House Price Prediction 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?**

   ANSWER:

   Optimal Value of Alpha for Ridge and Lasso are:
   - Ridge: 0.21
   - Lasso: 0.00013

   With these values of alpha, the most important predictor is:

   - Ridge: GrLivArea
   - Lasso: GrLivArea

   When we double the values of alpha i.e. (0.42 for Ridge and 0.00026 for Lasso) following changes can be seen in the model:

   I. Regularization is increased hence the model becomes simpler
   II. Model becomes more generalizable
   III. Bias of the model increases
   IV. Number of features in Lasso regularization decreases

   With alpha value is doubled, the most important predictor remains same:

   - Ridge: GrLivArea
   - Lasso: GrLivArea

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:

   Optimal value of lambda for Ridge and Lasso is 0.25 and 0.00013 respectively. I'll choose Lasso over Ridge as the final model for predicting the house prices.

   Ridge: 148 features, R2 Score: 0.93
   Lasso: 72 features, R2 Score: 0.91

   In Ridge, we have 148 features with an R-Squared value around 0.93 whereas Lasso has around half of the features in comparison to Ridge model with an R-Squared value around 0.91. This indicates that nearly half of the features in the model regularized using Ridge are not much significant predictors for the Sale Price of the house. These extra features are increasing the model complexity though not adding much to the predictive power. Hence, model regularized using Lasso is a better model as it has half of the number of features and having a decent predictive power. Also, the business would prefer to have a model built with lesser number of features and having decent predictive power which we can see for Lasso model.

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:

The top five features from the original Lasso model were:

1. GrLivArea (Above ground Living Area)
2. Condition2_PosN (Proximity to various places such as off-site feature-park, greenbelt, etc.)
3. OverallQual_9 (Overall Material and finish of the house is Excellent)
4. MSZoning_RL (General zoning classification of the sale as Low Density Residential Zone)
5. LotArea (Total Lot Size of the house)

The new **Lasso model excluding the above top five features** resulted in the following top five features:

1. 1stFlrSF (1^{st} Floor Area)
2. TotalSqFt (Total Area = 1^{st} Floor Area + 2^{nd} Floor Area)
3. GarageCars (Size of garage in car capacity)
4. OverallQual_3 (Overall Material and finish of the house is Fair)
5. BsmtCond_Po (Basement condition is Poor)

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:

A model is considered as robust and generalizable when the model is having less variance i.e., on changing the training data the model coefficients should not change a lot. In other words, model should not overfit the training data. To make sure the model is robust and generalisable we try to reduce the chances of model overfit by applying regularization on it. Regularization can be done using following methods:

1. Lasso Regularization (L1)
2. Ridge Regularization (L2)

By regularization I mean, we can add a penalty term to the cost function such that the model cost function is penalized for high values of coefficient. The model built using this penalty term or the regularization term has feature coefficients close to zero and is less complex and hence the chances of overfitting is also reduced.

Though, a regularized model is less prone to overfit but it has lesser accuracy than the original model. This is because after regularization the resultant model is much simpler than the original one and smoother. So, a model with high degree polynomials will become a model with lower degree polynomial. This results in the increased distance of prediction line from the actual data points which is why the regularized model which is robust and generalizable has lesser accuracy. Also, as we know from Bias-Variance Trade-Off that when the variance of the model decreases the bias increases. Similarly, in the case of regularized model the variance is reduced and hence the bias is increased. For test accuracy (accuracy on unseen data), for a regularized model the train and test accuracy will remain close to each other whereas for a non-regularized model the model could overfit and have comparatively less accuracy on unseen data.