Assignment A1: Classification

Student Name	Yash Mistry		Student No	218612723
Problem attempted	Complex Model 80-100%	Simple Model 40-79%	Student Id	ydmistry
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	30	24	18	12	6	0
Model Evaluation	30	24	18	12	6	0
Brief Comments	(see below)					Total 0 to 100

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.

Note: We will severely penalise the final submission when the partial submission is late or missing.

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)

Aim

Airbnb AI has approached us into figuring out three things with their data set. The dataset has provided us with a various attribute to work with. Attributes such as name, id , property type , it geo location , whether a location for rent is licenced or not etc. We will be working mainly with neighbourhood, price and review section of the data attribute.

Expectation

We must figure out three things and predict three things with efficiency. Which rental have the attracts the most review in the NYC suburbs. Which neighbourhood have the best rentals. If there are no review to a rental property such as home / entire house/shared room are attractive or even economical viable?

With the help of the data provided we have to predict with a good model to ensure the business will prosper on the rental property.

In the extended work process to boost our business development we can figure out if the dataset with attributes such as review which are not present or are not available, they economically viable?

Working

We must work our way into to figuring out what kind of problems and what kind of attributes we will face and clean up during the process.

Factors such as neighbourhood, the price of the rentals, the even the reviews will come under consideration when it comes to figuring out the solutions of our problems

With this process, we will benefit without understanding what kind of neighbourhood are most attractive. We will even find out what kind of rental attracts majority of reviews.

Once we figure out the main role of the different attributes to the given dataset and what role it plays in deciding if the rental is attractive or not.

In our other business complication we will use a different attributes and even give two or more process with different attributes to understand if any listing with no reviews are economically viable or not and even try to figure out the sense of what it might be caused by.

Benefit

With the help of the predictions we can figure out if the rentals which doesn’t have any review good or bad are they economical for the property owner or not. we can even figure out a way for good business altogether. We should even figure out if the rentals are good because of the neighbourhood or the price or just based on reviews the property has.

Data exploration (one page)

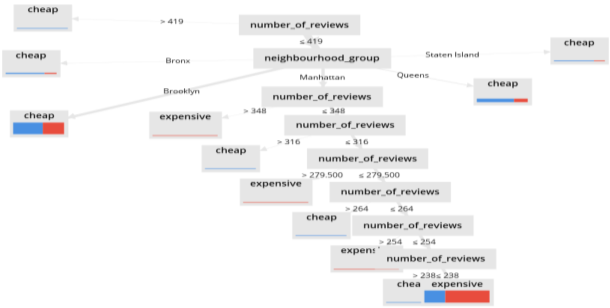


Fig1.1

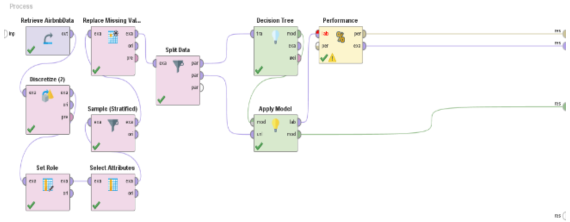


Fig1.2

To look at what price point the rentals at Airbnb are well worth it, we need to check that which of the rentals are in attractive range to the people with price point and reviews point. To figure out what type of rental attracts more in the data we have to do more process. Places such as Brooklyn and Manhattan have the most reviews. But with more processing we can figure out the rental situation of the property and its attractiveness. From the figure 1.2 we can see all the process we have applied for the predictions. Only the attributes such as room type, neighbourhood, price is taken into consideration.

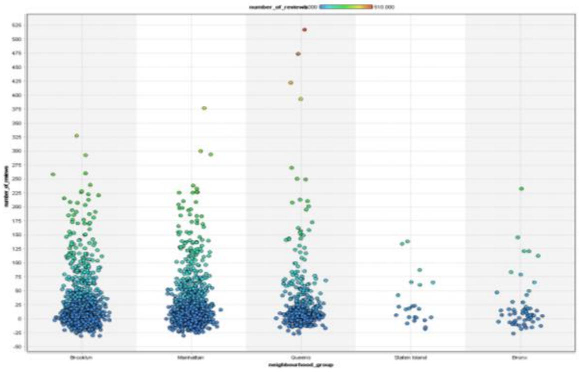


Fig1.3

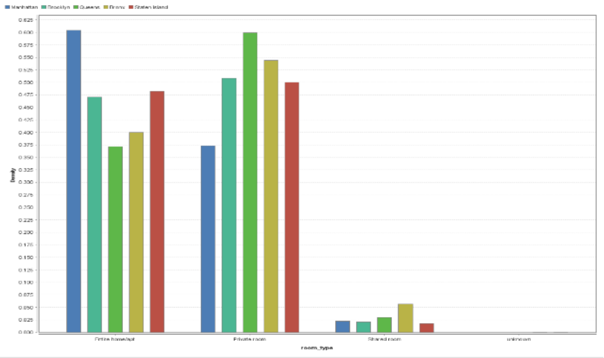


Fig1.4

In figure 1.3, it shows that how many reviews are submitted into the given house and their neighbourhood. It helps us into knowing why and how many reviews are needed to check if any of the house are economically viable or not. As we can see from the figure 1.4 there are lot of private rooms and entire house in numerous neighbourhood groups. But, there are few shared rooms and even rooms that are unknown or doesn’t fit into the other criteria as well. We have to distinguish between them as to have a clear view of our predictions.

Executive solution statement (one page)

Aim

In different neighbourhood we have on our dataset which of the rental property which does have any reviews are economical viable.

Simple Model

From the Fig 1.1 and 1.3 we can see that most of the reviews are in the neighbourhood location of Brooklyn, Manhattan, Queens. Second comes Bronx with very less Staten island. In this we will have to make our process free of any replacing values and even find if there are any duplicates of any data. Attributes such as filter and replace missing values with average or even null value (0) can be replaced. It can help us to find a better solution and a reliable predictive model for our business approach to queries.

From the process we can see that most of the rentals who are from Brooklyn, Manhattan, Queens are much more worth the price given their room type. Staten island has the least amount of housing property and least amount of reviews due to very less property on rent. In the figure 1.4 we can see the number of houses with their respective neighbourhood. In this we can see that Manhattan has a lot of entire house and even private room. Same as Manhattan, queens have same ratio of entire and house and private rooms. With the help of the dataset we know we have around 49000 entries and in that we can figure out how much room type is in which neighbourhood.

From figure 1.2 we can deduce a lot of information from our predictions and even a lot of attributes that we must fix in order to have a better prediction. Right now, we have a accuracy of only ~65%.

From our Thesis for our business problem, Manhattan, Brooklyn and at some property even queens is a very attractive neighbourhood.

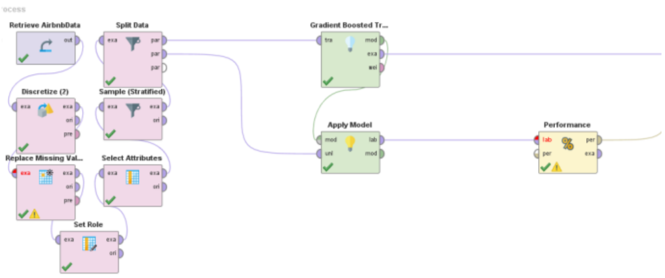


Fig 3.1

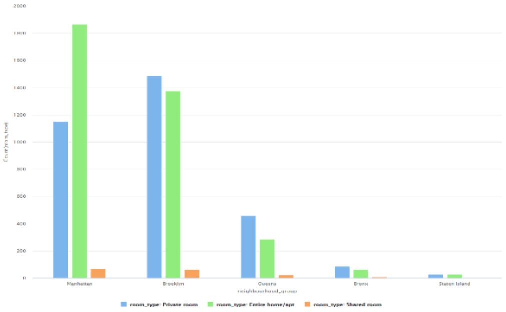


Fig 3.2

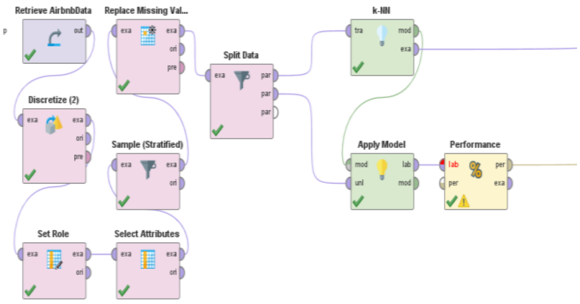
In the simple model stated above in figure 3.1. We have got some various results, which can help us with business decisions.

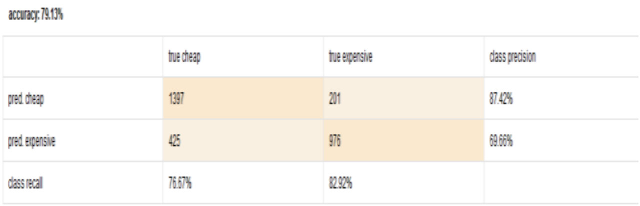
As we can see from the figure 3.2 the amount of house and house type in Brooklyn and queens Manhattan are very large. But as we can see further that their number of reviews are also lot. We can safely say that a lot of attractive rentals are in this area as these neighbourhood have a lot for rentals and property of different kind. Room type such as entire house and even private rooms are available at large in these areas which makes it a very attractive locality amount Airbnb dataset.

In our thesis and predictive model for the previous business question, it is safe to say the Manhattan and Brooklyn are also most suitable property neighbourhood to get a review. A lot of attributes come into play. Attributes such as neighbourhood and even price range to the property type is also better in these neighbourhood.

Data Preparation (one page)

K-NN model



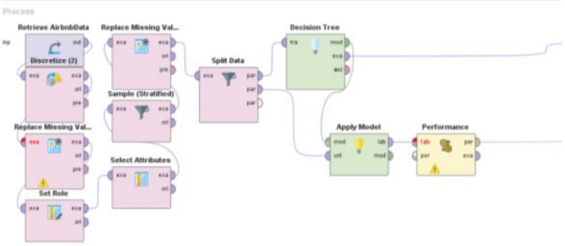


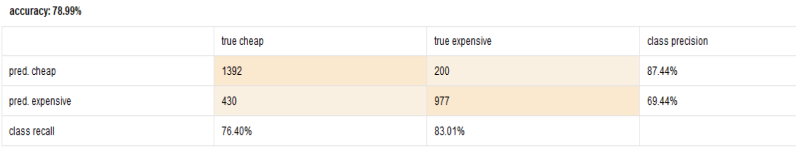
accuracy: 79.13%

	true cheap	true expensive	class precision
pred. cheap	1397	201	87.42%
pred. expensive	425	976	69.66%
class recall	76.67%	82.92%	

K-NN model stands for the nearest of the neighbour. It means that it will calculate the nearest neighbour (K value) or data points as per the value of k. between any two calls of data, k Value can be in odd value as it can reduce confusion. Calculated by measuring the accuracy for that value meaning k value representing maximum accuracy value is selected for the modelling. Also, k value can only be odd in order to avoid confusion between two classes of data. The accuracy with my K-NN model was low at 65% then it increased to 70% but no further than that. At this point we can now turn to another model which is decision tree and gradient tree model

Decision trees.



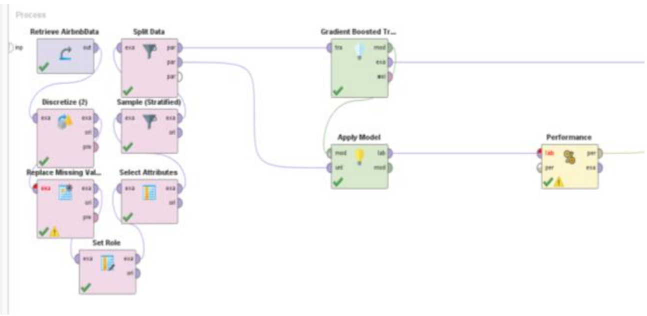


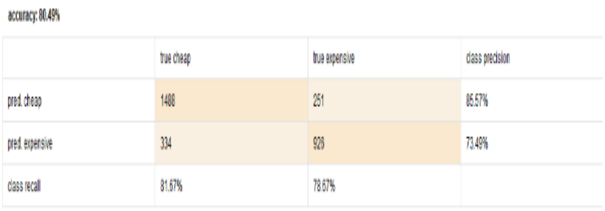
accuracy: 78.99%

	true cheap	true expensive	class precision
pred. cheap	1392	200	87.44%
pred. expensive	430	977	69.44%
class recall	76.40%	83.01%	

Decision Tree is a model which can take both numerical and nominal attributes and provides a tree representation of the model. Likewise, in k-NN model 70% of the data used for training is fed to Decision Tree and remaining 30% of the data is used for testing the model. The output of decision tree is fed to the Apply Model whose output is fed to Performance which calculates accuracy and kappa value of the model. With decision trees my prediction was a bit better than K-NN model with accuracy of 78.99 and kappa at 0.574

Gradient Boosted Trees.





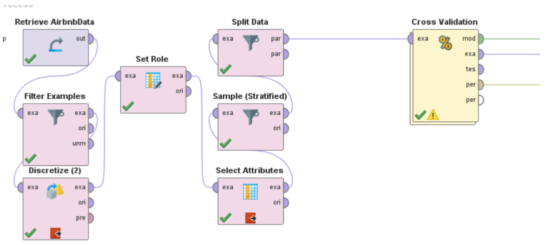
accuracy: 80.49%

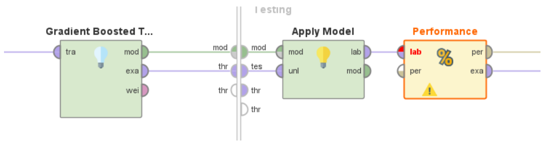
	true cheap	true expensive	class precision
pred. cheap	1488	251	85.57%
pred. expensive	334	926	73.49%
class recall	81.67%	78.67%	

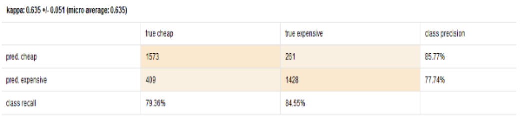
In Gradient boosted tree. Our predictions model was having a slight better accuracy and kappa too. Accuracy with 80%. In our gradient boosted trees we could see and can get help regarding the our Question 3. Which is any property without any reviews economically viable. In my process I have used this process because it can used better for our prediction model and can give us a better predictive model for our business problem. The model can be trusted as any data that can be put through this data while/may perform in the same range and with around same prediction altogether as well.

Model Development (one page limit)

Deployment

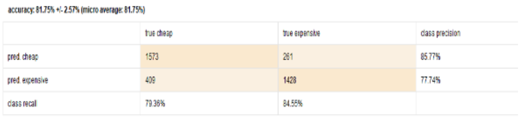






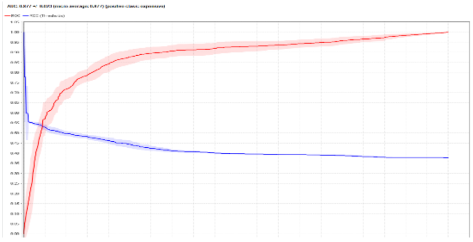
kappa: 0.635 +/- 0.051 (micro average: 0.635)

	true cheap	true expensive	class precision
pred. cheap	1573	261	85.77%
pred. expensive	409	1428	77.74%
class recall	79.36%	84.55%	



accuracy: 81.75% +/- 2.57% (micro average: 81.75%)

	true cheap	true expensive	class precision
pred. cheap	1573	261	85.77%
pred. expensive	409	1428	77.74%
class recall	79.36%	84.55%	



In our business deployment of our prediction model, we have chosen what is essential to our business problem. For this problem we need / have used attributes such as neighbourhood, property price, and even reviews. Reviews per month and even number of total reviews to the rental property as well. With our filtration process we can see if any property which any review does not have is economically viable i.e. is it making any money. In our deployment process, we have decided to make it into two process with <u>cross validation method.</u> With attributes such as minimum number of nights and even price can be used to speculate much further, as to catch the cause as what might cause shortage of review of not being of economical viability.

We a lot of filtration we have got a reputable process accuracy and kappa values. My accuracy with my predictive process is *81.75%* and kappa is at *0.635* which is found to be very reputable and can be worked with at certain extent. with this base model we can see which of the property has got less reviews and are they economical viable. The model such as set role and select attributes have been a lot of help as they are used to put all the attributes to their right places for a better predictive model of our data set.

Model Evaluation (one page)

Aim

This part of my report is to explain the performance of developed classification models.

Working

From our business predictive model with a reputable accuracy and kappa we can see that a lot of business problem can be addressed with any new situation or even new business management lookouts. From our business model we can see that a lot of reviews which are no present or are very negligible must be used . From our data prediction model, with good reputable accuracy we have seen a very less business to the property rentals with less review. Any property such as entire house and even private or shared rooms are very less likely to be economically viable to the business.

Business which run/ handle a property in a posh location with a very large margin in price range can be tricky to be solved as the locality place a big role in financing a business viable. This inquisition can be solved if there are any booking to the rental place. That can be located by the attribute such as *minimum number of nights to stay*. Some customer would only like to live at a rental property for a short period of time and is not looking a long term. Neighbourhood and even price range can be a big affair in the customers dealing with rentals properties and choosing it. The reviews are a very critical part of rental property in Airbnb listing as they can decide if the economical viable or not. In most of the cases listing which did not have any review did not come across as economically viable.

Using this business model in the future can be very constructive with a little tweaking for a desired outcome for any business-related complications. Neighbourhood such as Brooklyn and even Manhattan had some of the listing/ property rentals in the dataset with no reviews. Using this business process is / can be reliable to some extent. With accuracy at just 81% it is usable and can be shaped into much better proc cess and can be improved. But at some occasion, if we have some new business complications some tweaking/alter some process with different attributes can be adjusted.

Using *Rapid Miner* software, we can easily figure out many more concept or business concerns. With the help of boxplots and even scatter plots which is available to us in Rapid miner with a lot of ease can help through a lot of tough business material.