Assignment A2: Estimation

Student Name	YASH AHUJA		Student No	219608443
Problem attempted	Complex Model 80-100%	Simple Model 40-79%	Student Id	ahujaya
Place "Yes" in one only	YES	?	*Do not attempt a complex model unless you can complete a simple model first!*	

Partial Submission	Exceptional	Very Good	Good	Acceptable	Improve	Unaccept.
Exec Problem	5	4	3	2	1	0
Data Exploration	10	8	6	4	2	0

Final Submission	Exceptional	Very Good	Good	Acceptable	Improve	Unaccept.
Exec Solution	5	4	3	2	1	0
Data Preparation	20	16	12	8	4	0
Model Development	20	16	12	8	4	0
Model Evaluation	20	16	12	8	4	0
Model Application	20	16	12	8	4	0

Brief Comments

Read these notes

These and the following notes are trying to help you!
Read the rubric on how the report content is going to be assessed!
Your partial submission may not be perfect but has to reflect a genuine effort.
We expect your partial submission sections to be improved for the final submission.
We will not look at your partial submission until we mark the final submission.
We will assess the final submission and its mark stands.
However, we will deduct marks if the quality of the partial submission is poor.

Do not attempt a complex model unless you can complete a simple model first!
If you cannot formulate a complex problem, you will not get extra points for other complex criteria.
Use the font already used in the template, i.e. Arial 10 (and not MyTiniestFont 2).
If any submission aspects could only be determined by running the process, the marks will be severely reduced.

Note: If it is not in this report, it does not exist and does not get marked!

So, we will not check your RapidMiner scripts to check anything that was missing from the report.
Any part which carries points but is missing in the report gets zero marks.
We expect consistency between the report and RapidMiner scripts, so...

Note: Anything reported that cannot be substantiated
by RapidMiner scripts will be marked as zero.

It means that we will check the RapidMiner scripts when in doubt or even just curious.

Total

0 to 100

Executive problem statement (one page)

Problem Statement

To gain insights into the New York City rental properties and determine if there are any trends in price and customer satisfaction level. We also need to identify which kind of rentals receive what type of satisfaction level. Moreover, we need to predict the likely satisfaction level of the new rentals.

Question A:

After analysing the records, we found that the average customer satisfaction level does not differ across geo-location with all the borough groups such as Manhattan, Brooklyn, Bronx, Staten Island and Queens having approximately equal satisfaction level (see figure 1). Interestingly, the average customer satisfaction level is same across all the rental types as well. However, the average price differs significantly across these borough groups (see figure 3). The average price for Manhattan location is higher than all other locations for all types of rentals including entire house, private room and shared room. We can see that average price does not have significant impact on customer satisfaction level. Hence, an analysis is required to determine the factors which govern the customer satisfaction level so that Airbnb can improve their services for the current and as well as future rental projects in New York City.

We need to determine the properties of the rentals with low and high satisfaction level which can help us predict the satisfaction level of the new rentals. Also, we need to determine the properties of groups of rentals with different customer satisfaction levels. The insights about customer satisfaction level can be determined by analysing various other properties such as as number of bedrooms, minimum nights stayed, the number of occupants allowed in a rental, rental type, rental geo-location, its neighbourhood and borough group. It will help the top management of New York City Airbnb in optimising their services. These insights will help them to improve services in those areas by identifying the properties on which high satisfaction level depends. Moreover, it will also help them understand the needs of their clients and customers so that they can take necessary actions to meet them. The predicted satisfaction level of the new rentals will give them an edge by knowing the customer satisfaction levels ahead of time and the management can make changes or implement new procedures accordingly. At the end, it will benefit the clients and customers of the company by having improved services in those areas or rentals with low satisfaction levels. And these insights will also help in the future projects of the company.

Data preparation (one page)

Label Attribute: Overall Customer Satisfaction level as label as it is the attribute with discrete values which is being predicted as we want to determine the likely satisfaction level of the new rentals.

Predictors: All the attributes which help in the prediction of the label attribute are called predictors. After selecting numerical attributes and setting customer satisfaction as label and using correlation matrix, we found the weights of the numerical attributes. Higher the weight, the attribute is more likely to be relevant in predicting the label. Since, host id and room id have lowest weights, they are removed. Final attributes selected include top 7 numerical attributes from figure 2 along with nominal attributes such as room type, borough group, neighbourhood group.

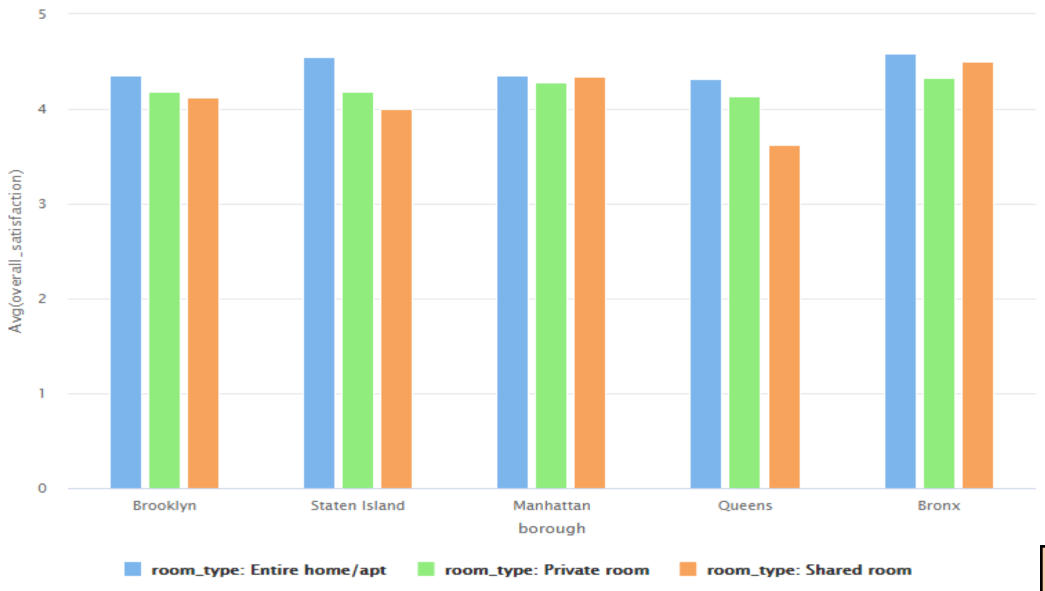
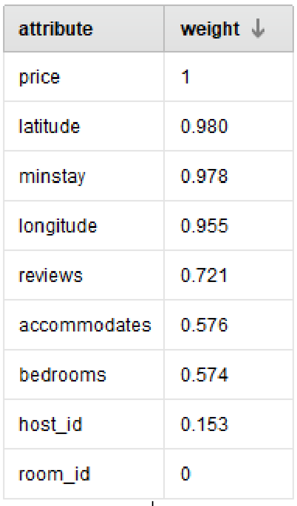


Figure 1 Trends in Customer Satisfaction



attribute	weight ↓
price	1
latitude	0.980
minstay	0.978
longitude	0.955
reviews	0.721
accommodates	0.576
bedrooms	0.574
host_id	0.153
room_id	0

Figure 2 Weights of attributes

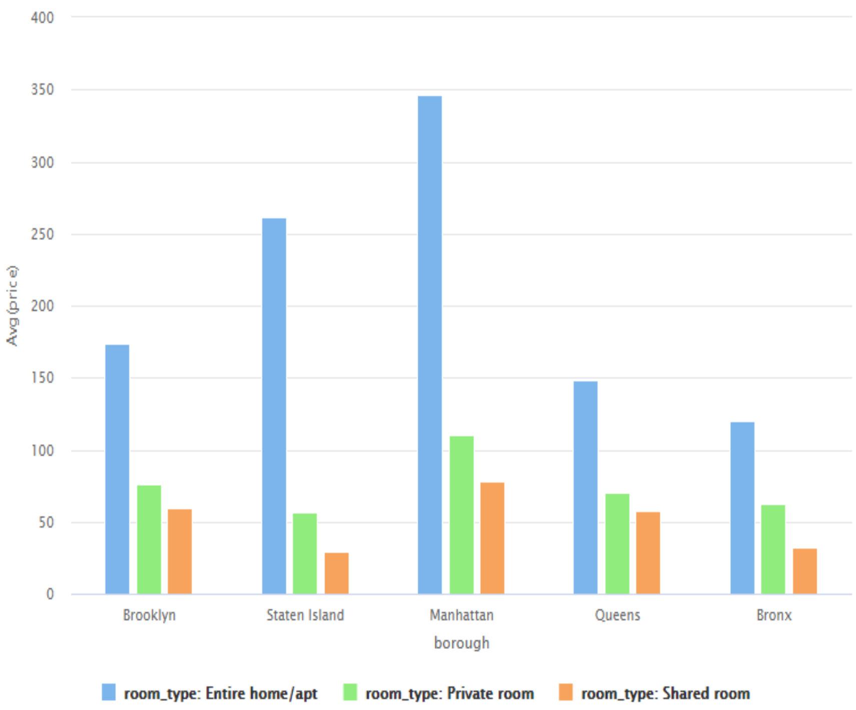
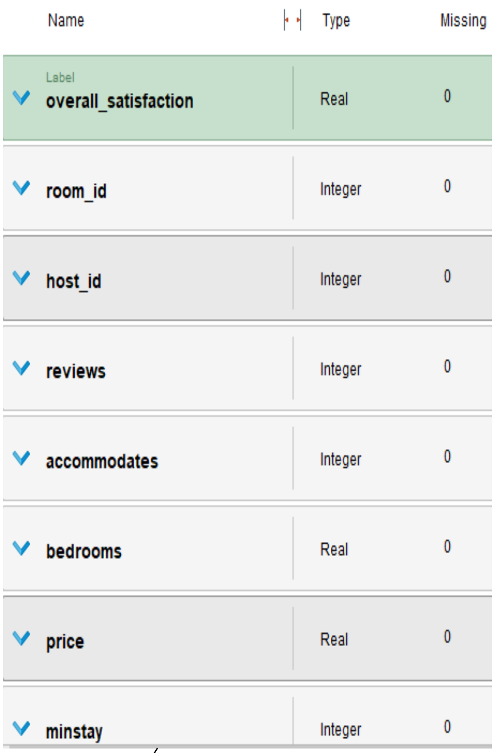


Figure 3 Trends in Price



Name	Type	Missing
Label overall_satisfaction	Real	0
room_id	Integer	0
host_id	Integer	0
reviews	Integer	0
accommodates	Integer	0
bedrooms	Real	0
price	Real	0
minstay	Integer	0

Figure 4 Missing values Imputed by using k-NN imputation model and duplicates are not identified in the example set by Remove Duplicates Operator

Executive solution statement (one page)

Question B: After analysing the 882,000 listings of Airbnb NYC provided by Airbnb AI, I have come to conclusion that the overall customer satisfaction is approximately equal and high for majority of the listings (see figure 5). After, dividing the whole group of listings into five groups, I found that all the other properties of rentals such as number of bedrooms, minimum nights stayed, the number of occupants allowed in a rental, rental type, rental geo-location, its neighbourhood and borough group do not have significant impact on the overall customer satisfaction level. As we can see that four out of five groups have same customer satisfaction level, even though the groups differentiate significantly in various properties such as price, number of occupants allowed, room type and borough group (see figure 6). As we can see, the price of group 0 rentals is on average 55.17% lower in contrast to the price of group 2 rentals which is on average 55.80% higher (see figure 6). But both the groups (0 and 2) have high customer satisfaction level (see figure 6). Moreover, the number of occupants allowed for group 3 is on average 44.92% lower and it is on average 45.38% higher for group 4 rentals (see figure 6) However, both the groups of rentals, group 3 and 4 have high customer satisfaction level (see figure 6).

Only one group, group 1 has low customer satisfaction level, which is on average 99.75% lower (see figure 6). Moreover, it is interesting that the number of reviews in this group is also on average 94.91% lower (see figure 6), which proves my claim that majority of the customers give high ratings to the rentals, but since the number of reviews for the group is missing, it is showing low satisfaction level, which might not be the actual case.

Question C: The customer satisfaction level for the new rentals is coming out to be approximately 4.5 for majority of rentals (see figure 20). The predicted customer satisfaction level for the new rentals is high. Moreover, the customer satisfaction level across all the borough groups is also approximately equal (see figure 20).

Recommendations: With the analysis, it is evident that properties such as number of bedrooms, minimum nights stayed, the number of occupants allowed in a rental, rental type, rental geo-location, its neighbourhood and borough group do not have significant impact on the overall customer satisfaction level. This is most likely because majority of customers usually give high ratings to rentals as they might be busy or just wanted to be free with the feedback process while checking out of the rental. Hence, Airbnb NYC need to implement some new strategies for getting honest feedback from customers in order to get a correct overview of the performance of their rentals. Strategies include online feedback survey which customer can fill as per their convenience. As a result, Airbnb can make improvements in areas where they have genuinely low customer satisfaction level. This will benefit both the customers as well as the Airbnb NYC as satisfied customers will lead to higher revenues and profits in the future.

Data Preparation and Exploration (first page of a two-page section)

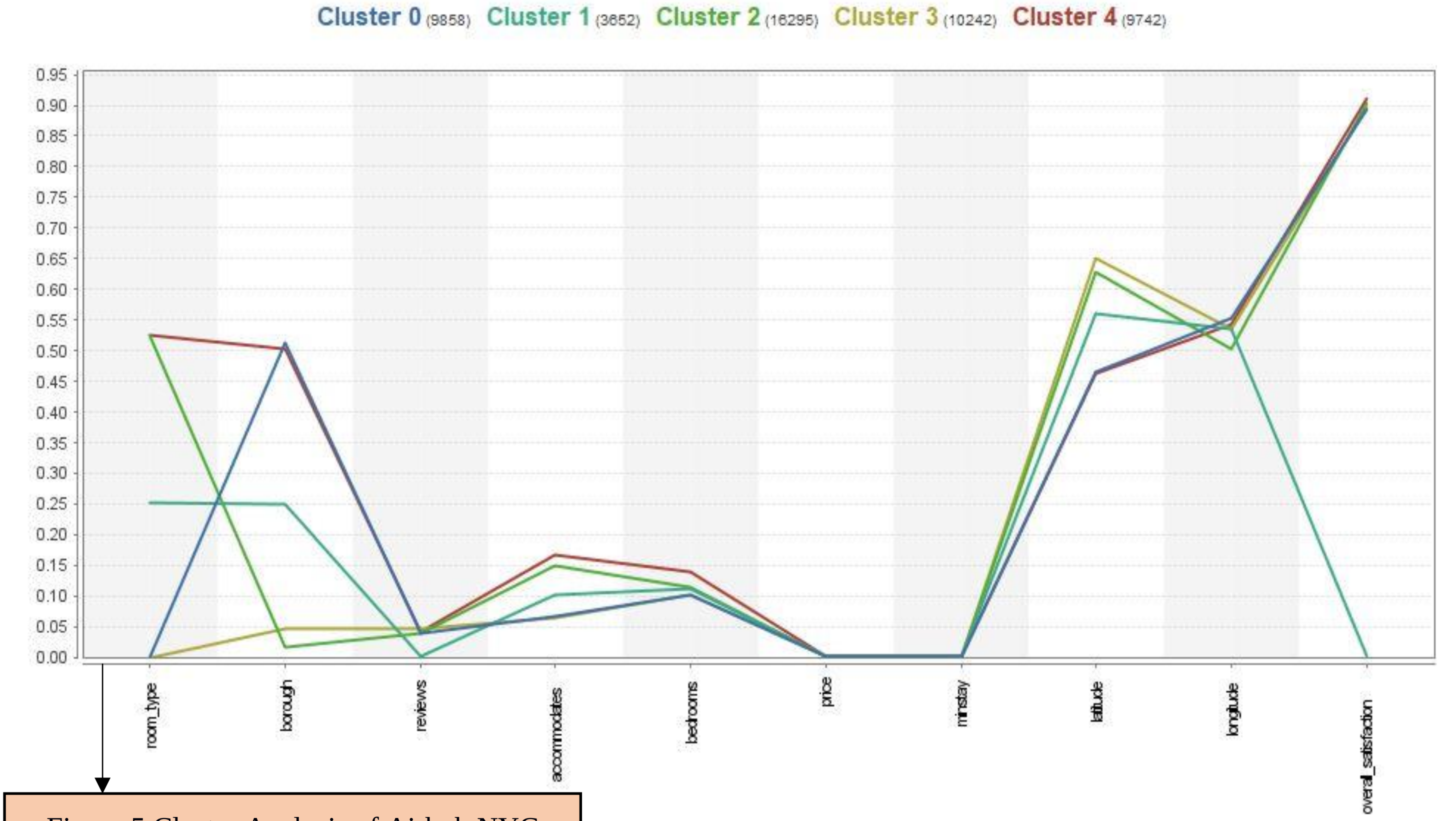
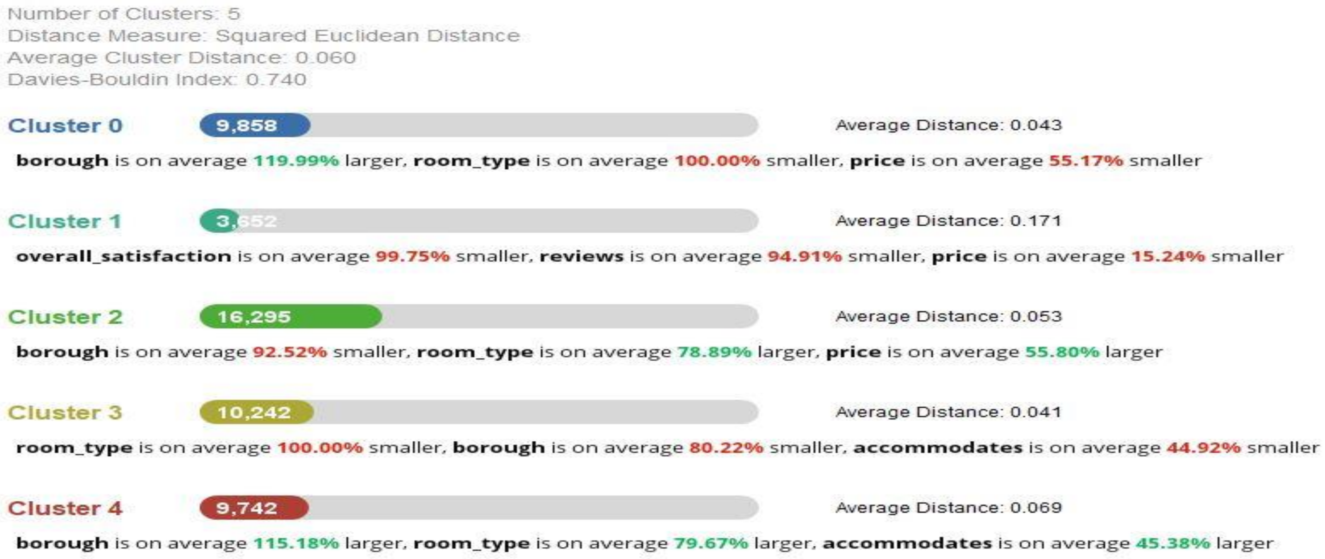


Figure 5 Cluster Analysis of Airbnb NYC



Number of Clusters: 5
Distance Measure: Squared Euclidean Distance
Average Cluster Distance: 0.060
Davies-Bouldin Index: 0.740

Cluster 0 9,858 Average Distance: 0.043

borough is on average 119.99% larger, **room_type** is on average 100.00% smaller, **price** is on average 55.17% smaller

Cluster 1 3,652 Average Distance: 0.171

overall_satisfaction is on average 99.75% smaller, **reviews** is on average 94.91% smaller, **price** is on average 15.24% smaller

Cluster 2 16,295 Average Distance: 0.053

borough is on average 92.52% smaller, **room_type** is on average 78.89% larger, **price** is on average 55.80% larger

Cluster 3 10,242 Average Distance: 0.041

room_type is on average 100.00% smaller, **borough** is on average 80.22% smaller, **accommodates** is on average 44.92% smaller

Cluster 4 9,742 Average Distance: 0.069

borough is on average 115.18% larger, **room_type** is on average 79.67% larger, **accommodates** is on average 45.38% larger

Figure 6 Overview of Clusters (Groups)

Final Cluster Model Performance Post Optimization

Within Sum of Squares (WSS)	0.232
Davies Bouldin	-0.740

Figure 7 Final Cluster Model Performance at clustering (k-means) k=5

Description: WSS is the total distance of data points from their respective cluster centroids, low WSS score signifies cluster cohesion which is the goal of data clustering.
Davies Bouldin is lowest near zero when clustering produces low intra cluster distances (high similarity).

Data Preparation and Exploration.

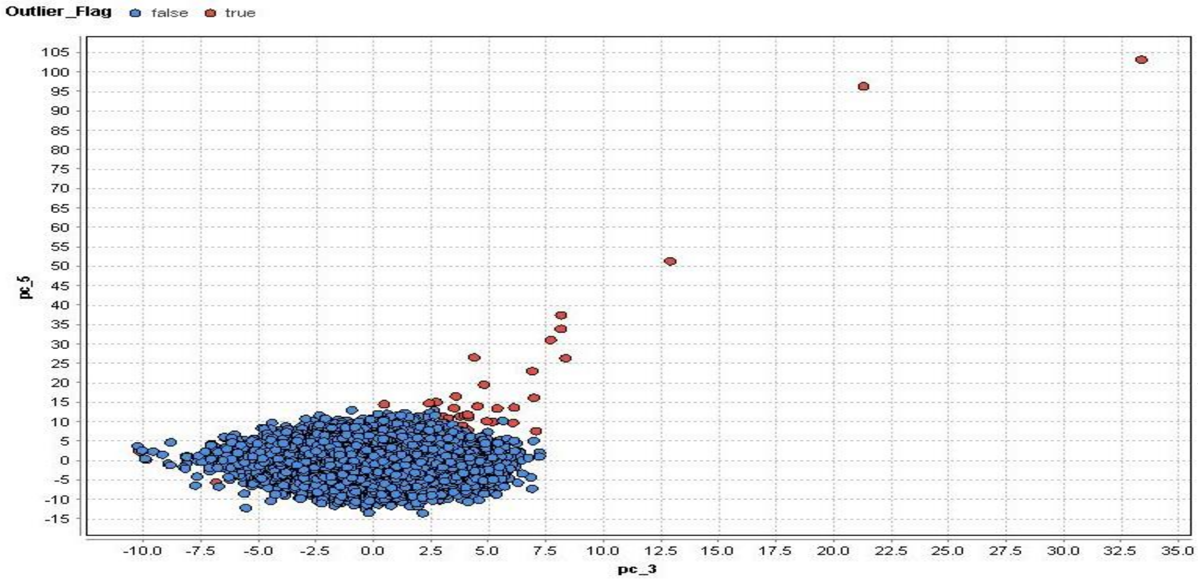


Figure 8 Anomaly Visualisation (PCA)

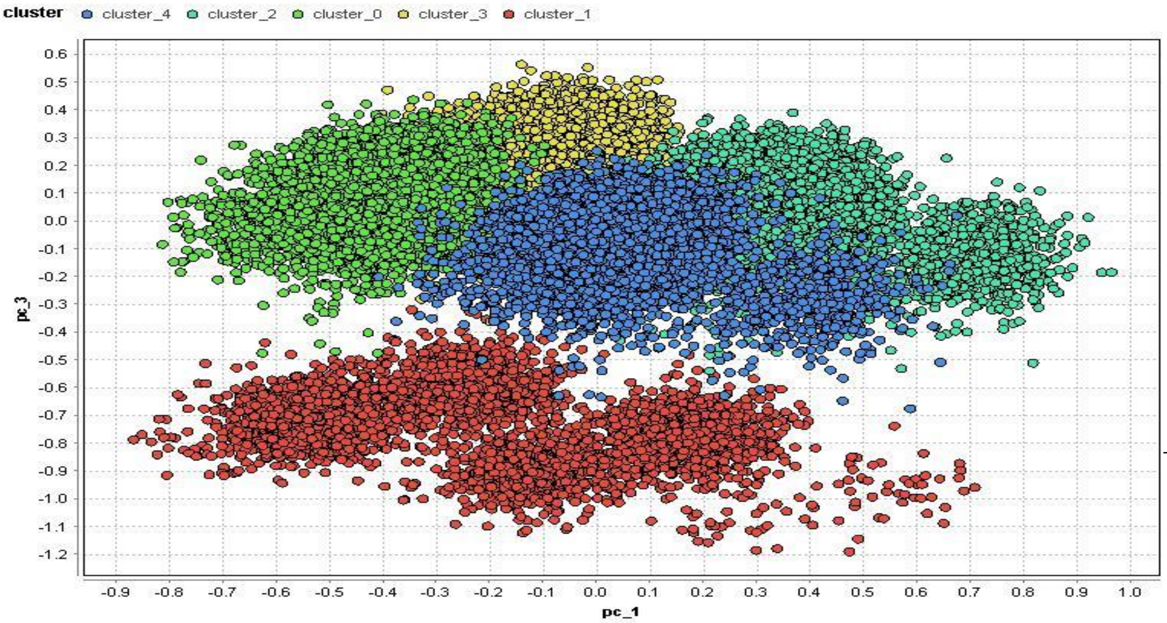


Figure 9 Cluster Visualisation (PCA) after dealing with anomalies using k-NN Global Anomaly Score, filtering out examples with high outlier scores

Description: On optimising Cluster model, we can see that model has sharp elbow point at k=12 where WSS=0.187 after which WSS is gradually decreasing, but due to financial feasibility for our clients to investigate the groups, 12 groups would be large hence we choose k=5 at which WSS is also low (0.232) and is optimum value for our client.

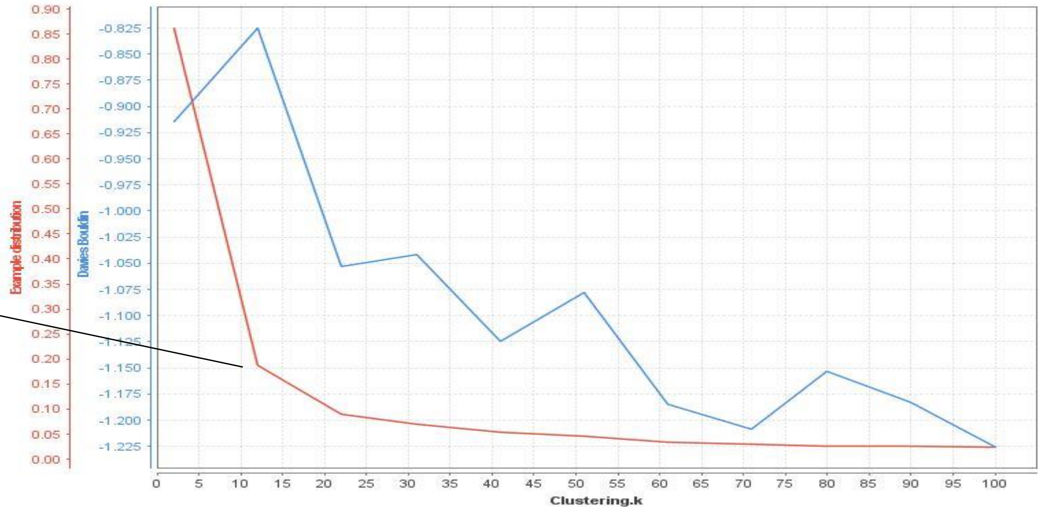
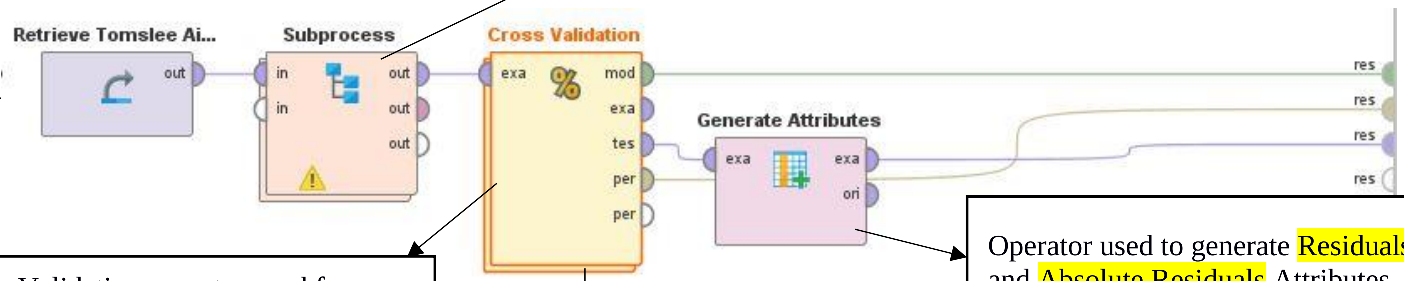


Figure 10 Cluster Optimisation

Model Development (one page)

Figure 11 Model Development using Decision

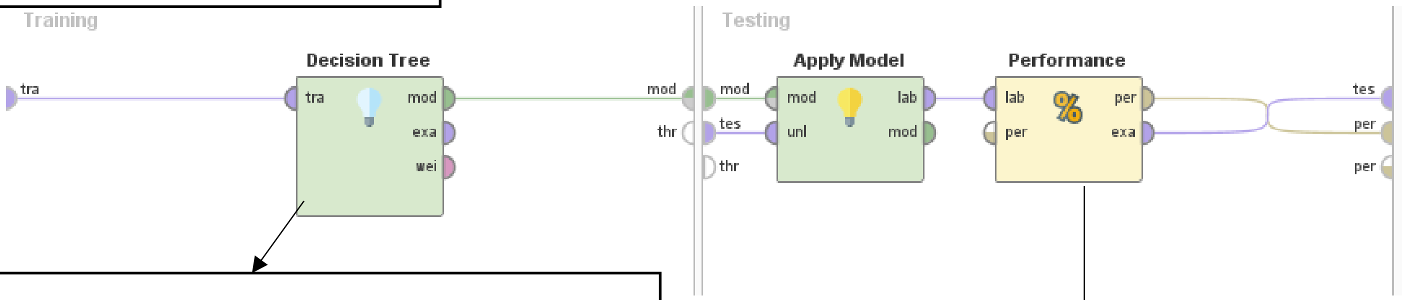
Operator used to create a subprocess within the main process comprising various operators



Cross Validation operator used for performing k=10-fold cross validation to training data in which 9 folds used for training and 1-fold used for validation (testing)

Operator used to generate Residuals and Absolute Residuals Attributes

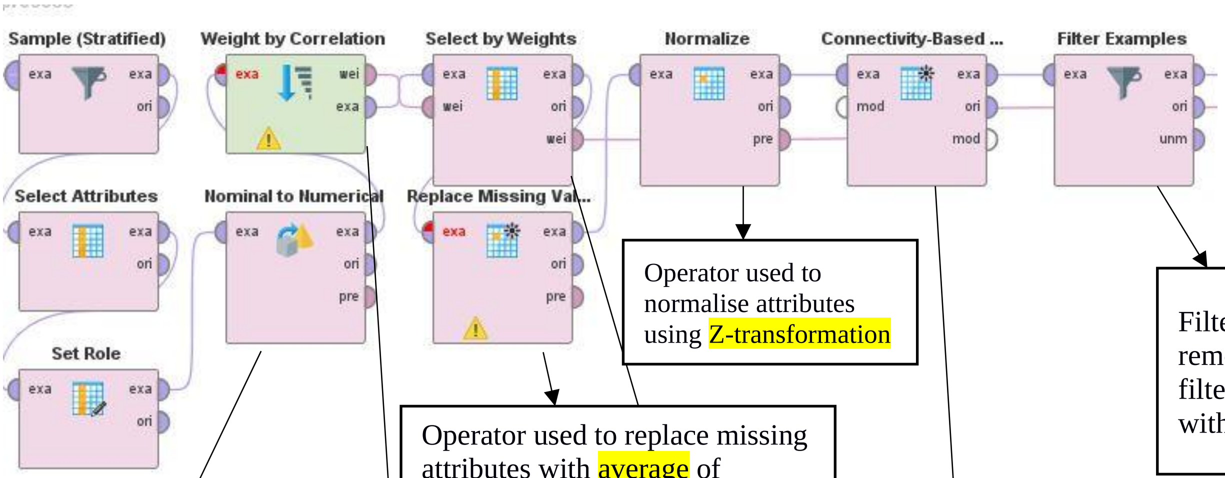
Figure 12 Operators within Cross Validation



Decision Tree operator used to perform regression (predicting customer satisfaction) with maximal depth =10 using least square criterion

Operator used to determine regression model performance using root mean square error, absolute error & correlation performance metrics

Figure 13 Operators within Subprocess



Operator used to normalise attributes using Z-transformation

Filter operator used to remove outliers by filtering out attributes with outlier score < 2

Operator used to replace missing attributes with average of available attributes

Operator used to convert nominal attributes such borough, room type into numerical attributes for performing regression using unique integers coding

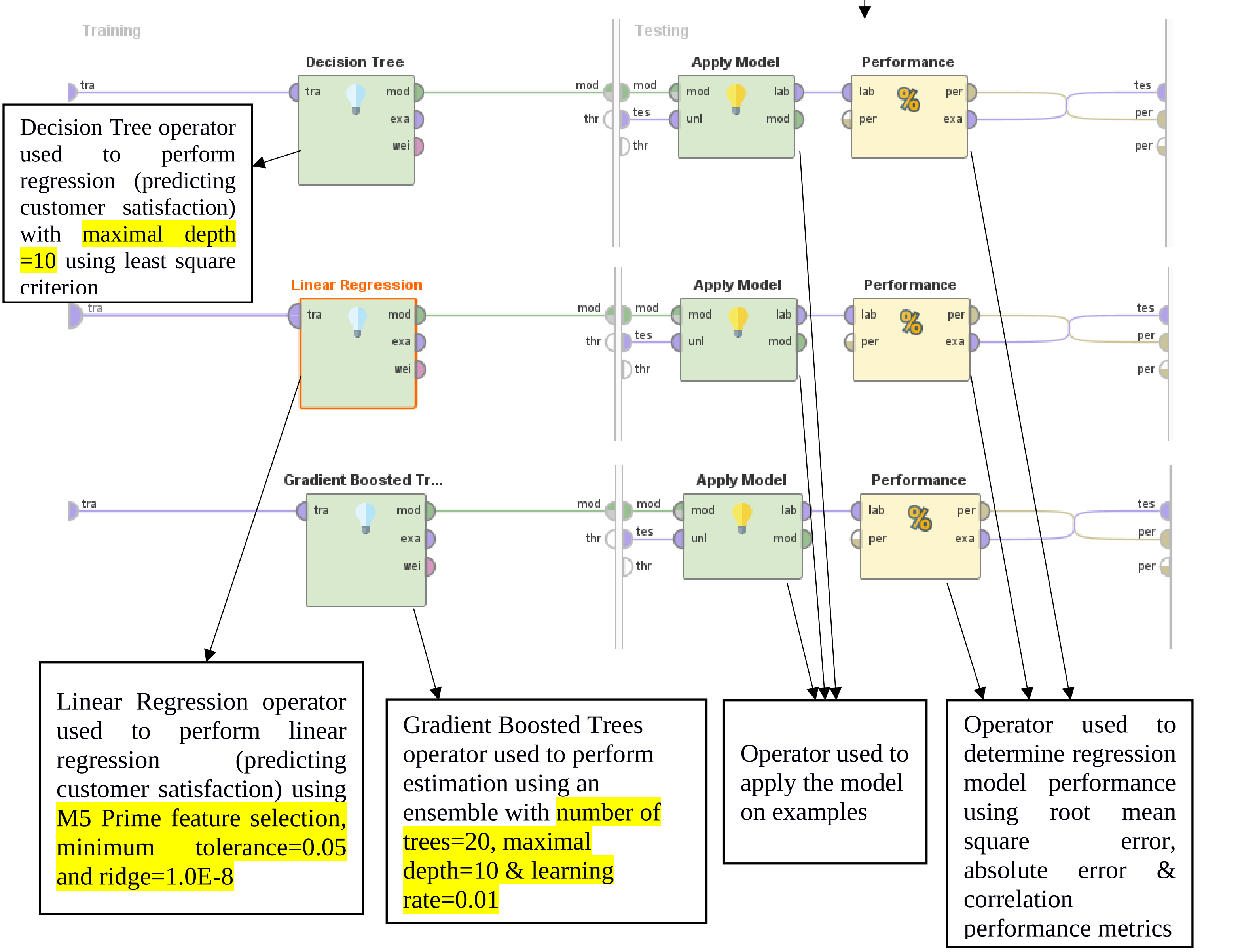
Operator used to weight attributes based on their correlation value with label (predictor)

Selecting attributes based on their weights. Top k=9 attributes selected

Connectivity Based Outlier Factor Operator used to detect anomalies using nearest neighbours, k=10

Model Evaluation and Optimisation

Figure 14 Estimating customer satisfaction with different models



Decision Tree operator used to perform regression (predicting customer satisfaction) with maximal depth =10 using least square criterion

Linear Regression operator used to perform linear regression (predicting customer satisfaction) using M5 Prime feature selection, minimum tolerance=0.05 and ridge=1.0E-8

Gradient Boosted Trees operator used to perform estimation using an ensemble with number of trees=20, maximal depth=10 & learning rate=0.01

Operator used to apply the model on examples

Operator used to determine regression model performance using root mean square error, absolute error & correlation performance metrics

Figure 15 Performance Comparison of Models

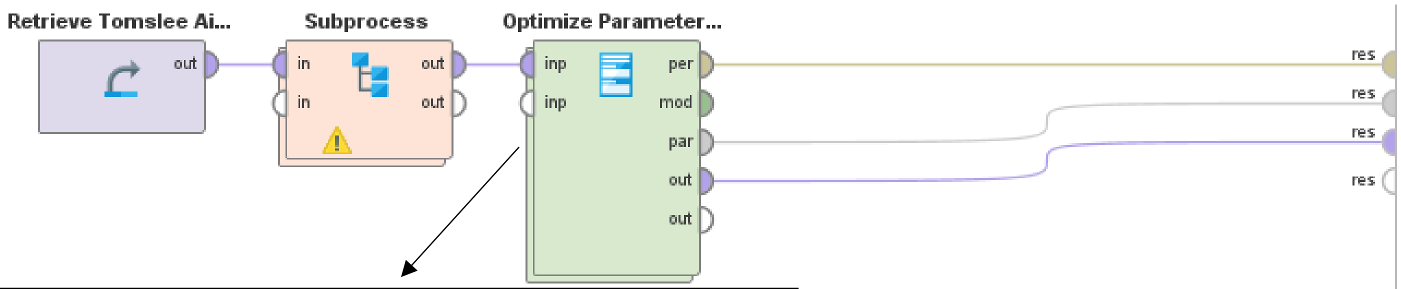
Performance Comparison of Models (Decision Tree, Linear Regression and Gradient Boosted Trees)

Model Name	Root Mean Square Error (RMS)	Absolute Error (MAE)	Correlation (r^2)	Prediction Average (Average Customer Satisfaction Level)
Linear Regression	1.174	0.652	0.201	4.209
Decision Tree	0.360	0.220	0.954	4.209
Gradient Boosted Trees	1.000	0.551	0.956	4.209

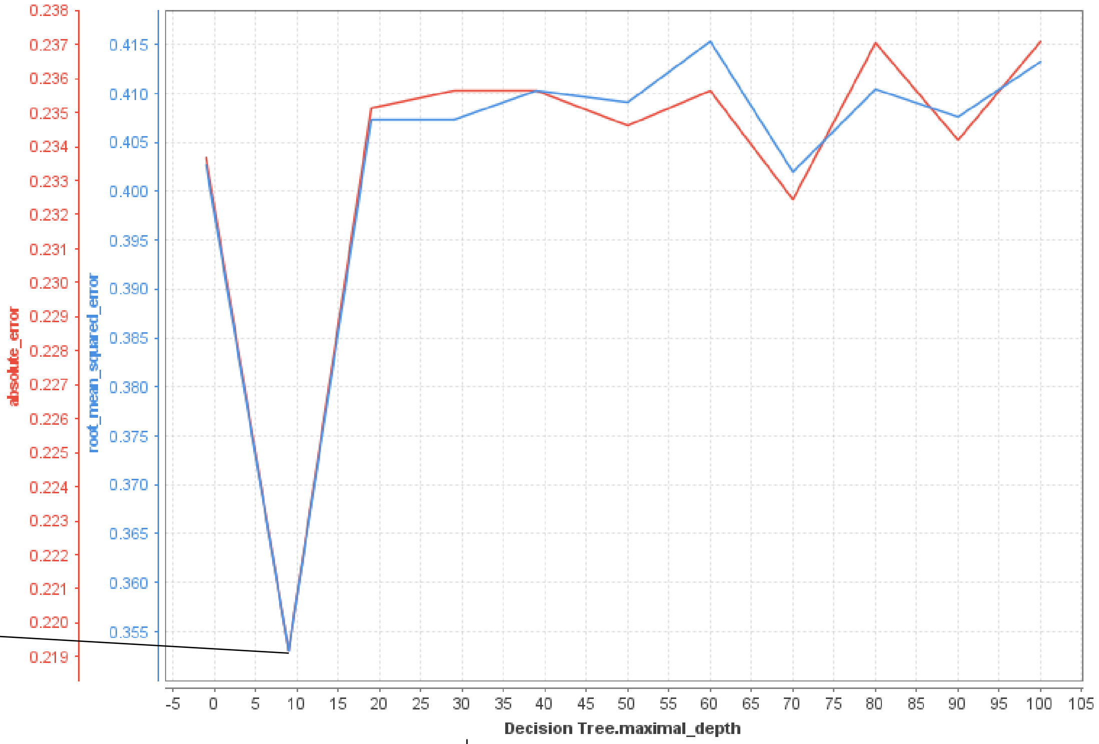
Best Model Chosen is Decision Tree since it has least root mean square error and least mean absolute error and high correlation value

Model Evaluation and Optimisation

Figure 16 Optimization of best-chosen model (Decision Tree)



Optimize Parameters Grid operator used to optimise the selected decision tree model with decision tree maximal depth ranging from -1 to 100



The best performance is of model with maximal depth=9 as root mean square error and mean absolute error is at the lowest at this tree height. RMS=0.353 and MAE=0.219

Figure 17 Evaluating model performance on different maximal depth (tree height)

Root Mean Square Error (RMS)	Mean Absolute Error (MAE)	Correlation (r^2)	Prediction Average (Customer Satisfaction)
0.353	0.219	0.955	4.209

Figure 18 Performance Metrics of Optimised Model (Decision Tree)

Model Application (one page)

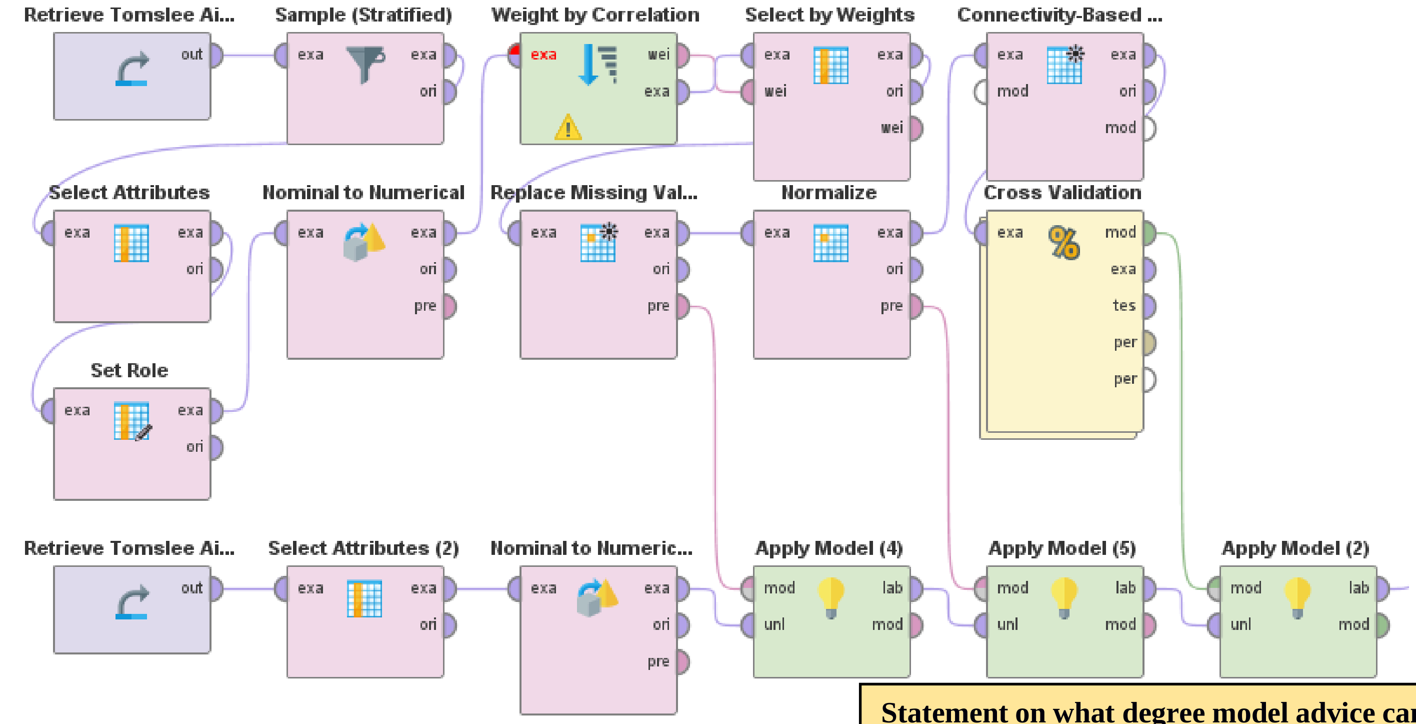


Figure 19 Model Application on new data at maximal depth=9 (Decision Tree)

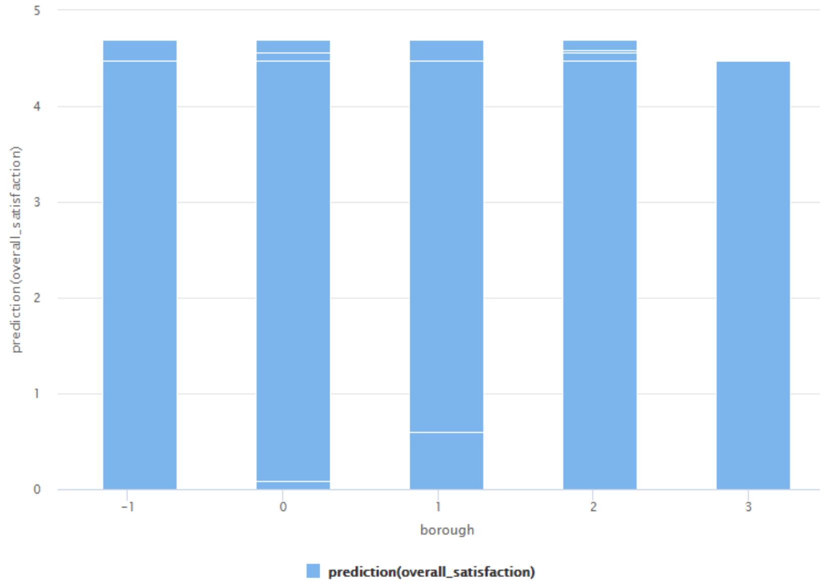
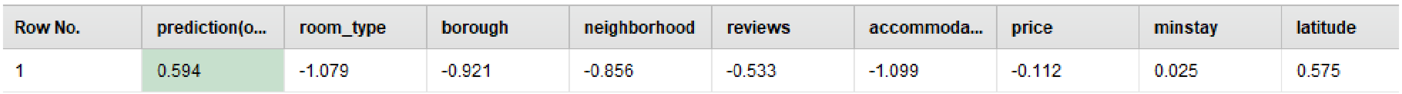


Figure 20 Average Customer Satisfaction level of New Data (Across Different Borough Groups)

Statement on what degree model advice can be trusted: As we saw in data exploration part that the training data had missing and extreme values and it violates the assumptions of regression models such as decision tree, linear regression. Moreover, as we saw in clustering analysis that clusters with contrasting prices and accommodation capacity, different room type and borough group had equal and high customer satisfaction level. And as the model is built on same attributes, its advice of satisfaction level is coming out to be approximately 4.5 which is high and approximately same across different borough groups and it may change if other attributes are included in the model which can have impact on customer satisfaction level.

Result Interpretation: As we can see that while doing cluster analysis on training data we found that there was one cluster which had less number of reviews got very low satisfaction level and as our model is built on training data and this single listing on which our model is applied has less number of reviews (-0.533, normalised value). Hence, its predicted satisfaction level is low.



Row No.	prediction(o...	room_type	borough	neighborhood	reviews	accommoda...	price	minstay	latitude
1	0.594	-1.079	-0.921	-0.856	-0.533	-1.099	-0.112	0.025	0.575

Figure 21 Applying model on single listing