Assignment A2: Estimation

Student Name	**Abhishek Vinod Rajput**		**Student No**	219266543
Problem attempted	**Complex Model 80-100%**	**Simple Model 40-79%**	**Student Id**	avra@deakin.edu.au
Place "Yes" in one only	**Yes**	**?**	***Do not attempt a complex model unless you can complete a simple model first!***	

Executive problem statement

The Business problem which AirbnbAI is trying to solve is predict and evaluate the quality of the Listings provided by them on their website based on several characteristics or attributes. This is achieved using a metric called the overall satisfaction. In other words AirbnbAI aims at predicting the overall satisfaction metric of the property using Rapidminer. The overall customer satisfaction lies in the range 0 to 5 where 0 defines dissatisfied Customer and 5 defines completely satisfied customer. Such customer satisfaction rating are great index for customer to understand about the property since most customer prefer properties with high customer satisfaction and having more number of properties with high customer rating will help Airbnb in keeping their quality high, therefore the need to predict the customer satisfaction for each property becomes of high importance. To predict this metric, a number of features of the property such as the location, price, number of bedrooms, the number of people the property can accommodate are taken into consideration.

Customers are asked to review their stay at the property on a scale of 0 to 5 and the average of these reviews is considered as the overall satisfaction value for the property of the property. It is estimated that 70% of the customers provide review for the property after the stay. The overall satisfaction rating provides Airbnb with a quantifiable measure to check how a property is performing as well as understanding how different attributes characteristics are responded by the customers and therefore Airbnb will be able identify as to which type of property are to be highlighted in what particular area.

This prediction of Overall Satisfaction will help AirbnbAI to understand the factors affecting the customer's perception. This will help in understanding if the customers are price sensitive or is just the service provided considered the important factor for an enjoyable stay or is it a combination of the two. This prediction can help Airbnb make inform strategies for reputation management, public relations and understanding public view points. Therefore saving time and helping them in decision making. This study based on clustering of similar features will help Airbnb understand the reason for satisfaction or dissatisfaction of customers and therefore help understand the services and strategies which have worked over those which have not been successful.

Figure 2 and Figure 3 in Data Preparation section show the prices to be directly affected by the number of bedroom and number of people accommodating in the property. The prices increase consistently as the bedroom increases or the number of people increase.

Figure 4 shows the t overall rating was consistently above 4 for all regions in New York but suddenly dropped after Jan 2020. Airbnb needs to investigate the reasons behind such a sudden change in the customer satisfaction and possibly try to correct all the wrongs. And from Figure 5 it can be inferred that a lot of the rating provided by the customers are good and only receive low ratings when the customers is extremely dissatisfied with the services.

Data Preparation

Attributes	overall_satisfacti...	price
room_type = Private room	0.010	-0.134
room_type = Entire home/apt	0.028	0.169
room_type = Shared room	-0.070	-0.074
borough = Manhattan	0.005	0.108
borough = Queens	-0.033	-0.056
borough = Brooklyn	0.023	-0.078
borough = Staten Island	-0.029	0.004
borough = Bronx	-0.031	-0.029
reviews	-0.025	-0.114
price	0.017	1
Accomodate	-0.070	0.198
Bedroom	-0.002	0.209
room_id	-0.038	0.043
overall_satisfaction	1	0.017

Figure 1 Correlation Matrix

The Corelation Matrix shows a strong positive relationship between price and Bedroom (+0.209) as well as between Price and Accommodate Attribute (+0.198). None of the attribute show significant relationship with Overall satisfaction (less than 0.1) or multicollinearity properties with any other attribute.

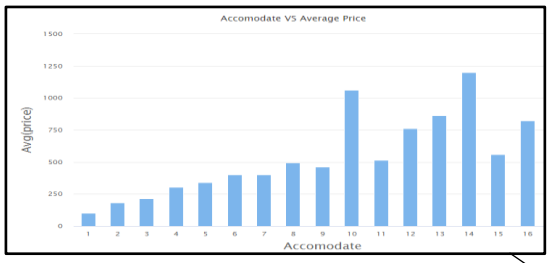

Figure 1 Bar chart for Accommodates vs Price

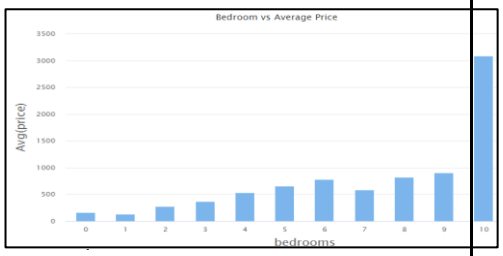

Figure 3 Bar Chart Bedrooms vs Price

The Bar graphs shows the relationship between accommodates and price and Bedrooms vs Average Price. The direct relation Between the Attributes can see in the graph as the prices of the property is directly proportional to the number of people living in the property and the Number of bedrooms in the property.

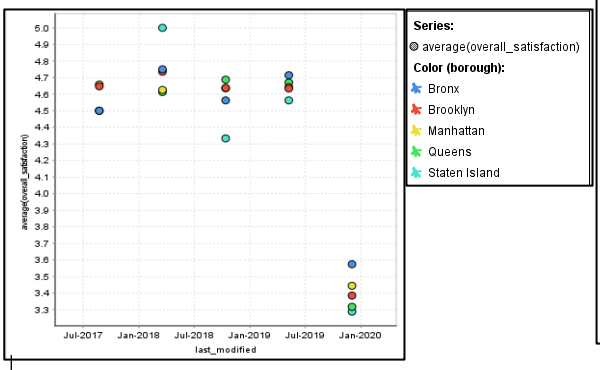

Figure 4 Changes in Overall satisfaction over Time

The Chart shows the overall satisfaction to be varying over time and has gone low in Jan 2020 irrespective of the location.

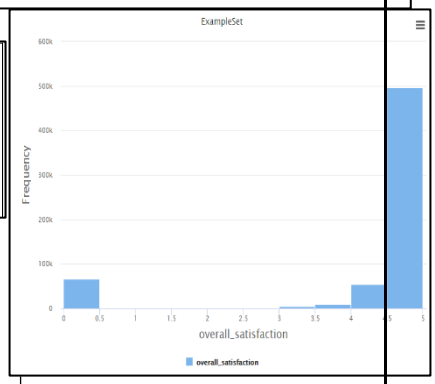

Figure 5 Histogram of overall satisfaction

The histogram shows the customer satisfaction rating to be highly skewed towards high satisfaction rating and negligible amount low rating scores. The average customer rating is 4.168 which is towards the higher side.

room_id	Integer	0	Min 105	Max 17732388	Average 6377426.916
host_id	Integer	280	Min 43	Max 120885348	Average 19136270.848
room_type	Polynominal	318	Least Shared room (26181)	Most Entire home/apt (468253)	Values Entire home/apt (468253), Private r
borough	Polynominal	0	Least Staten Island (4731)	Most Manhattan (443553)	Values Manhattan (443553), Brooklyn (348
neighborhood	Polynominal	0	Least Rossville (1)	Most Williamsburg (77971)	Values Williamsburg (77971), Bedford-Stuy
reviews	Integer	5721	Min 0	Max 414	Average 13.218
accommodates	Integer	37492	Min 1	Max 16	Average 2.732
bedrooms	Real	53195	Min 0	Max 10	Average 1.143
price	Real	0	Min 0	Max 140000	Average 165.018
minstay	Integer	167337	Min 1	Max 444443	Average 3.666
latitude	Real	0	Min 40.500	Max 40.912	Average 40.732
longitude	Real	0	Min -74.241	Max -73.702	Average -73.959
last_modified	Date time	0	Earliest date May 10, 2017 9:56 PM	Latest date Mar 15, 2020 10:11 PM	Duration 1039d 23h 15m 0s
overall_satisfaction	Real	250359	Min 0	Max 5	Average 4.168

Figure 6 Attributes and their characteristics

As seen in Figure 6 most of the data contain missing values. Attributes like accommodates, bedrooms and reviews can be imputed using aggregation function by grouping the data based on room_Id and reviews and taking the median value and replacing the missing value with the median value. Few examples have no median values for the attributes and hence need to be imputed using the KNN model. The main objective of the project is to predict the overall satisfaction hence it needs to assign as label. The model to be used requires the attributes to be converted to numerical hence the attribute neighbourhood needs to eliminated since it has 242 distinct values. The test data does not contain any value for minstay and last modified hence the attribute mainstay and last modified should also be eliminated. The examples with missing room type values should be eliminated.

Executive Solution Statement

In order to keep their quality high understand the market perception of its customers a model was developed for AirbnbAI to automatically estimate the "Customer Satisfaction" of new rental properties (See Figure 24 and Figure 25).

The Identified and segmented the data based on some features (Borough, room type, bedrooms, accommodates). The system suggests the factors like price have no effect on the customer satisfaction i.e. the customer satisfaction is price insensitive and it is the service that the customers are looking forward to. (See Figure 10 and the discussion followed)

- The model suggests the properties private properties in Brooklyn and Manhattan have low rating i.e. below 1.0. Since price is not an important factor, the house owners should pay special attention towards services provided by them preferably towards the cleanliness aspect since cleanliness is the one factors which results in low ratings.
- Entire houses in all locations often receive high ratings (4.1 or more) or moderate ratings (1.1 to 4.0).
- Shared houses do not show any trend and can receive any rating.

Different strategies were applied based on different estimation techniques and compared with each other to identify the best model which gives the most reliable results, as discussed in model development and Model Evaluation section below (Table 2). The model is able to predict the overall satisfaction with 92.6% accuracy. The other 7.4% data can be accurately explained using different factors which not present in the data.

Of the 803 examples the Model estimates around 550 examples with highly customer satisfaction and around 50 examples with low customer satisfaction and the rest were identified as moderate satisfaction see Figure 26.

AirbnbAI could work on understanding what factors affect the ratings most and possibly understand the reasons for the low ratings. Airbnb could ask its property owners to provide all the entertainment facilities, toiletries and possibly aim to provide any support if needed by the guests. This will help Airbnb in achieving greater revenue at the same time ensuring the quality of the listing remains good.

Data Preparation and Exploration

The data after cleaning and removing all missing and unwanted attributes is then used to check and remove any anomaly if present in the data and then clustering is done on the data to find any relationship between attributes and how similar items can be grouped together. The Data is normalised using range transform, ranging from 0 to 5 and passed to KNN global Anomaly operator in order to detect Outliers. The KNN global anomaly calculates outlier score based on K nearest neighbours implementation using Mixed Measure Technique. Selecting the value of K as 10 the model was run and the outlier's scores exceeding 1.0 were flagged and eliminated from the data so that they don't affect the clustering and the further estimation and the outlier Graph was visualised using SVD.

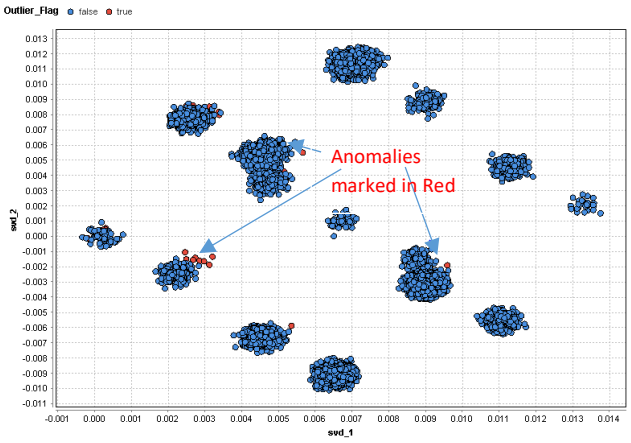

Figure 7 Anomalies based SVD1 and SVD2

The Figure 7 shows the outliers in the data and we can see that most outliers are outside of clusters. A 3D scatter plot can be generated in order to check that most of the data re not anomalies. After eliminating all the anomalies the data is then sent to the clustering operator which uses a KNN model to group all similar examples form the data set.

A cluster Model was generated in order segment the data and an optimiser was used in order to understand the best k value which can be used clustering operator giving minimum Davis Bouldin value and sum of squares result. The results of the optimiser parameter are as follows.

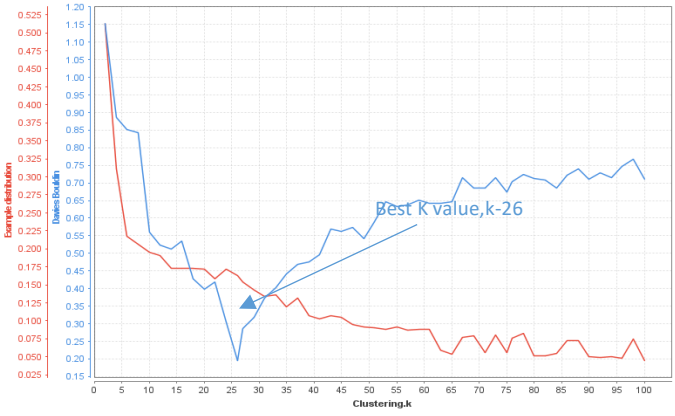

Figure 8 Clustering Optimiser Result

The best results for clustering were obtained when K=26 with Davis Bouldin value = 0.196, i.e. 26 number of unique clusters were identified by the model.

On conducting the K means clustering model for K =26 the results obtained were as shown in Figure 8.

Figure 9 explains how the examples are spread in all the clusters. The Cluster Centroid Chart for all the clusters si shown in the Figure 11.

Cluster Model

```
Cluster 0: 724 items
Cluster 1: 5232 items
Cluster 2: 3048 items
Cluster 3: 2098 items
Cluster 4: 145 items
Cluster 5: 463 items
Cluster 6: 700 items
Cluster 7: 302 items
Cluster 8: 136 items
Cluster 9: 762 items
Cluster 10: 371 items
Cluster 11: 108 items
Cluster 12: 126 items
Cluster 13: 32 items
Cluster 14: 27 items
Cluster 15: 28 items
Cluster 16: 22 items
Cluster 17: 417 items
Cluster 18: 124 items
Cluster 19: 141 items
Cluster 20: 27 items
Cluster 21: 37 items
Cluster 22: 75 items
Cluster 23: 3159 items
Cluster 24: 16 items
Cluster 25: 27 items
Total number of items: 18347
```

Figure 9 Cluster Result

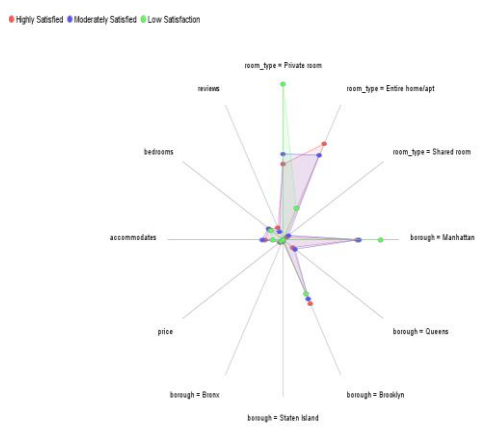

Figure 10 Cluster Web Chart

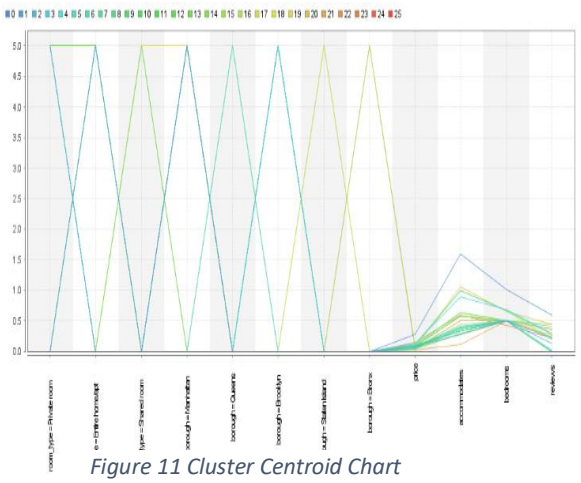

Figure 11 Cluster Centroid Chart

A web chart as shown in Figure 10. was generated in order to understand the relationship between the attributes and to understand the segmentation of the data. On visualising the web chart it was known that properties in boroughs like Bronx, Staten Island and queens have no relationship to the customer satisfaction similarly room type shared room do not contribute majorly towards Overall satisfaction and hence there is no trend. Prices also did not play a part when customer satisfaction was considered. The rest of the attributes were used and the web chart is as shown in Figure. Analysing the web chart shows that most of the low satisfaction rating were from Private rooms majorly from Manhattan. While for customers to rate a property as moderate there were no special features and could be located anywhere or can be of any room type. Most positive customer Satisfaction were received from the Entire House room type and these houses were located in Brooklyn. Attributes like bedrooms and accommodates although play some role but do not have a distinction between them and could not be used for segmentation analysis.

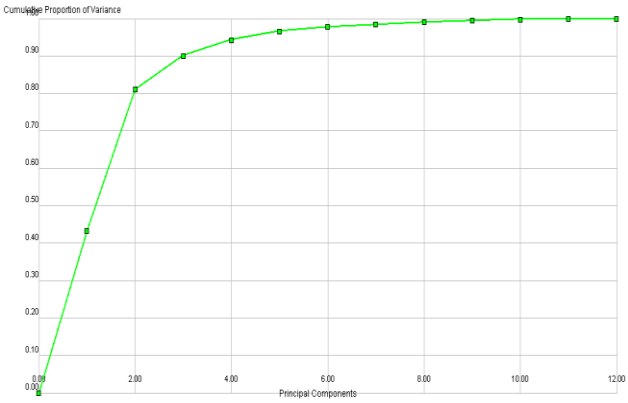

Figure 12 PCA Cumulative

The PCA cumulative Graph consisting of all the all Eigen values for each PCA is shown in Figure 12. The PCA graph for the cluster shows that PC1 , PC2 and PC3 are able to explain around 92% of variation in data. The rest PCs explain very less variance.

Support for answer to Question B:

The customer satisfaction ratings of new property or new listings in New York can be segmented and the factors responsible for the ratings can be measured using the above model created. The information about the location and type can of house can help in predicting the rating the house can get, which is independent of the price of the listing.

Model Development

In order to clean the Data the following process was implemented in Rapidminer. The Data consists of 882,000 rows with some rows containing errors and missing values.

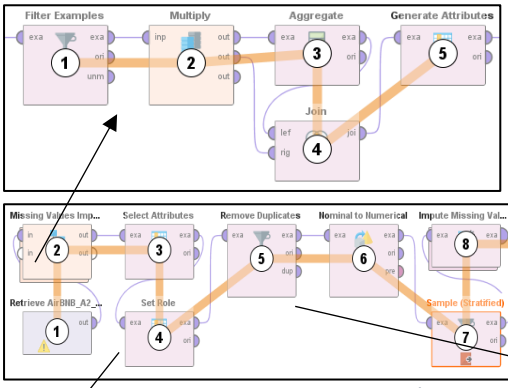

Figure 13 Pre-processing in RapidMiner

From Figure 13 , all rows with missing overall satisfaction columns are removed so that we have clean data to create a model to predict the same. At the same time there are 308 rows with missing room type which less than 0.01% of the total data hence are removed from the dataset. The data contains missing bedrooms and accommodates values for some rows but contains these values for the same roomID in some other rows. Hence the Data is grouped on RoomID and Room type and the most occurring values for bedrooms and accommodates are taken for the RoomID group. Two new Attributes are created to replace the missing bedrooms and accommodates with the mode value of the group and this new attributes shall be used for further analysys

Neighborhood is not selected as predictor since we are estimating using linear regression and such nominal data with 242 unique values would not any meaningful information to the regression. Minstay is excluded since the test data which is used for testing does not have the attribute. Host_ID and last_Modified are excluded as well. Rest all other attributes are selected.

AirbnbAI wants to predict the overall satisfaction and hence is set as label. The room_ID is set as ID and should not be used for any prediction Duplicate values form the data set is removed based on Room_id, Room_type , price and reviews so that all possible unique combinations are present. The data is used for regression hence nominal data is converted to numerical using dummy coding. A sample of 10000 data is used for the model and all the missing reviews are imputed using the KNN model.

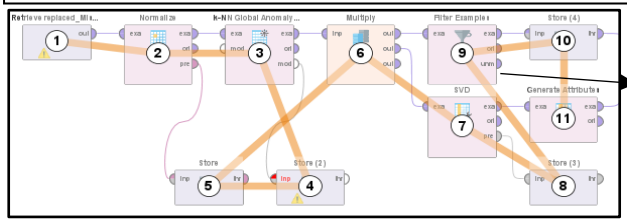

Figure 14 RM process for Anomaly Detection

The data is then used to check for anomalies, The normalize operator is used to normalize the attributes using z Transform and KNN Global anomaly operator is used to find the outliers using outlier scores and examples with outlier score above 1 are flagged as outlier and filtered. The outliers are visualized using SVD operators by reducing its dimensions

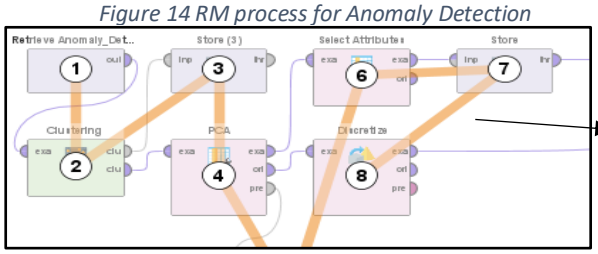

Figure 15 RM Process for Clustering

The Anomaly detected and filtered data is then passed through the clustering operator after find its optimal k value (k=26) and the output is then passed through PCA to visualize and reduce its dimensionality and in order to visualize the clustering result discretize operator is used to discretize Overall satisfaction into highly satisfied, highly disappointed and moderate rating.

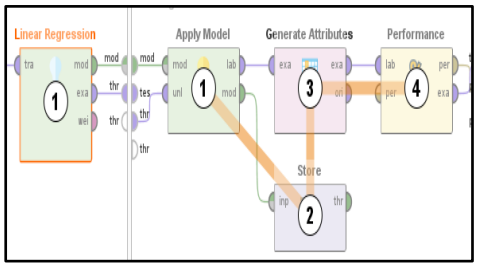

Figure 16 Linear Regression Model

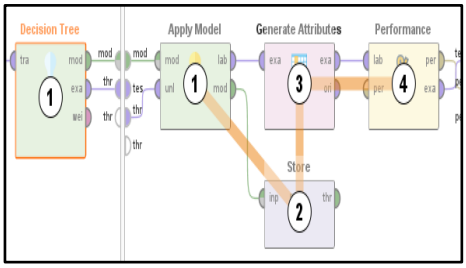

Figure 17 Decision Tree Model

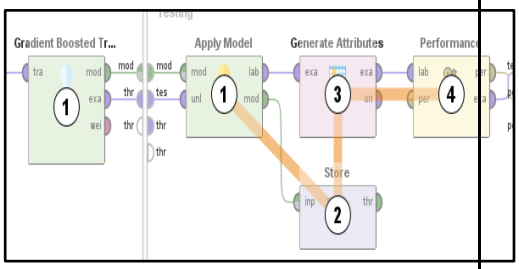

Figure 18 Gradient Boosted Trees Model

The PCA data was used to generate the linear regression model, Decision tree model and Gradient Boosted tree was used as an ensemble model inside a cross validation operator. The linear regression model develops a relationship with each model to find the result and finding the approximate coefficient to be multiplied with the variables in order to estimate the Label and hence proper ridge value needs to select which is done using the optimizer grid parameter. The model was optimized to find the ridge which gives the minimum Root Mean Squared Error, Mean absolute error and maximum correlation value. The ridge was set to 0 accordingly. The decision tree uses a splitting rule to divide the values into trees to estimate a label and hence proper tree depth needs to select. The Ensemble model Gradient Boosting Relies on previous results to get the best possible results for the next model in order to minimize the error by progressively improving the results. The values of number of trees and depth need to be optimized.

Model Evaluation and Optimisation

The models generated using for Regression, decision tree and Gradient Boosted Trees were evaluated by splitting the data into test and train and Cross validation in order to reduce the RMSE value. The cross validation trains the model better when the number of folds is set at 10 and hence all the three models have been used with cross validation in order to estimate the overall satisfaction.

Model Optimisation:

All three models were passed through optimiser parameter in order to understand the optimal parameters which give lowest RMSE, MAE values at the same time increasing the correlation and R2 values.

In order to optimise the linear regression model we need to select the ridge value which gives the best results. Similarly the depth of tree is optimised in decision tree and for the gradient boosted tree the depth as well as the number of trees is optimised and the results for the optimisation are as follows.

iteration	Linear ...	root... ↑	absolut...	correlation	squared_correlation
1	0	1.432	1.234	0.639	0.408
2	5	1.432	1.234	0.639	0.408
3	10	1.432	1.235	0.639	0.408
4	15	1.432	1.235	0.639	0.408
5	20	1.432	1.235	0.639	0.408

Figure 19 Linear Regression Optimiser Results

iteration	Decisio...	root_me... ↑	absolute_error	correlation	squared_corr...
6	51	0.459	0.194	0.969	0.938
7	61	0.459	0.194	0.969	0.938
4	31	0.459	0.194	0.969	0.938
9	80	0.459	0.194	0.969	0.938

Figure 20 Decision Tree Optimiser Results

iteration	Gradien...	Gradien...	root... ↑	absolut...	correlat...	square...
66	100	51	0.830	0.756	0.962	0.926
99	100	80	0.830	0.756	0.962	0.926
33	100	21	0.833	0.759	0.961	0.924
65	90	51	0.886	0.821	0.962	0.926
98	90	80	0.886	0.821	0.962	0.926

Figure 21 Gradient Boosted Optimiser Results

Model	Optimising Parameter	Optimised Value
Regression	Ridge	0
Decision Tree	Maximal Depth	51
Gradient Boosted	Maximal Depth, Number of tress	51, 100

Table 1 Model Optimiser Results

The model was run on the optimsed value using vross validation with K fold 10 and the results were tabulate as follows.

Model	RMSE	MAE	Corelation	R2
Regression	1.437 +/- 0.010	1.239 +/- 0.009	0.636 +/- 0.007	0.404 +/- 0.009
Descision Tree	0.498 +/- 0.079	0.213 +/- 0.044	0.962 +/- 0.014	0.926 +/- 0.026
Gradient Boosted	0.943 +/- 0.012	0.837 +/- 0.008	0.927 +/- 0.006	0.926 +/- 0.026

Table 2 Model Optimiser Results

The above table clearly shows the Descision tree model to be the best model to predict the customer satisfaction.

Model Evaluation:

The Descision Tree model was run and the values were estimated for the training data and the residuals and absolute residuals were calculated.

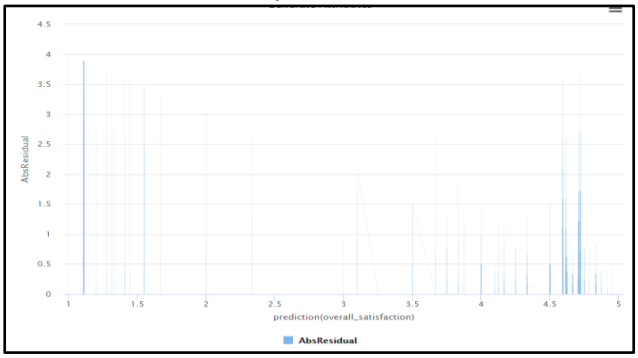

Figure 21 Residual Plot

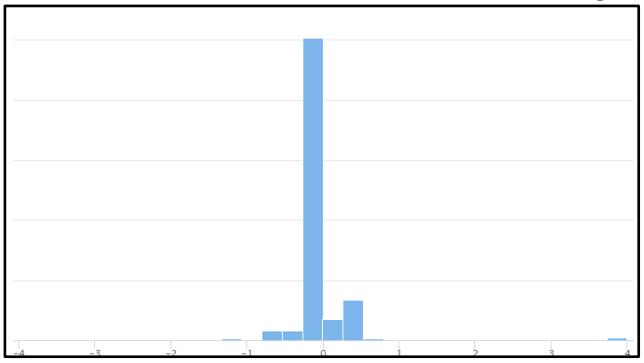

Figure 22 Histogram of Residual

Figure 21 and Figure 22 show the residual plots to be scattered all over and does not show any trends or pattern meaning the model is fit for prediction at the same time the histogram for residual is normally distributed meaning the model is able to predict both the high and the low satisfaction equally. The Predicted overall satisfaction and the given overall satisfaction are almost similar as seen in the Figure 23.

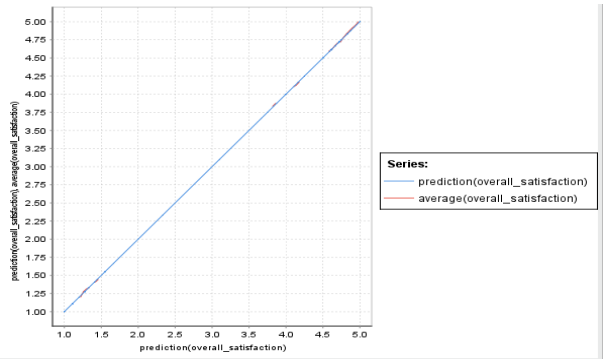

Figure 23 Predicted vs Observed Plot

Since the decision tree model has the least RMSE value and the highest correlation value, The AirbnbAI team can use the model to predict and estimate the overall satisfaction rating for their newer listings within the range of 0.498 +/- 0.079 with a predictive power of 92.6% given that information about the price,Room type, borough,accomodates and bedrooms are given i.e the data is similar to the training data. The models could be stored and retrieved later to be operated on newly listed properties to predict the customer ratings for the same.

Model Application

In order to apply the developed process to new data, care needs to taken so the new data is similar the data used to create the model. The models could be stored and retrieved later to be operated on newly listed properties to predict the customer ratings for the same.

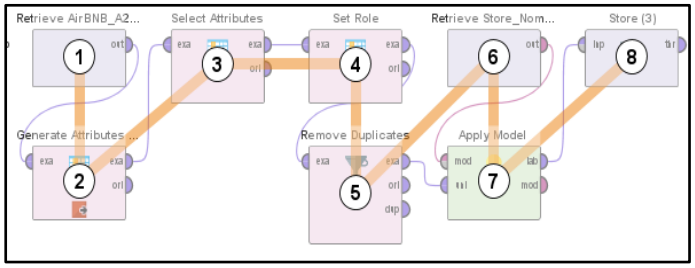

Figure 24 Test Data Preprocessing

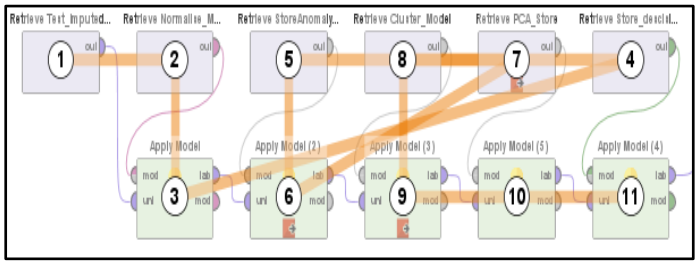

Figure 25 Test Data Model Preparation

The Model for Test can be generated as shown in Figure 24 and Figure 25. The Test data needs to preprocessed in a similar way, but anew attribute needs to be generated "overall satisfaction" with a dummy value since the test data does not contain the attribute. The newly generated atrribute is set as label. The stored nominal to numerical model is retived and applied. Similarly Normalise model, KNN global anomaly model, clustering model, PCA model and the descision tree model are retrived and apllied. This process will give the predictions for the customer satisfaction.

An applying the above steps on the new data the results obtained were as follows

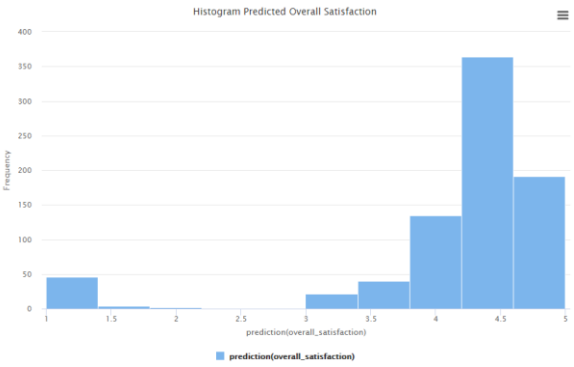

Figure 26 Histogram of Test Data Prediction

The results obtained were very similar to the training data where most of the reviews were above 4.0 and most of the low ratignwere rated as 1.0. Out of the 803 data, 50 properties were rates as 1.0 or less and 555 properties were rated 4.2 or above with an average rating of 4.2. The reason for which is similar to as discussed in section 2, people mostly review the property when they like it and is most rated high and rarely the properties are rated the bad.

The results are 92.6% accurate when all the predictors are present but inconsistent when any of the predictors is absent.

Support to answer Question C :

Using this model, the AirbnbAI can predict the overall satisfaction of new listings in New York 0.498 +/- 0.079 with a predictive power of 92.6% if all the attributes are present.