University of Illinois Urbana-Champaign

Department of Statistics

Real Estate Valuation

Group - 14
Sharvi Tomar (stomar2)
Shubahm Mehta (mehta45)
Brianna Grace Suits (bgsuits2)

Contributions

Sharvi: Multiple Linear Regression, Regression Trees, Report

Shubham: Exploratory Data Analysis, Random Forest, Report

Brianna: Introduction, Ridge Regression, Report

Contents

Chapter 1: Introduction

The real estate market is constantly changing and seemingly unpredictable, but with the help of data analytics, the most influential factors can be extracted to best fit a model that will predict the valuation of real estate. The purpose of this report is to analyze a real estate valuation dataset specifically from Taiwan by using predictive modeling that takes into consideration the basics of supply and demand pricing while also incorporating the influence of various factors such as the transaction date, house age, distance to MRT stations, proximity of convenience stores, and geographical location.

The data came from the article *"Building real estate valuation with comparative approach through case-based reasoning"* written by I-Cheng Yeh and Tzu-Kuang Hsu, and was collected from the Sindian District of New Taipei City, Taiwan between August 2012 and July 2013. This dataset, though limited in scope and application, offers a glimpse of how data analysis and predictive modeling can be used in other areas to cover a wider range of real estate markets.

1.1 Goal

In this project, our objective is to correctly predict the Price per Unit Area of the houses in Sindian Dist., New Taipei City, Taiwan. The project also compares the performance and interpretations of the results of different machine learning models.

1.2 Approach & Evaluation Criteria

Exploratory Data analysis was done to get a good understanding of the data. Then, different machine-learning algorithms- Multiple Linear Regression, Ridge Regression, Regression Trees and Random Forest were used to predict the Price per unit area of the houses. Root Mean Squared Error(RMSE) on test data was chosen as the evaluation criteria to compare different models.

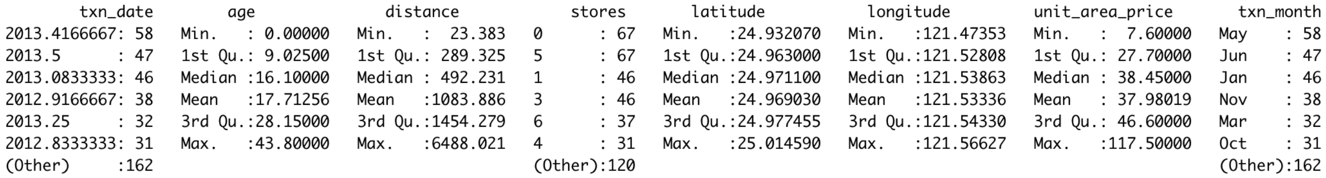
Chapter 2: Exploratory Data Analysis

2.1 Data Dimensions & Description

There are 414 observations in the data. We extracted the month information and added it as a predictor in the dataframe with the name 'txn_month'. The "store" predictor has only 11 unique values. Thus, we decided to keep the number of convenient stores as categorical. Similarly, the transaction date and transaction month had 12 unique values, which corresponds to each month from Aug, 2012 to July, 2013. Thus, we kept them as categorical.The class for other variables is self-intuitive. We also found that there were no missing values in the data. My final data for analysis has 7 predictors and 1 response. The below table summarizes the description of the variables.

Variable	Class	Description
txn_date	categorical	transaction date between August 2012 and July 2013
age	numerical	house age at the time of transaction (in years)
distance	numerical	distance to the nearest MRT station (in meters)
stores	categorical	the number of convenience stores in the living circle on foot (as an integer)
latitude	numerical	the geographical coordinate of latitude
longitude	numerical	the geographical coordinate of longitude
unit_area_price	numerical	house price of the unit area (10000 New Taiwan Dollar/Ping; 1 Ping = 3.3 square meters)
txn_month	categorical	transaction month extracted from txn_date

2.2 Numerical Summary



```
        txn_date        age             distance           stores       latitude          longitude        unit_area_price     txn_month
 2013.4166667: 58   Min.   : 0.00000   Min.   :  23.383   0      : 67   Min.   :24.932070   Min.   :121.47353   Min.   :  7.60000   May    : 58
 2013.5      : 47   1st Qu.: 9.02500   1st Qu.: 289.325   5      : 67   1st Qu.:24.963000   1st Qu.:121.52808   1st Qu.: 27.70000   Jun    : 47
 2013.0833333: 46   Median :16.10000   Median : 492.231   1      : 46   Median :24.971100   Median :121.53863   Median : 38.45000   Jan    : 46
 2012.9166667: 38   Mean   :17.71256   Mean   :1083.886   3      : 46   Mean   :24.969030   Mean   :121.53336   Mean   : 37.98019   Nov    : 38
 2013.25     : 32   3rd Qu.:28.15000   3rd Qu.:1454.279   6      : 37   3rd Qu.:24.977455   3rd Qu.:121.54330   3rd Qu.: 46.60000   Mar    : 32
 2012.8333333: 31   Max.   :43.80000   Max.   :6488.021   4      : 31   Max.   :25.014590   Max.   :121.56627   Max.   :117.50000   Oct    : 31
 (Other)     :162                                         (Other):120                                                            (Other):162
```

Figure 2.1: Numerical Summary of data

- The data is from Aug '12 to July '13 with transactions in every month. Thus, duration of transactions is 1 year
- The age of purchased houses is from 0 to 43.8 years, with mean age of 17.7 years
- The number of convenient stores range from 0 to 10, with discrete integer values
- Average unit area price of the house is 37.9 thousands New Taiwan Dollar/Ping

2.3 Graphical Analysis

Firstly, we wanted to understand how the number of transactions vary across each of the predictors. We started with transaction month and stores.From figure 2.2, we noticed that the number of transactions across every month vary, with highest in May - 2013 and lowest in Jul - 2013. From figure 2.3, we noticed that most houses have either 0 or 5 convenience stores in the living circle. There are few houses (10) which have 10 convenience stores in the living circle.

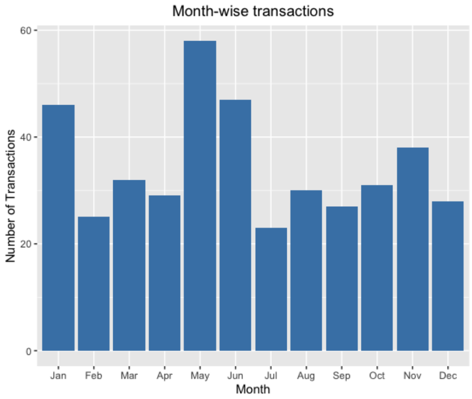


Figure 2.2: Month-wise transactions

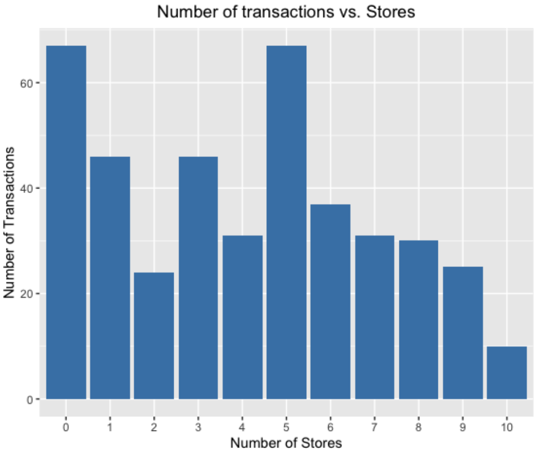


Figure 2.3: Transaction vs. Stores

Then, we identified how transactions vary across the numerical predictors like age, distance and the response 'Unit Area Price'. From the density plots and histograms as shown in figure 2.4, we can observe that Distance,Unit Area price are right skewed. From the boxplots, we found that distance has many outliers, unit area price has only 2-3 outliers and age did not have any outliers.

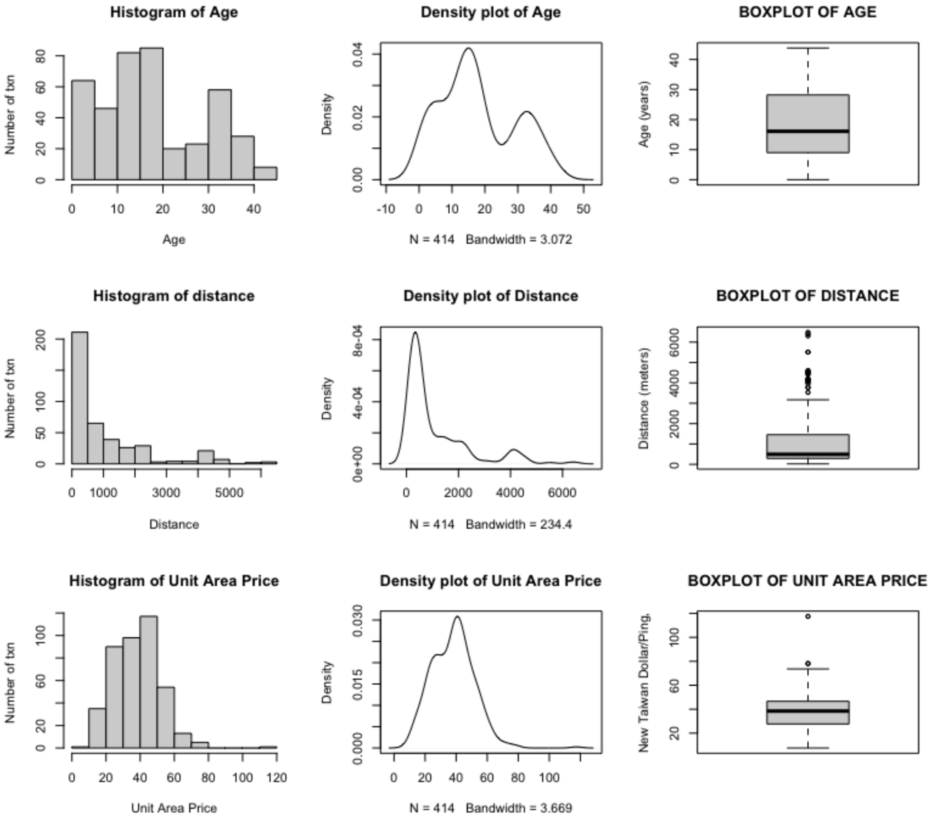


Figure 2.4: Univariate Analysis of Distance, Age and Unit Area Price

Secondly, we wanted to understand how each of the predictors affect the price using graphical summaries. We observed , from figure 2.5, that houses which are closer to the nearest MRT station have higher prices and this inverse relationship between price and distance is very strong. We also noticed that houses which are new have higher prices.However, this inverse relationship between price and distance is weak as can be seen from figure 2.5.

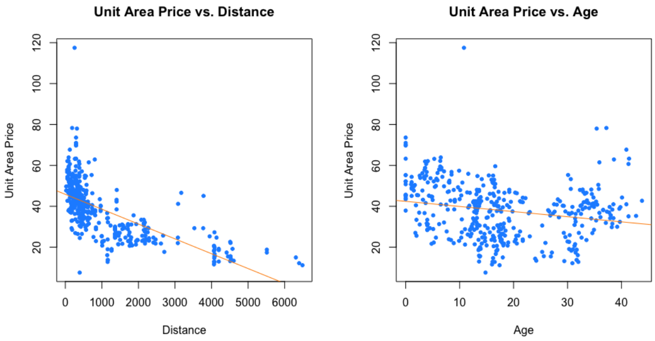


Figure 2.5: Unit Area Price vs. Distance,Age

Next, we evaluated how predictors like Number of Stores and Transaction Month affect the Unit area price of the houses. We can observe from figure 2.6, that the median unit area price does not vary significantly by the month.However, median unit area price varies significantly by the number of stores (figure 2.7). In general, higher the number of stores, higher is the unit area price. We also noticed that the Unit Area price shows maximum variation in the month of July 2013 and minimum variation in the month of Jan 2013.

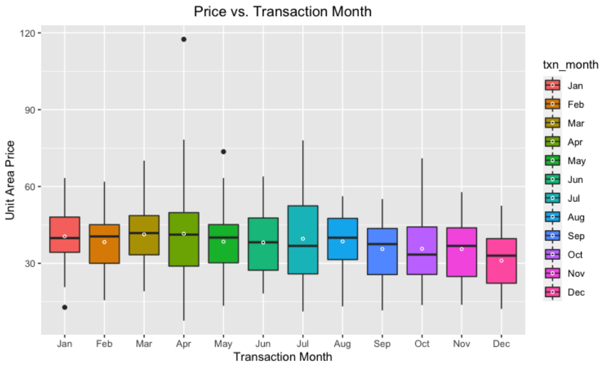


Figure 2.6: Unit Area Price vs. Transaction Month

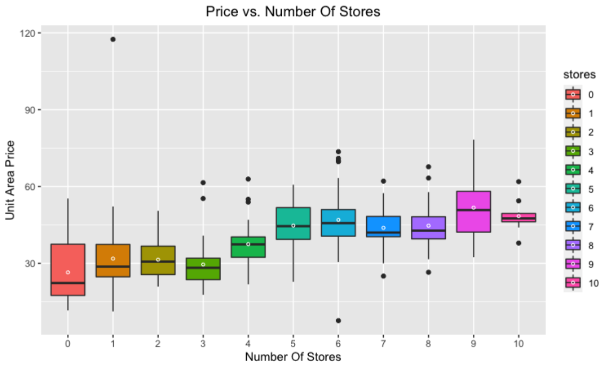


Figure 2.7: Unit Area Price vs. Number of stores

Then, we wanted to check how the geographical location of the houses influence their unit area price. We divided the house prices into 3 categories based on price per unit area: Low(<30), Mid(30-60) and High(60 and above)

From the geographical plot in figure 2.8, we can see that many houses which are clustered together have high,medium unit area price whereas the houses which are on the outskirts have less unit area price. Thus, location of the houses seem to play an important role in determining the unit area price of the house.

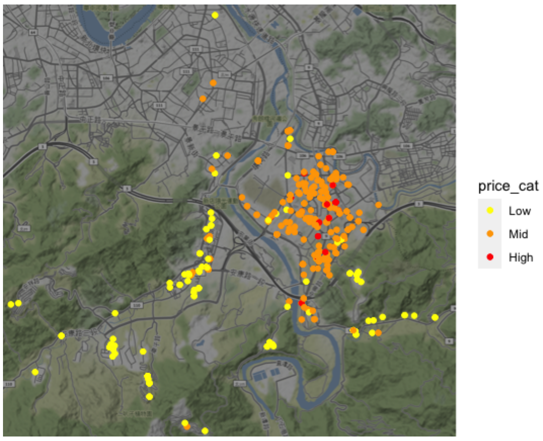


Figure 2.8: Price Category vs. Location of houses

From the correlation plot (figure 2.9) of numerical predictors, we observed the following:

- Unit area price shows a moderate/high negative correlation with the distance (-0.67). Thus, as distance from the nearest MRT station increases, unit area price decreases
- Unit area price is shows a low negative correlation with the age (-0.21). Thus, an older house generally will have low unit area price
- Unit area price shows a moderate positive correlation with both latitude(0.55) and longitude (0.52).Thus, location of the house plays an important role in determining its

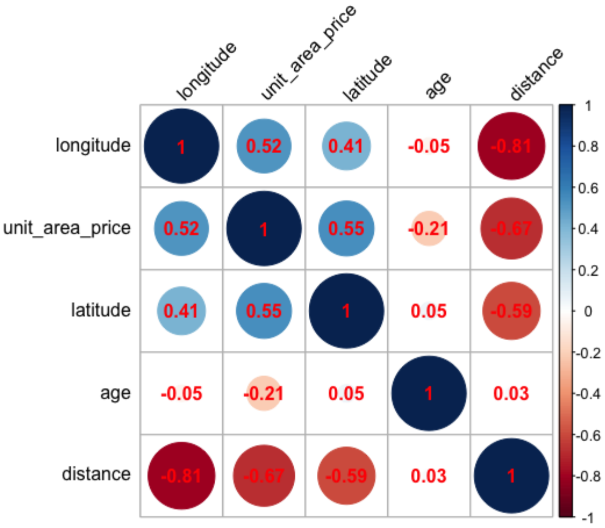


Figure 2.9: Correlation Matrix

Chapter 3: Methods

Train and Test Split

We decided to split the data into train and test, with 80% of the data in the training set. We trained the data on the training set and evaluated the RMSE on the testing set.

3.1 Multiple Linear Regression(MLR)

3.1.1 MLR-1

We started with a simple regression model (MLR1) considering all 7 predictors on training data. This model was able to explain 61.2 % variance in the unit area price.

MLR 1: unit_area_price ∼ txn_date + age + distance + stores + latitude + longitude + txn_month

Training Error : 71.45 (10000 New Taiwan Dollar/Ping)
Testing Error : 64.28 (10000 New Taiwan Dollar/Ping)

Model Diagnostic & Interpretation:

- We noticed that all levels in txn_month had NA values. This was because the transaction month is perfectly correlated with transaction date, leading to a rank deficient matrix. Thus, we decided to drop the transaction month for the future MLR model.
- Model Diagnostics: We conducted model diagnostics to check the assumptions of the linear regression.

  1. Constant Variance: Using Breusch-Pagan test , we found that the assumption of constant variance of residuals is being violated.
  2. Normality: Using Shapiro Wilk test, we found that the assumption of normality of residuals is being violated.
  3. Independence: Using the Durbin-Watson Test, we found that assumption of independence of residuals was not being violated.
  4. Linearity: We did residual vs predictor plot to check this. We found that the variable distance unusual patterns in residual vs distance plot
  5. From the box-cox plot, we identified that the response 'unit_area_price' needed a logarithm transformation which would help us deal with normality and linearity. From the residual vs predictor plot, we concluded that the predictor distance needed logarithmic transformation.

- Unusual observations: As the regression model results are greatly affected by unusual observations, we test for high-leverage points, outliers and influential observations. Observation 221 is an influential observation. The observations 271,114,221,48,127,149 are outliers. They need to be removed and the model should be re-fitted.
- Multicollinearity: Using Variation Inflation (VIF) values, we noticed that none of the predictors had VIF greater than 10. Thus, we concluded that the predictors are not strongly correlated with each other.

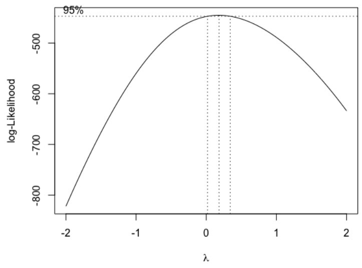


Figure 3.1: MLR-1: Box-cox

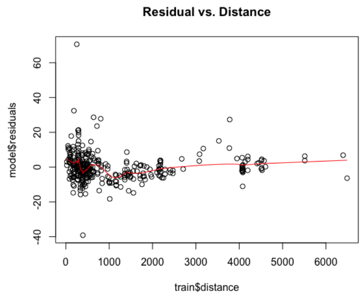


Figure 3.2: MLR-1: Residual vs. Distance

3.1.2 MLR-2

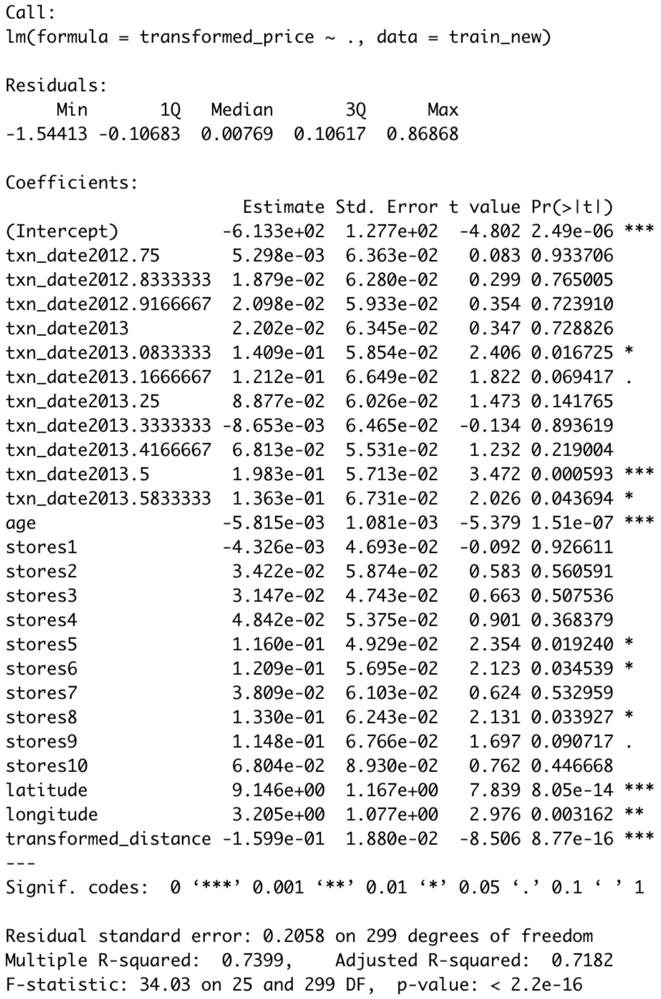
Next we re-fit the model after incorporating inferences from above model i.e. remove outliers, apply logarithmic transformation on the unit area price and distance, drop the 'txn_month' we fitted the linear regression model (MLR 2). This model was able to explain 74% variation in the unit area price.

MLR 2: log(unit_area_price) $\sim$ txn_date + age + log(distance) + stores + latitude +longitude

Training Error : 7.01 (10000 New Taiwan Dollar/Ping)
Testing Error : 7.31 (10000 New Taiwan Dollar/Ping)

Model Diagnostic & Interpretation:

- We again checked for the assumptions of linearity, independence, constance variance in MLR 2 and found that this did not violate these assumptions. However, it violated the assumption of normality.
- We also noticed that all the variables were significant at 5% level. Thus we decided to keep the variables 'txn_date', 'age','stores','latitude','longitude' and 'log(distance)' in the final MLR model.
- The most significant variables are log(distance), latitude and age. Both log(distance) hence, distance and age negatively influence the unit area price with the former having greater affect as can be seen in Figure 3.3. We had concluded the same from EDA.
- Both latitude and longitude are significant predictors. This suggests location play an important role in influencing the unit area price of the house. This was also reflected in Figure 2.8
- All stores (except store1) have a positive affect on the price in comparison to store0(baseline). This suggests that having at least 2 stores increases the unit area price as compared to having 0 stores



```
Call:
lm(formula = transformed_price ~ ., data = train_new)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.54413 -0.10683  0.00769  0.10617  0.86868 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -6.133e+02  1.277e+02  -4.802 2.49e-06 ***
txn_date2012.75      5.298e-03  6.363e-02   0.083 0.933706    
txn_date2012.8333333 1.879e-02  6.280e-02   0.299 0.765005    
txn_date2012.9166667 2.098e-02  5.933e-02   0.354 0.723910    
txn_date2013         2.202e-02  6.345e-02   0.347 0.728826    
txn_date2013.0833333 1.409e-01  5.854e-02   2.406 0.016725 *  
txn_date2013.1666667 1.212e-01  6.649e-02   1.822 0.069417 .  
txn_date2013.25      8.877e-02  6.026e-02   1.473 0.141765    
txn_date2013.3333333 -8.653e-03 6.465e-02  -0.134 0.893619    
txn_date2013.4166667 6.813e-02  5.531e-02   1.232 0.219004    
txn_date2013.5       1.983e-01  5.713e-02   3.472 0.000593 ***
txn_date2013.5833333 1.363e-01  6.731e-02   2.026 0.043694 *  
age                 -5.815e-03  1.081e-03  -5.379 1.51e-07 ***
stores1             -4.326e-03  4.693e-02  -0.092 0.926611    
stores2              3.422e-02  5.874e-02   0.583 0.560591    
stores3              3.147e-02  4.743e-02   0.663 0.507536    
stores4              4.842e-02  5.375e-02   0.901 0.368379    
stores5              1.160e-01  4.929e-02   2.354 0.019240 *  
stores6              1.209e-01  5.695e-02   2.123 0.034539 *  
stores7              3.809e-02  6.103e-02   0.624 0.532959    
stores8              1.330e-01  6.243e-02   2.131 0.033927 *  
stores9              1.148e-01  6.766e-02   1.697 0.090717 .  
stores10             6.804e-02  8.930e-02   0.762 0.446668    
latitude             9.146e+00  1.167e+00   7.839 8.05e-14 ***
longitude            3.205e+00  1.077e+00   2.976 0.003162 ** 
transformed_distance -1.599e-01  1.880e-02  -8.506 8.77e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2058 on 299 degrees of freedom
Multiple R-squared:  0.7399,    Adjusted R-squared:  0.7182 
F-statistic: 34.03 on 25 and 299 DF,  p-value: < 2.2e-16
```

Figure 3.3: MLR-2: Model Summary

3.2 Ridge Regression

Penalised regression models help to overcome the overfitting problem and also handle multi-collinearity in data. Since ridge estimators still depend on the least squares minimization technique, they are susceptible to outliers, much like the OLS estimator.
We used data without outliers and with log transformed ‘distance’ and ’unit area price. We also included variable ‘txn_month’ which we had dropped while MLR-2 since ridge regression is able to handle multicollinearity issues which MLR was incapable to do.

Ridge Regression: ***log(unit_area_price)* ∼ *txn_date* + *age* + *log(distance)* + *stores* + *latitude* +*longitude* + txn_month**

Training Error : 7.92 (10000 New Taiwan Dollar/Ping)
Testing Error : 8.75 (10000 New Taiwan Dollar/Ping)

Model Interpretation:

- We can infer (from figure 3.6 and 3.7) that both age and log(distance) and hence, distance negatively influence the unit area price of the house. We observed the same from the final MLR model
- All stores (except store1) have a positive affect on the price in comparison to store0(baseline). This suggests that having at least 2 stores increases the unit area price as compared to having 0 stores. We observed the same conclusion from the final MLR model
- We can not infer which predictors are significant, as no p-values are given by the ridge model. However, we were able to conclude the significant predictors from the final MLR model
- Testing error is higher in ridge by 1.44 (10000 New Taiwan Dollar/Ping) as compared to the final MLR. This suggests that MLR gives better predictions than the ridge regression in our data.

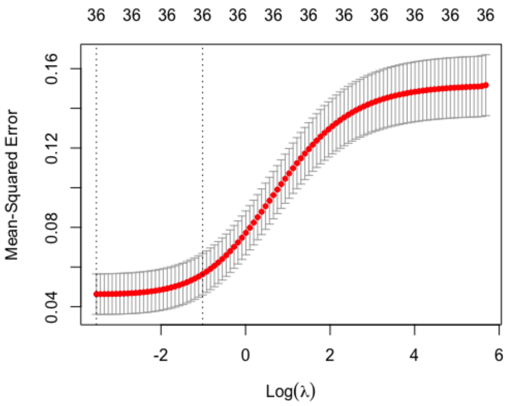


Figure 3.4: Cross Validation vs. log(λ)

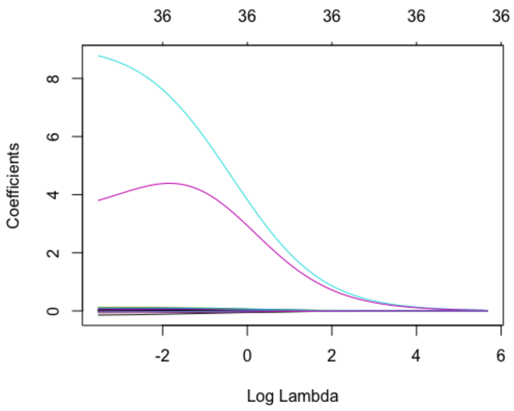
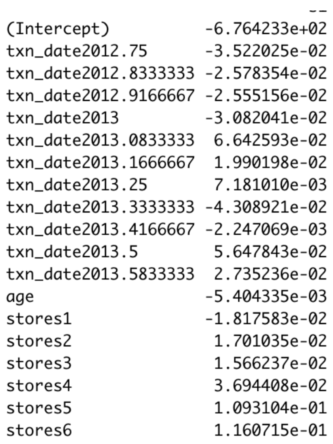
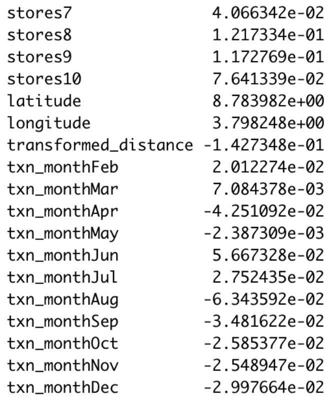


Figure 3.5: Ridge Regression Coefficients as function of λ



```
                                 --
(Intercept)         -6.764233e+02
txn_date2012.75     -3.522025e-02
txn_date2012.8333333 -2.578354e-02
txn_date2012.9166667 -2.555156e-02
txn_date2013        -3.082041e-02
txn_date2013.0833333  6.642593e-02
txn_date2013.1666667  1.990198e-02
txn_date2013.25       7.181010e-03
txn_date2013.3333333 -4.308921e-02
txn_date2013.4166667 -2.247069e-03
txn_date2013.5        5.647843e-02
txn_date2013.5833333  2.735236e-02
age                 -5.404335e-03
stores1             -1.817583e-02
stores2              1.701035e-02
stores3              1.566237e-02
stores4              3.694408e-02
stores5              1.093104e-01
stores6              1.160715e-01
```

Figure 3.6: Ridge Regression Coefficients



```
stores7              4.066342e-02
stores8              1.217334e-01
stores9              1.172769e-01
stores10             7.641339e-02
latitude             8.783982e+00
longitude            3.798248e+00
transformed_distance -1.427348e-01
txn_monthFeb         2.012274e-02
txn_monthMar         7.084378e-03
txn_monthApr        -4.251092e-02
txn_monthMay        -2.387309e-03
txn_monthJun         5.667328e-02
txn_monthJul         2.752435e-02
txn_monthAug        -6.343592e-02
txn_monthSep        -3.481622e-02
txn_monthOct        -2.585377e-02
txn_monthNov        -2.548947e-02
txn_monthDec        -2.997664e-02
```

Figure 3.7: Ridge Regression Coefficients

3.3 Regression Trees

We used regression trees to train the data. Regression trees are robust to outliers and give good interpretations. Firstly, we built a regression tree using all the variables (keeping all of them as untransformed). We fitted a 5 fold cross-validation CART model and pruned the tree by setting the cp (complexity parameter) corresponding to the minimum xerror from the cptable. In case of regression tree-1, the cp value was set to 0.01409231. We observed the following RMSE values using the first decision tree:

Regression Tree: unit_area_price ∼ txn_date , age, distance, stores, latitude, longitude, txn_month

Training Error : 7.29 (10000 New Taiwan Dollar/Ping)
Testing Error : 7.97 (10000 New Taiwan Dollar/Ping)

Model Interpretation:

- Variable 'distance' is used for the first split in the tree and hence is the most important variable in determining the unit area price. We can notice from tree nodes that the avg. price of houses with the distance>=827 is less(25) while for distance < 827 is more(45).
- We observe the last split is being made on the variable 'store' and hence, it is does not help to assess the unit area price as well as other predictors do.
- We notice that 'txn_month' and 'txn_date' are not present as tree nodes, hence they are not important to predict the unit area price.

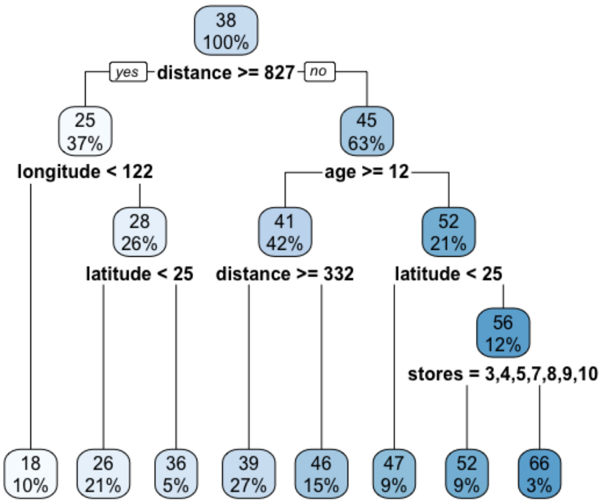


Figure 3.8: Regression Tree Model

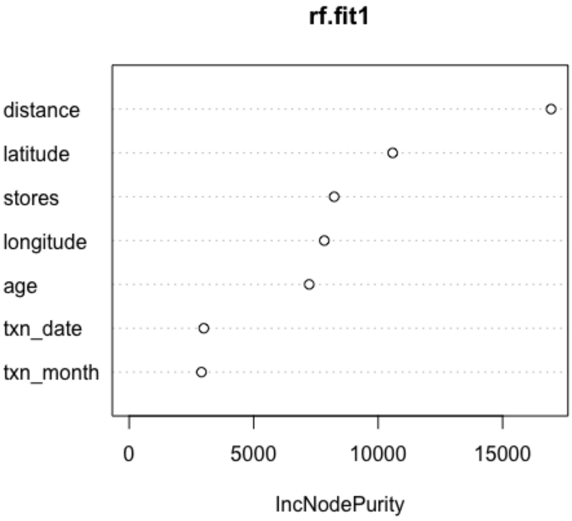


Figure 3.9: Random Forest: Variable Importance

3.4 Random Forest

A random forest algorithm trains many decision trees on various sub-samples of the data (creating a forest) and uses them as a classifier/regressor. It uses averaging to improve the predictive accuracy and control over-fitting. We built a random forest using all the variables. We built the random forest using 50 trees and 2 features at random were considered by the model at each split.

Random Forest: unit_area_price ∼ txn_date , age, distance, stores, latitude, longitude, txn_month

Training Error: 7.82 (10000 New Taiwan Dollar/Ping)
Testing Error: 6.90 (10000 New Taiwan Dollar/Ping)

Model Interpretation:

- Random forest gives us the best results in terms of test errors, so it empirically proves that the ensemble learning outperforms individual tree model.
- From Figure 3.9, we can observe that distance plays a very important role in affecting the unit area price of the houses
- Latitude (location), number of stores, longitude (location) and age of the house are the next most important predictors which influence the unit area price of the house.
- Transaction date and transaction month are least important predictors in the RF. Thus, they do not play an important role in deciding the unit area price as per the random forest model.

Chapter 4: Results and Conclusion

* Testing Errors reported in the table below are in 10000 New Taiwan Dollar/Ping

Model	Testing Errors	Cons	Pros
Multiple Linear Regression	7.31	Sensitive to outliers, multi-collinearity	Great interpretability- gives magnitude, direction of variable affect as well as variable significance in terms of of p-values, Good predictions
Ridge Regression	8.75	Sensitive to outliers, Does not provide any information on variable significance	Handles multicollinearity
Regression Tree	7.97	Leads to overfitting, requires pruning	Great interpretability, robust to outliers
Random Forest	6.90	Limited interpretability	Great predictions, robust to outliers

- Random Forest gives the best testing error of 6.90 (10000 New Taiwan Dollar/Ping) on the test data followed by Linear Regression, Regression Tree and Ridge Regression in order

- Decision trees and Linear regression give best interpretation. One can easily find the average price using the decision tree chart in figure 3.8. Linear regression and ridge regression clearly tells the magnitude and direction of the influence of each variable on price, whereas Random forest only suggests which variable played in an important role in splitting but can't suggest the magnitude and direction of the influence of a predictor on price

- Every model suggested that Distance to the nearest MRT station is the most important predictor that influences the Unit Area Price of the houses. After Distance; location of the house, age of the house and number of convenience stores affect the price with Transaction Date and Transaction Month as the least important predictors that affect the Unit Area price

- The closer the house is to the nearest MRT station, the newer its construction, and greater the number of stores in its living circle, higher will be the unit area price of the house. Houses situated in certain geographical regions have higher prices as compared to houses in other locations.

Our project aimed to predict the price per unit area using the Real Estate Valuation dataset. As it was a regression problem, we decided to use Multiple Linear Regression, Ridge Regression , Regression Tree, and Random Forest models. The best model out of all aforementioned is the Random Forest, since it is an ensemble model, and it outperforms any of the individual models. However, the Random Forest Model is harder to interpret in comparison with the individual models.

We believe that the knowledge which we got from this project could be applied to real industrial problem-solving. It helped us to look into the course material more generally and better understand the practical applications of statistical analysis.