

# Relatório Detalhado de Análise de Modelagem: Predição do Aluguel Estimado

Análise baseada no notebook `03_Predição_Aluguel_Estimado.ipynb`

14 de outubro de 2025

## Resumo

Este relatório apresenta uma análise aprofundada do processo de modelagem realizado para prever o valor do “Aluguel Estimado”. A análise cobre desde o pré-processamento dos dados até a avaliação do modelo final e a verificação de suas suposições estatísticas.

## 1 Resumo Executivo

O projeto teve como objetivo criar um modelo de regressão para prever o “Aluguel Estimado” com base em características quantitativas e qualitativas de domicílios.

- **Metodologia:** A abordagem foi bem estruturada, iniciando com um pré-processamento cuidadoso que incluiu a codificação de variáveis categóricas (**One-Hot Encoding**) e a padronização (**StandardScaler**) das variáveis numéricas e da variável alvo. Um processo robusto de **seleção de características progressiva (forward selection)** foi implementado para comparar o desempenho da **Regressão Linear** e do **Random Forest Regressor** com um número crescente de variáveis (de 1 a 14).
- **Resultados da Seleção:** A análise comparativa demonstrou que o **Random Forest** consistentemente superou a **Regressão Linear**, atingindo um  $R^2$  de **0.60** com 10 variáveis, ponto em que os ganhos de performance começaram a diminuir. Com base nisso, um conjunto ótimo de 10 características foi selecionado.
- **Modelo Final e Avaliação:** Curiosamente, embora o **Random Forest** tenha sido superior na seleção, o modelo final treinado foi uma **Regressão Linear** utilizando as 10 características identificadas. Este modelo, avaliado em um conjunto de teste (30% dos dados), alcançou um **Coefficiente de Determinação ( $R^2$ ) de 0.54**, indicando que explica aproximadamente 54% da variância no aluguel estimado.
- **Verificação de Suposições:** A análise aprofundada das suposições do modelo linear final revelou pontos críticos:
  - **Suposições Atendidas:** Ausência de multicolinearidade (fatores VIF baixos) e independência dos resíduos (estatística Durbin-Watson próxima de 2).
  - **Suposições Violadas:** Os testes estatísticos (Shapiro-Wilk e Breusch-Pagan) e a análise gráfica **rejeitaram fortemente as suposições de normalidade e homocedasticidade dos resíduos**.

- **Conclusão e Recomendações:** O processo de engenharia e seleção de características foi excelente. No entanto, as violações significativas das suposições do modelo linear final comprometem sua robustez estatística e a confiabilidade das inferências. Recomenda-se fortemente a exploração de modelos alternativos que não dependem dessas suposições (como o próprio `Random Forest`) ou a aplicação de transformações na variável alvo (ex: logarítmica) para tentar corrigir as violações e construir um modelo preditivo mais preciso e confiável.

## 2 Análise Detalhada da Metodologia

### 2.1 Pré-processamento de Dados

O tratamento inicial dos dados foi executado seguindo boas práticas de mercado:

1. **Separação de Variáveis:** As colunas foram corretamente divididas em quantitativas (`quanti_cols`) e qualitativas (`quali_cols`), permitindo a aplicação de técnicas de pré-processamento distintas e adequadas para cada tipo.
2. **Codificação de Categóricas:** Foi utilizado o método `pd.get_dummies` com o parâmetro `drop_first=True`. Esta é uma excelente escolha, pois converte as variáveis categóricas em um formato numérico enquanto previne a multicolinearidade perfeita.
3. **Padronização (Scaling):** O `StandardScaler` foi aplicado tanto às variáveis numéricas (`quanti_cols`) quanto à variável alvo (`Aluguel Estimado`). A padronização das features é fundamental para modelos de regressão, garantindo que todas as variáveis contribuam de forma equilibrada.

### 2.2 Seleção de Características e Modelos (Forward Selection)


Esta foi a etapa mais robusta do notebook. Um método de seleção progressiva foi implementado para identificar o melhor subconjunto de variáveis.

- **Modelos Comparados:** Foram testados dois algoritmos distintos: `Linear Regression` e `RandomForestRegressor`.
- **Processo Iterativo:** O código iterou de  $k = 1$  a  $k = 14$ , construindo e avaliando modelos para cada valor de  $k$ .
- **Métricas de Avaliação:** As métricas  $R^2$  (Coeficiente de Determinação) e  $SQE$  (Soma dos Quadrados dos Erros) foram armazenadas para cada modelo.

## 3 Análise dos Resultados da Seleção

Os resultados, compilados no `resultados_df` e visualizados nos gráficos, são muito informativos.

- **Comparação de Desempenho:** O `Random Forest Regressor` (linha laranja) apresentou um desempenho consistentemente superior à `Regressão Linear` (linha azul) em ambas as métricas.



grafico\_r2\_sqe.png

Figura 1: Gráfico de  $R^2$  e SQE vs. Número de Características ( $k$ ) para ambos os modelos.

- **Ponto de Saturação (Elbow Point):** O gráfico de  $R^2$  vs.  $k$  mostra um “cotovelo” claro. Para o Random Forest, o ganho de  $R^2$  é significativo até  $k = 10$  ( $R^2 \approx 0.60$ ). Após esse ponto, os ganhos são marginais, indicando um excelente equilíbrio entre complexidade e poder preditivo.
- **Seleção Final das Características:** A escolha das 10 características identificadas pelo Random Forest foi uma decisão acertada. As variáveis selecionadas foram:
  1. Qtd de banheiros exclusivos
  2. Rendimento mensal mínimo geral (R\$)
  3. A água é aquecida por energia elétrica?\_Sim
  4. Material do telhado\_Outro material

5. Material do telhado\_Telha sem laje de concreto
6. Material do piso\_Cimento
7. Tipo de escoadouro sanitário\_Rede geral, ...
8. A água é aquecida por lenha ou carvão?\_Sim
9. A água é aquecida por gás?\_Sim
10. Tipo do domicílio\_Casa

## 4 Análise do Modelo Final: Regressão Linear

- **Performance no Conjunto de Teste:**
  - $R^2 = 0.54$ : O modelo final foi capaz de explicar **54%** da variância do “Aluguel Estimado” no conjunto de teste.
  - **MSE = 0.46**: O Erro Quadrático Médio de 0.46 está na escala padronizada. A Raiz do Erro Quadrático Médio (RMSE) seria  $\sqrt{0.46} \approx 0.68$ , indicando que o erro médio das previsões é de aproximadamente 0.68 desvios padrão.

## 5 Verificação Detalhada das Suposições da Regressão Linear

### 5.0.1 Homocedasticidade (Variância Constante dos Erros)

- **Análise Gráfica:** O gráfico de “Resíduos vs. Valores Previstos” mostra um padrão claro de **cone**, onde a dispersão dos resíduos aumenta à medida que o valor previsto aumenta.
- **Teste de Breusch-Pagan:** O teste confirmou a suspeita visual, com um **p-valor de 0.0000**, rejeitando a hipótese nula de homocedasticidade.
- **Consequência:** A presença de **heterocedasticidade** torna os erros padrão do modelo não confiáveis, invalidando testes de hipótese.

### 5.0.2 Normalidade dos Resíduos

- **Análise Gráfica e Teste de Shapiro-Wilk:** Tanto o histograma quanto o teste estatístico (com **p-valor de 0.0000**) rejeitaram a hipótese de que os resíduos seguem uma distribuição normal.
- **Consequência:** A **não-normalidade dos resíduos** afeta a validade dos intervalos de confiança e testes de hipótese.

### 5.0.3 Independência dos Resíduos

- **Análise Gráfica e Teste de Durbin-Watson:** O gráfico de resíduos por ordem não mostrou padrão, e a estatística de Durbin-Watson foi de **1.9927** (próximo de 2.0).

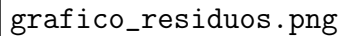


Figura 2: Gráfico de Resíduos vs. Valores Previstos, mostrando heterocedasticidade.

- **Consequência:** A suposição de independência dos resíduos **foi atendida**.

#### 5.0.4 Ausência de Multicolinearidade

- **Fator de Inflação de Variância (VIF):** Todos os valores de VIF foram baixos (o maior foi 1.39), bem abaixo do limite de 5.
- **Consequência:** A suposição de ausência de multicolinearidade **foi atendida**.

## 6 Síntese Geral e Recomendações

### Pontos Fortes:

- Processo de pré-processamento e seleção de características bem executado.
- Comparação robusta entre diferentes modelos.
- Análise de suposições do modelo linear extremamente completa.

**Pontos de Melhoria e Recomendações:** As violações das suposições de **homocedasticidade** e **normalidade** são os pontos mais críticos. Sugere-se:

1. **Transformar a Variável Alvo:** Aplicar uma **transformação logarítmica** (ex:  $\text{np.log1p}(y)$ ) na variável alvo para estabilizar a variância e normalizar os resíduos.

2. **Utilizar o Modelo Random Forest:** Este modelo não exige as suposições lineares e já demonstrou desempenho superior. Seria a escolha mais lógica para o modelo final.
3. **Explorar Regressão Robusta:** Modelos como `HuberRegressor` são menos sensíveis a outliers e violações de suposições.