House Prices - Advanced Regression Techniques

Guilherme Yukio Iasunaga

2023-11-17

1.Varificação dos valores faltantes (missings)

Primeiramente verifiquei os missings presentes na base de dados, tanto de treino quanto de teste. Com base nas descrições de cada variável, presente no arquivo disponibilizado *data description*, algumas variáveis categóricas possuem uma classe 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 de fato uma classe.

Após a substituição desses valore 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 missing, respectivamente.

2. Feature Engeneering

Nesta etapa criei algumas variáveis, assim como normalizei as numéricas e criei variáveis dummies para as categóricas.

2.1 Variáveis Criadas

1) *total area*: representa o tamanho total da casa, sendo que somei as seguintes variáveis originais: TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, GarageArea, OpenPorchSF, WoodDeckSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea e LotArea.

2) *total bathroom*: representa o número total de banheiros presentes na casa, sendo a soma de: BsmtFullBath, BsmtHalfBath, FullBath, HalfBath e TotRmsAbvGrd

3) *house year old*: representa a idade da casa, sendo a diferença entre: YrSold e YearBuilt

3. Correlação das variáveis numéricas com a variável target

Depois da imputação dos missings, 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

Logo a baixo, estão as correlações das variáveis criadas:

Variáveis	Correlação
total bathroom	0.670909
total area	0.617849
house year old	-0.650120

4. Ajuste do modelo

4.1 Ridge

4.2 Lasso

4.3 Elastic Net

Stepwise (forward e backward)