Write-up – Project on Property Price Prediction

By: Hassaan Ayub

Data Preparation in Excel

The first step was data cleaning, where extreme values were replaced with mean values of that attribute, null value were replaced with the mean values for numeric variable and mode value for the factor variables. The dataset was a combination of categorical (factor) variables and numeric variables. Since the regression was run in excel, so all variables had to be brought I numeric form. Thus, major task of the assignment was to bring all the variables in workable form.

As excel has a limit that it cannot handle more than 16 variables at a time, thus different approaches were taken to deal with different factor variables. Those factor variables which had few attributes, those were made as dummy variables which were treated as binary variables where the digit 1 represented true and 0 represented 0. For instance, the factor variable Area ID had only three types of attributes, so two dummy variables were made out of these, where each true represented belonging to that Area ID, while false in both represented belonging to third Area ID.

In a different approach, a factor variable where the number of attributes were high, each attribute was assigned a number and a key was created that explained which number represents which type of apartment. In this case, for the variable property type, the key used was the following;

(1 =Flat, 2 =Gymnasium, 3 = hotel apartment, 4= hotel rooms, 5 = office, 6 = shops, 7 = show rooms, 8 = sized partition, 9 = unknown)

Likewise, many similar were combined into one factor variable as named them as “others”. For instance, the factor variable “number of rooms” had numeric as well as different types as mentioned, so those “duplex, loft, penthouse, studio, non-specified” were aggregated as “Others” and were assigned a number of 99.

Another variable (lagging amount paid) was introduced in the dataset to capture the shifting trends in the prices of apartments.

Training & Testing Split

From the total dataset, 80% of the dataset was split as the training dataset and the remaining 20% dataset was taken in testing, making sure there was no overlap. The splitting was done using the variable Area ID, where 20% dataset was extracted from each of the three Area IDs randomly.

Regression on Training

Using data analysis function on excel, regression was run on the training dataset and co-efficient of all the variables were used in the working of testing dataset to calculated the predicted value of Y (Amount Paid).

Working on Testing Dataset to find Predicted Amount Paid

Using all the co-efficient, intercept and the testing dataset, the Predicted Amount Paid was calculated. Those calculated values were compared with the original values of Y (Amount Paid), and the Delta (Error) was calculated. Likewise, mean absolute percentage error (MAPE), Median Absolute Percentage Error (Median APE) and Correlated between original Amount paid and predicted Amount Paid was calculated.

Conclusion

The R square value of Regression on Training Dataset was 74% which means that 74% of the variations in the dataset were explained by the variables in our model. The Mean Absolute Percentage Error (MAPE) was 22.65%, Median Absolute Percentage Error was 22% and the Correlated between original Amount Paid and Predicted Amount Paid was 86%.

Data Preparation in R Programming

The first step was data cleaning, where extreme values and null value were replaced with mean values of that attribute. The dataset was a combination of categorical (factor) variables and numeric variables. Hence, for factor variables where there were various different categories, so a pivot table was used to determine the number of occurrence of each category, and then a benchmark was set (subjectively) so that all above the benchmark were treated as categories of that factor variables and the rest were accumulated in a single category of "Others". This technique was done for variables; "Land ID", "Floor Number" and "Number of Rooms" and the purpose was to lower the number of variables that were fed into R. Similarly, for a factor variable where the number of count for any given category was very few, so those rows were deleted completely so as to avoid the errors rising due to distribution of that dataset into any one of training or testing dataset. For instance, for the variable "Type of Property", Gymnasium category appeared once and Sized Partition Category appeared twice, so deleting these three rows from the dataset of 36000+ rows does not impact the results.

The rows for which sale date was not available were deleted from the dataset, and then the data set was sorted based on sale date. Then, an additional variable of lagging sales variable was introduced to capture the trend of changing prices.

As excel has a limit that it cannot handle more than 16 variables at a time, thus the linear model was run on R to deal with different factor variables.

Code used in R

```
data <- read.csv("CleanedData.csv", header = T, stringsAsFactors = T)
str(data)
data$Area.ID. <- as.factor(data$Area.ID.)
data$Building.ID <- as.factor(data$Building.ID)
str(data)


train <- data[1:2965,]
test <- data[2966:3656,]
model <- lm(Amount.Paid ~ . , train, family = "binomial")
```

```
model2 <- step(model)
trainingpreds <- model$fitted.values
testingpred <- predict(model2, test, type = "response")
allpreds <- c(trainingpreds, testingpred)
write.csv(allpreds , file = "Property Price Prediction.csv")
```

Training & Testing Split

From the total dataset, 80% of the dataset was split as the training dataset and the remaining 20% dataset was taken in testing, making sure there was no overlap. The splitting was done using the code in R.

Working on Predicted Values output from R

Using the output file generated by R, the predicted amount price paid was used as Yi (Predicted Y) and that was compared with the actual amount paid (actual Y). the Delta (Error) was calculated. Likewise, mean absolute percentage error (MAPE), Median Absolute Percentage Error (Median APE) and Correlated between original Amount paid and predicted Amount Paid was calculated.

Conclusion

The Mean Absolute Percentage Error (MAPE) was 29.13%, Median Absolute Percentage Error was 18.27% and the Correlated between original Amount Paid and Predicted Amount Paid was 85%.