TECH CHALLENGE 1 - FIAP

IA para DEVS, turma 5

Desenvolvimento de um modelo preditivo de regressão do valor dos financiamentos do Programa Minha Casa Minha Vida do Governo Federal*.

Um estudo para embasar o lançamento da nova faixa do programa social do governo federal, a faixa 4.

*Contratos do MCMV-Financiado com FGTS (dados analíticos) – Apresenta dados sobre os contratos de financiamento do FGTS, nas faixas de renda do MCMV, ao patamar de pessoa física.

Equipe:
João Helton RM364239
João Almeida Furtado Neto RM364164
Murilo Polli RM364642
Rafael Pinheiro RM363960
Ricardo Honorato RM364026

1. Links:

a. Descrição do Trabalho:
b. Github: https://github.com/ricardohonorato/FIAP-IADEVS-TC1
c. Video: https://www.youtube.com/watch?v=hsimJHc0Mk0

2. Problemática:

Devido a necessidade de expandir o Programa Minha Casa Minha Vida para faixas de rendas maiores, com essa necessidade o governo anunciou a criação de uma nova faixa de renda, a faixa 4.

Conforme notícia:
https://agenciabrasil.ebc.com.br/economia/noticia/2025-05/caixa-comeca-oferecer-minha-casa-minha-vida-para-classe-media

Para termos essa ampliação do programa, deve-se ter um estudo de viabilidade e uma projeção de impacto na economia e no orçamento.

Utilizando o “chapéu” do governo, estamos propondo um modelo de regressão para que possamos estimar o impacto orçamentário para essa nova faixa e ver a análise de viabilidade.

Usamos o arquivo “**dados_abertos_FGTS_ANALITICO_202505.csv”**
que tem tamanho de 160 megabytes compactado utilizando o .rar, já descompactado tem 1,122 gigabytes e possui 6.911.513 linhas.

3. Dicionário de dados:

Col	Nome	Descrição	Tipo	Atributos
1	data_referencia	Data de referência da geração dos dados.	Texto	
2	cod_ibge	Código de 6 dígitos do município, segundo a Tabela de Códigos de Municípios do IBGE, excetuando-se o dígito verificador.	Texto	

3	txt_municipio	Nome do Município.	Texto	
4	mcmv_fgts_txt_uf	UF em que o município se localiza.	Texto	
5	txt_regiao	Região em que a UF se localiza.	Texto	
6	data_assinatura_financiamento	Data da contratação do financiamento.	Data	
7	qtd_uh_financiadas	Quantidade de unidades habitacionais contratadas.	Número	
8	vlr_financiamento	Valor total das operações de crédito realizadas entre o agente financeiro e os mutuários.	Número	
9	vlr_subsidio_desconto_fgts	Valor total do desconto concedido pelo FGTS (Fundo de Garantia do Tempo de Serviço) na operação de crédito para abatimento do valor a ser financiado pelo mutuário.	Número	
10	vlr_subsidio_desconto_ogu	Valor total do desconto concedido pelo OGU (Orçamento Geral da União) na operação de crédito para abatimento do valor a ser financiado pelo mutuário.	Número	
11	vlr_subsidio_equilíbrio_fgts	Valor total do desconto concedido pelo FGTS(Fundo de Garantia do Tempo de Serviço) na operação de crédito para diminuição da taxa de juros a ser aplicada ao financiamento.	Número	
12	vlr_subsidio_equilíbrio_ogu	Valor total do desconto concedido pelo OGU (Orçamento Geral da União) na operação de crédito para diminuição da taxa de juros a ser aplicada ao financiamento.	Número	
13	vlr_compra	Descreve o valor pedido pelo vendedor, normalmente o mesmo que o valor de garantia.	Número	
14	vlr_renda_familiar	Representa o valor de renda familiar do mutuário/contratante, utilizada para a contratação	Número	

		do financiamento.		
15	txt_programa_fgts	Descreve o programa vinculado à área responsável pela aplicação dos recursos contratados.	Texto	Apoio à Produção ; Carta de Crédito Associativo; Carta de Crédito Individual ; Faixa Estendida; PróCotista
16	num_taxa_juros	Representa o índice da taxa de juros inicial vinculada ao empréstimo contratado, em forma percentual	Número	
17	txt_tipo_imovel	Descreve se o imóvel é Novo ou Usado	Texto	Novo; Usado
18	bln_cotista	Descreve se o mutuário/contratante é titular de conta vinculada do FGTS ou não, no formato (S ou N)	Texto	S; N; X
19	txt_sistema_amortizacao	Descreve o sistema de amortização do contrato de financiamento.	Texto	price; sac; sacre
20	dte_nascimento	Data de nascimento do beneficiário.	Data	
21	txt_compatibilidade_faixa_renda	Descreve a faixa de renda do beneficiário conforme o enquadramento normativo no Programa Minha Casa Minha Vida.	Texto	Faixa 1; Faixa 2; Faixa 3; FORA MCMV/CVA
22	txt_nome_empreendimento	Nome do empreendimento, quando disponível.	Texto	

4. Determinando tamanho da amostra:

Antes de extrair a amostra, você precisa saber quantas linhas são necessárias para que ela seja "estatisticamente relevante". Dado isso, imagine se tivessemos 6 milhões de linhas teriamos o seguinte cenário:

- **Tamanho da População (N):** 6.000.000
- **Nível de Confiança:** Geralmente 95% (mas pode ser 99%). É a probabilidade de a amostra refletir a população real.
- **Margem de Erro:** Geralmente 5% (mas pode ser menor, como 3% ou 1%). É o quanto os resultados da sua amostra podem desviar dos resultados reais da população.

Para uma população tão grande (6 milhões), o tamanho da amostra se estabiliza. Usando uma calculadora de tamanho de amostra online ou a fórmula:

$$n = \frac{Z^2 \times p \times (1-p)}{E^2}$$

Onde:

- Z é o Z-score (1.96 para 95% de confiança)
- p é a proporção estimada (use 0.5 para o pior caso, que maximiza o tamanho da amostra)
- E é a margem de erro (0.05 para 5%)
- **Para 95% de confiança e 5% de erro:** Você precisará de aproximadamente **385** amostras.
- **Para 95% de confiança e 3% de erro:** Você precisará de aproximadamente **1067** amostras.
- **Para 95% de confiança e 1% de erro:** Você precisará de aproximadamente **9600** amostras.

Escolha o tamanho (`n`) com base na precisão que você necessita. Vamos usar `n = 1067` como exemplo.

5. Análise dos Resultados:

Abaixo temos os resultados e suas análises:

Utilizamos as seguintes variáveis:

1) Valor do Financiamento (vlr_financiamento): Valor total das operações de crédito realizadas entre o agente financeiro e os mutuários.

2) Valor de Compra (vlr_compra): Descreve o valor pedido pelo vendedor, normalmente o mesmo que o valor de garantia.

3) Valor da Renda Familiar (vlr_renda_familiar): Representa o valor de renda familiar do mutuário/contratante, utilizada para a contratação do financiamento

Utilizamos uma amostra de 100 mil linhas de todo o universo (quase 7 milhões de registros) para fazer essa análise. Abaixo algumas estatísticas detalhadas e interpretações.

```
--- Estatísticas Detalhadas para: vlr_financiamento ---
  Medidas de Tendência Central:
    Média: 100450.20
    Mediana: 96000.00
    Moda: [112000.0]


  Medidas de Dispersão:
    Desvio Padrão: 41040.33
    Variância: 1684308772.69
    Amplitude (Range): 299993.00
    Intervalo Interquartil (IQR): 49224.73


  Medidas de Forma:
    Assimetria (Skewness): 0.82
    Curtose (Kurtosis): 1.55


  Percentis Específicos:
    Percentil 10: 54000.00
    Percentil 90: 152000.00
----------------------------------------
```

```
--- Estatísticas Detalhadas para: vlr_compra ---
  Medidas de Tendência Central:
    Média: 135215.97
    Mediana: 131000.00
    Moda: [145000.0]

  Medidas de Dispersão:
    Desvio Padrão: 56828.40
    Variância: 3229467415.28
    Amplitude (Range): 766550.00
    Intervalo Interquartil (IQR): 75265.67

  Medidas de Forma:
    Assimetria (Skewness): 0.74
    Curtose (Kurtosis): 1.58

  Percentis Específicos:
    Percentil 10: 68500.00
    Percentil 90: 209176.00
----------------------------------------
```

```
--- Estatísticas Detalhadas para: vlr_renda_familiar ---
  Medidas de Tendência Central:
    Média: 2738.31
    Mediana: 2300.22
    Moda: [1600.0]

  Medidas de Dispersão:
    Desvio Padrão: 1370.38
    Variância: 1877931.35
    Amplitude (Range): 7800.00
    Intervalo Interquartil (IQR): 1474.52

  Medidas de Forma:
    Assimetria (Skewness): 1.45
    Curtose (Kurtosis): 1.94

  Percentis Específicos:
    Percentil 10: 1517.72
    Percentil 90: 4753.68
----------------------------------------
```

```
Testes de Normalidade:

Coluna: vlr_financiamento
  -> Estatística do Teste: 11239.948925398769
  -> p-valor: 0.0
  -> A distribuição não é normal (rejeita-se a hipótese nula).

Coluna: vlr_compra
  -> Estatística do Teste: 10124.62035762497
  -> p-valor: 0.0
  -> A distribuição não é normal (rejeita-se a hipótese nula).

Coluna: vlr_renda_familiar
  -> Estatística do Teste: 22670.879672641902
  -> p-valor: 0.0
  -> A distribuição não é normal (rejeita-se a hipótese nula).
```

faixa_etaria	proportion
Sem Data de nascimento cadastrada	41.593777
Adultos plenos (31-45 anos)	31.628097
Jovens (15-30 anos)	15.474077
Adultos seniores (45 anos ou mais)	9.066262
Idosos (Acima de 60 anos)	2.237786

dtype: float64

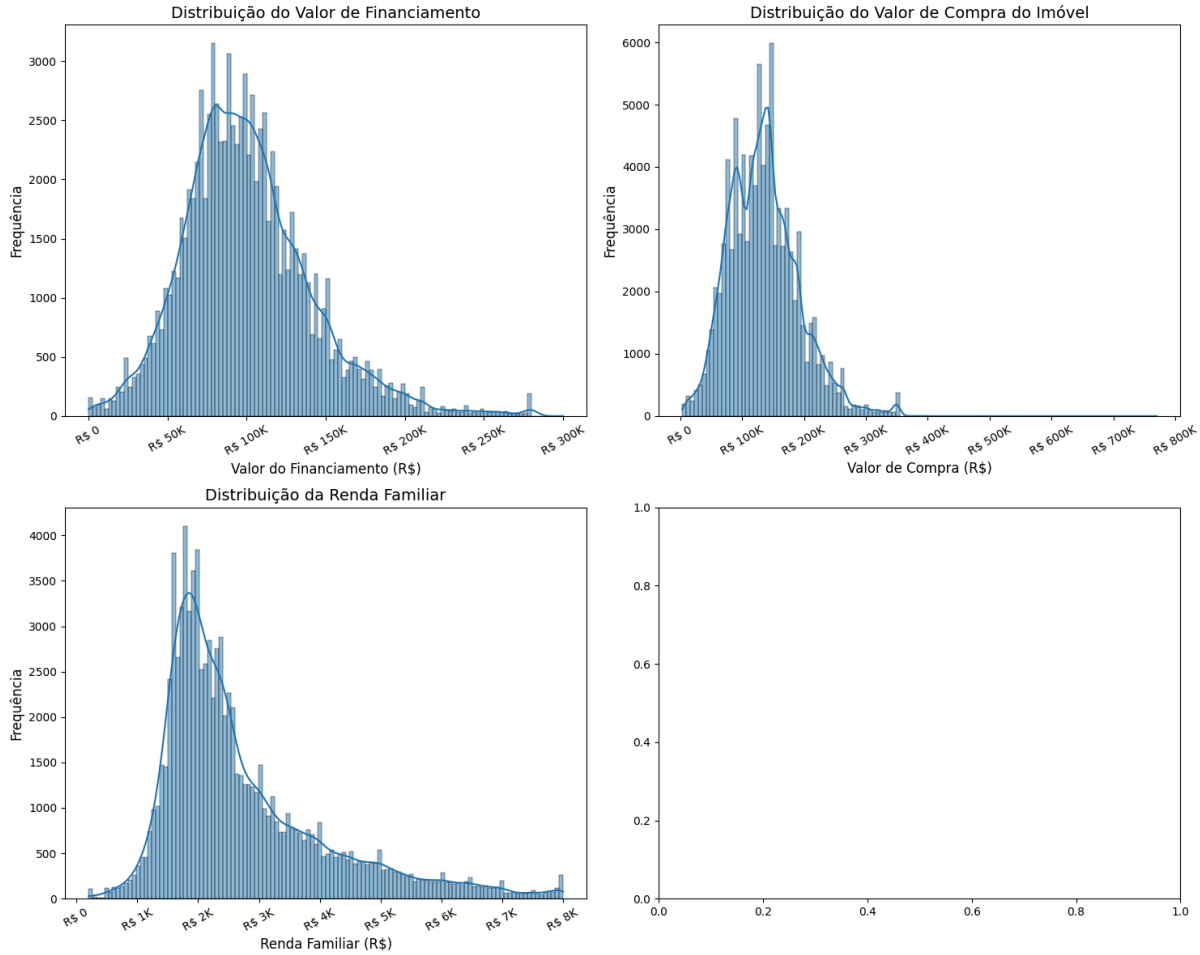
Conforme a figura acima, temos alguns contratos que não tem data de nascimento cadastrada, mas a maioria dos contratos são feitos com adultos em fase laboral já estabilizada.

```
Distribuição por faixa de renda familiar:
faixa_renda_familiar
Entre 1 e 2 salários minimos    56782
Entre 2 e 3 salários minimos    16340
Menos que um salário minimo      9303
Entre 3 e 4 salários minimos     6994
Maior que 4 salários minimos     3530
Name: count, dtype: int64
```

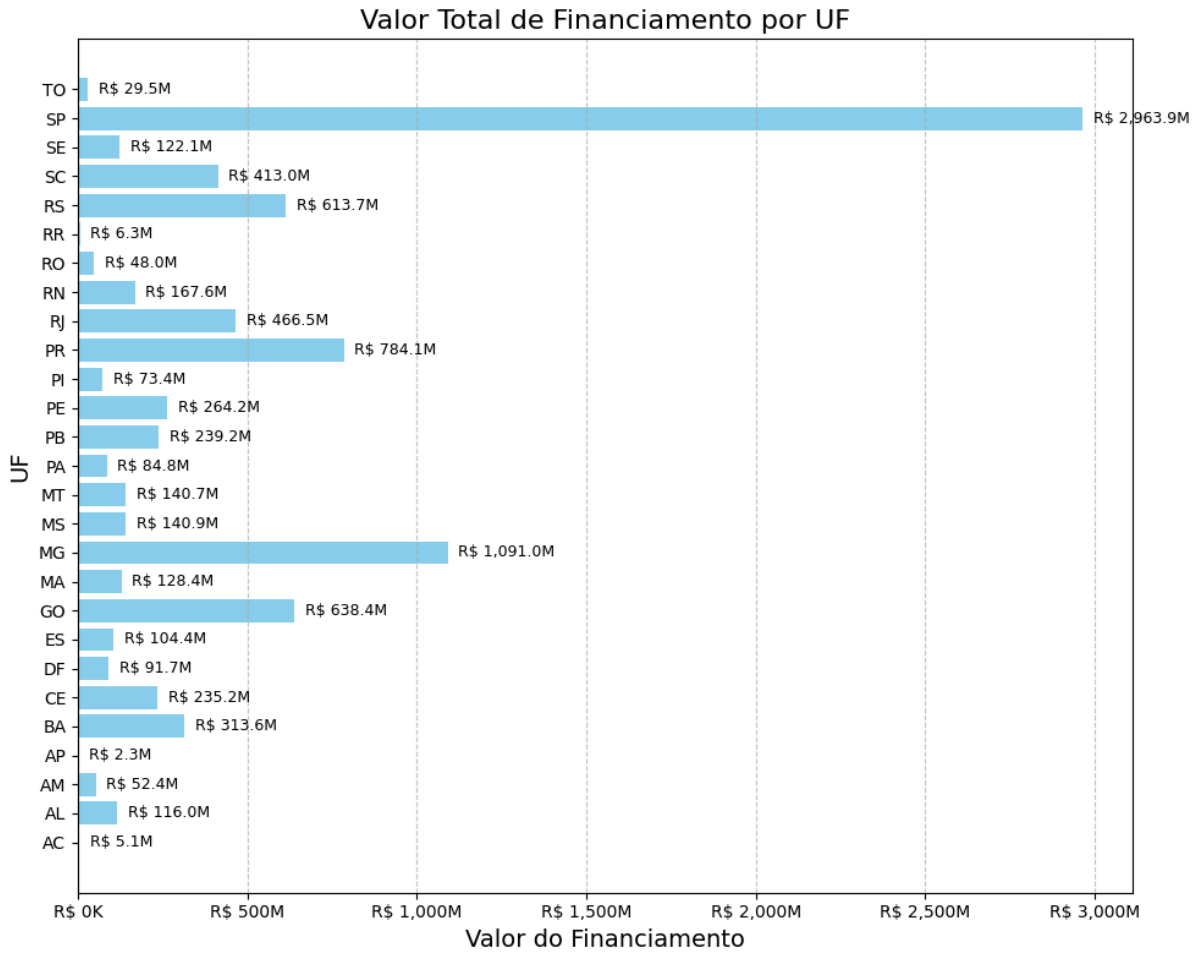
Conforme acima, das 100 mil linhas, mais da metade da renda familiar está entre 1 a 2 salários mínimos e se considerarmos até 3 salários temos mais de 72% das contratações.

Utilizamos uma amostra de 10.000 linhas, inicialmente para desenvolvermos o nosso trabalho. Obtivemos uma acurácia de 85%, ao aumentarmos a nossa amostra para 50.000 linhas tivemos uma melhora para 88 a 90%. Essa variação ocorreu por conta dos dados. Ao aumentarmos para 100.000 linhas o modelo superou 92%.

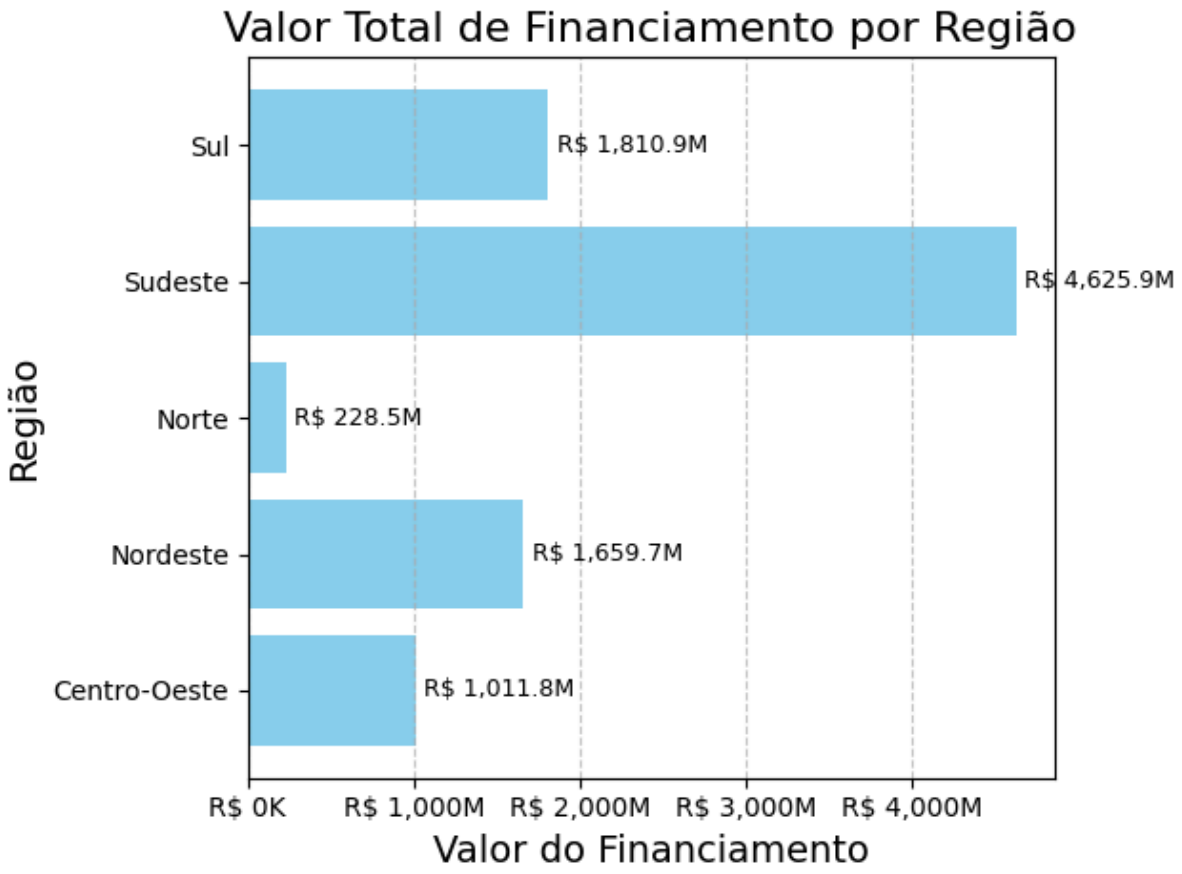
Com isso identificamos que a seleção da amostra e o percentual da divisão dos dados para testes e treinamento afetam bastante os resultados dos modelos.

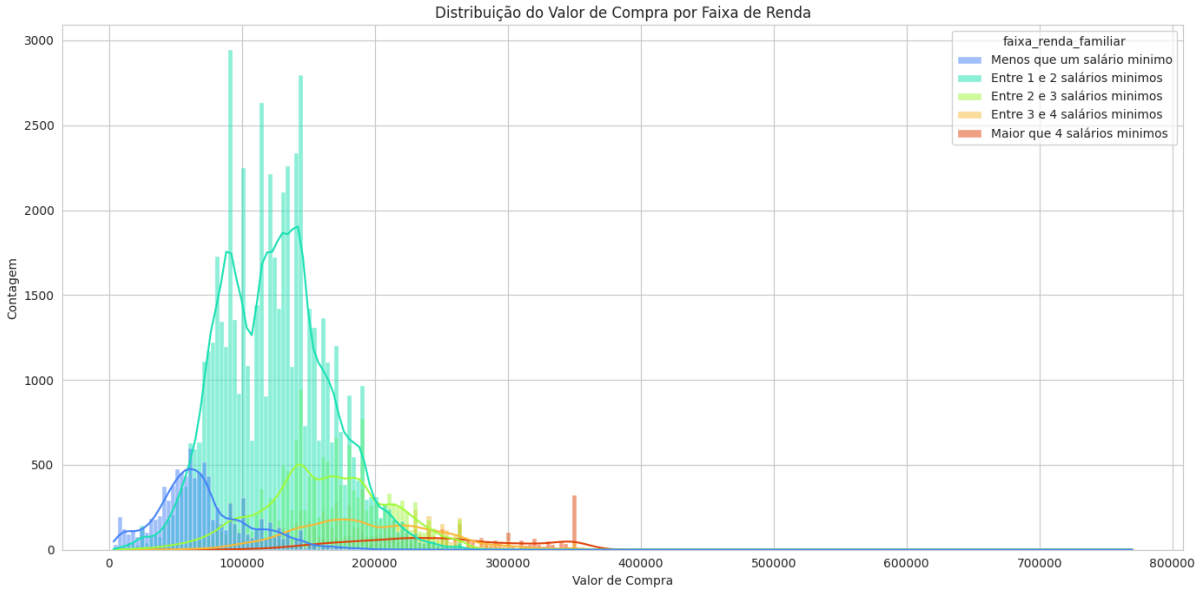


Podemos notar que há uma assimetria nas distribuições para a esquerda, o que denota menores valores de compra e financiamento. Isso corresponde à expectativa do programa do governo, uma vez que os imóveis financiados são, em sua grande maioria, destinados a pessoas de baixa renda.

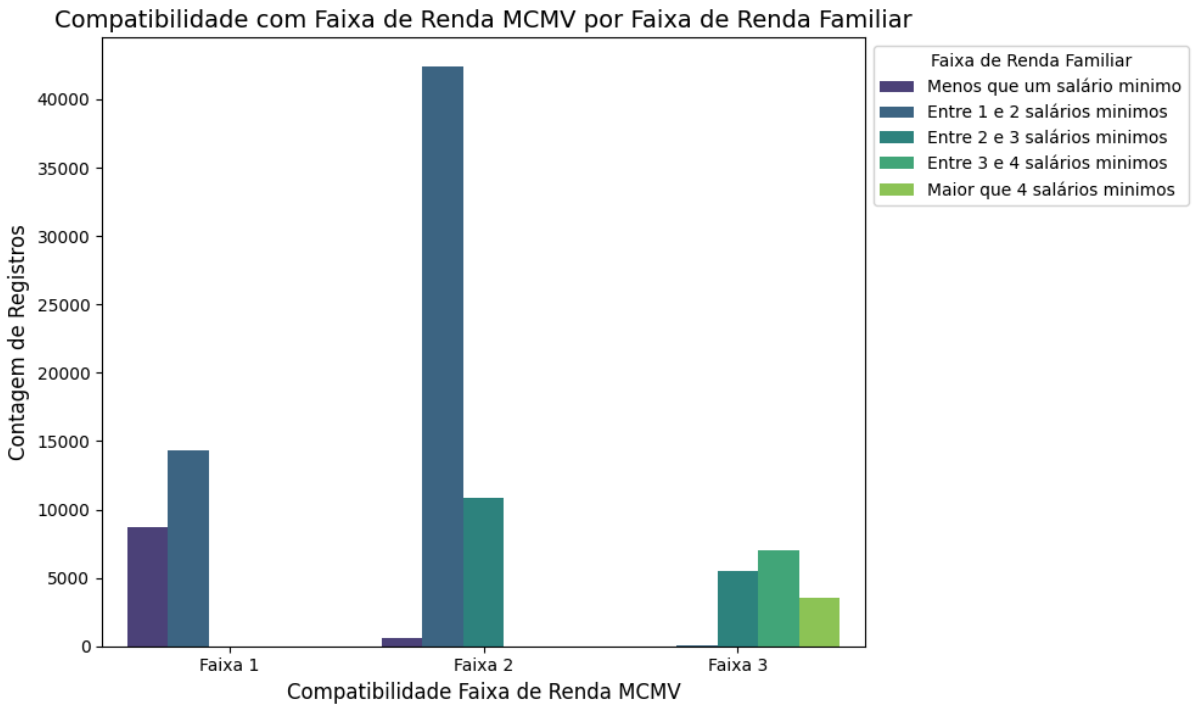


Podemos notar que os estados de São Paulo, Minas Gerais e Paraná se apresentam como outliers em uma visualização boxplot do valor financiado agregado pela UF. Entretanto, eles não são outliers dado que esses valores são reais e não erros cadastrais.

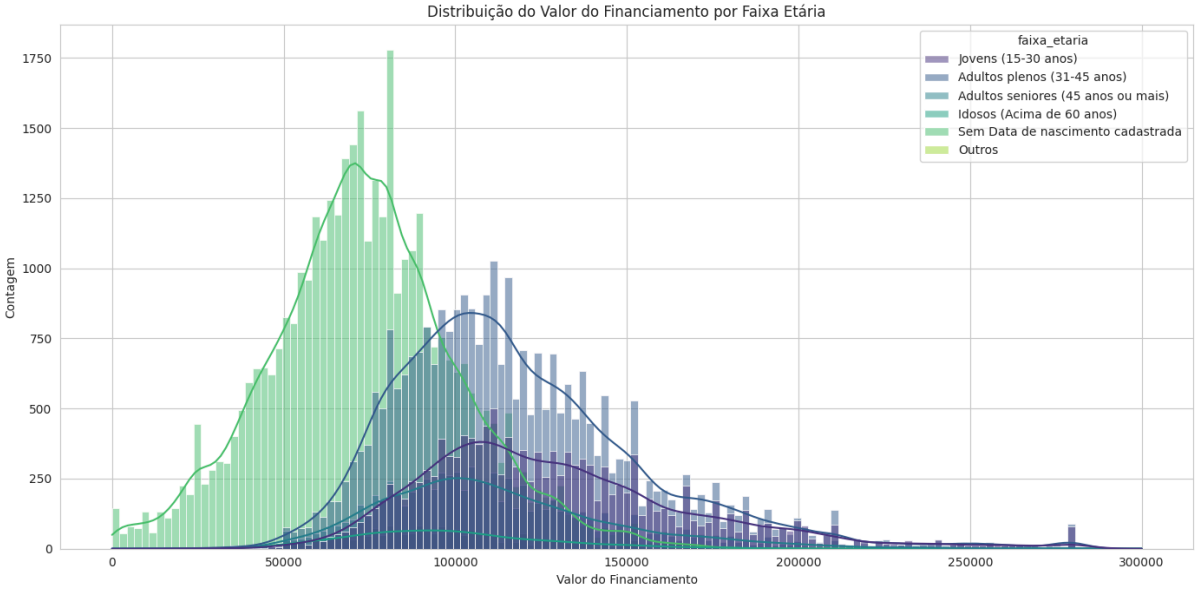




Observamos que o programa de financiamentos pelo programa Minha Casa Minha Vida está atingindo majoritariamente pessoas com renda familiar de até 3 salários mínimos sendo a significativa maioria concentrada em que recebe entre 1 e 2 salários.



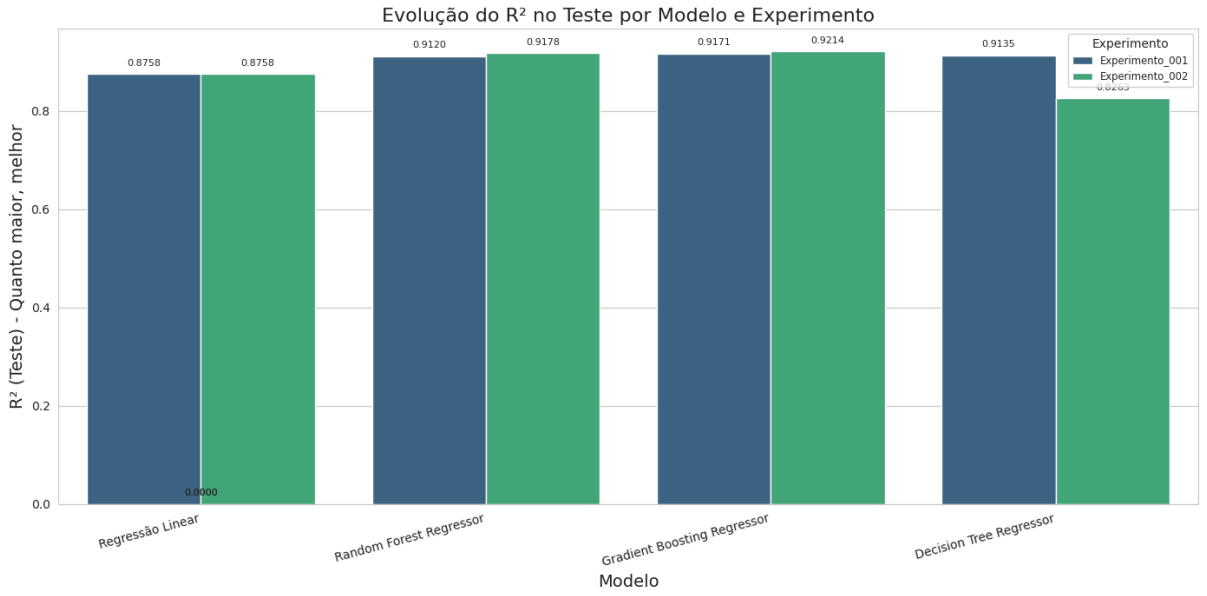
Nosso último gráfico mostra que a maioria dos financiamentos estão na faixa 2 e que a cada nova faixa criada um novo grupo de pessoas é englobado pelo programa. Apesar da faixa 2 ser a mais representativa, as faixas 1 e 3 permitem a cobertura tanto das classes menos favorecidas quanto das pessoas da classe média que ainda não possuem seu imovel.

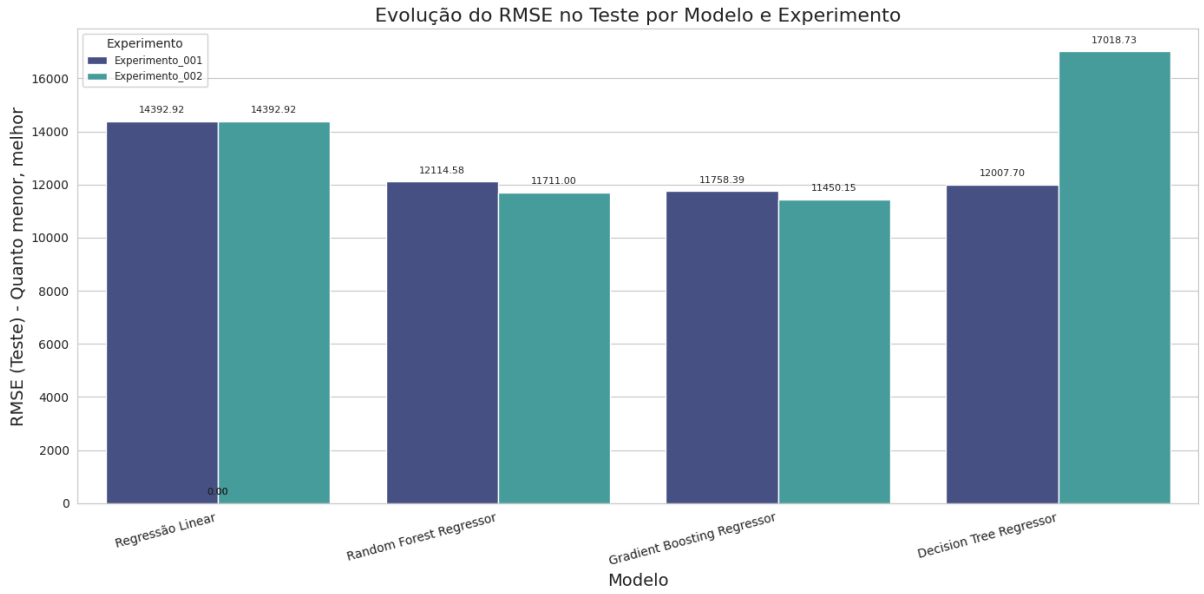


Existe um grande volume de financiamento cujo beneficiário não tem data de nascimento cadastrada, possivelmente por falha de cadastro ou pelo beneficiário não saber a informação. Eliminando esses dados percebe-se que a maioria dos financiamentos são realizados na faixa de adultos plenos onde eles já estão estabilizados no mercado de trabalho e ainda novos para cobrir todo o prazo de financiamento. Os adultos seniores e os

idosos mostram que o número de contemplações do programa diminui com a idade, talvez também devido ao tempo médio de financiamento ser de 30 a 35 anos e a expectativa de vida do brasileiro ser na casa dos 75 anos.

Abaixo segue a matriz de correlação:





Pelos gráficos, constatamos que a configuração dos hiperparâmetros melhorou o resultado obtido pelo Gradient Boosting Regressor.

6. Desafios Futuros/ Insights:

- ☐ Finops;
- ☐ Criar esteira devops (ingestão, processamento e armazenamento);
- ☐ Aumentar a quantidade de dados, chegando a usar toda a série temporal e observar os resultados;
- ☐ Fazer mais análises e inferências cruzando dados:
  - ☐ Do SCR, nível de endividamento das pessoas do programa;
  - ☐ Fazer uma análise dos indicadores do programa (PGI); (https://dados.gov.br/dados/conjuntos-dados/minha-casa-melhor)
  - ☐ Analisar o impacto na economia do programa, por exemplo saber o impacto desse programa na construção civil, no mercado de material de construções e quantos empregos gerados pelo programa;
  - ☐ Utilizar dados do open finance;
- ☐ **Impacto de usar SMART CONTRACT;**
- ☐ Ingestão de dados online X batch;
- ☐ Uso de data lake;
- ☐ Usar o conceito de data mesh;

7. Fonte/Referências:

Programa:
https://www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/habitacao/programa-minha-casa-minha-vida/

Dicionários:
https://www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/habitacao/programa-minha-casa-minha-vida/arquivos/Dicionarios_SNH_2025_04_14.pdf

Dados:
https://www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/habitacao/programa-minha-casa-minha-vida/bases-de-dados-do-programa-minha-casa-minha-vida