Predictive Analytics Assignment

Shruti Kirtikumar Deolekar

19200705

- **Exploratory Data Analysis:**

1. **Using a boxplot, histogram and summary. Describe the distribution of the sales price of the houses.**

- The summary shows that mean of the price data is greater than the median, hence the data distribution of price is right-skewed (positively skewed). Furthermore, plotting the histogram and drawing lines it can be graphically seen that the data distribution for price is bimodal.

2. **Convert all the categorical variables to factors. Using the summary and a boxplot describe how sales prices vary with respect to the number of bedrooms, bathrooms, garage size and school.**

- In the given data, we can term the variables Bed, Bath, Garage and School as categorical variables. While variable School is already a factor, I have converted variables Bed, Bath and Garage to factors and below are the observations:

Price v/s Bed:

- The data distribution seems to be symmetric with 3 bedrooms whereas with 4 bedrooms it is slightly skewed positively, with 2 possible outliers for 4 bedrooms which conveys that the price is unusually high for two individual samples of data when the number of bedrooms is high.
- The data distribution is highly skewed negatively when number of bedrooms is 2 or 5 and we can observe a single entry in the dataset with six number of bedrooms.

Price v/s Bath:

- There is a huge variation in the data distribution as the number of bathrooms change, so we can observe skewness in the data. However, the data stands symmetric when there is only a single bathroom in the house.

Price v/s Garage:

- From the statistics obtained, we can make a statement that the price is likely to increase with an increase in number of garages from 0 to 2, all being skewed negatively.
- The distribution stands symmetric when the number of Garages is 3 and an outlier is observed in the distribution which conveys that for an observation in the data with no garages available, the price is unusually high.

Price v/s School:

- Houses near the Alexandra College are found to be cheapest where as the ones near the High School appear to be the most expensive ones
- There is an outlier each for school as the St. Louis High School and the St. Mary’s College stating the price is unusually high for that particular observation.

3. **Using the summary, correlation and the pairs plots discuss the relationship between the response sales price and each of the numeric predictor variables.**

- There isn't any high correlation observed of Price with any of the numeric variables (Size, Lot and Year). All the numeric variables are weak-positively correlated to Price as Size (0.2014), Year (0.1541) and Lot (0.2442).

- **Regression Model:**

1. **Fit a multiple linear regression model to the data with sales price as the response and size, lot, bath, bed, year, garage and school as the predictor variables. Write down the equation for this model.**

- I have rescaled the numerical variables before fitting the regression model and following is the equation obtained:

  Price = β0 + β1 Lot + β2 Size + β3 Year + β4 Bath1.1 + β5 Bath2 + β6 Bath2.1 + β7 Bath3 + β8 Bat3.1 + β9 Bed3 + β10 Bed4 + β11 Bed5 + β12 Bed6 + β13 Garage1 + β14 Garage2 + β15 Garage3 + β16 SchoolHigh + β17 SchoolNotreDame + β18 SchoolStLouis + β19 SchoolStMarys + β20 SchoolStratford + ϵ

2. **Interpret the estimate of the intercept term β0.**

- We have obtained the intercept term β0 as **376.1016** which conveys that for a house with mean Lot value, mean Size value, mean Year value, 2 Beds, 1 Bath, no (or zero) garage and near Alexandra College the price would be 376.1016 thousand Euros (376101.6 €).

3. **Interpret the estimate of βsize the parameter associated with floor size (Size).**

- The price of house increases by **59.4503** thousand Euros per thousand increase in the floor size.

4. **Interpret the estimate of βBath1.1 the parameter associated with one and a half bathrooms.**

- The price of house increases by **135.8983** thousand Euros if it has one and half bathrooms.

5. **Discuss and interpret the effect the predictor variable bed on the expected value of the house prices.**

- If the number of beds in the houses increases (after 2) then there is a significant decrease in the price of the house.

6. **List the predictor variables that are significantly contributing to the expected value of the house prices.**

- The variables significantly contributing to the price of the house are:
  Lot, Size, Bed, Bath, Garage and School

7. **For each predictor variable what is the value that will lead to the largest expected value of the house prices.**

- Before fitting the model, I had rescaled the data for numeric variables, following are the values:
  Lot: 6.98667
  Size: 0.924507
  Year: 34.9333
  Bath: The Price value is 135.8983 corresponding to Bath1.1 (level = 1.1)
  Bed: The Price value is same as that in the intercept corresponding to Bed2 (level = 2)
  Garage: The Price value is 18.2435 corresponding to Garage3 (level = 3)
  School: The value is 113.2774 corresponding to schoolHigh (level = High)

8. **For each predictor variable what is the value that will lead to the lowest expected value of the house prices.**

- Before fitting the model, I had rescaled the data for numeric variables, following are the values:
  Lot: -3.0133
  Size: -0.531493
  Year: -65.06667
  Bath: The Price value is same as that in the intercept corresponding to Bath1 (level = 1.0)
  Bed: The value is -238.2609 corresponding to Bed4 (level = 4)
  Garage: The Price value is same as that in the intercept corresponding to Garage0 (level = 0)
  School: The value is same as that in the intercept corresponding to schoolAlex (level = Alex)

9. **By looking at the information about the residuals in the summary and by plotting the residuals do you think this is a good model of the expected value of the house prices.**

- The data can be seen equally distributed towards either side of zero, majorly lying between 50, -50 which indicate there is a large difference in the observed and the estimated value of the response variable. The model can be considered as a good fit as 51.25% of variation in Price is explained by the predictor variables.

10. **Interpret the Adjusted R-squared value.**

- The adjusted R-squared value is 0.5125. It conveys that 51.25% of variation in Price is explained by the predictor variables in the model thus the model can be considered as a good fit.

11. **Interpret the F-statistic in the output in the summary of the regression model.**

- We get F-statistic value as **4.942**
  Hypothesis test:
  H0: β1 = β2 = β3 = β4 = β5 = β6 = β7= β8 = β9 = β10 = β11 = β12 = β13 = β14 = β15 = β16 = β17 = β18 = β19 = β20 = 0
  Ha: At least one of [β1, ………… , β20] ≠ 0

  p-value = 1.265e-06 < 0.05

Thus, we can say that at least one of [β1, ………… , β20] is non-zero and hence we reject the null hypothesis H0.

- **ANOVA:**

1. Compute the type I ANOVA table. Interpret the output.

Hypothesis test for variable Lot:

H0: β1 = 0
H1: β1 ≠ 0
F-statistic is: 9.1767
p-value is 0.003729 which is < 0.05
Hence, we reject H0 and can say that variable Lot is significant.

Hypothesis test for variable Size:

H0: β2 = 0
H1: β2 ≠ 0
F-statistic is: 5.6498
p-value is 0.020964 which is < 0.05
Hence, we reject H0 and can say that variable Size is significant.

Hypothesis test for variable Year:

H0: β3 = 0
H1: β3 ≠ 0
F-statistic is: 2.6715
p-value is 0.107872 which is > 0.05
Hence, we fail to reject H0 and can say that variable Year is insignificant.

Hypothesis test for variable Bath:

H0: β4 = β5 = β6 = β7 = β8 = 0
H1: At least one of [β4, β5, β6, β7, β8] ≠ 0
F-statistic is: 4.2760
p-value is 0.002345 which is < 0.05
Hence, we reject H0 and can say that variable Bath is significant.

Hypothesis test for variable Bed:

H0: β9 = β10 = β11 = β12 = 0
H1: At least one of [β9, β10, β11, β12] ≠ 0
F-statistic is: 2.8458
p-value is 0.032393 which is < 0.05
Hence, we reject H0 and can say that variable Bed is significant.

Hypothesis test for variable Garage:

H0: β13 = β14 = β15 = 0
H1: At least one of [β13, β14, β15] ≠ 0
F-statistic is: 3.0245
p-value is 0.0037179 which is < 0.05

Hence, we reject H0 and can say that variable Garage is significant.

Hypothesis test for variable School:

H0: $\beta16 = \beta17 = \beta18 = \beta19 = \beta20 = 0$
H1: At least one of [$\beta16, \beta17, \beta18, \beta19, \beta20$] $\neq 0$
F-statistic is: 7.9020
p-value is 1.153e-05 which is < 0.05
Hence, we reject H0 and can say that variable School is significant.

2. **Which predictor variable does the type 1 ANOVA table suggest you should remove the regression analysis?**

- After analysis of the ANOVA table, we get to know that the variable Year is insignificant and should be removed from the regression analysis.

3. **Compute a type 2 ANOVA table comparing the full model with all the predictor variables to the reduced model with the suggested predictor variable identified in the previous question removed.**

   Hypothesis test:
   H0: $\beta3 = 0$
   H1: $\beta3 \neq 0$
   F-statistic is: 2.7064
   p-value is 0.1057 which is > 0.05
   Hence, we fail to reject H0 and can say that variable Year is insignificant.

- **Diagnostics:**

1. **Check the linearity assumption by interpreting the added variable plots and component plus residual plots. What effect would non-linearity have on the regression model and how might you correct or improve the model in the presence of non-linearity?**

- By looking at the AV plots we can say that there exists a linear relationship between all the predictor variables individual with the variable Price.
- However, the component plus residual plot gives us a better insight of the linearity which indicate a line in colour pink that models the residuals of our predictor and a dashed line in colour blue which shows how the best fit would look like. The pink line appears curved relative to the blue dashed line for the variable Year, so we can say that we non-linearity exists between Price and Year. We also have issues with linearity between Price and Lot but the relationship between Price and Sie is strongly linear. For the categorical variables, a box plot is plotted and we can say that the linearity is altered with varying skewness.
- The effect of non-linearity accounts to inconsistent and biased prediction estimates and can fail to give you proper expected results. Non linearity can be corrected by transformation of the model expression or performing segmentation or with the usage of splines and polynomials.

2. **Check the random/i.i.d. sample assumption by carefully reading the data description and computing the Durbin Watson test (state the hypothesis of the test, the test statistic and p-value and the conclusion in the context of the problem). What are the two common violations of the random/i.i.d. sample assumption? What effect would dependant samples have on the regression model and how might you correct or improve the model in the presence of dependant samples?**

- Hypothesis Test for DW Test:
  H0: Autocorrelation does not exist
  H1: Autocorrelation exists

  DW Statistic: 1.614157
  The p-value is 0.034 so the hypothesis of autocorrelation in rejected and the observations cannot be classified as independent.

- Violations: Heteroskedasticity, Inefficiency/bias
- Corrections: Time Series Analysis, Mixed effect Models

3. **Check the collinearity assumption by interpreting the correlation and variance inflation factors. What effect would multicollinearity have on the regression model and how might you correct or improve the model in the presence of multicollinearity.**

- The VIF of our numeric predictor variables is close to 1 that means correlation is not present to a greater extent and the variance of bj is not much inflated.
- Multicollinearity effects the precision of the estimates we achieve from our regression model and lessens the significance of the p-values.
- Multi collinearity can be reduced by removing highly correlated predictors from the model and by using methods that cut the number of predictors to a smaller set of uncorrelated components such as Partial Least Square Regression (PLS), Principal Component Analysis (PCA), Ridge regression

4. **Check the zero conditional mean and homoscedasticity assumption by interpreting the studentized residuals vs. fitted values plots and the studentized residuals vs. predictor variable plots. What effect would heteroscedasticity have on the regression model and how might you correct or improve the model in the presence of heteroscedasticity**.

- The data can be seen equally plotted towards either side of zero, so we can say that no major heteroscedasticity exists.
- The effect of heteroscedasticity accounts to biased results which can be corrected by using Weighted Least Square Method.

5. **Check the Normality assumption by interpreting the histogram and quantile-quantile plot of the studentized residuals. What effect would non-normality have on the regression model and how might you correct or improve the model in the presence of non-normality.**

- From our qq plot we can see a near to 45-degree line which conveys that our distribution is equal, similar conclusions result from the histogram.
- Non normality may result in incorrect values of critical t and F-tests which can be corrected by using transformations, interactions or a different model.

- **Leverage, Influence and Outliers:**

1. **What is a leverage point? What effect would a leverage point have on the regression model?**

- A leverage point is a one with unusual X-value.
- Effect: Model statistics such as R-square, SSE etc. are affected and minor effect can be seen on the regression coefficients as well. The fit of the model can be altered if leverage points are too high.

  The following are the leverage points after observing leverage plot of our model: 1, 2, 3, 4, 5, 6, 7, 9, 15, 20, 21, 22, 28, 31, 32, 33, 34, 35, 36, 37, 39, 41, 42, 43, 44, 46, 47, 49, 50, 51, 52, 53, 56, 57, 58, 63, 66, 69, 71, 72, 73, 74, 76

2. **What is an influential point? What effect would an influential point have on the regression model?**

- An influential point is a one with unusual Y -value along with an unusual X-value. High leverages cases are usually potentially influential and their removal can cause a large change in the fit, thus they should be examined for their influence.
- It influences the regression model in its direction by creating a major impact on the model coefficients.

  The following are the influence points after observing influence plot of our model: 30, 44, 47, 21

3. **What is an outlier? What effect would an outlier have on the regression model? How would you correct the outliers?**

- An outlier is an observation where the response does not correspond to the model fitted to the bulk of the data.
- The outliers might affect the regression estimates. We should check data entries, investigate for unexpected irregularities and decide to exclude/include after checking it's influence.
- In our outlier test, the column for Bonferroni-p is NA, thus we can say that there are no outliers in our data.

- **Expected value, CI and PI:**

1. **Plot the observed house prices, their expected vale (fitted value), confidence intervals (in red) and prediction intervals (in blue). Looking at this plot is this model providing a good estimate of the house prices.**

- As observed in the plot, most of the values lie inside the interval, hence we can conclude that our model is a good fit.

**********************END************************