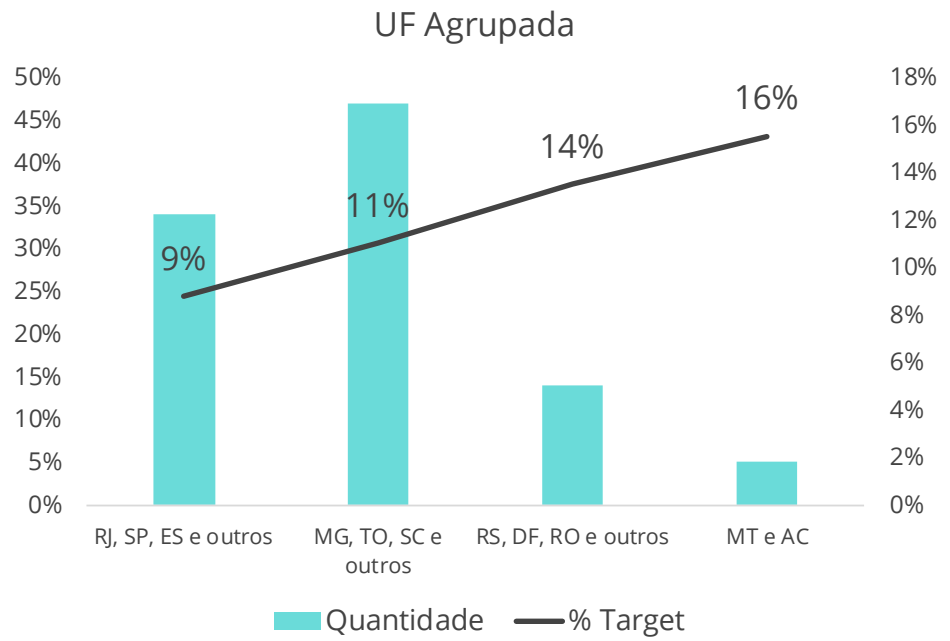


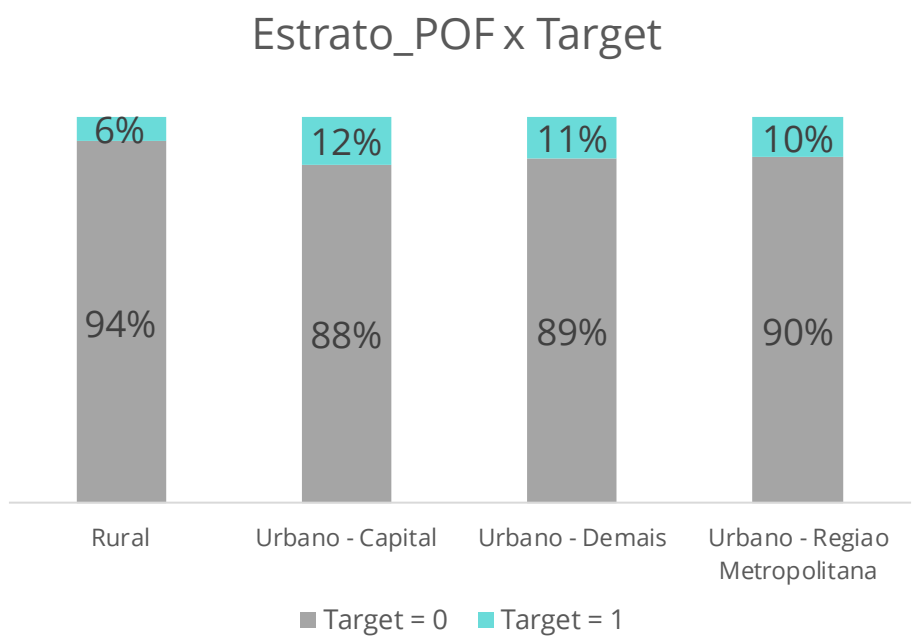
Medida	Valor
Mínimo	12
1º Quartil	12
Mediana	12
Média	**13**
3º Quartil	**13**
Máximo	16

A quantidade de anos de estudo possui distribuição assimétrica acentuada á direita, com a presença de outliers superiores, sendo que 50% dos entrevistados tem 12 anos de estudo e 75% possuem 13 anos de estudo.

Análise detalhada

ANÁLISE EXPLORATÓRIA | BIDIMENSIONAL





LEGENDA

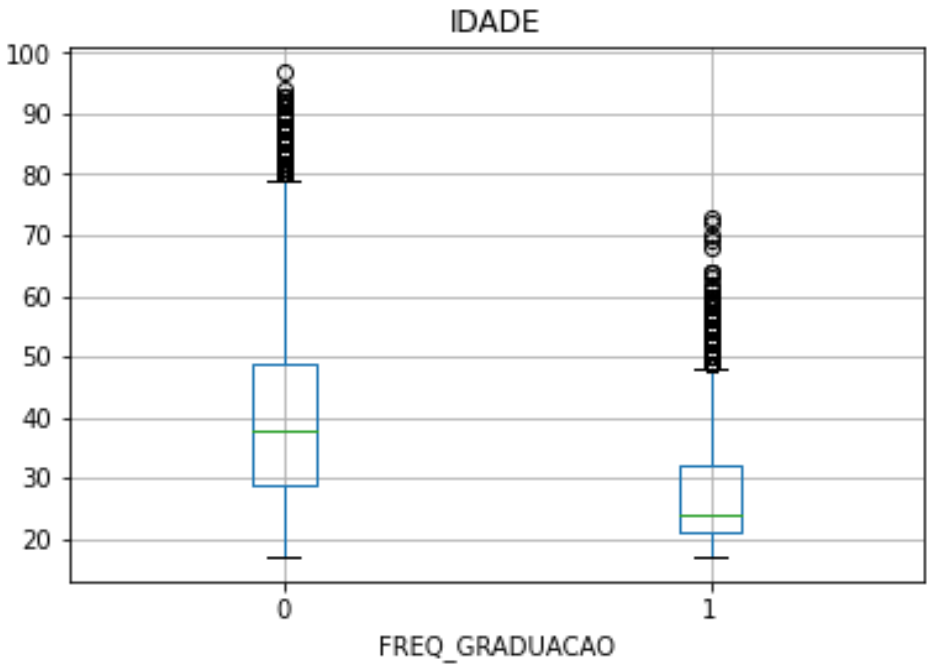
- **Covariável parece explicar bem a resposta**
- **Covariável parece explicar pouco a resposta**
- **Covariável parece não explicar a resposta**

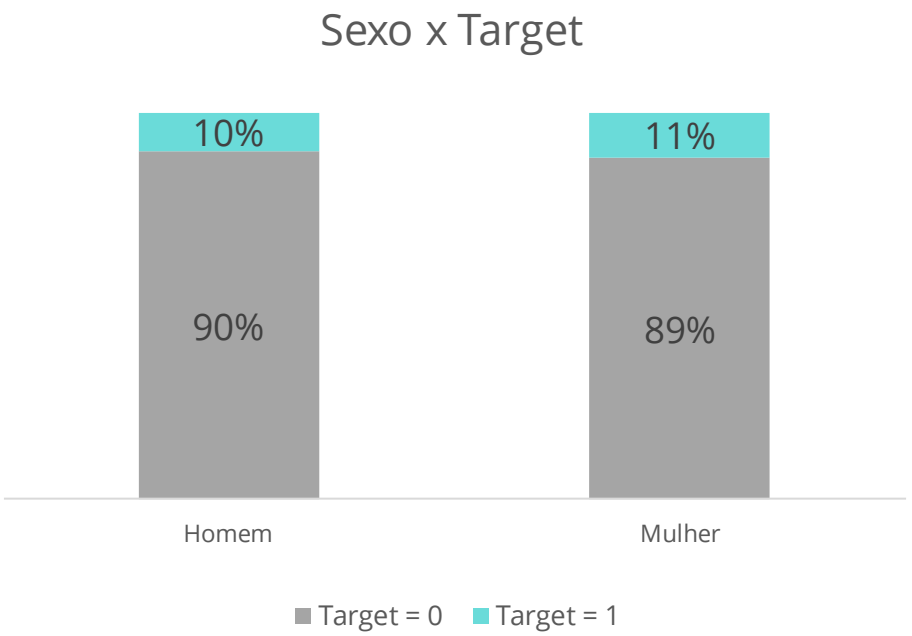
As Ufs foram agrupadas de acordo com o percentual do target, demonstrando que o % de target tem relação com o fator regional

O Estrato_POF parece discriminar muito pouco, pois todas suas categorias possuem percentual do target próximo à média da base de 11%, com exceção da categoria “Rural” que tem menor percentual

Análise detalhada

ANÁLISE EXPLORATÓRIA | BIDIMENSIONAL





LEGENDA

- **Covariável parece explicar bem a resposta**
- **Covariável parece explicar pouco a resposta**
- **Covariável parece não explicar a resposta**

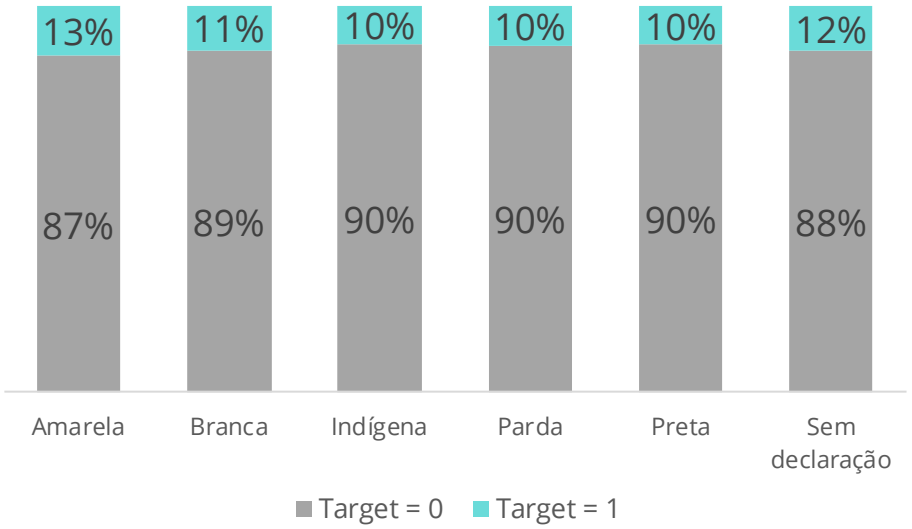
A idade parece explicar bem o target, podemos ver que pessoas mais jovens são as que frequentam uma graduação

O Sexo parece não discriminar o target dado que as duas categorias possuem percentual do target próximo à média da base de 11%,

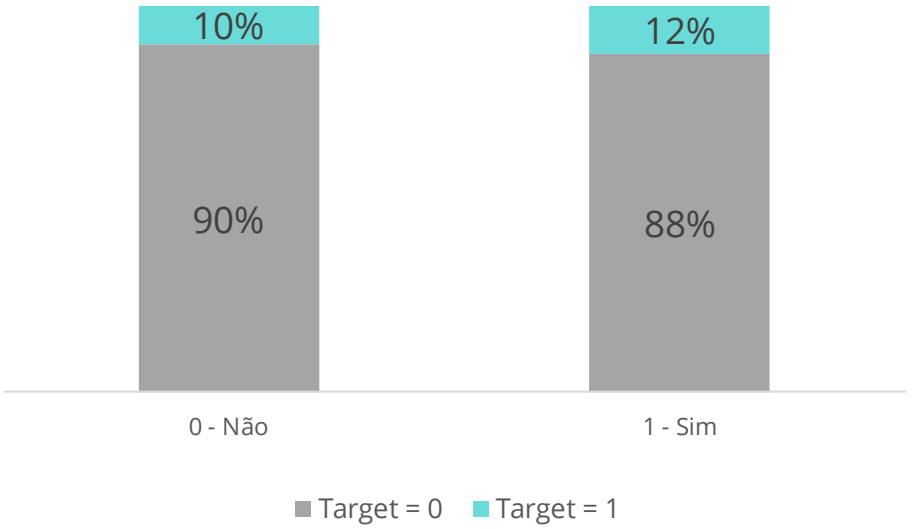
Análise detalhada

ANÁLISE EXPLORATÓRIA | BIDIMENSIONAL

Cor_Raca x Target



Tem_Plano_Saude x Target



LEGENDA

- **Covariável parece explicar bem a resposta**
- **Covariável parece explicar pouco a resposta**
- **Covariável parece não explicar a resposta**

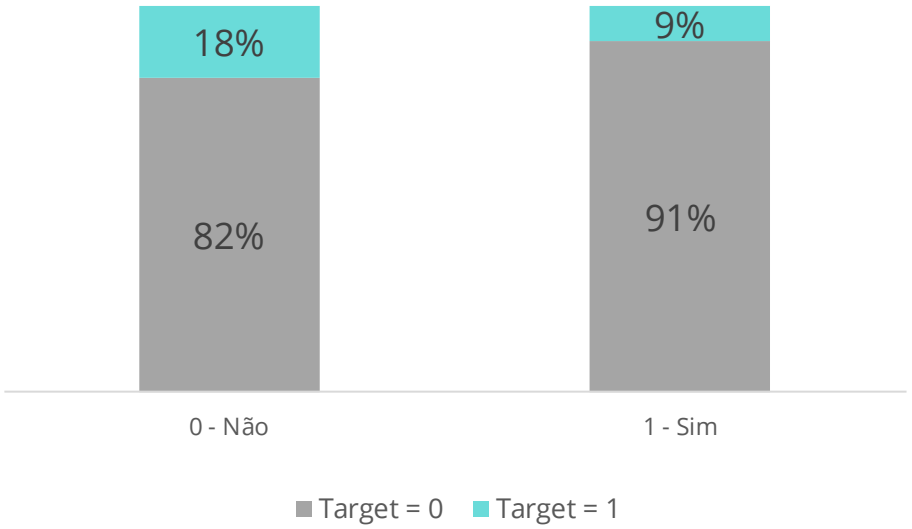
A Cor_raca parece discriminar muito pouco, pois todas suas categorias possuem percentual do target próximo à média da base de 11%, com exceção da categoria “Amarela” que tem percentual um pouco maior

Tem_Plano_Saude parece não discriminar o target dado que as duas categorias possuem percentual do target próximo à média da base de 11%

Análise detalhada

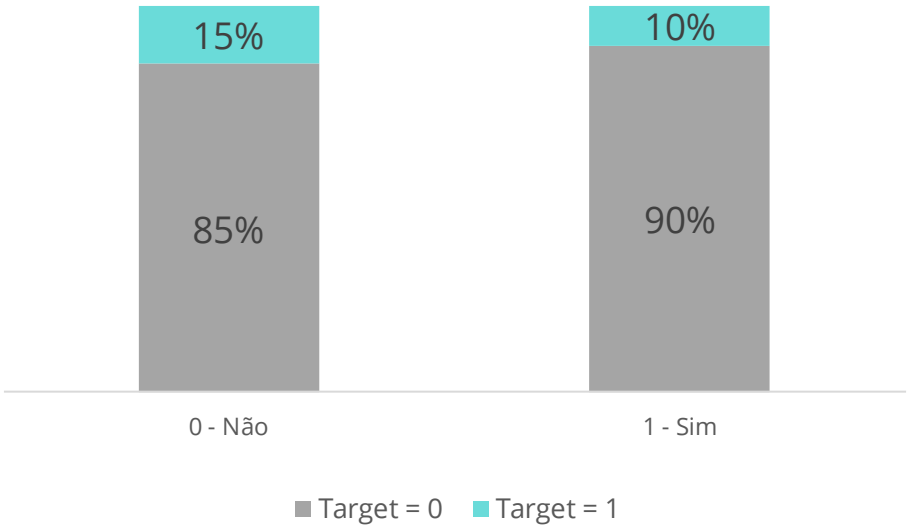
ANÁLISE EXPLORATÓRIA | BIDIMENSIONAL

Trabalhou_Ult_12M x Target



Trabalhou_Ult_12M parece discriminar bem o target, podemos ver que pessoas que não trabalharam nos últimos 12 meses tem maior percentual do Target

Gastos_Sem_Renda x Target



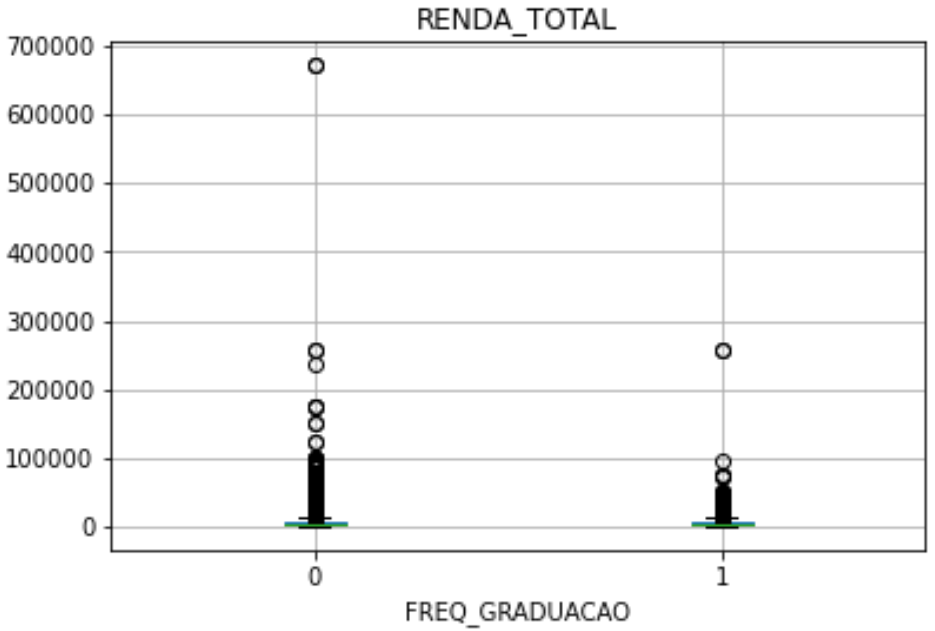
Gastos_Sem_Renda parece discriminar bem o target, podemos ver que pessoas que não tem gastos sem renda tem maior percentual do Target

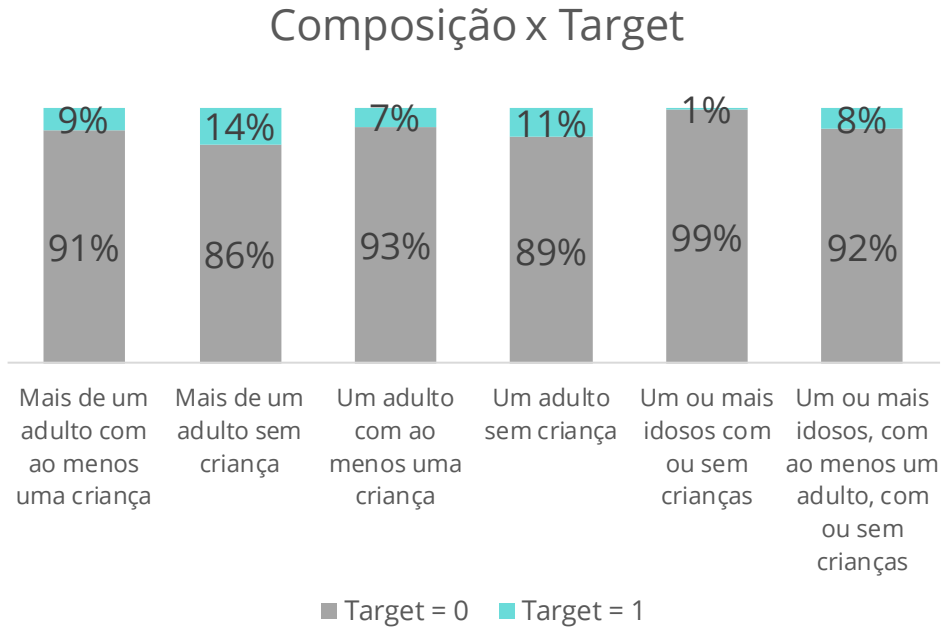
LEGENDA

- **Covariável parece explicar bem a resposta**
- **Covariável parece explicar pouco a resposta**
- **Covariável parece não explicar a resposta**

Análise detalhada

ANÁLISE EXPLORATÓRIA | BIDIMENSIONAL





LEGENDA

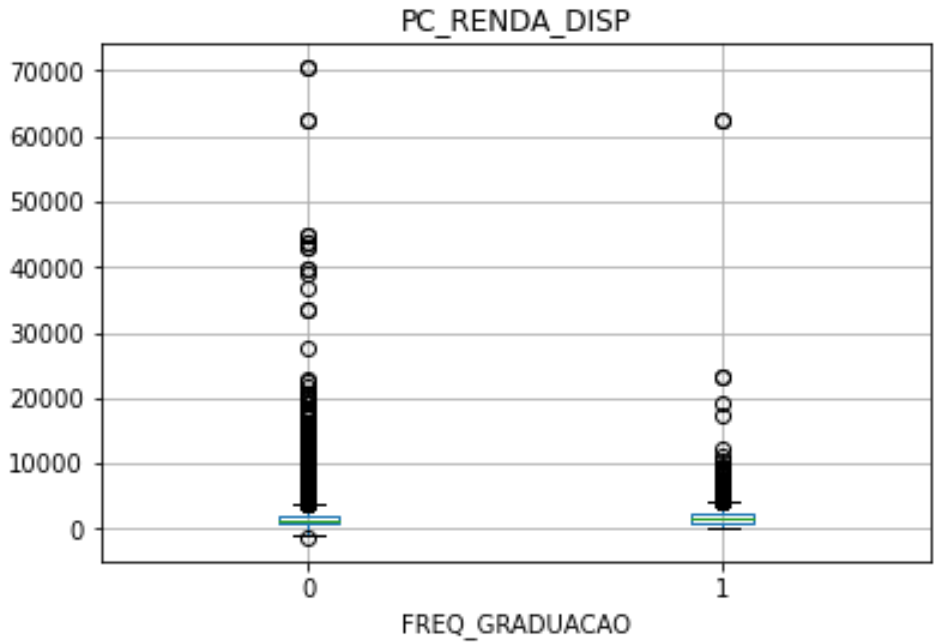
- **Covariável parece explicar bem a resposta**
- **Covariável parece explicar pouco a resposta**
- **Covariável parece não explicar a resposta**

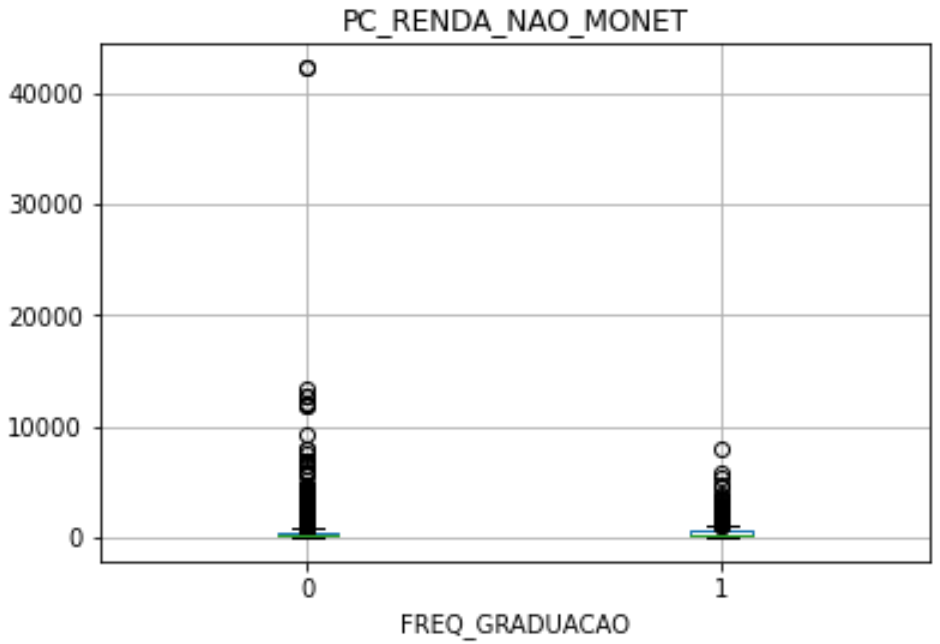
Renda_Total não parece discriminar bem o target, o que podemos ver é que quem não frequenta graduação possui maior outlier

Composição familiar parece discriminar bem o target, podemos ver os percentuais de cada categoria variam muito desde “Mais de um adulto sem criança” com 14% e “Um ou mais idosos com ou sem crianças” com 1%

Análise detalhada

ANÁLISE EXPLORATÓRIA | BIDIMENSIONAL





LEGENDA

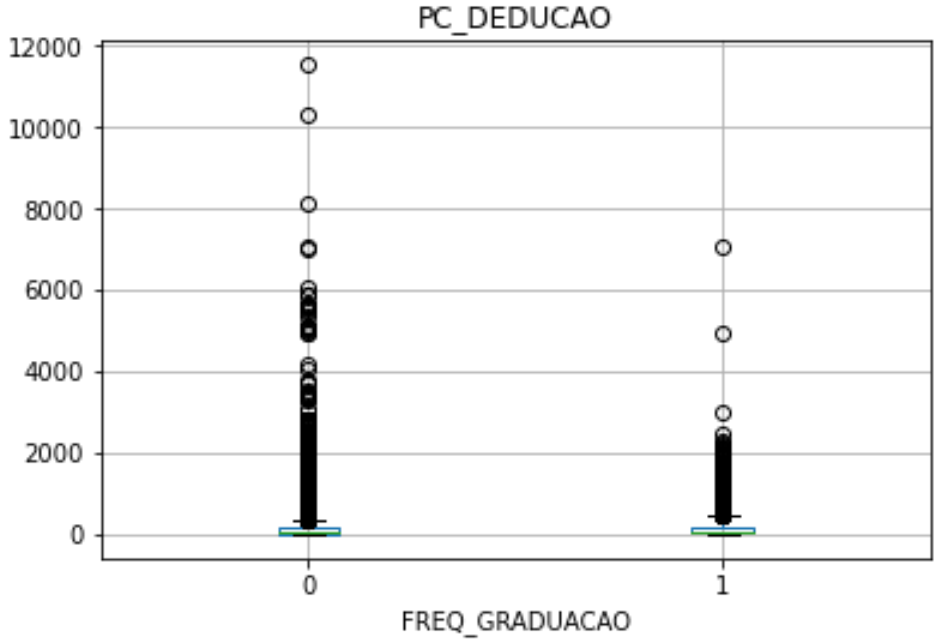
- **Covariável parece explicar bem a resposta**
- **Covariável parece explicar pouco a resposta**
- **Covariável parece não explicar a resposta**

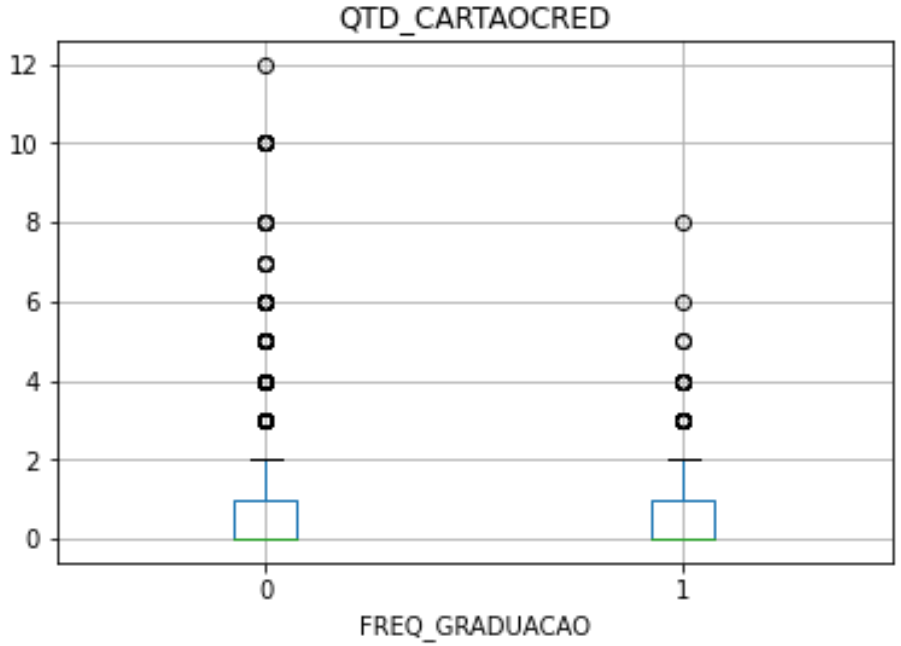
PC_Renda_Disp parece discriminar pouco o target, vemos que quem frequenta graduação possui mediana um pouco maior.

PC_Renda_Nao_Monet parece discriminar pouco o target, vemos que quem frequenta graduação possui mediana um pouco maior.

Análise detalhada

ANÁLISE EXPLORATÓRIA | BIDIMENSIONAL





LEGENDA

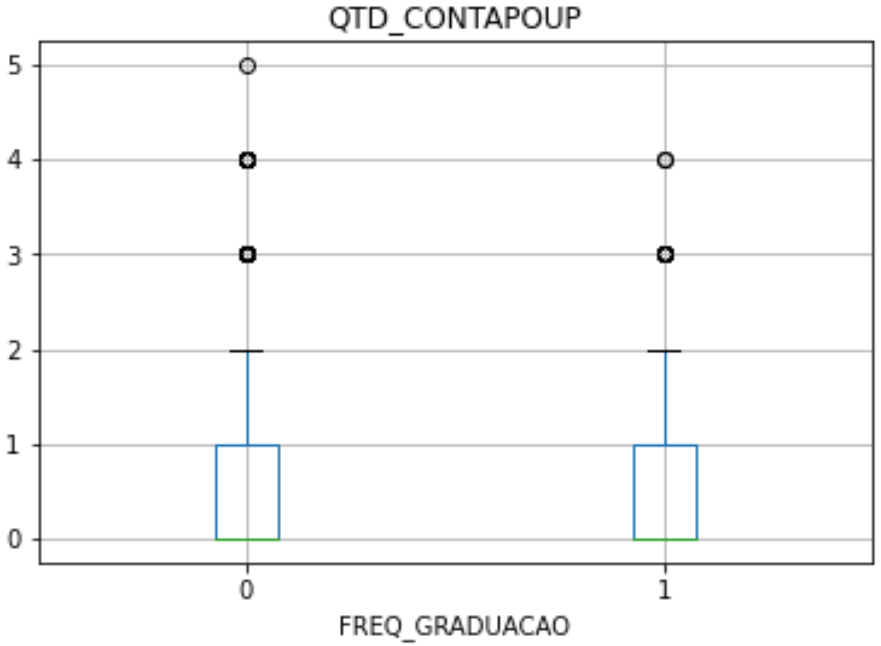
- Covariável parece explicar bem a resposta
- Covariável parece explicar pouco a resposta
- Covariável parece não explicar a resposta

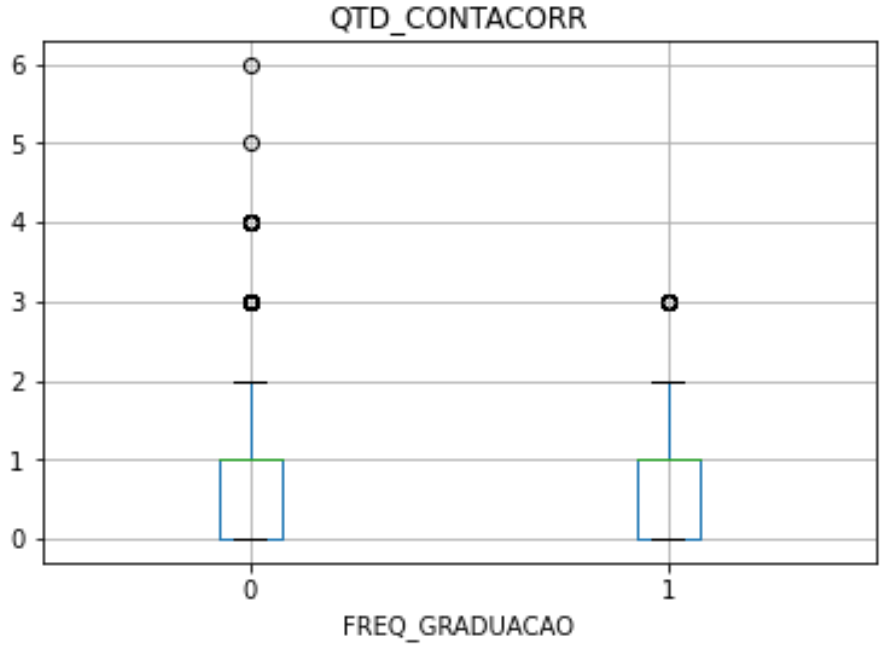
PC_Deducao não parece discriminar bem o target, o que podemos ver é que quem não frequenta graduação possui maior outlier

Qtd_Cartaocred não parece discriminar bem o target, o que podemos ver é que quem não frequenta graduação possui maior outlier

Análise detalhada

ANÁLISE EXPLORATÓRIA | BIDIMENSIONAL





LEGENDA

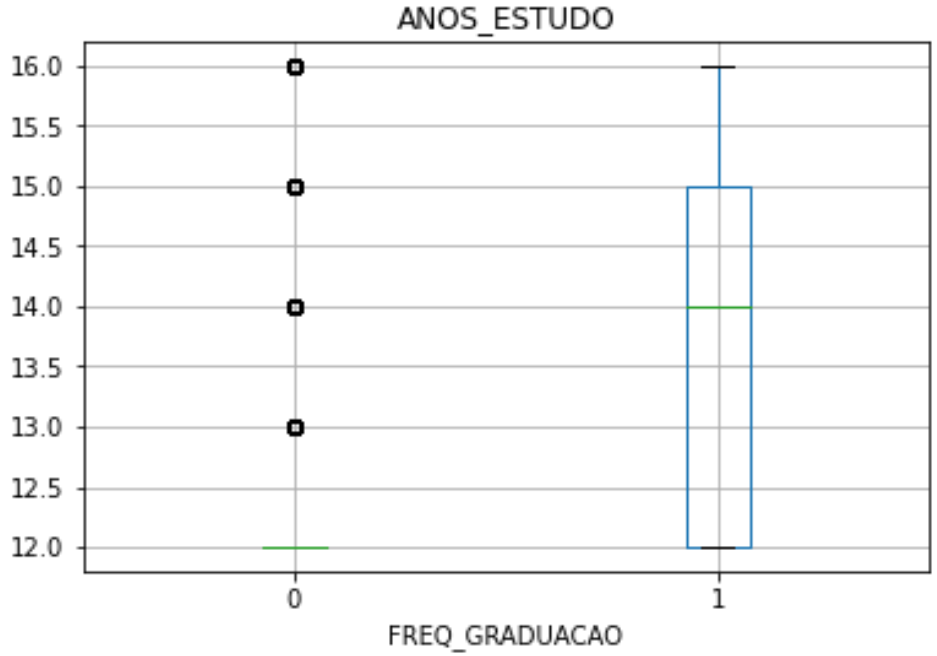
- **Covariável parece explicar bem a resposta**
- **Covariável parece explicar pouco a resposta**
- **Covariável parece não explicar a resposta**

Qtd_Contapoup não parece discriminar bem o target, o que podemos ver é que quem não frequenta graduação possui maior outlier

Qtd_Contacorr não parece discriminar bem o target, o que podemos ver é que quem não frequenta graduação possui maior outlier

Análise detalhada

ANÁLISE EXPLORATÓRIA | BIDIMENSIONAL



LEGENDA

- **Covariável parece explicar bem a resposta**
- **Covariável parece explicar pouco a resposta**
- **Covariável parece não explicar a resposta**

Anos_Estudo parece discriminar bem o target, quem frequenta um curso superior possui mais anos de estudo

Variáveis removidas – Detalhe 1

Variáveis só vem preenchidas, ou quase, quando a variável resposta é 0

		Freq_Graduacao	
Frequência		0	1
Ja_Freq_Escola	0	0,00	0,00
	1	1,00	0,00

		Freq_Graduacao	
Frequência		0	1
Conc_Curso_Ant	0	0,08	0,00
	1	0,92	0,00

		Freq_Graduacao	
Frequência		0	1
Curso_Mais_Elevado_Ant	Antigo científico, clássico, etc. (médio 2º ciclo)	0,04	0,00
	Doutorado	0,00	0,00
	Educação de jovens e adultos – EJA do ensino médio ou supletivo do 2º grau	0,03	0,00
	Especialização de nível superior (duração mínima de 360 horas)	0,02	0,00
	Mestrado	0,00	0,00
	Regular do ensino médio ou do 2º grau	0,75	0,00
	Superior – graduação	0,16	0,00

Variáveis removidas – Detalhe 2

Variáveis só vem preenchidas, ou quase, quando a variável resposta é 0

Frequência		Freq_Graduacao 0	Freq_Graduacao 1
Conc_1Periodo_Curso_Ant	Curso não classificado em séries ou anos	0,00	0,00
	Não	0,01	0,00
	Sim	0,99	0,00

Frequência		Freq_Graduacao 0	Freq_Graduacao 1
Ult_Periodo_Conc_Curso_Ant	Décimo	0,02	0,00
	Décimo primeiro	0,00	0,00
	Décimo segundo	0,00	0,00
	Nona(o)	0,00	0,00
	Oitava(o)	0,04	0,00
	Primeira(o)	0,01	0,00
	Quarta(o)	0,05	0,00
	Quinta(o)	0,01	0,00
	Segunda(o)	0,02	0,00
	Sexta(o)	0,01	0,00
	Sétima(o)	0,01	0,00
	Terceira(o)	0,82	0,00

Variáveis removidas – Detalhe 3

Variáveis utilizadas na construção da variável resposta

- Tipo_Curso
- Vai_Na_Escola

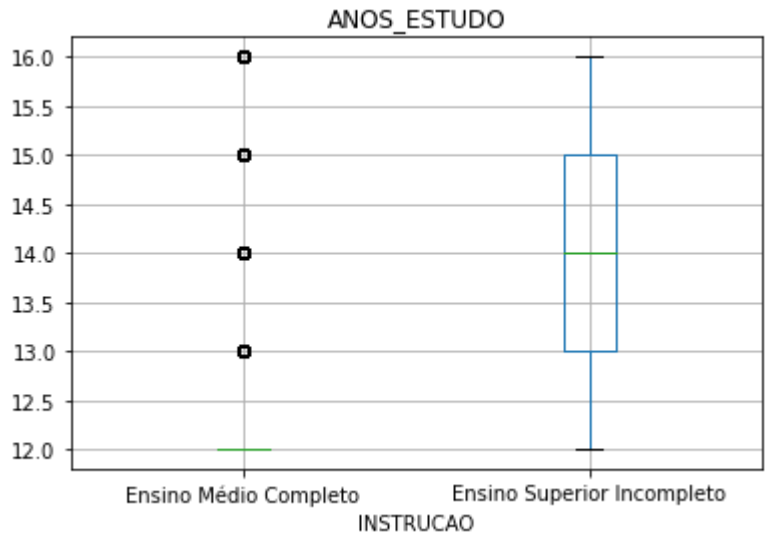
Variáveis removidas porque não entendi o significado

- Peso
- Peso_Final

Variável "Tipo_Situacao_Reg" removida porque é um resumo da variável "Estrato_Pof"

		Tipo_Situacao_Reg	
Frequência		Rural	Urbano
Estrato_Pof	Rural	0,10	0,00
	Urbano - Capital	0,00	0,30
	Urbano - Demais	0,00	0,43
	Urbano - Regiao Metropolitana	0,00	0,18

Variável "Instrucao" removida porque tem relação com a variável "Anos_Estudo"



Variáveis removidas – Detalhe 3

Variáveis removidas porque já temos a variável IDADE

- Dia_Nasc
- Mes_Nasc
- Ano_Nasc

Variáveis removidas porque é de identificação dos entrevistados

- Cod_Upa
- Num_Dom

- Cod_Upa
- Num_Dom

Outros Motivos

Variável	Motivo
Grau_Parentesco	Variável removida porque não faz diferença qual o grau de parentesco do entrevistado com a pessoa de referência da unidade de consumo
Morador_Presente	Variável removida porque não fará diferença saber se o morador estava presente na hora da entrevista
Sabe_Ler_Escrever	Variável removida porque todos na base sabem ler e escrever