Instituto Federal de Goiás – IFG – Câmpus Goiânia

Especialização em Inteligência artificial aplicada

PROJETO FINAL DO MÓDULO 2

PREVISÃO DE PREÇO ATRAVÉS DE MODELO DE REGRESSÃO DE TRANSAÇÕES IMOBILIÁRIAS

Heuller César Gomes
Lara Souza Ribeiro Vargas e Aragão

Professores: Dr. Lucas de Almeida Ribeiro
Dr. Gustavo de Assis Costa

20 Dezembro 2023

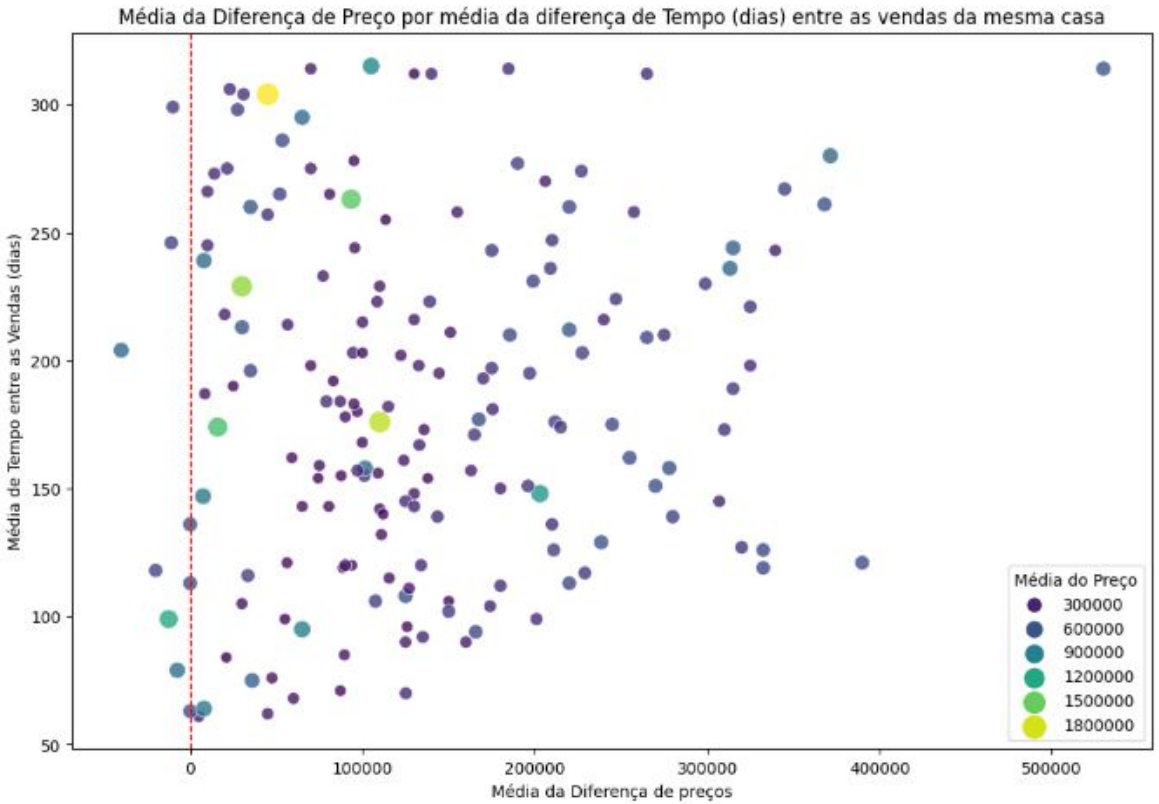
SUMÁRIO

- Objetivos
- Pré-processamento e engenharia de atributos
- Lasso
- XGBoost
- Catboost
- Comparação dos Modelos
- Conclusão

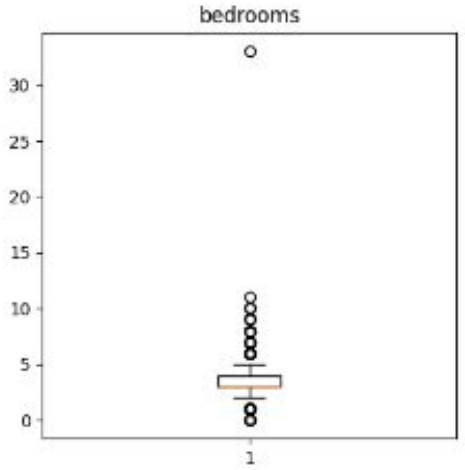
OBJETIVOS

- Geral:
  - ❏ Prever o preço das vendas imobiliárias do dataset
- Específicos:
  - ❏ Fazer pré-processamento e engenharia de atributos
  - ❏ Aplicar diferentes algoritmos (Lasso, Xgboost e Catboost)
  - ❏ Comparar os resultados

PRÉ-PROCESSAMENTO



Apenas a venda mais recente



Substituído pela mediana

```
print(sales.shape)
print(sales_sem_duplicatas.shape)
```

```
(21613, 21)
(21436, 21)
```

ENGENHARIA DE ATRIBUTOS

One hot encoding - Datetime	
Mês	11
Dia da semana	6

One hot encoding zipcode	
Zipcode	69

Duas colunas acrescentadas	
Tem porão	1
Foi renovada	1

PRÉ-PROCESSAMENTO

- Formato final:
  - ❑ Lasso e Xgboost:

```
X.shape
```

```
(21436, 106)
```

  - ❑ Catboost

Warning

Do not use one-hot encoding during preprocessing. This affects both the training speed and the resulting quality.

```
X_catboost.shape
```

```
(21436, 23)
```

PRÉ-PROCESSAMENTO

- Data scaling:
  - ❑ Standard Scaler

$$z = \frac{x - \mu}{\sigma}$$

LASSO - least absolute shrinkage and selection operator

- Por que Lasso?

  - Regressor de encolhimento
    - Penalização dos coef com alto grau de correlação
    - Seleciona as variáveis mais importantes

  - ↓ a complexidade do modelo (dimensionalidade)

  - Lida com multicolinearidade

LASSOCV

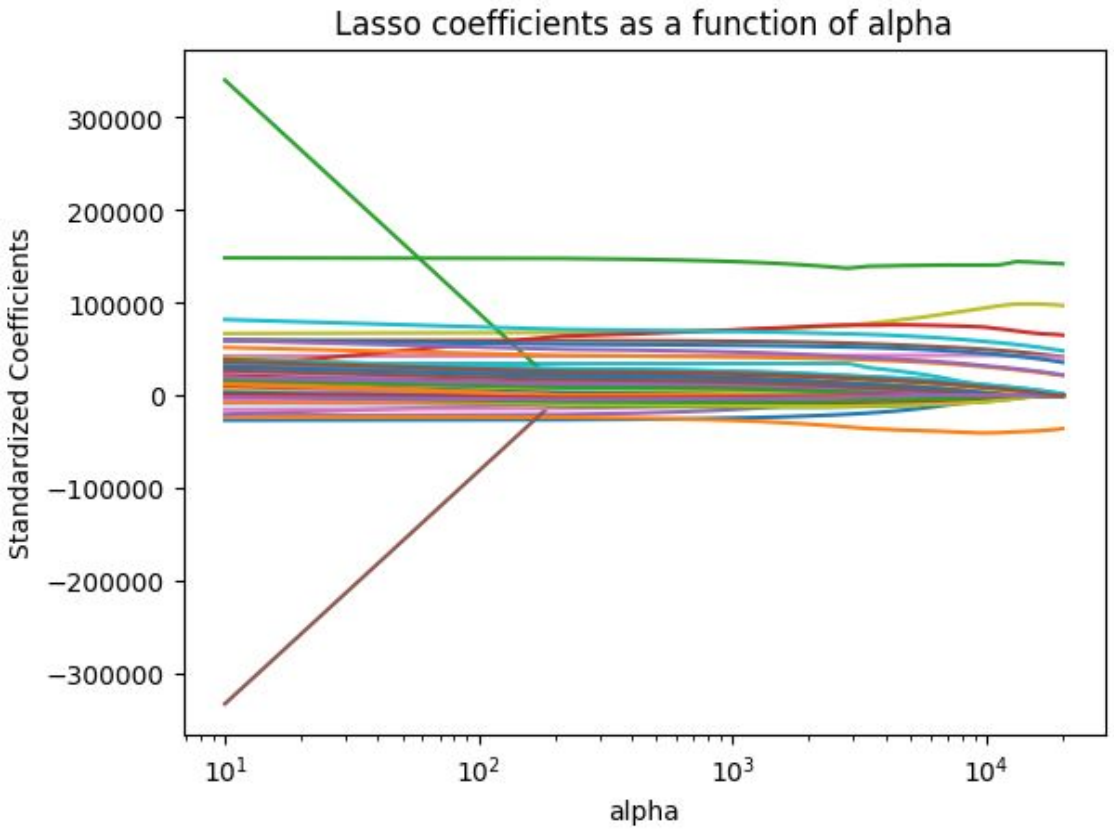
```
model_lasso_cv.alpha_
```

```
255.7545083453325
```

```
Alpha: 255.7545083453325
Mean Squared Error (MSE) for each fold: [2.68100885e+10 2.39713810e+10 3.27414894e+10 2.41984770e+10
 2.28434007e+10]
Mean MSE across folds: 26112967331.657173
```

Aumentando α

- Simplifica o modelo
- Resulta em menor variância, mas maior viés.



LASSO
Melhor alpha

- Features:

❑ Restaram 97

```
Coeficiente: 0.0, Característica: sqft basement
Coeficiente: 0.0, Característica: renovada
Coeficiente: 0.0, Característica: mes 9
Coeficiente: 0.0, Característica: dia_semana_3
Coeficiente: 0.0, Característica: zipcode 98034
Coeficiente: -0.0, Característica: zipcode 98038
Coeficiente: -0.0, Característica: zipcode_98108
Coeficiente: -0.0, Característica: zipcode 98148
Coeficiente: -0.0, Característica: zipcode_98177
```

❑ Mais importantes:

```
Coeficiente: 147702.19559853763, Característica: sqft_living
Coeficiente: 70970.96897309019, Característica: zipcode_98004
Coeficiente: 67627.96433005745, Característica: grade
Coeficiente: 64794.00594573946, Característica: lat
Coeficiente: 59021.668353969806, Característica: waterfront
Coeficiente: 54884.885060303925, Característica: zipcode_98039
Coeficiente: 49311.490493835285, Característica: zipcode_98112
Coeficiente: 42629.96717734995, Característica: zipcode_98040
Coeficiente: 42456.135806751685, Característica: view
Coeficiente: 33942.70688783101, Característica: sqft_above
Coeficiente: 28218.754572606325, Característica: zipcode_98105
Coeficiente: 28170.227101659253, Característica: zipcode_98119
Coeficiente: -26395.194540826804, Característica: bedrooms
```

LASSO- Best alpha

10 Execuções

```
----- MSE -----
Média: 26069539644.058945
Melhor encontrado: 25993389492.45284
Pior encontrado: 26121424156.749252
Desvio Padrão: 35033282.526388854

----- RMSE -----
Média: 161460.6442575371
Melhor encontrado: 161224.655349152
Pior encontrado: 161621.23671333928
Desvio Padrão: 5918.892001581787

----- R^2 -----
Média: 0.8077047652433889
Melhor encontrado: 0.8085401612247913
Pior encontrado: 0.8072080835941616
Desvio Padrão: 0.0004359682410903478
```

Média das médias
Melhor média da métrica encontrada

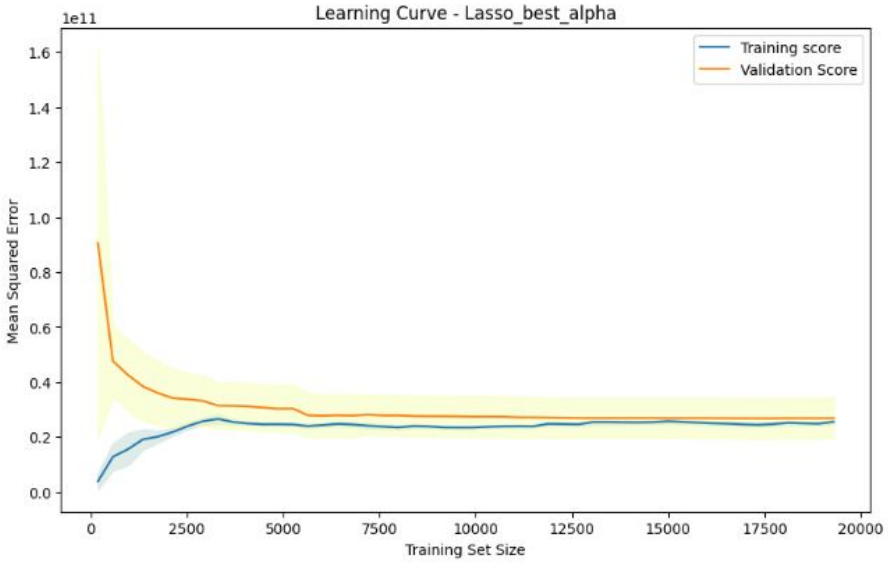
1 Execução

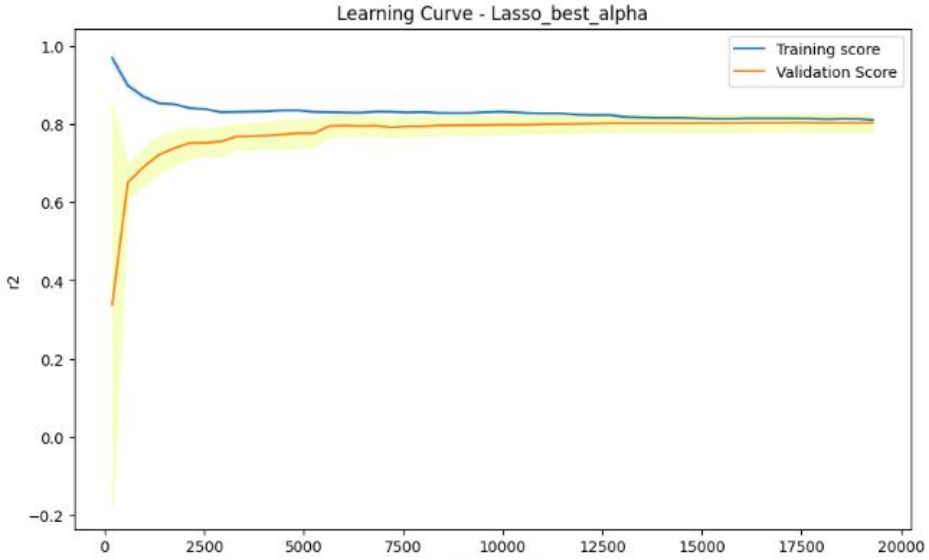
```
----- MSE -----
Média: [2.60762273e+10]
Melhor encontrado: [1.97452634e+10]
Pior encontrado: [3.33509608e+10]
Desvio Padrão: 4241193785.0180798

----- RMSE -----
Média: [161481.35271229]
Melhor encontrado: [140517.84006421]
Pior encontrado: [182622.45435489]
Desvio Padrão: 65124.448443100686

----- R^2 -----
Média: [0.80736039]
Melhor encontrado: [0.82399198]
Pior encontrado: [0.78620012]
Desvio Padrão: [0.01167604]
```

Média da execução
Melhor métrica por fold (10)

LASSO- Best alpha

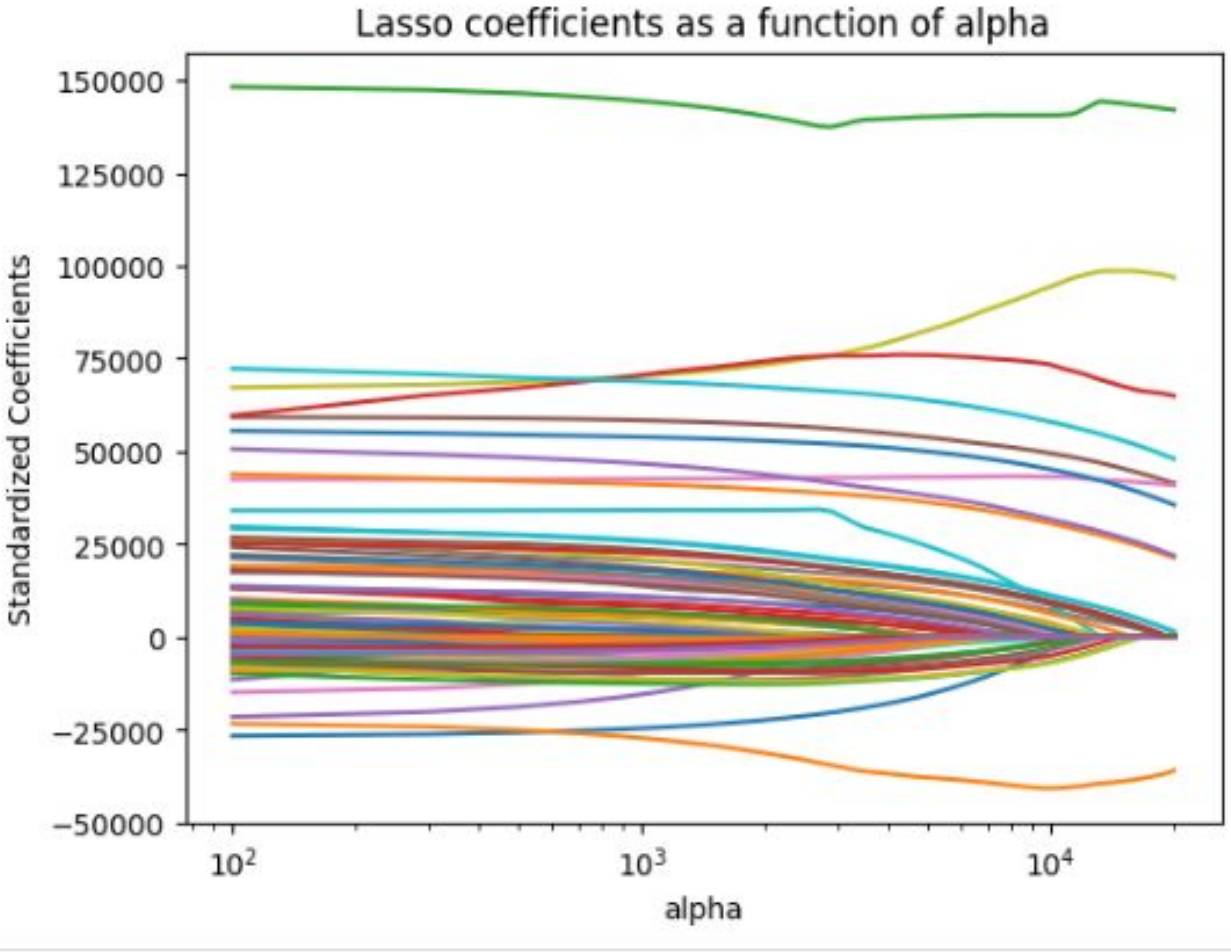
Learning Curve - MSE e R^2






```
[learning_curve] Training set sizes: [  192   582   972  1362  1752  2141  2531  2921  3311  3700  4090  4480
  4870  5260  5649  6039  6429  6819  7208  7598  7988  8378  8768  9157
  9547  9937 10327 10716 11106 11496 11886 12276 12665 13055 13445 13835
 14224 14614 15004 15394 15784 16173 16563 16953 17343 17732 18122 18512
 18902 19292]
```

LASSOCV



- Features:
  - ❑ Restaram 35

LASSO
alpha 10000

- Features:

❑ Restaram 35:
Zerou mês, dias da semana e muitos zipcodes

```
Coeficiente: 0.0, Característica: sqft_lot
Coeficiente: 0.0, Característica: floors
Coeficiente: 0.0, Característica: sqft_basement
Coeficiente: 0.0, Característica: yr_renovated
Coeficiente: -0.0, Característica: long
Coeficiente: 0.0, Característica: sqft_living15
Coeficiente: -0.0, Característica: sqft_lot15
Coeficiente: 0.0, Característica: renovada
Coeficiente: -0.0, Característica: tem_porao
```

❑ Mais importantes:

```
Coeficiente: 140556.20195049015, Característica: sqft_living
Coeficiente: 94430.37784864858, Característica: grade
Coeficiente: 73293.34756486684, Característica: lat
Coeficiente: 57874.45537203412, Característica: zipcode_98004
Coeficiente: 49428.18988739083, Característica: waterfront
Coeficiente: 45093.29514132832, Característica: zipcode_98039
Coeficiente: 43129.42352397838, Característica: view
Coeficiente: -40813.38132870279, Característica: yr_built
Coeficiente: 31653.002737472947, Característica: zipcode_98112
```

LASSO- alpha 10^4

10 Execuções

```
----- MSE -----
Média: 31110078095.188896
Melhor encontrado: 31031074699.08326
Pior encontrado: 31163158123.466225
Desvio Padrão: 42131428.87023587

----- RMSE -----
Média: 176380.492388441
Melhor encontrado: 176156.39272840274
Pior encontrado: 176530.8984950403
Desvio Padrão: 6490.872735637009

----- R^2 -----
Média: 0.7706594151245298
Melhor encontrado: 0.7718257684610813
Pior encontrado: 0.7699382461337724
Desvio Padrão: 0.0006273695480550562
```

1 Execução

```
----- MSE -----
Média: [3.11274344e+10]
Melhor encontrado: [2.33558823e+10]
Pior encontrado: [3.9903865e+10]
Desvio Padrão: 5271473538.4402895

----- RMSE -----
Média: [176429.6868269]
Melhor encontrado: [152826.31410074]
Pior encontrado: [199759.51794553]
Desvio Padrão: 72604.91401028095

----- R^2 -----
Média: [0.77021391]
Melhor encontrado: [0.78950252]
Pior encontrado: [0.74211433]
Desvio Padrão: [0.01545429]
```

XGBoost

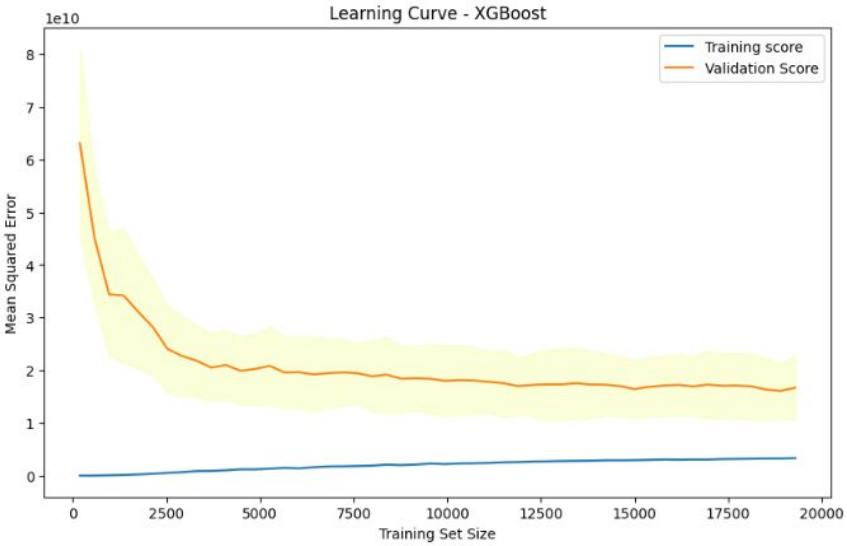
10 Execuções

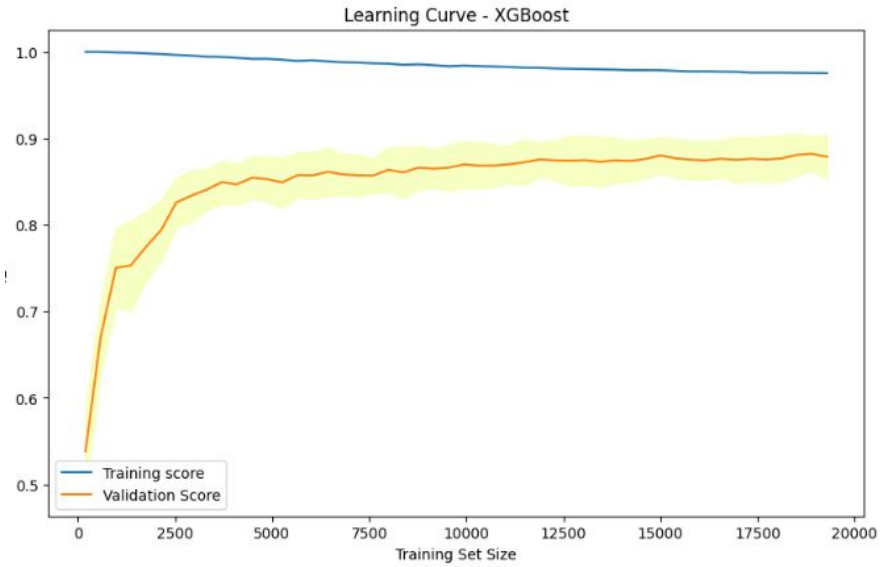
```
----- MSE -----
Média: 15544259905.234043
Melhor encontrado: 14540372021.089132
Pior encontrado: 16283449148.572483
Desvio Padrão: 483364556.8164797

----- RMSE -----
Média: 124676.62132586865
Melhor encontrado: 120583.46495722012
Pior encontrado: 127606.61874907775
Desvio Padrão: 21985.553366164786

----- R^2 -----
Média: 0.8853879727961473
Melhor encontrado: 0.8921653209573094
Pior encontrado: 0.879592269382972
Desvio Padrão: 0.003562989520561595
```

1 Execução

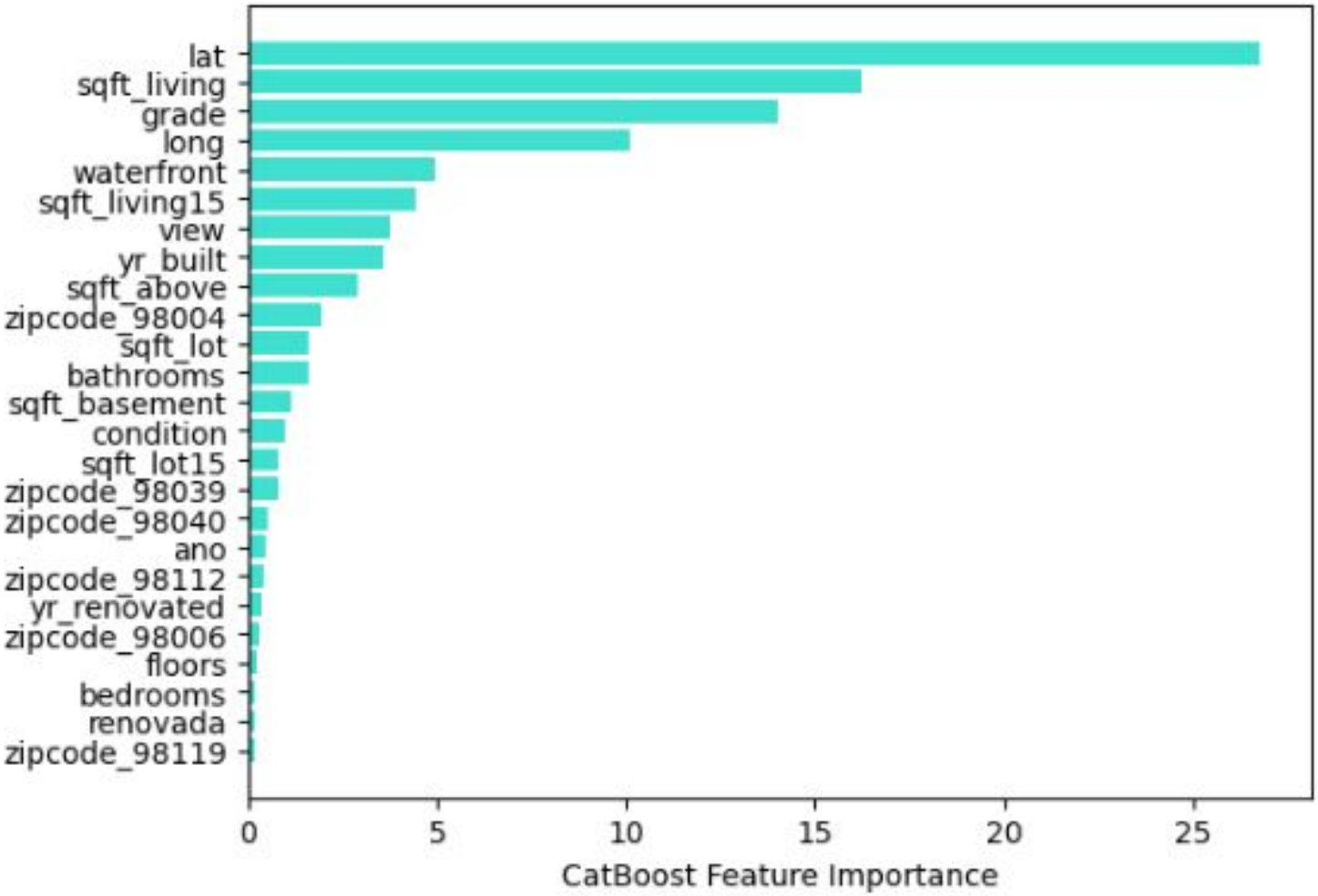
```
----- MSE -----
Média: [1.62834491e+10]
Melhor encontrado: [1.15029531e+10]
Pior encontrado: [2.35179139e+10]
Desvio Padrão: 4436894089.070637

----- RMSE -----
Média: [127606.61874908]
Melhor encontrado: [107251.82074642]
Pior encontrado: [153355.51458534]
Desvio Padrão: 66610.01493071922

----- R^2 -----
Média: [0.87991252]
Melhor encontrado: [0.91307071]
Pior encontrado: [0.83537088]
Desvio Padrão: [0.0250387]
```

XGBoost


Learning Curve - XGBoost
1e10
Training score
Validation Score
Mean Squared Error
Training Set Size


Learning Curve - XGBoost
Training score
Validation Score
Training Set Size

CATBoost


lat
sqft_living
grade
long
waterfront
sqft_living15
view
yr_built
sqft_above
zipcode_98004
sqft_lot
bathrooms
sqft_basement
condition
sqft_lot15
zipcode_98039
zipcode_98040
ano
zipcode_98112
yr_renovated
zipcode_98006
floors
bedrooms
renovada
zipcode_98119
0
5
10
15
20
25
CatBoost Feature Importance

CATBoost

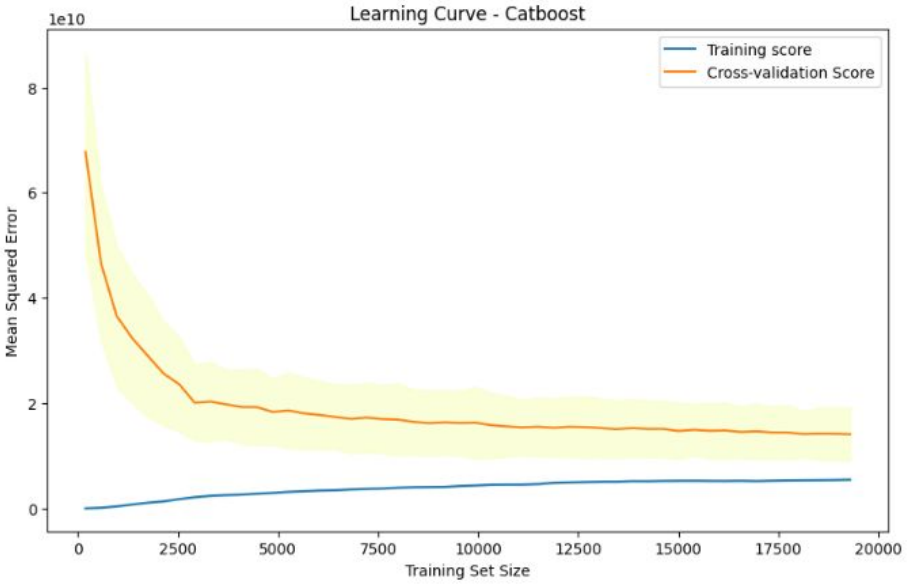
10 Execuções

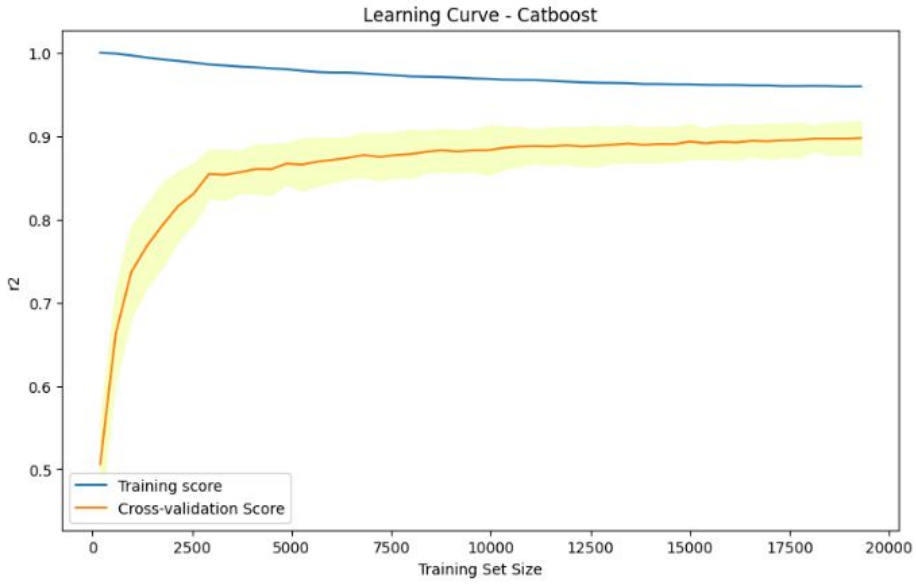
```
----- MSE -----
Média: 13325290504.536999
Melhor encontrado: 12963737514.863361
Pior encontrado: 13538847605.364302
Desvio Padrão: 206759060.68374214

----- RMSE -----
Média: 115435.22211412337
Melhor encontrado: 113858.40994350554
Pior encontrado: 116356.55377057324
Desvio Padrão: 14379.118911941097

----- R^2 -----
Média: 0.9018769741395148
Melhor encontrado: 0.9045269844843651
Pior encontrado: 0.8995426732941338
Desvio Padrão: 0.0014566023678535117
```

1 Execução

```
----- MSE -----
Média: [1.35327708e+10]
Melhor encontrado: [9.93978371e+09]
Pior encontrado: [1.90999188e+10]
Desvio Padrão: 2772347541.9632955

----- RMSE -----
Média: [116330.43811431]
Melhor encontrado: [99698.46393002]
Pior encontrado: [138202.45585568]
Desvio Padrão: 52653.08672778164

----- R^2 -----
Média: [0.90017196]
Melhor encontrado: [0.91686563]
Pior encontrado: [0.87663848]
Desvio Padrão: [0.01171248]
```

CATBoost


Learning Curve - Catboost
1e10
Training score
Cross-validation Score
Mean Squared Error
Training Set Size
0
2
4
6
8
0
2500
5000
7500
10000
12500
15000
17500
20000


Learning Curve - Catboost
Training score
Cross-validation Score
r2
Training Set Size
0.5
0.6
0.7
0.8
0.9
1.0
0
2500
5000
7500
10000
12500
15000
17500
20000

Comparação dos Modelos

Comparativo dos modelos, considerando 10 execuções

Modelo	Média RMSE	Desvio RMSE	Média R^2	Desvio R^2
Lasso - best alpha	161.460,64	5.918,89	0,80770	0,00044
Lasso - alpha 10^4	176.380,49	6.490,87	0,77066	0,00063
XGBoost	124.676,62	21.985,55	0,88539	0,00356
CATBoost	115.435,22	14.379,12	0,90188	0,00146

CONCLUSÃO

- Abordagens.
- Resultados.
- Trabalhos futuros.

OBRIGADO!