House Prices - Advanced Regression Techniques

Guilherme Yukio Iasunaga

2023-11-19

1. Correlação das variáveis numéricas com a variável target

Calculei a correlação entre as variáveis numéricas originais e a variável target, resultando nessas maiores correlações absolutas:

Variáveis	**Correlação**
OverallQual	0.809829
GrLivArea	0.731310
GarageCars	0.690711
YearBuilt	0.652682
GarageArea	0.649379
FullBath	0.635957
TotalBsmtSF	0.602725
TotalBsmtSF	0.602725
GarageYrBlt	0.593788
1stFlrSF	0.575408
YearRemodAdd	0.571159
TotRmsAbvGrd	0.532586
Fireplaces	0.519247
OpenPorchSF	0.477561
LotArea	0.456461
MasVnrArea	0.421309
LotFrontage	0.409076
WoodDeckSF	0.353802
HalfBath	0.343008
BsmtFinSF1	0.301871
2ndFlrSF	0.293598

2.Varificação dos valores faltantes (missings)

Primeiramente verifiquei os dados faltantes presentes na base de dados, tanto de treino quanto de teste. Com base nas descrições de cada variável, dispostos no arquivo *data description*, algumas variáveis categóricas possuem uma classe específica NA, indicando a falta daquela variável, por exemplo na *GarageType*, o valor NA indica a falta de garagem no estabelecimento. Além disso, algumas dessas variáveis também possuem alguma variável numérica relacionada a elas.

Assim, verifiquei as observações em que possui NA na categórica e 0 na numérica, indicando que de fato o valor NA é uma classe. Com isso, boa parte dos valores missings das variáveis categóricas eram, levando em consideração essa abordagem, uma classe.

Apenas para as colunas *Alley* e *Fence* que não possuem uma variável numérica relacionada que eu substitui os valores faltantes por um valor que representasse a falta dessas características na casa.

Após a substituição desses valores por um outro que represente a classe NA, preenchi o restante dos valores faltantes com a moda e a mediana para as variáveis categóricas e numéricas, respectivamente.

3. Dados e transformação

Nesta etapa padronizei os dados numéricos pois as variáveis não estavam na mesma escala, atrapalhando e piorando o ajuste do modelo. Já para as variáveis categóricas criei variáveis dummies a fim de representar cada classe de cada variável no modelo ajustado.

4. Ajuste do modelo

Para o ajuste do modelo utilizei o método da validação cruzada, 5 folds com 3 repetições, a fim de encontrar os melhores parâmetros.

4.1 Ridge

Para o modelo Linear com regularização L1 (Ridge), segundo vários ajustes, o melhor que se encaixou aos dados foi com um λ igual a 54.99.

4.2 Lasso

Para o modelo Linear com regularização L2 (Lasso), segundo vários ajustes, o melhor que se encaixou aos dados foi com um λ igual a 79.24.

4.3 Elastic Net

Para o modelo Linear com regularização Elastic Net, segundo vários ajustes, o melhor que se encaixou aos dados foi com um λ igual a 0.92, considerando um α de 0.5.

4.4 Gradient Boosting

Para o modelo de Gradient Boosting, usando o *GridSearchCV()*, a fim de procurar os melhores parâmetros para o mdoelo, o melhor que se encaixou aos dados foi com uma taxa de aprendizagem de 0.1, sendo calculadas 100 árvores de decisão com profundidade máxima 5 camadas.

4.5 Floresta Aleatória

Para o modelo de Floresta Aleatória, usando o *GridSearchCV()*, a fim de procurar os melhores parâmetros para o mdoelo, o melhor que se encaixou foi com 150 árvores com no máximo 15 camadas de profundidade,

4.6 Comparação do erro de validação (MSE)

O erro fora da amostra foi calculado pelo Erro Quadrático Médio (MSE). Os valores mostrados na tabela abaixo são raízes desses valores para ficar mais claro o resultado.

Modelo	Erro fora (de validação)
Ridge	28305.380071
Lasso	27980.655668
ElasticNet	30752.624133
GradientBoostingRegressor	27300.156311
RandomForestRegressor	30070.379012

Com isso, o modelo final escolhido foi Gradiente Descendente.

4.7 Observação:

Aqui vale ressaltar que neste relatório não inclui os modelos de stepwise pois ao ajustar e calcular o erro fora da amostra ele teve um valor muito acima dos demais. Assim, inclui apenas os mais relevantes.

5. Score do Kaggle

Após a submissão no Kaggle, meu score resultou em 0.14158.