Assignment Part-I

1-Which variables are significant in predicting the price of a house

Based on both Lasso and Ridge regression models, the following variables are significant in predicting the price of a house:

OverallQual

GrLivArea

GarageArea

OverallCond

Neighborhood_StoneBr

Neighborhood_NridgHt

YearBuilt

Exterior1st_BrkFace

MSSubClass_30

Neighborhood_Crawfor

FullBath

YearRemodAdd

CentralAir

BldgType_Twnhs

The variables effectively describe house prices, with r2 scores of 0.8947 (Lasso) and 0.8912 (Ridge), explaining 89% of price variance. MAE values of 0.0914 (Lasso) and 0.0934 (Ridge) show close predictions to actual prices. These significant variables help the company make informed property investment decisions for high returns.

2- How well those variables describe the price of a house.

The important variables identified by Lasso and Ridge models effectively describe house prices. Both models' r2 scores and MAE values suggest that the variables are useful in predicting house prices. The company can use these variables for property investment decisions.

Assignment Part-II

1- What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Best alphas: Ridge Regression: 8; Lasso Regression: .0007. Doubling alpha affects Ridge and Lasso models by increasing the penalty term, which shrinks beta values and simplifies the model. This results in higher bias and lower variance, potentially causing underfitting. Conversely, decreasing alpha reduces the penalty, leading to lower bias and higher variance, which might cause overfitting. The optimal alpha value is determined through hyperparameter tuning.

Top 10 features with beta values from Ridge after using alpha= 16

OverallQual 0.229066

GrLivArea 0.161305

GarageArea 0.124397

OverallCond 0.123903

2ndFlrSF 0.121972

Neighborhood_StoneBr 0.117965

1stFlrSF 0.106791

FullBath 0.094897

Exterior1st_BrkFace 0.092420

Neighborhood_NridgHt 0.086085

Top 10 features with beta values from Lasso after using alpha= .0012

GrLivArea 0.372094

OverallQual 0.324650

GarageArea 0.139289

OverallCond 0.137807

Neighborhood_StoneBr 0.132524

Neighborhood_NridgHt 0.104472

YearBuilt 0.097228

Exterior1st_BrkFace 0.096810

MSSubClass_30 -0.089377

Neighborhood_Crawfor 0.087917

So, after using double the value of alpha, the most important variable: In Ridge model: OverallQual (Rates the overall material and finish of the house) In Lasso model: GrLivArea (Above grade (ground) living area square feet)

2- <u>You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?</u>

For specific use cases, the choice of model depends on the requirements. Lasso is ideal when feature selection is necessary and there are many variables, while Ridge Regression is preferable when large coefficients need to be avoided. The Razor principle states that models should not be excessively complex. Although both models normalize beta coefficients, Lasso has fewer features, resulting in a simpler model with similar performance. As such, we should select the simpler model, and in this case, Lasso is the final choice.

3- <u>After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?</u>

The five most important predictor variables in the Lasso model after excluding the top 5 variables are:

GarageArea: 0.139289

OverallCond: 0.137807

Neighborhood_StoneBr: 0.132524

Neighborhood_NridgHt: 0.104472

YearBuilt: 0.097228

4- <u>How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?</u>

To create a robust and generalizable model, it should be complex enough to learn data patterns in the training dataset but not too complex that it learns noise as well. Overfitting can be identified by comparing model performance on training and testing datasets. A model with high bias and low variance is underfitting, while a model with low bias and high variance is overfitting. Normalization and Lasso can help to reduce the complexity of the model and overcome overfitting. A robust model should have low bias and low variance, achieved by finding a trade-off between them. Additionally, a robust model should have similar accuracy on both training and testing datasets.