Surprise Real Estate – Advanced Regression Assignment

Question-1

Question

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 alpha:

- The optimal alpha for **Ridge** is **7**
- The optimal alpha for **Lasso** is **0.0001**

Changes after doubled alpha:

After fitting the models using alpha = 14 and alpha = 0.0002 respectively both models lost some of their predictive powers.

For the **Ridge** model:

- The **R2 on training** set went down **from 0.892 to 0.882**
- The **R2 on test** set went down **from 0.873 to 0.865**

Obviously the model makes more errors (more biased):

- The - square root of - **Mean Squared Error**:
  - on the training set went up **from 0.001274 to 0.001401**
  - on the test set went up **from 0.001728 to 0.001842**

For the **Lasso** model:

- The **R2 on training** set went down **from 0.895 to 0.889**
- The **R2 on test** set went down **from 0.881 to 0.880**

Obviously the model makes more errors (more biased):

- The - square root of - **Mean Squared Error**:
  - on the training set went up **from 0.001242 to 0.001309**
  - on the test set went up **from 0.001617 to 0.001634**

The changes are intuitive. Higher-than-optimal alpha over-penalizes the model, forcing it to be less complex, more generic. That results in a simpler-than-optimal model that is more biased. As a result it fits worse on both train and test sets.

Most important predictors:

There is no one unique definition for 'most important predictor' The most used definition says the most important predictor is the one having a coefficient with the highest absolute value. I go with this definition.

- Most important predictor for **Ridge** is: **Overall Quality**
- Most important predictor for **Lasso** is: **Gross Living Area**

Question-2

Question

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

The optimal value for alpha for ridge regression is 7, for lasso it is 0.0001. The **best** model is the **Lasso trained on** the dataset with the **full feature set** (result of EDA/Data Preparation phases).

It performs better on both the training and the test sets than any other models built in this exercise. It has the **highest R2 values** for both **train** and **test** sets. It produces the **lowest errors** on both train and test sets.

The **performance** of the model is **similar** on both **train** and **test** sets, there is no significant drop in the model performance from train to test. This indicates that there is no overfitting and the **high R2** value shows a **nice overall fit**.

The residual analysis shows that the assumptions of linear regression are met. All in all, this is a nice and solid model for predicting the real estate prices.

Question-3

Question

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 **original top 5** features of the Lasso model (with coefficients of the highest absolute values) are:

- Overall Quality
- Gross Living Area
- 2nd Floor SF
- 1st Floor SF
- Neighborhood_StoneBr

As we can see the price is determined mainly by the **quality**, the **size** and the **location** of the property. Our **intuition** or 'expertal judgement' **confirms** this conclusion.

After removing these variables and re-training the model on the remaining features the **new top 5 features are:**

- Overall Condition
- HalfBath
- FullBath
- BsmtHalfBath
- Condition1_PosA

Now this is fascinating!

1st of all we removed the variable **Overall Quality**. Its place is now taken by the **Overall Condition** that is also a kind of qualitative descriptor of the real estate! And in this sense these features are similar.

We removed **features** which represented the **size** of the property. Their places are taken by the **number of** Half and Full **Bathes**. Again, the bigger the house the more bathrooms it has on average. So somehow again the predictors still indicate the size of the house.

It tells us the story that whatever the actual variables are, on a very high level the price is influenced by:

- size

- quality

- locaton

Question-3

Question

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

Models that are simpler are usually more robust and more generalizable. That's because more complex models tend to "memorize" the datapoints instead of learning general rules and relationship (the real-life model that describes the behavior). As a result, they perform excellently on the train set but they perform bad on the unseen dataset. Also, the model changes significantly if there is a change in the training dataset.

There are multiple ways for making the model simpler, more robust and generalisable:
We can use **regularization** where the regularization term tries to force the model to use lower coefficients.
We can **restrict the model complexity** using other measures. It heavily depends on the type of the model we use. In case of regression for example we can **limit the number of features** used. In case of decision tree, we **can limit the number of nodes**, the **number of levels** etc.
Not only the model but the data preparation side can also contribute to more robust models. Continuous **variables can be binned** for example so that change in the training data results in smaller change in the model. Outliers can impact the robustness of the model therefore outliers have to be removed if they distort the model itself.

But all this comes with a price. By increasing the stability and robustness of the model we lose some of its predictive power as the model becomes more biased. Whether we use fewer features, lower coefficients, bin the continuous variables… we chose to lose a small bit of information. It results in lower variance and higher bias.

It shows the limits of the entire process. If we oversimplify the model, it will be of no use for us. One must find the perfect balance between bias and variance!