CS6301.004 Special Topics in Computer Science – Practical Aspects Of Data Science

Name of Students: Prashant Kumar

Brijesh Kumar

Satyam Rameshbhai Bhikadiya (SXB180124)

Number of free late days used: \_\_\_\_\_\_\_\_0\_\_\_\_\_\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

The dataset we have used is “house price advanced regression techniques”.

**Data Preprocessing:** Initially for the column having numeric value in the dataset with null values greater than 2 %, we have replaced all the null values with the mean values of the column.

**Analysis of each of the features against predicted variables:** We have first ignored non-integer columns so that we can find correlation matrix easily.

**Correlation matrix:**

A close up of a piece of paper

Description automatically generated

**Figure 1. Correlation Matrix**

From this, we have removed those variables whose correlation with Y (i.e. SalePrice) is less than 0.5. This leaves us with the following variables:

OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, GrLivArea, FullBath, TotRmsAbvGrd, GarageCars, GarageArea

From the correlation matrix we can say that, OverallQual (i.e. Over all material and finish Quality) is highly correlated with the Sales Price.

|  |
| --- |
| > cor(dataset$OverallQual,dataset$SalePrice)  [1] 0.7909816 |
|  |
| |  | | --- | |  | |

A picture containing photo

Description automatically generated

**Figure 2. Correlation Plot between OverallQual and SalePrice**

**Regression Diagnostics:** Next, as suggested, we found the regression diagnostics for each variable.

A screenshot of a cell phone

Description automatically generated

Now by looking at the above diagnostics, we further filter some of the variables from the above variable set. As we know that, a low p-value indicates a highly statistical significance (i.e. stronger evidence that we can reject the null hypothesis) of independent variable (predictors) with the Response.

Thus, based upon the p-value. Following are the variables left with us to fit the linear model:

OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, GrLivArea, GarageCars

From the above diagnostics, we can also say that the model predicted is very good by looking on F-statistics (546.8 on 9 and 1450 Degree of Freedom) and Adjusted R2 (0.771).

The plot for the same is shown below:

A screenshot of a cell phone

Description automatically generated

**Figure 3. Residuals vs Fitted (Model)**

A screenshot of a social media post

Description automatically generated

**Figure 4. Normal Q-Q Plot (Model)**

A close up of a map

Description automatically generated

**Figure 5. Scale-Location (Model)**

A close up of a map

Description automatically generated

**Figure 6. Residual vs Leverage (Model)**

Now, when we look at the regression diagnostics of a new model (Model 2) with the following predictors:

OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, GrLivArea, GarageCars

A screenshot of a cell phone

Description automatically generated

We observe, that the p-value remain same for both the model and there is very less significant change in the Residual Standard error and Adjusted R2 value. So, we can consider this model to predict the SalePrice

The plot for the new model (Model 2) is shown below:

A screenshot of a cell phone

Description automatically generated

**Figure 7. Residuals vs Fitted (Model2)**

A screenshot of a social media post

Description automatically generated

**Figure 8. Normal Q-Q Plot (Model2)**

A screenshot of a cell phone

Description automatically generated

**Figure 9. Scale-Location (Model2)**

A close up of a map

Description automatically generated

**Figure 10. Residual vs Leverage (Model2)**

**Selection of Predictors:**

We have used stepAIC function to select the best optimum variable for the model. We have used stepAIC in both forward and reverse direction. stepAIC function also removes Multicollinearity if it exists.

A screenshot of text

Description automatically generated

The summary for stepAIC is as follow:

A screenshot of a cell phone

Description automatically generated

After performing stepAIC we are left with following predictors:

OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, GrLivArea, FullBath, GarageCars, GarageArea

From the above regression diagnostics, we can observe that step.model has a low residual standard error than the full model i.e. (model). The p-value remains same for both step.model and full model. The step.model has a better Adjusted R2 value than full model.

**Best Model Selection:**

Let us compare the three models i.e. Full model (model), model containing with less p-value (Model 2) and step.model (model obtain after performing stepAIC)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models** | **Residual Standard Error** | **Multiple R-squared** | **Adjusted R-squared** | **F-statistics** | **p-value** |
| Model | 38020 on 1450 degrees of freedom | 0.7724 | 0.771 | 546.8 on 9 and 1450 DF | < 2.2e-16 |
| Model2 | 38100 on 1453 degrees of freedom | 0.7709 | 0.77 | 815 on 6 and 1453 DF | < 2.2e-16 |
| Step.Model | 38000 on 1451 degrees of freedom | 0.7724 | 0.7711 | 615.5 on 8 and 1451 DF | < 2.2e-16 |

Now, from the above table we can observe that step.model has the lowest Residual Standard error of 38000 on 1451 degree of freedom and Adjusted R2 value of 0.7711. The p-values of all the three models are same.

Hence, based upon the Residual standard error and Adjusted R2 value we can say that step.model is the best model.

**ANOVA testing for stronger Analysis:**

Step.model is better than full model, can also be proved by using ANOVA.

A screenshot of a cell phone

Description automatically generated

The above result shows a non-significant result (p=0.9703). Thus, we should reject model and stick with the step.model

Hence, by using ANOVA testing we proved that step.model is a better model than full model

Below is the plot for step.model:

**A screenshot of a cell phone

Description automatically generated**

**Figure 11. Residuals vs Fitted (step.model)**

**A screenshot of a social media post

Description automatically generated**

**Figure 12. Normal Q-Q Plot (step.model)**

**A close up of a map

Description automatically generated**

**Figure 13. Scale-Location (step.model)**

**A close up of a map

Description automatically generated**

**Figure 14. Residual vs Leverage (step.model)**

**Some of the observations are as follows:**

1. OverallQual vs SalePrice shows gradual increase in SalePrice with increase in OverallQual.

A close up of a mans face

Description automatically generated

**Figure 15. OverallQual vs SalePrice**

Summary for OverallQual vs SalePrice:

A screenshot of a cell phone

Description automatically generated

1. YearBuilt vs SalePrice shows a slight increase in the SalePrice with increase in YearBuilt

A picture containing room

Description automatically generated

**Figure 16. YearBuilt vs SalePrice**

Summary for YearBuilt vs SalePrice:

A screenshot of a cell phone

Description automatically generated

1. YearRemoAdd vs SalePrice shows increase in SalePrice with increase in YearRemoAdd

A close up of a mans face

Description automatically generated

**Figure 17. YearRemodAdd vs SalePrice**

Summary for YearRemodAdd vs SalePrice:

A screenshot of a cell phone

Description automatically generated

1. TotalBsmtSF vs SalePrice shows, increase in variable gives increase in SalePrice upto some point and then a sudden decrease

A close up of a mans face

Description automatically generated

**Figure 18. TotalBsmtSF vs SalePrice**

Summary for TotalBsmtSF vs SalePrice:

A screenshot of a cell phone

Description automatically generated

1. GrLivArea vs SalePrice also shows positive slope increase graph.

A screenshot of a cell phone

Description automatically generated

**Figure 19. GrLivArea vs SalePrice**

Summary for GrLivArea vs SalePrice:

A screenshot of a cell phone

Description automatically generated

6) GarageCars vs SalePrice also show a positive slope of graph

A close up of a white wall

Description automatically generated

**Figure 20. GarageCar vs SalePrice**

Summary for GarageCar vs SalePrice:

A screenshot of a cell phone

Description automatically generated

7) FullBath vs SalePrice shows increase in SalePrice with increase in FullBath

A picture containing photo, sitting, table, white

Description automatically generated

**Figure 21. FullBath vs SalePrice**

Summary for FullBath vs SalePrice:

A close up of text on a white background

Description automatically generated

8) GarageArea vs SalePrice shows a positive slope in the graph upto certain point then decreases

A close up of a mans face

Description automatically generated

**Figure 22. GarageArea vs SalePrice**

Summary for GarageArea vs SalePrice:

A screenshot of a cell phone

Description automatically generated