# Regularized Regression Model Assignment - ML2

EPGP in ML and AI - Apr 2023

# Problem Statement - Part 1

US-based housing company named Surprise Housing has decided to enter the Australian market. Requirement is to build a regression model using regularisation in order to predict the actual value of the prospective properties and to make an investment decision.

Following are the expected outcome:

- Finding which variables are significant in price prediction for a house
- Determining how well these variables describe the price of a house
- Optimal value determination of lambda for Ridge and Lasso regression.

# Problem Statement - Part 2

Following questions are required to be answered:

- 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?
- 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 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?
- 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?

# Solution

# Part - I

Covered as part of the Python code file provided in the repository.

## As evident, considering both Ridge and Lasso Regression

## Following are the most significant variables affecting the Sale Price

- Ground Living Area
- Overall Quality
- Basement Finished Area
- Total Basement Area

# Questions and Answers (Assignment Part - 2)

# Questions and Answers (Assignment Part - 2)

**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?**

For Ridge regression, optimum alpha value is 500. For Lasso regression, optimum alpha value is 0.01.

The effect of doubling the alpha value will be same for both ridge and lasso, the slope of the regression line will reduce and it will become horizontal. For significantly higher value of alpha can cause model underfitting.

Metrics with different alpha values:

# Questions and Answers (Assignment Part - 2)

| Metric | Ridge (Alpha = 500) | Lasso (Alpha = 0.01) | Ridge (Alpha = 1000) | Lasso (Alpha = 0.02) |
|---|---|---|---|---|
| R2 Score (Train) | 0.942896 | 0.943294 | 0.927688 | 0.930261 |
| R2 Score (Test) | 0.850395 | 0.843671 | 0.843659 | 0.835052 |
| RSS (Train) | 53.718419 | 53.344456 | 68.025130 | 65.604174 |
| RSS (Test) | 75.351742 | 78.338283 | 78.744619 | 83.079752 |
| MSE (Train) | 0.230394 | 0.229591 | 0.259265 | 0.254610 |
| MSE (Test) | 0.416679 | 0.425940 | 0.425957 | 0.437525 |

**There is no change in the predictor value for Lasso.**

**For Ridge, Kitchen Quality and Garage Car Capacity becomes significant variables. This may be due to higher comparative increase in the alpha value for Ridge which goes from 500 to 1000.**

# Questions and Answers (Assignment Part - 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?**

We can choose the two model to make a comparison with each other.

For Ridge regression shrinkage penalty is sum of squares of model coefficients.

For Lasso regression shrinkage penalty is sum of absolute value of model coefficients.

As Lasso Regression reduced coefficients to zero which leads to feature selection. We can then compare model generated by Lasso with the model using Ridge.

# Questions and Answers (Assignment Part - 2)

**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?**

Assuming that incoming dataset does not have values for five most important variables. We can easily use the existing model to figure out the five most important predictor variables. Lasso model will shrink the coefficients for the earlier important predictor variables to zero, resulting in elimination of the old (important) predictor variables. The results will show the new set of important predictor variables.

# Questions and Answers (Assignment Part - 2)

By Lasso Regression

```
betas.sort_values(by=['Lasso'], ascending=False).head(5)
```

| | Linear | Ridge | Lasso |
|---|---|---|---|
| **GrLivArea** | -1.494287e+11 | 0.091348 | 0.329125 |
| **OverallQual_Excel** | -5.362439e+07 | 0.054137 | 0.108477 |
| **OverallQual_VExcel** | -3.441872e+07 | 0.062609 | 0.107266 |
| **BsmtFinSF1** | -1.044440e+11 | 0.056179 | 0.084370 |
| **TotalBsmtSF** | 9.982031e+10 | 0.069147 | 0.084120 |

The most important predictor variables for the given train dataset.

# Questions and Answers (Assignment Part - 2)

**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?**

To make the model robust and generalisable, we can split the data into train, validation and test. This will allow the model to be verified on entirely unseen data. We can also go for a simpler model to make it more generalisable. But this may result in reduction of accuracy as the we are trying to reduce bias which will result in increase in variation.

# Thank You