Nikhil Ambati
Final Exam

Initial Analysis or Data Splitting:

First, by analysing the data, I observed that data for Property 1 is in columns 1 to 22, and data for Property 2 is in columns 23 to 96. Based on this, I created three separate data sets and ran different models on each.

Once these data sets were created, I applied various regression techniques on Property 1, used classification techniques for Property 2, and performed same regression techniques on Property 3. Utilizing these methods, I successfully predicted the missing values.

We will first see the regression techniques for the property 1 with the detail's steps.

Property 1 Predication

1. Data Assessment

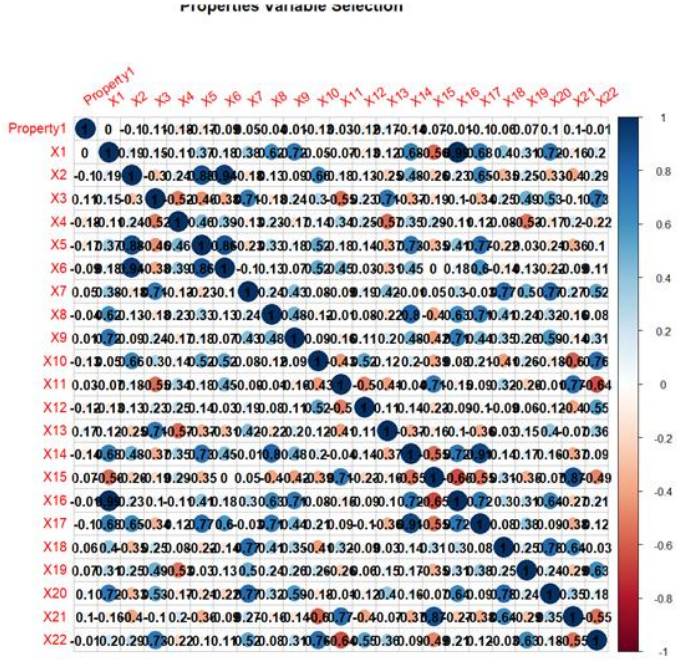
In the give data set I have found out below issues, so we have corrected them using different techniques.

a) Outliers
b) Null values
c) Evaluated and correct the correlation and non-correlation among the features.

2. Feature Selection

a) Correlation Matrix & Heat Map

In the feature selection part, I have checked the correlation between the features. This was done checking the correlation matrix and studied the heatmap. I identified strong positive correlations and negatively correlated and selected the feature as per that.

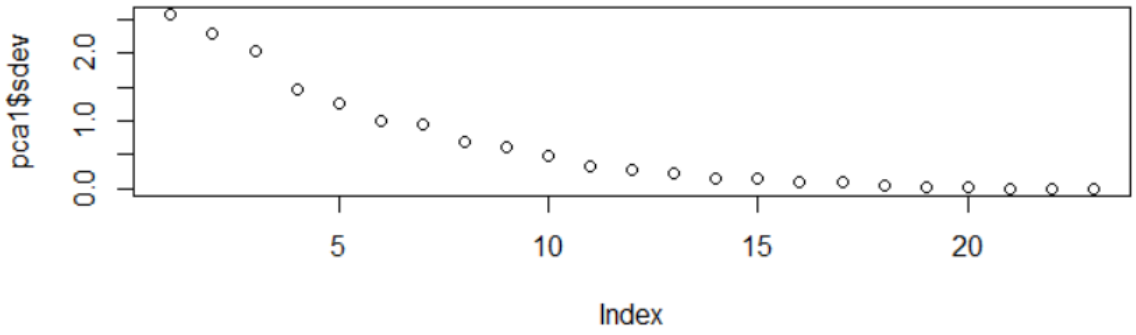


Nikhil Ambati
Final Exam

Looking at the above correlation matrix and the heat map I have Property1" **"X2",”X4”, X12", "X13","X19".**

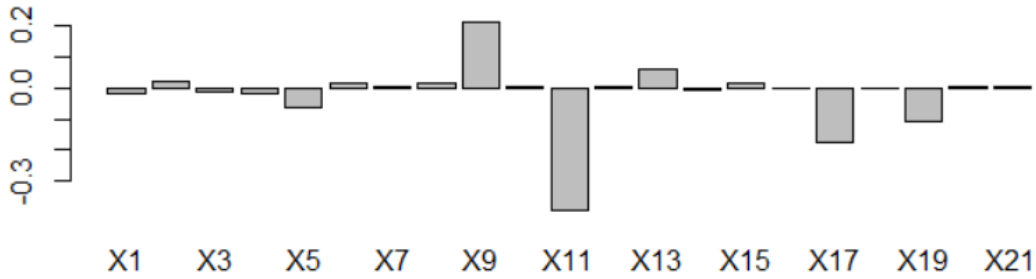
b) PCR (Principal Component Regression)

During the PCR analysis, an elbow bend was observed at 7 PCS, leading to the selection of 7 PCS.



c) Variable Importance in Projection (VIP PLOT)

Below is the VIP plot which I have got, and we could see that for X11, X9, X17 and X19 has good VIP scores. So, I consider them.



The final features selected from considering all analysis are following
Feature1","X2","X4",”X9”,"X11","X12","X13","X17".

3. Result after applying regression models.

Regression Model	RMSE train	RMSE test	RSQ
Linear Regression	0.9857	0.99601	0.0132
Multilinear Regression	0.98006	0.9499	0.0939
Gaussian Process Regression	0.759357	0.837382	0.2732
Random Forest Regression	0.533232	0.977832	0.5232

We could see the difference between RMSE train RMSE test are significant, which means the data is overfitted. Whereas the difference in MLR model is low.

So, I have **finalized MLR model and predicted** the below values.

```
> prediction_mlr
       1        2        3        4        5        6        7        8
131.2761 142.2807 128.6567 129.8774 131.0980 142.0537 136.5758 140.0497
>
```

4. Conclusion

Considering the predicted value are not too high, I find that the model is performing adequately. While some predictions are notably good, overall, my confidence in the model's performance is moderate.

Nikhil Ambati
Final Exam

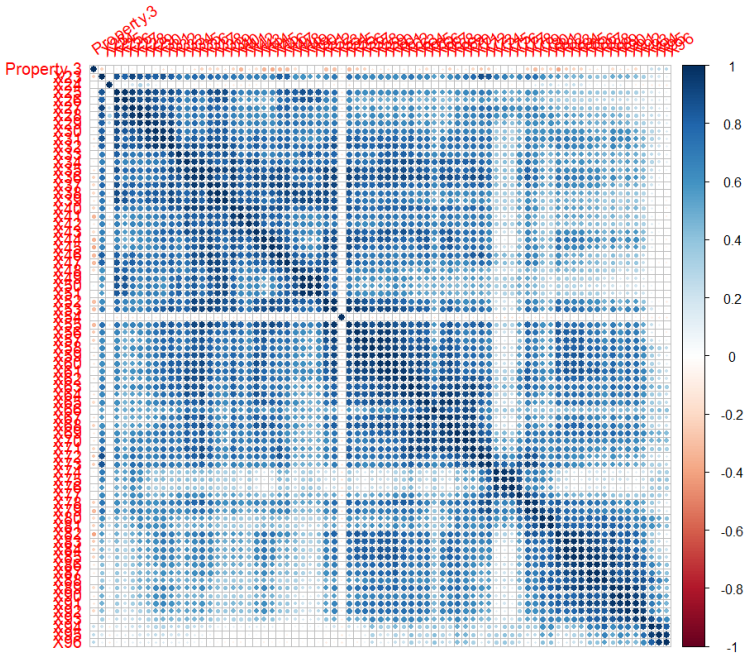
Property 3 Prediction using Classification.

1. Data Assessment

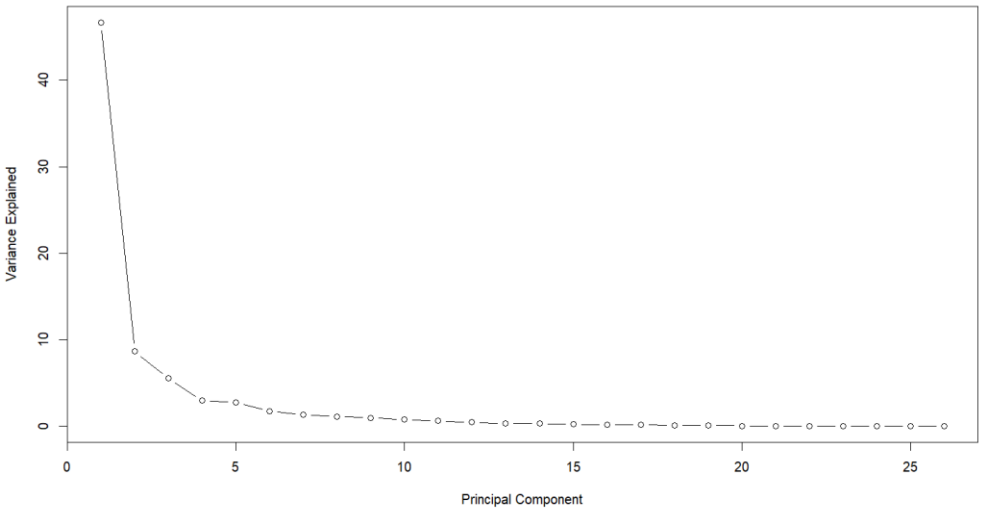
I have performed similar steps as for Property 1 to check the quality and outliers in the data. It seems we have less outliers as compared to Prop 1.

2. Feature Selection

In the feature selection part, I used VIP score and PCA plot to select the feature. As you can see from the below correlation matrix it's hard to decide.

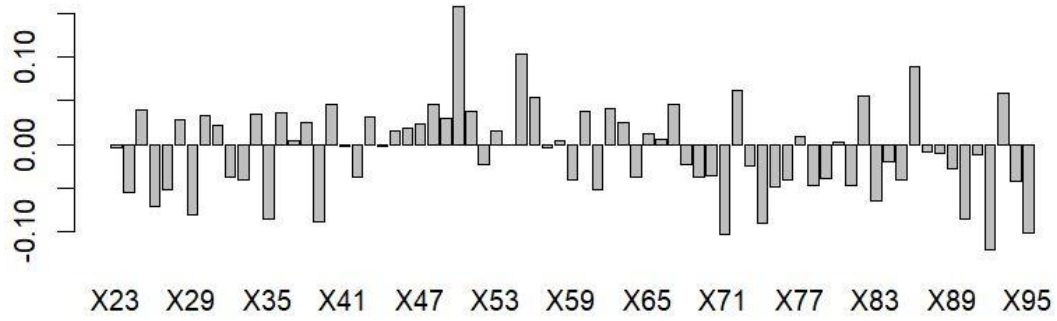


b) PCR (Principal Component Regression)



For the above we could see that we have elbow at 5, so selected 5 PC.

c) Variable Importance in Projection (VIP PLOT)



Based on the plot scores and the above correlation the finalized features were **50,55,71,92,95.**

3) Result after applying regression models.

Regression Model	RMSE train	RMSE test
Linear Regression	53.25005846	68.65436184
Lasso Regression	53.25053484	68.62512075
Ridge Regression	53.25007009	68.64914295
Support Vector Regression	68.91254828	66.50468694

We can that the difference between RMSE Train and Test for SVR is significantly low, so SVR is better. While other models have higher difference, which mean that the data is overfitting.

4) Predicted Value:

Below are predicted values using support vector regression.

	Property3
1	152.1822
2	176.7147
3	164.8929
4	166.2819
5	164.0638
6	154.0265
7	157.3730
8	145.1387

Nikhil Ambati
Final Exam

4. Conclusion

Considering the predicted values are not reasonably high, I find that the model is performing average. My confidence in the model's performance is moderate.

Nikhil Ambati
Final Exam

Property 2 Prediction using Classification.

I have followed same steps for feature selection and data analysis or correction as I did for Property 1. In this data I have performed below classification techniques:

Also I have modified the values of High as 1 and Values of Low as 0.

a) Support Vector Machine
b) Decision Tree
c) Logistic Regression

a) Support Vector Machine

```
> (misclass <- table(svm.pred_train, truth = train_svm$High))
              truth
svm.pred_train -1  1
             1  6 69
```

```
> (misclass <- table(svm.pred_test, truth = test_svm$High))
             truth
svm.pred_test -1  1
            1  3 16
```

b) Decision Tree

```
> with(train_dt,table(tree.pred_train,High))
               High
tree.pred_train No Yes
            No   1   9
            Yes  5  60
```

```
> with(test_dt,table(tree.pred_test,High))
              High
tree.pred_test No Yes
           No   2   0
           Yes  1  16
```

c) Logistic Regression

```
> (misclass <- table(glm.pred_train, truth = train_logit$High))
              truth
glm.pred_train  0  1
             1  6 69
```

Nikhil Ambati
Final Exam

```
> (misclass <- table(glm.pred_test, truth = test_logit$High))
             truth
glm.pred_test  0  1
            1  3 16
```

Predicted Values:

Using SVM:

Property2
High
High
High
High
High
High
High
High

Using Decision Tree Classifier:

Property2
High
High
High
High
High
Low
High
High

UsingLogisticsRegression

Property2
High
High
High
High
High
High
High
High

Nikhil Ambati
Final Exam

Conclusion:

Considering the predicted values all as high and the requirement was to be high so I feel that my model is performing good and my confidence is great.

Final Conclusion:
Based on the provided information, it seems that the predicted values for Property 1 and Property 3 do not meet the design requirements, while the values for Property 2 are high and meet the requirements. As a result, there is less confidence in the predictions for Property 1 and 3. But for Property 2 the values are high which meets design requirements. I am overall confident about all 3 models.

Nikhil Ambati
Final Exam