Advanced Predictive Models for Business

SUBMITTED BY: DHAVALA SHARMA

Executive problem statement

Airbnb is the one of top leading companies in providing rental accommodation to its customers through via internet and mobile app bookings. They provide rooms, houses, holidays homes and etc across the globe. Airbnb has approached our firm again to form a process for them so that they can predict the overall satisfaction of the customers of new rentals who are renting the New York city's Airbnb in the range of 0 to 5.The objective of creating this process and report is to study the trend of overall satisfaction for the Airbnb properties. And later provide insights and predictions about the overall satisfaction level of new rentals.

There are several factors that affect the overall satisfaction level for the Airbnb properties such as room type, neighbourhood, bedrooms, price, stay period, etc. The whole report can be split into three sections which are trend analysis, Grouping of Airbnb properties based on similar characteristics and overall satisfaction and predicting the satisfaction level of new Airbnb Rentals.

Trend analysis

In this part of report, we are going to study the trends of overall satisfaction over the period of three years. The aim is to see if there is any variation in overall satisfaction of properties over the period. On the same grounds we are also going study price of the rentals over the period to see if there is any variation or fluctuations in the cost of Airbnb Rentals.

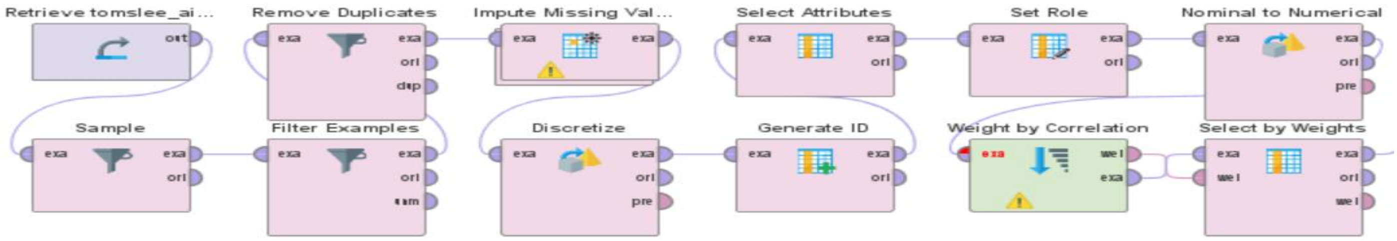
Grouping of Airbnb properties based on similar characteristics and overall satisfaction

The aim of this part of the report is to study the overall satisfaction level of different Airbnb properties based on several factors such room type which is later sub divided into single room, entire home or apartment. Other factors are neighbourhood of property, price of the rental, etc. Based on this groups are created which have their own unique characteristics. Satisfaction level of new properties which have similar characteristics to the properties belonging to a group can then be estimated.

Predicting the satisfaction level of new Airbnb Rentals

In this section we try to achieve to a better level of accuracy in estimating the likely level of satisfaction for new rentals by creating different models. Model with accuracy is then used to obtain prediction by applying it test data. Performance measures plays the key role in determining the efficiency of the predictions.

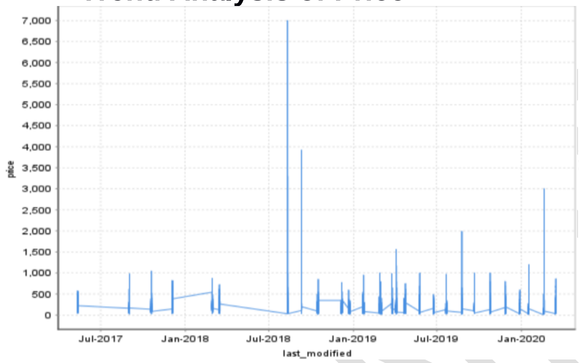
Data preparation

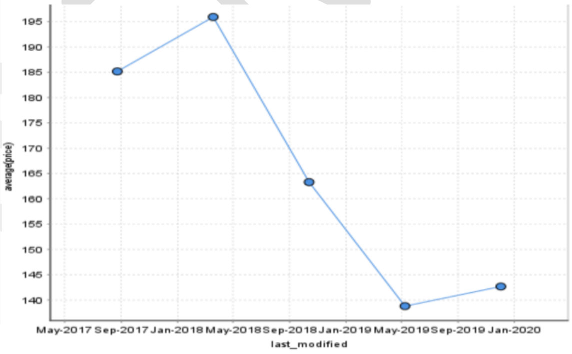


Airbnb provided us with 882,000 listings of data. I have taken a random sample of 8000 listings using Sample Operator. Filter example operator is used to filter out all the missing values from overall satisfaction. Multiple values for same room_ id was found. Thus, those entries that are common in last_modified and room_id are duplicates and are removed. Attributes have missing values. There are 4 attributes like that who have missing values. We are using Impute Missing Value operator in which k is 8, which means that it uses 8 nearest neighbours to predict the missing value in the attribute. Discretize Operator is used to change the overall satisfaction data to categorical data. Five categories are created. Generate Id operator is used to generate a unique Id. A unique id is created because there are multiple entries found for both room_id and host_id. In select attributes operator we consider all the attributes except room_id and host_id. In Set Role function, overall satisfaction is our label, price target role is prediction, last_modified target role is date and id as id. Weight by Correlation attribute is used to compute the weight of attributes in reference to our label attribute. Top 6 predictors which we have derived from weight by correlation are taken into account.

Visualisation and Analysis of Trends in Price and Overall Satisfaction

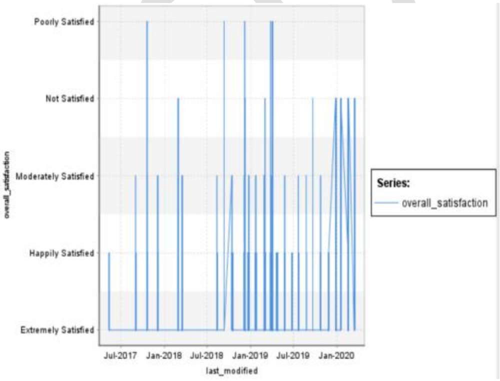
Trend Analysis of Price

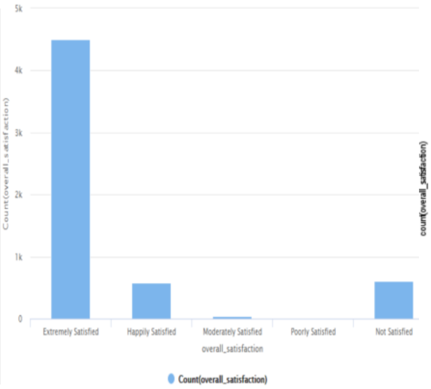


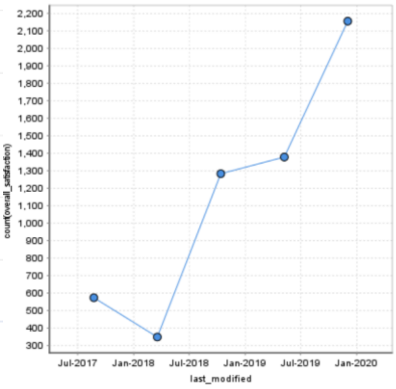


From the graph on the left we could figure out that there is a cyclical trend in price as there are multiple peaks. Apart from certain peak in July 2018, probable because the data contains outliers. The graph on right shows that there is a decrease in price, or I should there is a downward trend in cost from April 2018.

Trend Analysis of Overall Satisfaction







The graph on the extreme right shows that the count for overall satisfaction has increased. The graph in the middle shows that most of the values are in extremely satisfied category and least number can be found in poorly satisfied. From the first graph it can be seen that there are some outliers and from Jan 2020 overall satisfaction if quite often seen in the not satisfied range.

Executive solution statement

Multiple insights were created as there were several models were created and the model with best efficiency and accuracy was selected and its predictions are closer to reality in practical sense. Three models were created in order to generate the most accurate predictions. Important predictors which are going to affect the model were identified. Room type and borough are two key predictors. Based on those key predictors clustering was performed. Basically, groups were created based on similar characteristic properties and then the new rental property overall satisfaction was predicted by comparing the property characteristic with those property groups which matches their characteristics. One of the quality insights which was found that most of the Airbnb properties have their satisfaction level more than four. There is a very small number of count of people who have given lower level of satisfaction to Airbnb properties.

Process created while doing clustering analysis is used as foundation model while doing test on a sample test data provided. The test data contains the data of nearly 800 rental properties. The best model identified. It had an efficiency of 66.1%

The firm can use this information very effectively. As those properties which are getting lower level of satisfaction can be visited again and improvements should be made keeping in mind the overall satisfaction level of properties greater than four. They can compare properties and improve the quality of property lacking behind. It's not only a benefit for the customer but also its god for house owners as they can make improvements based on important feedback and increase the rating for their property.
The results of the clustering are:

- Private rooms in Manhattan and Brooklyn have more rental properties in medium satisfaction range.
- Rental properties in Queens and Brooklyn having medium number of rooms and less count of entire homes/apt type rental properties have lower level of satisfaction level.
- High satisfaction rate can be found in other room types except private room type rentals. They can be found in Manhattan and in very small region of Brooklyn.

Airbnb should publish this information that most of its customers are having satisfaction level of more than four on their booking and interaction platforms such as mobile apps and website. Most of the customers of Airbnb are students, travellers and back packers who are very keen in knowing the rating given by previous customers. If they show that the previous customers have given them good rating, they could attract a large audience that is vulnerable and can be attracted by positive reviews.

Airbnb can use this information in filtering out the properties which are not giving any kind of satisfaction to its customers. Properties and rentals having good feedback and overall satisfaction level can be promoted up the in the list, so that customers can be attracted more towards it. Properties having lower level of satisfaction level can be filtered out. Improvements can be made on them.

Airbnb can use our insights in order to find out the characteristics of a property which is going to affect the satisfaction level of its customers the most. Key characteristics should be identified so that they can be monitored and regularly audited. In order to maintain higher satisfaction level, those key characteristics should be taken care off.

Airbnb can use our insights in order to predict the satisfaction level of new rental properties by comparing it with those properties which have similar characteristics and have been in usage with the company. In short people have been reviewing it and feedbacks are provided. Those feedbacks can be used in predicting the satisfaction level of new properties. This gives the firm an early advantage to make amendments in the property to in order to make the experience of its customers more enjoyable.

Data Preparation and Exploration

Airbnb has provided us with the data sent of 882,000 records. The data set contains 14 different attributes which are in real, integer, date and polynomial type. Missing vales are found in 7 different unique attributes. The data set provided is uploaded in RapidMiner using Read CSV. After reading it I have stored it using the store function.

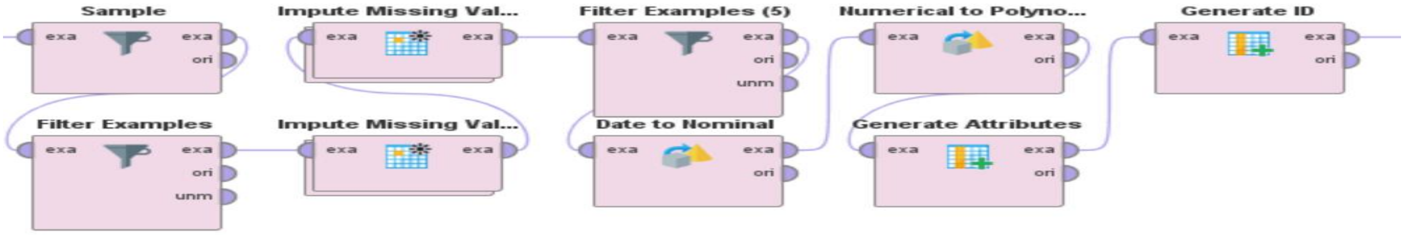


Figure showing Subprocess

A sample of 8000 is taken to create the model. Sampling make visualisation easier. I have used stratified sampling for higher precision. Filter Example Operator is used to filter out the missing vales in host_id,room_type and overall_satisfaction. Missing values found in the attributes effects the results of the model. We perform clustering based on overall satisfaction, therefore there should be no missing value. After that comes Impute Missing Value operator where k=8. It means that it predicts the missing values by looking at thy eight nearest neighbours. The first operator imputes the missing values of accommodates, mainstay and reviews. The second operator for imputing missing values is used for predicting real values in bedroom. Date was changed to nominal using Date to nominal operator. During our analysis it was found that room_id was getting repeated in some rows. To overcome this problem, Generate Id operator is used. A distinctive ID is created. It is created by combing room_id with the data modified. Generate attribute operator is used to in order to create a new attribute. In order to generate a new attribute, we changed date and numerical type to nominal type.

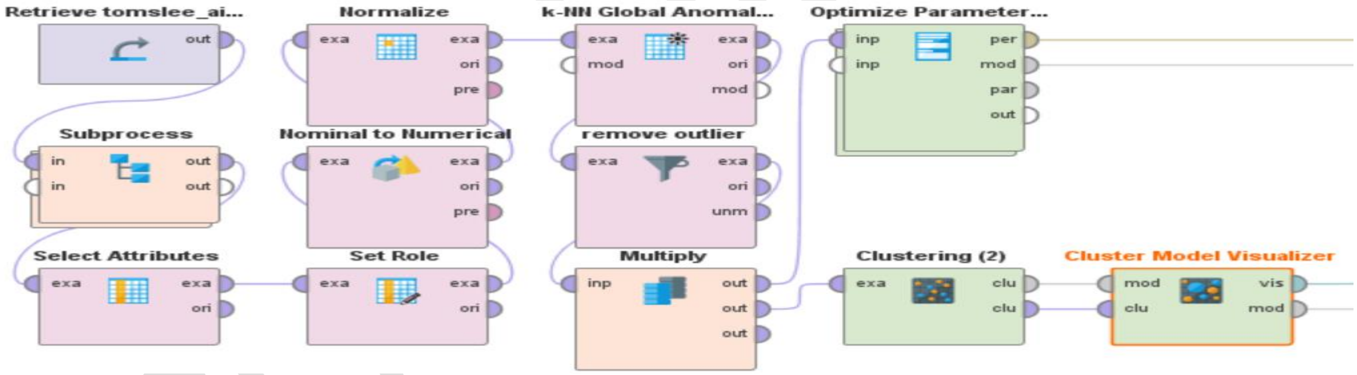
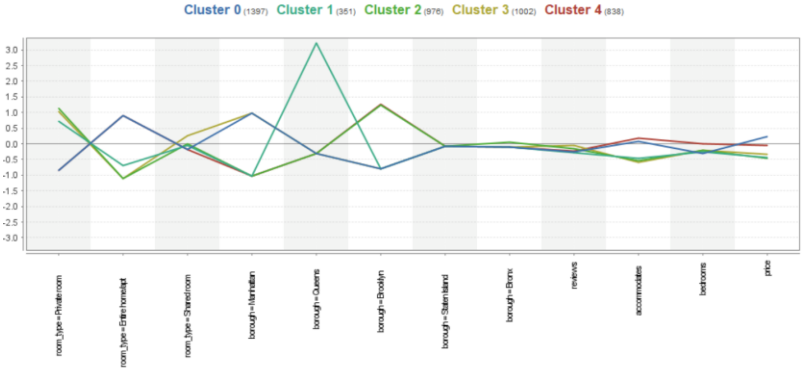


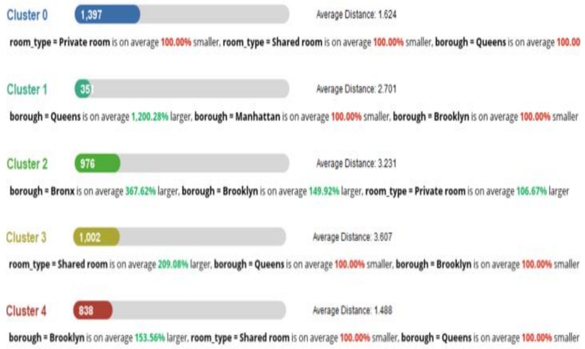
Figure: Cluster model Generation and visualisation

After subprocess we move on to Select Attributes operator. Here we include all the attribute important for clustering except 6 attributes which are host_id, latitude, longitude, mainstay, neighbourhood, and room_id. Set Role function is used after that where we define our label, which is overall satisfaction. Id's target tole is id and last_modified target role is date. In order to apply dummy coding on room type and borough, so that we can identify that whether it's private room, entire house or share accommodation, we use Nominal to Numerical operator. After that we use Normalize Operator, which uses Z-transformation to normalize every attribute except id and overall_satisfaction. Now we use k-NN Global Anomaly Score to identify and remove outliers. The operator uses 10 nearest neighbours to calculate the average distance. A higher score means that it an outlier. Remove outlier operator comes into action after identification of outliers. It removes all the outliers who have a higher score value. In remove outlier operator we set condition class to custom filter and then we add a filter saying that all the values(outliers) greater than 0.4 are filtered out.

After that we apply "Optimise Parameters" operator, so that we can apply different values of k and obtain the best k to be used in clustering model visualizer. We use Davies Bouldin as the performance measure to measure cluster distance performance and efficiency. Davies Bouldin should be close to zero. Ginicoefficient is used to measure the item distribution efficiency. Ginicoefficient close to 1 is generally preferred

Cluster Visualization




Cluster 0 1,397 Average Distance: 1.624

room_type = Private room is on average 100.00% smaller, **room_type = Shared room** is on average 100.00% smaller, **borough = Queens** is on average 100.00

Cluster 1 351 Average Distance: 2.701

borough = Queens is on average 1,200.28% larger, **borough = Manhattan** is on average 100.00% smaller, **borough = Brooklyn** is on average 100.00% smaller

Cluster 2 976 Average Distance: 3.231

borough = Bronx is on average 367.62% larger, **borough = Brooklyn** is on average 149.92% larger, **room_type = Private room** is on average 106.67% larger

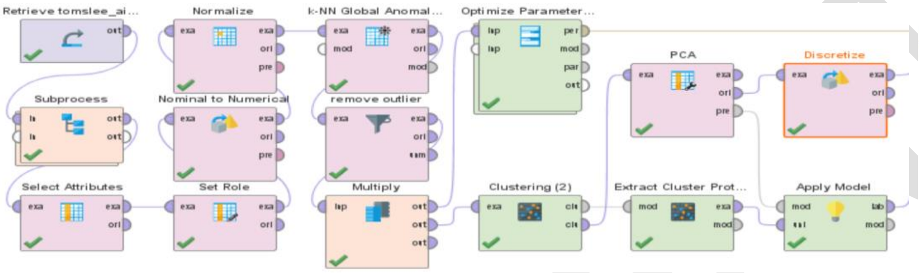
Cluster 3 1,002 Average Distance: 3.607

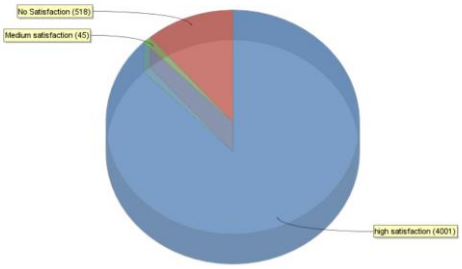
room_type = Shared room is on average 209.08% larger, **borough = Queens** is on average 100.00% smaller, **borough = Brooklyn** is on average 100.00% smaller

Cluster 4 838 Average Distance: 1.488

borough = Brooklyn is on average 153.56% larger, **room_type = Shared room** is on average 100.00% smaller, **borough = Queens** is on average 100.00% smaller

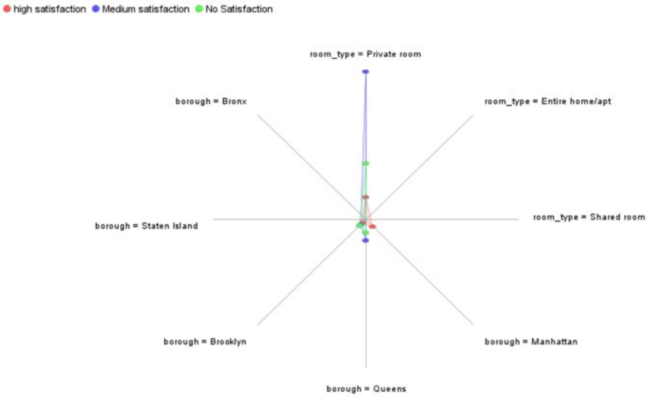
The best value of k, which we have obtained is 5. Davies Bouldin is -0.775. Gini coefficient is also close to 1. Five clusters are formed. They are made by using numerical attributes. Those attributes are dummy coded variable (borough and room type), price, reviews, bedrooms and accommodates. Each cluster has its own property. Each of them has its own distribution as well.





Figures showing clustering model with PCA operator and visualisation of output through pie chart and web chart.

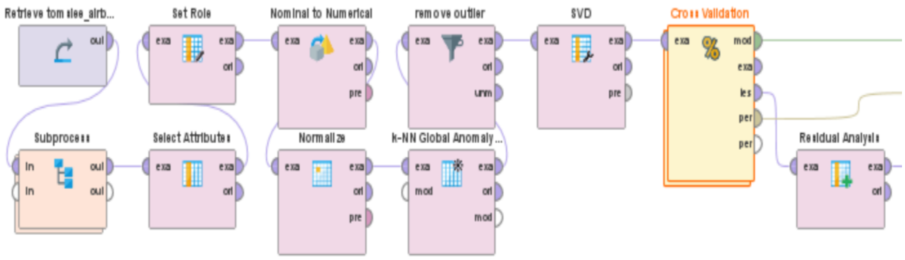
Visualisation is made possible using PCA operator. Principal Component Analysis operator is used for cluster visualisation. variance threshold in PCA is 0.95. Benefit of PCA is that it deletes any multi-collinearity in data. We also study if there is any overlapping of data. Extract Cluster Prototypes operator is used to extract the properties of cluster. Discretise operator is used to create three classes based on satisfaction level. It is important to create classes as it is essential to compare the classes based on overall satisfaction.



Following findings were generated from the pie and web chart.

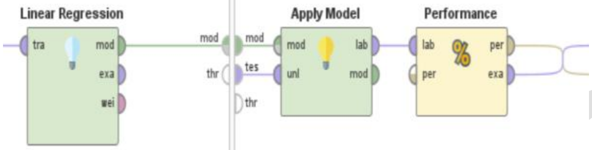
1) There are 4001 values in high satisfaction, 518 values in no satisfaction and 45 in medium satisfaction.
2) Rental Properties with large number private rooms have medium satisfaction level. Most of the values belong to queens.
3) Rental properties with very fewer private rooms have high level of satisfaction. It does not have an exact distribution but somewhat equal (visualisation). Values lies in Manhattan and Brooklyn.
4) Lower level of satisfaction can be found in those properties which have medium number of private rooms. Most of the values lie between Queens and Brooklyn.

Model Development



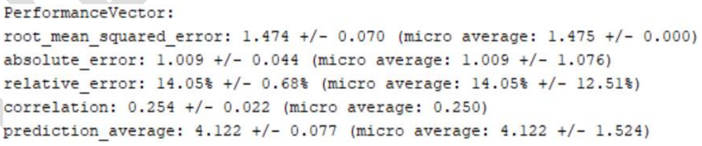
Airbnb wants to know about the overall satisfaction level of its new rental properties. So, in order to estimate and predict the overall satisfaction level of new rental properties we need to create an estimation model using the dataset (tomslee_airbnb_nyc_train) provided to us. For visualization we have done Residual Analysis. Initially I have applied a linear regression model and decision tree model in order to estimate the prediction for overall satisfaction.

In select attributes operator all the important attributes are identified and selected. Then we go to set role operator where we define our label as overall satisfaction, take id's target role as id and last_modified target role is date. Dummy coding is used to transform all the categorical variables to integers. This step is performed because linear regression can only work with numerical values. Therefore, a transformation was needed within the data set for linear regression and anomaly detection to work.in Normalize operator we normalise the values using Z-transformation method. K-NN Global Anomaly operator and remove outlier operator is used to identify the outlier and then for removing them. Singular Value Decomposition is because when we apply dummy coding, it increases the number of attributes in the dataset. SVD creates unique and independent vectors. SVD also reduces the data dimensionality. After that Cross-Validation operator is used. it optimises the performance of linear regression model. The performance measures is root mean squared error, absolute error, relative error and correlation. Due to curiosity I have also included some other performance measures. Residual analysis is done to generate residual and absolute residual values. Residual value is difference in between actual value and predicted value. Its basically predicts an in individual error value between actual value and the predicted value.

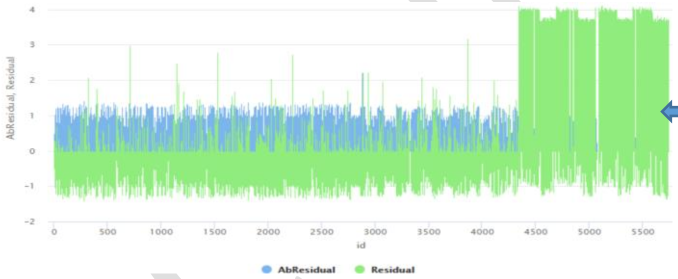


Linear regression model is applied. Feature selection is greedy with min tolerance of 0.05. I have used greedy as it takes the best of all predictors in estimation.

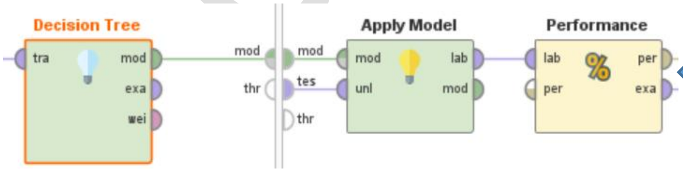
Correlation value should be close to 1 whereas we have received 0.254. efficiency of the model is low. Relative error is close to 14% and root mean squared error is 1.474.



```
PerformanceVector:
root_mean_squared_error: 1.474 +/- 0.070 (micro average: 1.475 +/- 0.000)
absolute_error: 1.009 +/- 0.044 (micro average: 1.009 +/- 1.076)
relative_error: 14.05% +/- 0.68% (micro average: 14.05% +/- 12.51%)
correlation: 0.254 +/- 0.022 (micro average: 0.250)
prediction_average: 4.122 +/- 0.077 (micro average: 4.122 +/- 1.524)
```

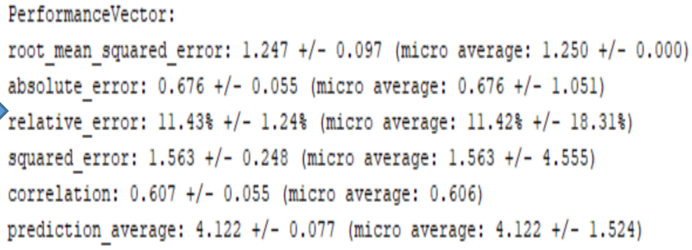


The Graph on the left show's distribution of residuals and absolute residuals. The values on right are away from zero. One fourth of the vales in the dataset are reason for such distribution. Because of this the model's efficiency is low. This can be one of the top reasons that we are not going to select this model.



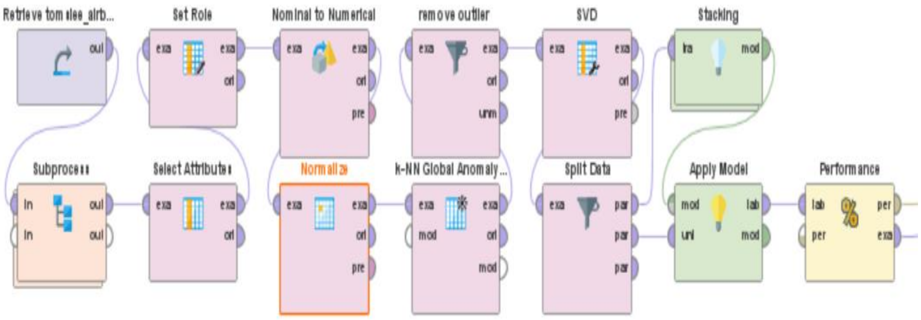
Decision tree model is created and applied. least square approach is adopted. Maximum depth is 10 and pre pruning is applied. Model is not over trained. Pre pruning is done in order to avoid formation of pure tree.

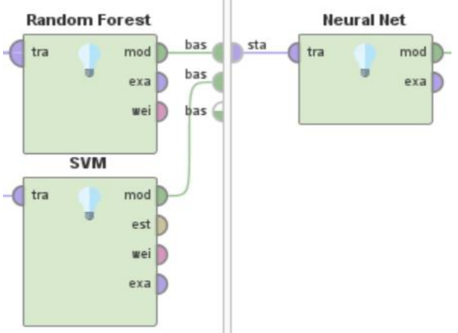
In comparison with linear regression model, decision tree correlation is 0.607 which is closer to 1 and better then linear regression. Relative error is 3% less then linear regression at 11.43%. Root mean squared value is 1.247. decision tree model is better and more efficient.



```
PerformanceVector:
root_mean_squared_error: 1.247 +/- 0.097 (micro average: 1.250 +/- 0.000)
absolute_error: 0.676 +/- 0.055 (micro average: 0.676 +/- 1.051)
relative_error: 11.43% +/- 1.24% (micro average: 11.42% +/- 18.31%)
squared_error: 1.563 +/- 0.248 (micro average: 1.563 +/- 4.555)
correlation: 0.607 +/- 0.055 (micro average: 0.606)
prediction_average: 4.122 +/- 0.077 (micro average: 4.122 +/- 1.524)
```

Model Evaluation and Optimisation





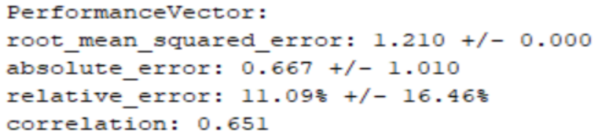
Above figure shows the stacking model. Above figure shows the stacking operator

Now comes evolution and optimisation. Now we need to optimise our model. To optimise our model, we can run multiply models at the same time. Stacking operator is applied for application of multiple models at the same time. Stacking is one of the most important operators in my view. Stacking is an ensemble method. It creates results and predicts after creating a combination of models. Initially all the models generate results on seventy percent of the data. Once the optimised model is crested, it Is tested on the thirty percent of the data set aside.

In the current scenario I have applied three models in stacking for model evolution and optimisation.

1) Random forest: In random forest we use least square method in order to estimate the predictions. Prepruning is applied. Number of trees have been limited to 40. Maximal depth is 10. We apply pre pruning to reduce the complexity. Model is being tested on a small set of data many times to generate the best results.
2) Support Vector: dot kernel type method is applied. It uses java algorithm for generating fast and efficient results.it creates a hyperplane or a set of hyperplanes. All points are plotted on a single plane.
3) Neural Net: in order to optimise the model more efficiently, we use another model on top of those model which acts as base models and have been already applied on the data set. The parameters specify the number of cycles to 200. The operator manipulates the weights of each connection in order to reduce error. The process is repeated, the number of times we specify N, which in this case is 200.

Performance vector of the stacking Model

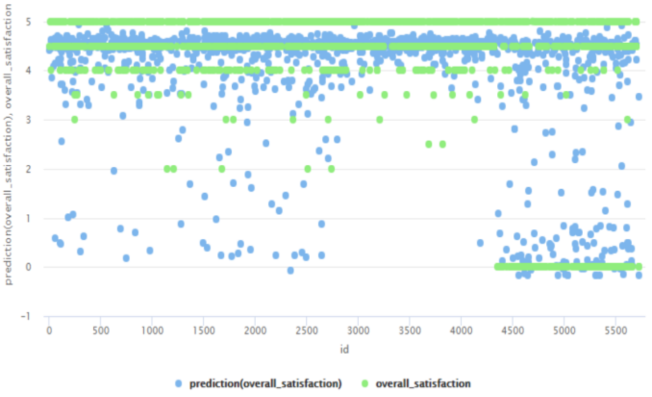


```
PerformanceVector:
root_mean_squared_error: 1.210 +/- 0.000
absolute_error: 0.667 +/- 1.010
relative_error: 11.09% +/- 16.46%
correlation: 0.651
```

Correlation for the new model is 0.651 Relative error is around 11.09%.
Absolute error is 0.667 and root mean squared for the combination of all the three model is 1.201

If we compare the performances of the decision tree model, linear regression model and stacking model the best results are generated by stacking model. The best results are generated by stacking model with an efficiency of 65 % which is the highest among all. Model optimisation is done in order to generate the best results with the best model identified. Model Optimisation is done on stacking model.

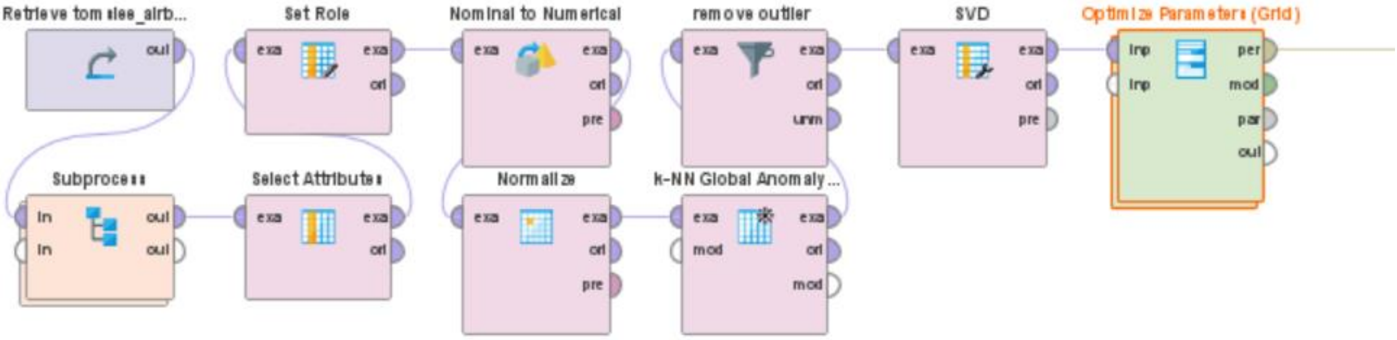
Visualisation of the Model



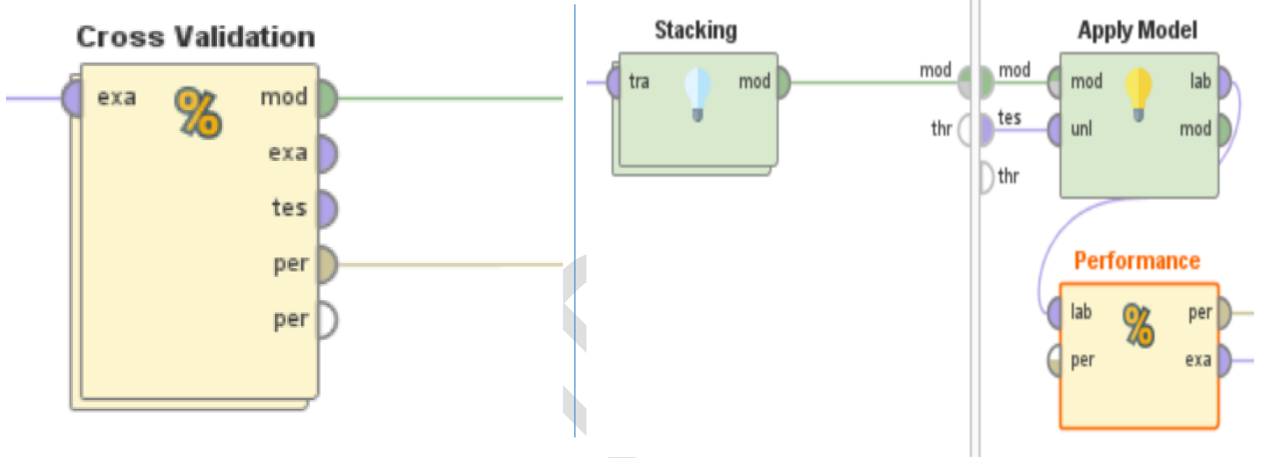
The graph on the left shows the data values of predicted (overall satisfaction) and overall satisfaction.
The graph will give us the actual difference between the actual values of overall satisfaction and the predicted value.

Model Optimisation

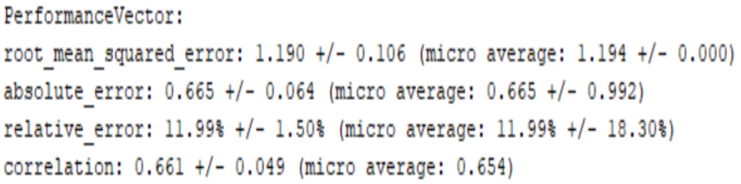
From earlier analyses, we discovered that out our stacking model gives us the best result if we compare it with our linear model and decision tree model.



Thea above predicted model is created and applied using RapidMiner on the dataset provided. In order to obtain the best estimated model, cross validation was optimised with the help of optimise parameters (Grid) operator. As the most efficient and accurate results were shown by stacking model, therefore we optimise the stacking model for estimating overall satisfaction.



Visualusation of Performnace Vector



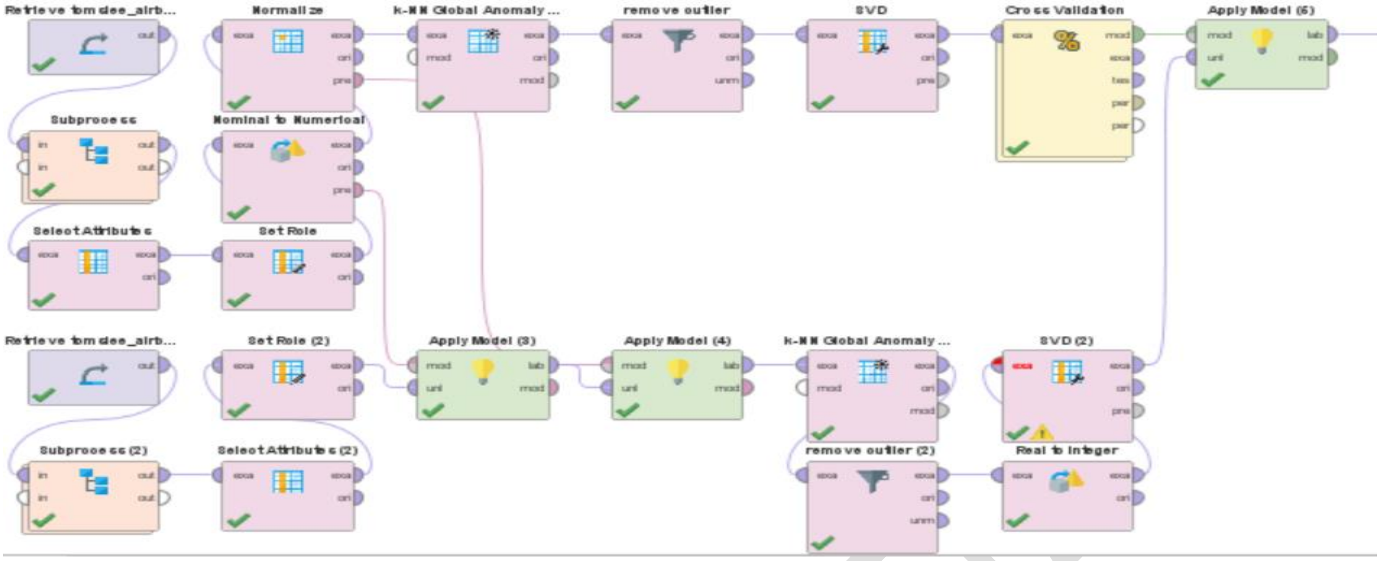
```
PerformanceVector:
root_mean_squared_error: 1.190 +/- 0.106 (micro average: 1.194 +/- 0.000)
absolute_error: 0.665 +/- 0.064 (micro average: 0.665 +/- 0.992)
relative_error: 11.99% +/- 1.50% (micro average: 11.99% +/- 18.30%)
correlation: 0.661 +/- 0.049 (micro average: 0.654)
```

The model has the highest correlation at 0.661. The model has an efficiency of 66.1% which is the highest among all.

Comparison table for all the Models Created

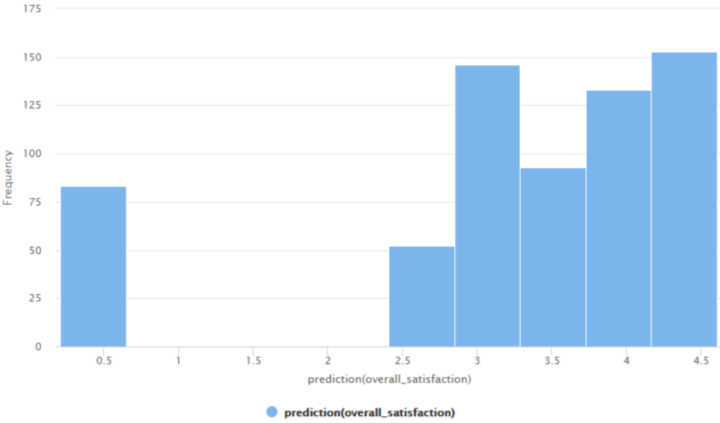
Model	RMSE	MAE	CORRELATION
Decision Tree Model- cross validation	1.247	0.676	0.607
Linear Regression Model - Cross Validation	1.474	1.009	0.254
Ensemble Model - Hold Out Method	1.201	0.667	0.651
Ensemble Model - Cross Validation	1.190	.665	**0.661**

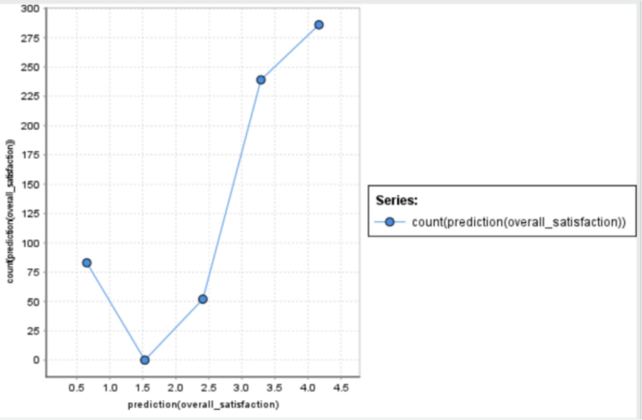
Model Application



The most efficient model which we have received in the past stage is now going to be tested on test data (tomslee_airbnb_nyc_test) provided to us. Initial cleaning of data is done in sub process operator. Non-essential attributes such as mainstay are not included in the analyses as all the values were missing. In select attribute operator we select all the important attributes. In set role operator we take id target role as id. Apply model operator is used. Dummy encoded values are passed, dummy coding works on the data values passed first. Apply model operator is applied again similarly normalised values are passed this time. Our next step is to identify outliers and remove they by using K-NN Global Anomaly operator and remove outlier operator. Then the value type is changed to integer using real to integer operator. Singular Value Decomposition (SVD) Operator is used to reduce the dimensionality of the data set. Cross validation operator is used to apply the model obtained in previous section using stacking Operator.

Visualisation of results and predicted values





- In total 660 vales are predicted.
- From the above the graphs it can been seen very clearly that most the rental properties have good level of satisfaction or a high level of satisfaction.
- From the graph on left it can be noted that there no rental properties that have satisfaction level of 1 or 2, which is good, because it clearly states that that some properties are bad and should be re furbished. It can be verified from the graph on the left as at rating 1.5 the count of overall satisfaction level is 0.
- High satisfaction level is achieved by a very large number of rental properties. It can be verified from the graph on the right as well. Certain peaks can be noticed at satisfaction level greater then 3 and at satisfaction level greater than 4.

  Near about 85% of data lies under higher satisfaction range. The model used is the best one with the efficiency of 66.1.% which is highest among all the models.