Name: Le Thuc Anh
Email: thucanhle149@gmail.com
Course: Executive PG Programme in Machine Learning & AI - March 2023
Module: Advanced Regression Assignment
Github link: https://github.com/1darknight/Houses-Price

Assignment - Part II

Question 1

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

The **optimal alpha value** for Ridge Regression is 6 and Lasso Regression is 100. If we choose to **double the value of alpha**, both Ridge and Lasso will have a *higher bias as there will be a higher regularization on coefficients* therefore, the coefficients will be lower than before.

Ridge Regression

The original most important predictors of Ridge Regression are :
1.***GrLivArea***: The area of living above ground (by square fit) is bigger, the higher the price of the house
2+3. ***OverallQual (Excellent & Very Excellent)***: the overall material and finish of the house is excellent will indicate a higher price for the house
4.***TotalBsmtSF***: The area of the basement (by square fit) is bigger, the higher the price of the house
5.***BsmtFinSF1***: The area of Type 1 basement that is finished, the larger the area the higher the price
6. ***GarageCars***: The more cars the garage could accommodate, the higher the price
In general, the bigger and better the house is, the higher the price as both area and quality of various parts of the house are the most important variables.

After doubling the alpha, the new 5 most important predictors are the same as before for Ridge Regression.

Lasso Regression

The original most important predictors of Lasso Regression are :
1.**Condition2**: Near positive off-site features, if there is any positive off-site features nearby, the house price is lower
2.***GrLivArea***: The area of living above ground (by square fit) is bigger, the higher the price of the house
*3+4.****OverallQual (Excellent & Very Excellent)***: the overall material and finish of the house is excellent will indicate a higher price for the house
5.***TotalBsmtSF***: The area of the basement (by square fit) is bigger, the higher the price of the house
6.**KitchenAbvGr**: The lower the number of kitchen on the ground, the higher the house price

After doubling the alpha, the new 4 most important predictors are the same as before except the 5th varibles:
1.**GrLivArea**
2.**Condition2**
*3+4.****OverallQual (Excellent & Very Excellent)***
*5.****TotalBsmtSF***
6. **GarageCars**: The higher the number of cars the garage could accommodate, the higher the house price

Question 2

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?

After training both models Ridge and Lasso Regression, I tested them on a validation dataset to see which of them perform better in an unseen dataset. Which resulted in Ridge doing better in an unseen dataset as below:

Metrics	Linear Regression	Ridge Regression	Lasso Regression

R2 Score (Train)	**0.963**	0.929	0.935
R2 Score (Test)	0.765	**0.887**	0.885
RSS (Train)	220,738,908,208.922	424,776,473,655.041	391,729,089,180.970
RSS (Test)	577,904,619,881.230	277,845,483,301.021	281,825,045,320.771
MSE (Train)	217,906,128.538	419,325,245.464	386,701,963.653
MSE (Test)	1,328,516,367.543	638,725,248.968	647,873,667.404

Ridge regression has a higher R squared in the test dataset with lower errors. While Linear Regression has overfitted the train dataset and Lasso Regression did worse in the test dataset in both R squared and errors.
So, I chose Ridge Regression to be the best model as they perform better in an unseen dataset.

Last note that both Ridge and Lasso is quite close in performance which means that our set of variables is quite equal in terms of feature importance the whole dataset is generally contributed to the predictions of the house price instead of some having a significant influence.

Question 3

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?

The most important features are removed from the dataset (both train and test), as the upcoming data do not have these available. So, I retrain the model without the top 5 most important variables and tested on the test set.

The results that **the model performs worse in both the train and test datasets** as the R squared is around 0.7 compared to 0.9 in the train and 0.88 in the test dataset before removing.

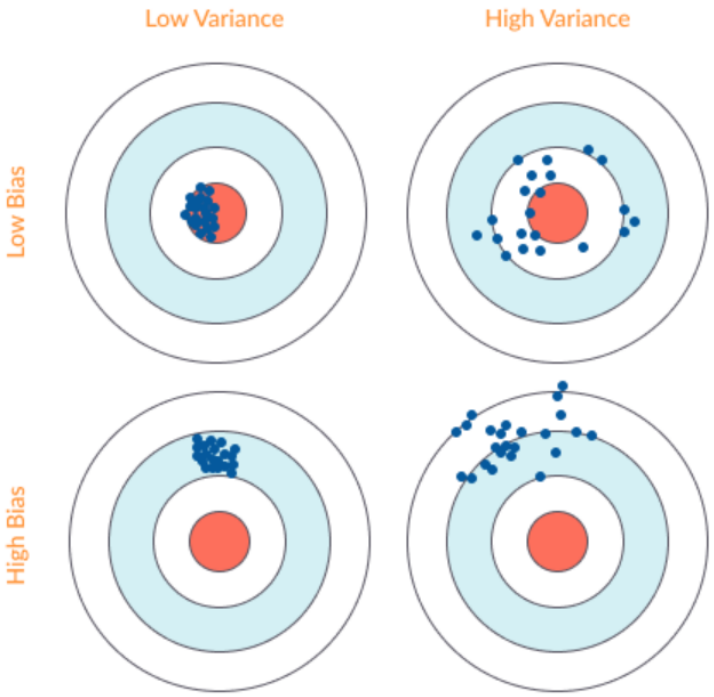
Here are the **new 5 most important features**:

- **BsmtQual**: The better the quality of the basement, the higher the house price
- **Neighborhood** (Neighborhood_NridgHt): The house in neighborhood of Northridge Heights tend to have a higher price
- **Fireplaces**: The higher the number of fireplaces, the higher the house price
- **Fullbath**: The higher the number of full bathrooms above ground, the higher the house price
- **BsmtFinType1** (***BsmtFinType1_GLQ***): The rating of the basement finished area is better, the higher the house price

Question 4

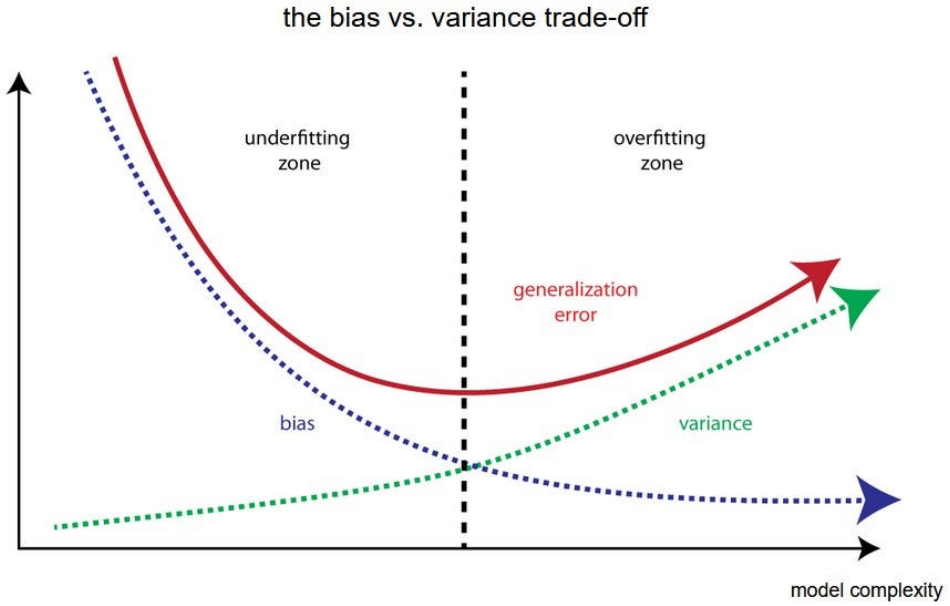
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

The robustness of a model is measured by the model's bias and the generalisation of a model is measured by the model's variance. There is a trade-off between these metrics as the model complexity varies. A simple model will have high bias and low variance and a complex model will have low bias and high variance.



The best model is trying to have both low bias and low variance so that the predictions will be accurate. To increase accuracy we have 2 options, reduce bias or reduce variance while

there is a tradeoff between both. In order to find the optimal point where both bias and variance are at their optimal, in the assignment, I perform Ridge and Lasso Regression.



With these two models, a little reduction in variance will significantly improve the accuracy of the whole model on the unseen dataset and avoid violation of the regression assumptions.