Problem Statement - 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 double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

Optimal value for alpha for Ridge: **6**.

Optimal value for alpha for Lasso: **0.001**.

After changing the alpha value to double of it, i.e. **12** and **0.002** for **Ridge** and **Lasso** respectively.

Ridge: R2 score of train data is drop from **0.935837** to **0.92891** and **0.889468** to **0.88756** for test.

Top Features:

- GrLivArea
- OverallQual_Excellent
- 1stFlrSF
- Neighborhood_StoneBr
- OverallQual_Very Good

Lasso: R2 score of train data is drop from **0.911898** to **0.892761** for train and **0.889251** to **0.87619** for test.

Top Features:

- GrLivArea
- OverallQual_Excellent
- Functional_Typ
- OverallQual_Very Good
- Neighborhood_Crawfor

If we compare both Ridge and Lasso, '**GrLivArea**' is most important predictor feature.

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?

Answer:

We will choose Lasso, as it works well when there's not too many significant parameters. It removes unwanted features even without impacting the model accuracy. Which makes model more generalized, simple and accurate.

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?

Answer:

Top 5 Feature in Lasso model before drop:

- GrLivArea
- OverallQual_Excellent
- Neighborhood_StoneBr
- OverallQual_Very Good
- Functional_Typ

After dropping above features and rebuilding, best alpha changed to '**0.0001**'

- After rebuilding R2 score changed from **0.911898** to **0.93876** for train.
- After rebuilding R2 score changed from **0.889251** to **0.87814** for test.

Top 5 Feature in Lasso model after drop:

- MSZoning_FV
- MSZoning_RL
- RoofMatl_WdShngl
- MSZoning_RH
- MSZoning_RM

Question 4.

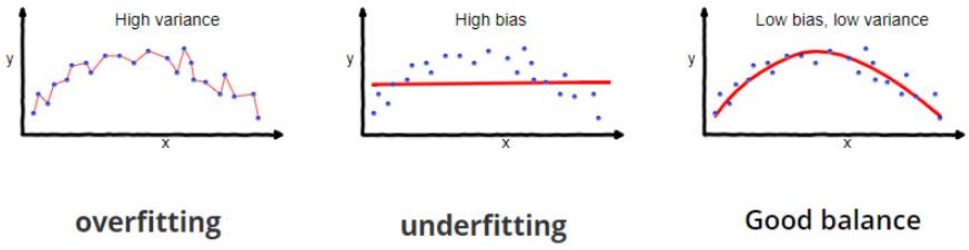
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

When building a machine learning model with a high-dimensional dataset, it is always advisable to start with a simple model, then we may add complexity as needed. During model evaluation, it is important to perform several tests to make sure the model is not capturing random effects in the dataset. We can have separate set of unseen test/validation data as well as techniques like K-fold cross validation can be helpful to overcome it.

Simple model helps in:

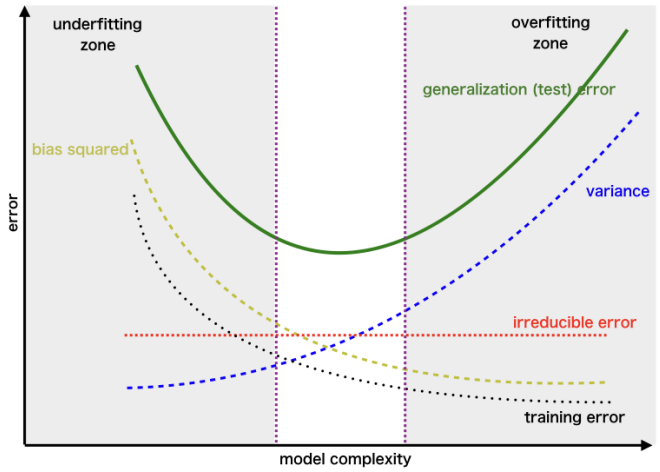
- **Prevents Overfitting**: A high-dimensional dataset having too many features can sometimes lead to overfitting (model captures both real and random effects).
- **Interpretability**: An over-complex model having too many features can be hard to interpret especially when features are correlated with each other.
- **Computational Efficiency**: A model trained on a lower-dimensional dataset is computationally efficient (execution of algorithm requires less computational time).



Bias Variance Tradeoff:

If a model is simple and have a smaller number of features, then it may have high bias and low variance, in contrast, if a model has huge number of features, then it may have low bias and high variance. So, as the bias increases variance decreases and vice-versa. So, we need to get a model which has low bias as well as low variance. That is why the trade-off is required.

Bias can be minimized by training with more data and variance can be reduced by using Ridge/Lasso regularization methods



An optimal balance of bias and variance would never overfit or underfit the model.

Therefore understanding bias and variance is critical for understanding the behaviour of prediction models.