



Universidade Federal do Mato Grosso

Instituto de Computação

House Prices - Advanced Regression Techniques

Trabalho de Mineração de dados

Felipe Borges de Lima Naufal
Felipe Nunes Costa
Hiago de Sousa Patrício
Leonardo Barbosa Almeida

Setembro de 2021

Introdução

O relatório a seguir mostra como se sucedeu a participação e quais técnicas foram utilizadas na competição *House Prices - Advanced Regression Techniques*. A competição fornece diversos dados técnicos sobre diversos imóveis como localização, condição da garagem, número de cozinhas etc, e pede para que utilizemos técnicas de regressão para que possamos prever qual o valor de venda de cada um deles.

Metodologia

Os dados de treinamento possuíam 80 colunas e 1460 linhas, e os de predição 79 e 1459. Após os pré-processamentos, os dados de treinamento foram divididos em 80% das linhas para treinamento e 20% para validação dos modelos.

Para o pré-processamento dos dados foram utilizados diversos métodos que seguem elencados abaixo:

1. Remoção de linhas que possuíam valores faltando, mas isso resultou em um dataset com 0 linhas, por isso não foi utilizado.
2. Remoção de colunas que possuíam valores faltando, seguido de One Hot Encoder.
3. Remoção das colunas categóricas que possuíam valores faltando, seguida de imputação nas colunas numéricas utilizando as estratégias listadas abaixo, após isso foi aplicado One Hot Encoder para as colunas categóricas restantes.
 - a. Média
 - b. Mediana
 - c. Mais frequente
4. Foi repetido o processo descrito no item 3, acrescentando colunas que indicam quais linhas sofreram imputação.

Após os pré-processamentos a quantidade de linhas e colunas são as informadas abaixo:

Tabela 1 - Formato dos dados de treinamento processados

Data format		
Data pre-processing	Number of rows	Number of columns
1 - Rows with missing values removed	0	79
2 - Columns with missing values removed	1460	149

3 - Imputation using mean strategy	1460	160
4 - Imputation with flag using mean strategy	1460	163

Os modelos utilizados para treinamento foram:

- *DecisionTreeClassifier*
- *DecisionTreeRegressor*
- *RandomForestRegressor*
- *XGBRegressor*

Todos os modelos foram treinados e realizaram previsões com cada um dos pré-processamentos listados acima, exceto o primeiro. Para os modelos *DecisionTreeClassifier* e *DecisionTreeRegressor* foram testadas diversas configurações iterativamente e as configurações com menor Mean Absolute Error (MAE) foram utilizadas para gerar as submissões. Para cada modelo foram registrados o score obtido na submissão das previsões à competição e o MAE obtido nos treinamentos.

Resultados

Na tabela de MAE podemos observar que os modelos *DecisionTreeClassifier* e *DecisionTreeRegressor* possuem os mesmos valores em todos pré-processamentos de dados, isso leva a crer que os scores serão os mesmos, mas posteriormente veremos que não.

Quanto menor o MAE, melhor, então os modelos que melhor performaram em todos os pré-processamentos foram o *XGBRegressor* em primeiro lugar, *RandomForestRegressor* em segundo, *DecisionTreeClassifier* e *DecisionTreeRegressor* empatados em terceiro.

Tabela 2 - Mean Absolute Error dos modelos com seus pré-processamentos de dados

Data pre-processing	Model			
	DecisionTree Classifier	DecisionTree Regressor	RandomForest Regressor	XGBRegressor
2 - Columns with missing values removed	25439.85	25439.85	18582.22	17859.92
3.a - Imputation using mean strategy	22547.71	22547.71	16973.00	14749.49
3.b - Imputation using median strategy	23243.03	23243.03	15399.70	13847.18
3.c - Imputation using most_frequent strategy	19874.28	19874.28	16327.23	15071.39
4.a - Imputation with flag using mean strategy	24037.00	24037.00	18476.38	17946.52
4.b - Imputation with flag using median strategy	26257.31	26257.31	18139.01	17054.43
4.c - Imputation with flag using most_frequent strategy	25896.91	25896.91	19218.59	18648.51
MAE mean	23899.44	23899.44	17588.02	16453.92
Better MAE	19874.28	19874.28	15399.70	13847.18
Worse MAE	26257.31	26257.31	18139.01	18648.51

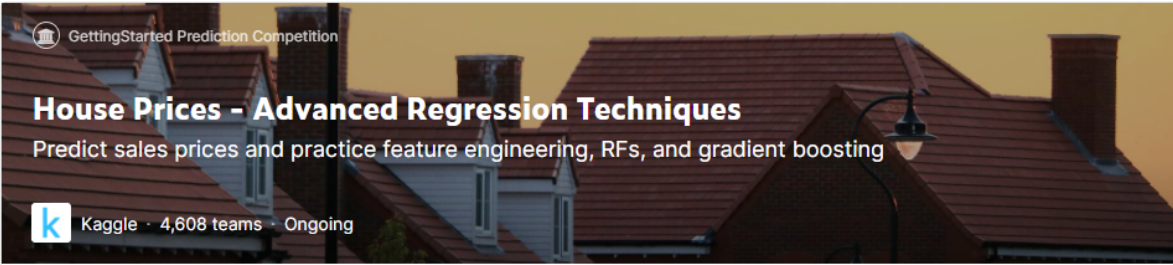
Na tabela de Scores podemos ver que os scores dos modelos *DecisionTreeClassifier* e *DecisionTreeRegressor* diferem apesar de seus MAE's serem os mesmos em todos os pré-processamentos, nesse caso os scores de *DecisionTreeRegressor* foram melhores.

A tabela de MAE não se mostrou coerente com os resultados obtidos na tabela de Scores quando analisamos um modelo de cada vez com seus pré-processamentos, por exemplo, o melhor resultado do modelo *DecisionTreeClassifier* segundo a tabela de MAE deveria ser com o pré-processamento 3.c mas acabou sendo o 2, isso acontece com o restante dos modelos. A ordem do melhor e segundo melhor modelo se mantém, em seguida ficam *DecisionTreeRegressor* e *DecisionTreeClassifier*.

Tabela 3 - Scores dos modelos com seus pré-processamentos de dados

Scores				
	Model			
Data pre-processing	DecisionTree Classifier	DecisionTree Regressor	RandomForest Regressor	XGBRegressor
2 - Columns with missing values removed	0.25518	0.20321	0.14905	0.14455
3.a - Imputation using mean strategy	0.27177	0.20261	0.15044	0.14099
3.b - Imputation using median strategy	0.26015	0.19232	0.14910	0.13875
3.c - Imputation using most_frequent strategy	0.26095	0.19306	0.15131	0.13983
4.a - Imputation with flag using mean strategy	0.27177	0.20261	0.15044	0.15293
4.b - Imputation with flag using median strategy	0.26015	0.19232	0.14910	0.13643
4.c - Imputation with flag using most_frequent strategy	0.31073	0.23519	0.18861	0.17199
Score mean	0.2701	0.20305	0.15544	0.14225
Better Score	0.25518	0.19232	0.14905	0.13643
Worse Score	0.31073	0.23519	0.18861	0.17199

Figura 1 - Score obtido com o pré-processamento 4.b e modelo XGBRegressor



GettingStarted Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 4,608 teams · Ongoing

Overview Data Code Discussion Leaderboard Rules Team My Submissions Submit Predictions ...

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
Imputation with flag using median stra...	just now	1 seconds	0 seconds	0.13643

Complete

[Jump to your position on the leaderboard ▾](#)

Conclusão

O pré-processamento dos dados ajudou a melhorar o score, mas foi uma pequena parcela, a escolha dos modelos foi mais determinante para obter melhores resultados. O resultado foi satisfatório, mas poderia ser um pouco melhor se os dados de treinamento estivessem completos, explorar outros modelos poderia ser uma alternativa para melhorar o score.

Apêndice

<https://www.kaggle.com/leonardoalmeida6/notebook-md?scriptVersionId=75996538>