Assignment A1: Classification				
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	30	24	18	12	6	0
Model Evaluation	30	24	18	12	6	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. 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)

Problem Statement

To gain insights into the New York City Airbnb rental properties and determine the neighbourhoods with most attractive Airbnb rentals and the type of rental properties which have most reviews. Furthermore, we need to determine the economic viability of the rentals with missing reviews.

We look forward to identifying the neighbourhoods in New York City with most attractive Airbnb rentals and the type of rentals with majority of reviews. The attractiveness depends upon several factors such as room type, number of reviews, reviews per month and price of the rental properties. The type of rentals with majority of reviews depends upon various factors such as neighbourhood group, price, reviews per month and number of reviews. The minimum nights stayed on the rental properties also contributes to the attractiveness and majority of reviews of rental properties. Moreover, the rentals which have more than one review per month are economically viable i.e. these rental properties are financially important to Airbnb. This project will benefit Airbnb in determining the neighbourhoods with average level attractiveness and less reviews. With the help of this data, Airbnb can improve their services in those areas and will also get to know the type of rentals which are not attracting majority of reviews and hence can enhance services of those rental types. The Airbnb clients which are the owners of these rental properties will get to know meaningful insights about their properties and can adjust according to the rental attractiveness of their properties.

After analysing the records of New York City Airbnb rental data, we concluded that neighbourhoods such as Manhattan and Brooklyn have the most attractive Airbnb rentals. Manhattan and Brooklyn have the greatest number of reviews, reviews per month across all the neighbourhoods of New York City (see Figure 1 and 6). Furthermore, Brooklyn has a little edge over Manhattan in terms of affordability as we can see by the distribution of average price of the rental properties across neighbourhoods (see Figure 2). However, the average number of minimum nights stayed in Manhattan rental properties is more than the Brooklyn rental properties (see Figure 5). Overall, Brooklyn and Manhattan rental properties stands out of all the neighbourhoods in terms of rental attractiveness because of the large number of reviews, reviews per month, minimum nights stayed and affordability (in case of Brooklyn).

As shown in Figure 4 most reviews are attracted by the entire house/apt type of rental properties. The factors such as price, reviews per month, minimum nights stayed, and number of reviews collectively contribute in determining the rentals with majority of reviews. Entire house rentals have the greatest number of reviews and reviews per month as compared to other rental types across majority of the neighbourhoods (see Figure 4). Moreover, the average number of minimum nights stayed in entire house rentals is more than other rental types across majority of neighbourhoods (see Figure 5). However, the average price of entire house rentals is also more than other rental types across all neighbourhoods (see Figure 2). Also, the kind of rentals which attracts most reviews also depends upon the neighbourhood group as shown in Figure 3 where Manhattan's entire house rentals has more reviews than Brooklyn's entire house rentals and Brooklyn's private room rentals have more reviews than Manhattan's private room rentals. Although, the average price of entire house rentals is more than other rental types but in conclusion, the entire house rentals especially in Manhattan have attracted the majority of reviews because of most number of reviews, minimum nights stayed and reviews per month as compared to other rental types such as private room or shared room.

Data exploration (one page)

Label Attribute: Reviews per month as label is supposed to be the attribute with discrete values which is being predicted and as we want to determine the economic viability of rentals with missing values of reviews per month, we will choose the reviews per month attribute as label.
Predictors: All the attributes which help in the prediction of the label attribute are called predictors. Number of reviews, minimum nights, availability, room type, neighbourhood group, price, name, host name, last review date, latitude, longitude are the attributes which are used as predictors as they are related to the label attribute reviews per month and hence will help in determining it.

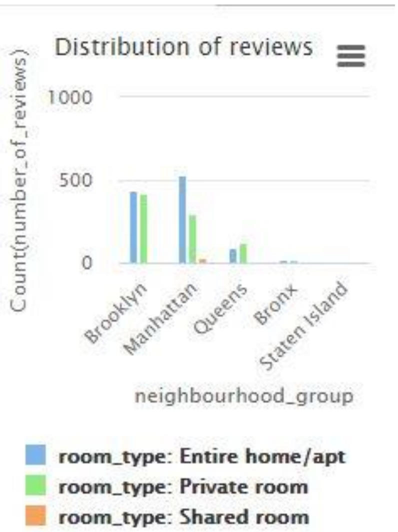


Figure 1 Distribution of number of reviews

Figure 3 Reviews across rental (detailed)

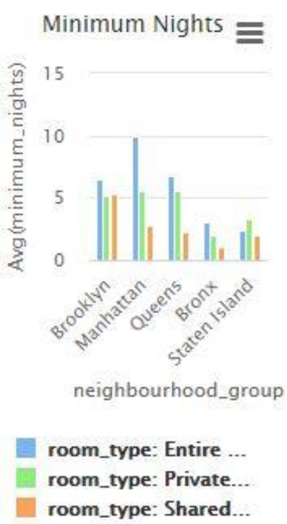


Figure 5 Distribution of minimum nights

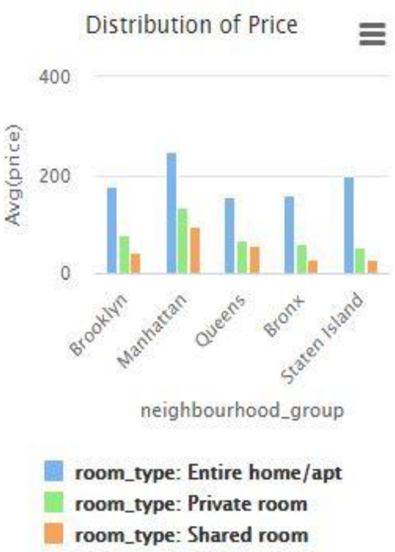


Figure 2 Distribution of Price

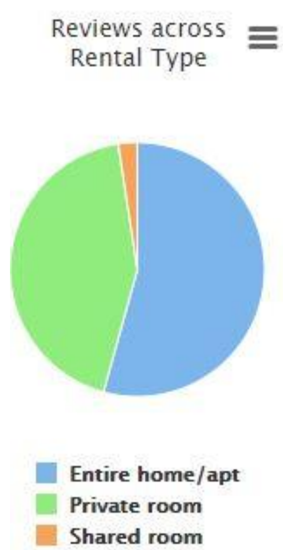


Figure 4 Reviews across Rentals

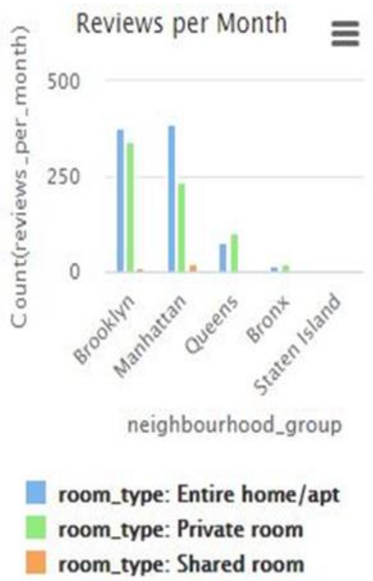


Figure 6 Distribution of reviews per month

Name	Type	Missing
name	Polynominal	16
host_name	Polynominal	21
last_review	Date	10052
reviews_per_month	Real	10052

Figure 22 Attributes with missing values

Missing Values
Attributes such as reviews per month, last review date, name, host name have missing values which can be replaced by average or imputing values from remaining values. Our model is built on gradient boosted trees and they work well with missing values so there is no need to replace them.

Executive solution statement (one page)

Solution Statement

After analysing the New York City Airbnb rentals with missing reviews, we concluded that around 36.44% of the total rentals with missing reviews are economically viable. Out of 10,052 rentals with missing reviews, around 3,663 rentals are economically viable whereas 6,389 rentals are economically unviable (see Figure 7). Thus, majority of the rentals with missing reviews are economically unviable. This provides an opportunity for New York City AirBnb to investigate these remaining 63.56% of the rentals with missing reviews. Manhattan and Brooklyn neighbourhood groups have the greatest number of economically viable rentals with missing reviews (see Figure 8). However, the economically unviable rentals in Manhattan are more than twice the number of viable rentals (see Figure 8). Bronx has approximately the same number of economically viable and unviable properties (see Figure). Furthermore, Queens has more unviable rentals as compared to the viable ones (see Figure 8).

The economic viability of the New York City rentals is related to various other factors such as availability of these rentals, minimum nights stayed at these rentals, host name, neighbourhood group and last review date of these rentals. The average number of minimum nights stayed in the viable rentals is approximately five more than those of unviable rentals (see Figure 10). Also, the average availability of the unviable rentals is approximately 46 days less than the viable rentals (see Figure 12). Therefore, analysing all these factors together, we were able to determine the economic viability of those rentals with missing reviews. Moreover, these factors such as neighbourhood groups, availability and minimum nights stayed are also related to each other (see Figure 9).

The results of this analysis can be used by New York City Airbnb to inform their clients who are the owners of these rental properties. It will help the clients to investigate the causes and take necessary actions. Both New York City AirBnb as well as their clients will benefit from this project. Also, the customers will also benefit from the improved services of rentals with low economic viability. Thus, the project will not only help in analysing the current operations of New York City AirBnb rentals but will also help in improving the services by predicting the viability of these rentals with missing reviews. It will also help in determining the relationships between various factors such as availability and minimum nights stayed on these rentals. The services can be improved by understanding the relationships between economic viability and availability and minimum nights stayed on these rentals and the relationships between availability, minimum nights stayed, neighbourhood group of these rentals.

Data Preparation (one page)

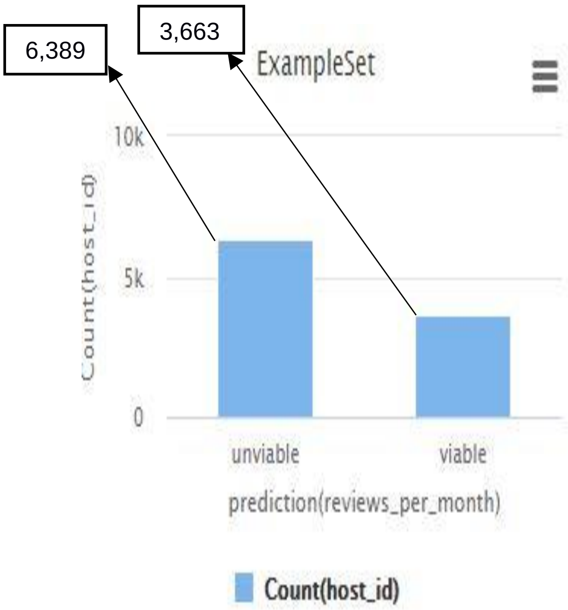


Figure 7 Distribution of economically viable and unviable rentals with missing reviews

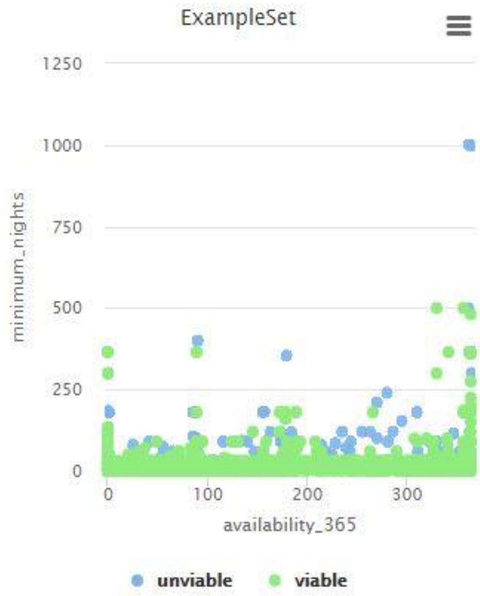


Figure 9 Relationship between availability & minimum nights across rentals with missing reviews

attribute	weight
name	0.989
last_review	0.343
host_name	0.334
availability_365	0.158
minimum_nights	0.062
host_id	0.038
neighbourhood	0.037
longitude	0.012
neighbourhood_group	0.009
latitude	0.002

Figure 11 Weights of predictor attributes are used to select the top 8 predictors in model

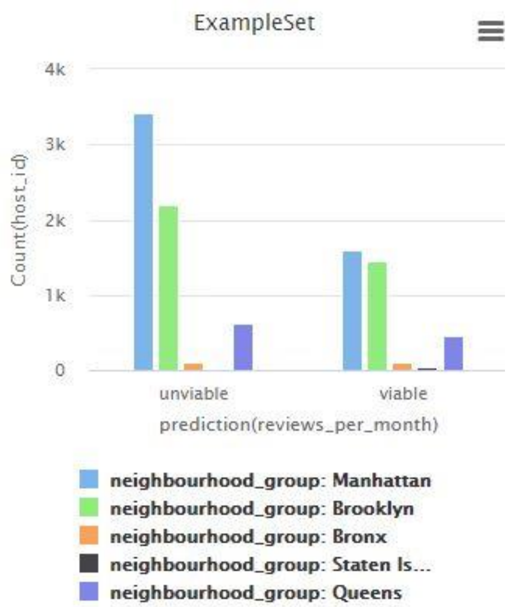


Figure 8 Distribution of viable & unviable rentals among neighbourhood groups

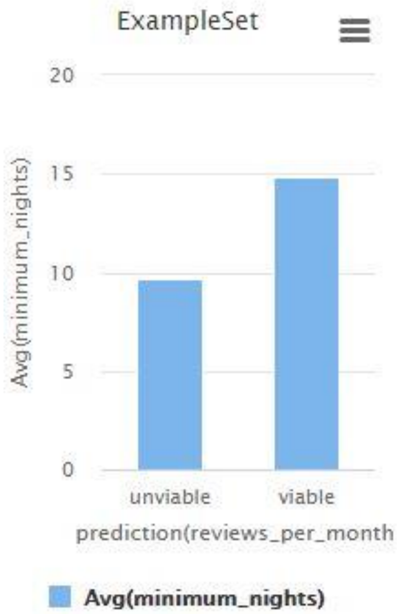


Figure 10 Average minimum nights stayed at viable & unviable rentals

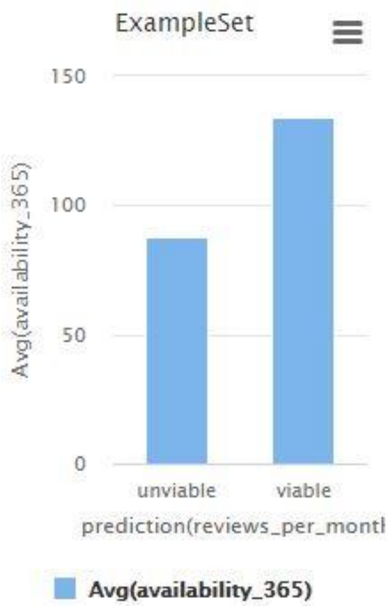


Figure 12 Average availability of viable & unviable rentals

Model Development (one-page limit)

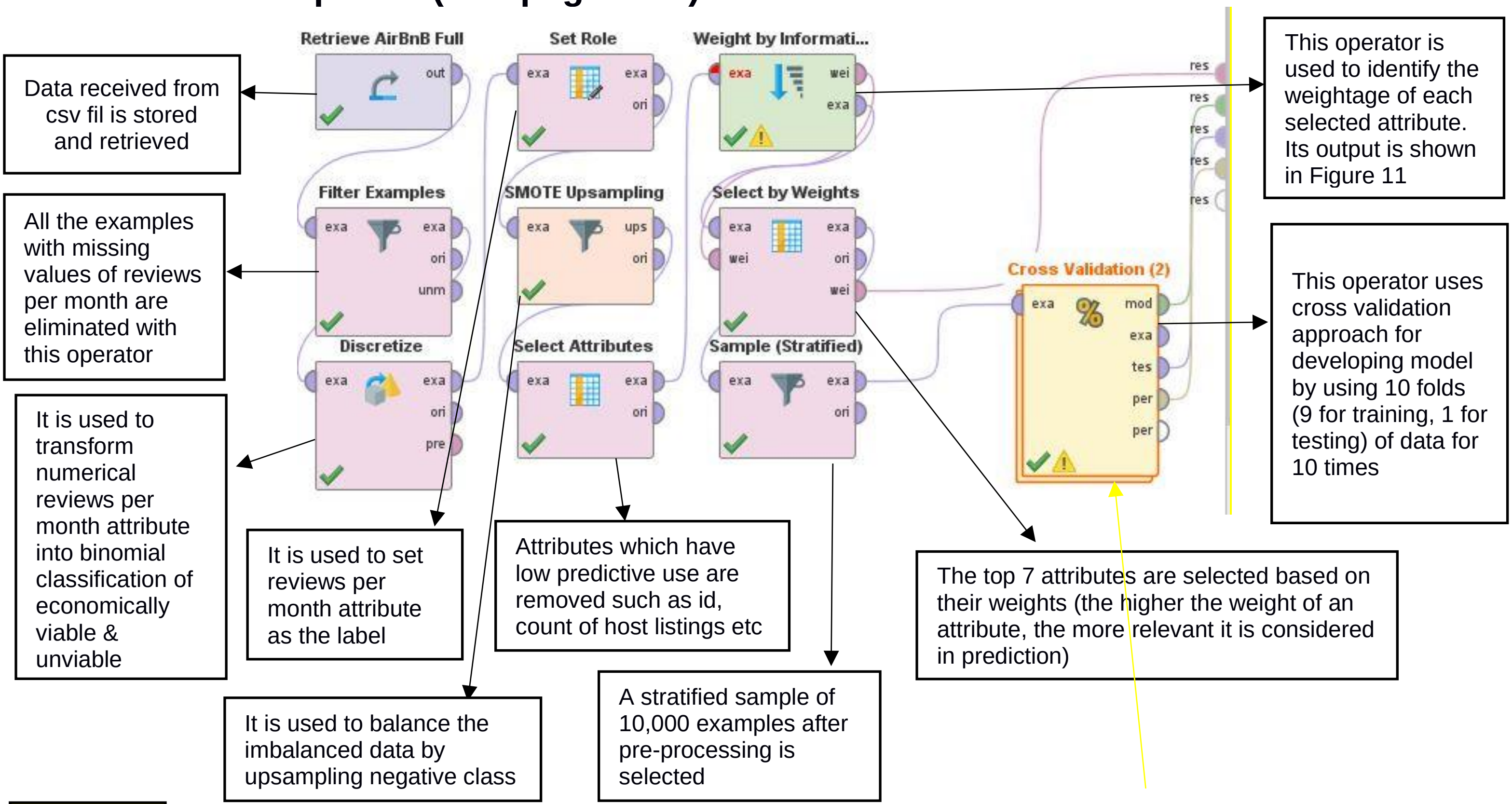
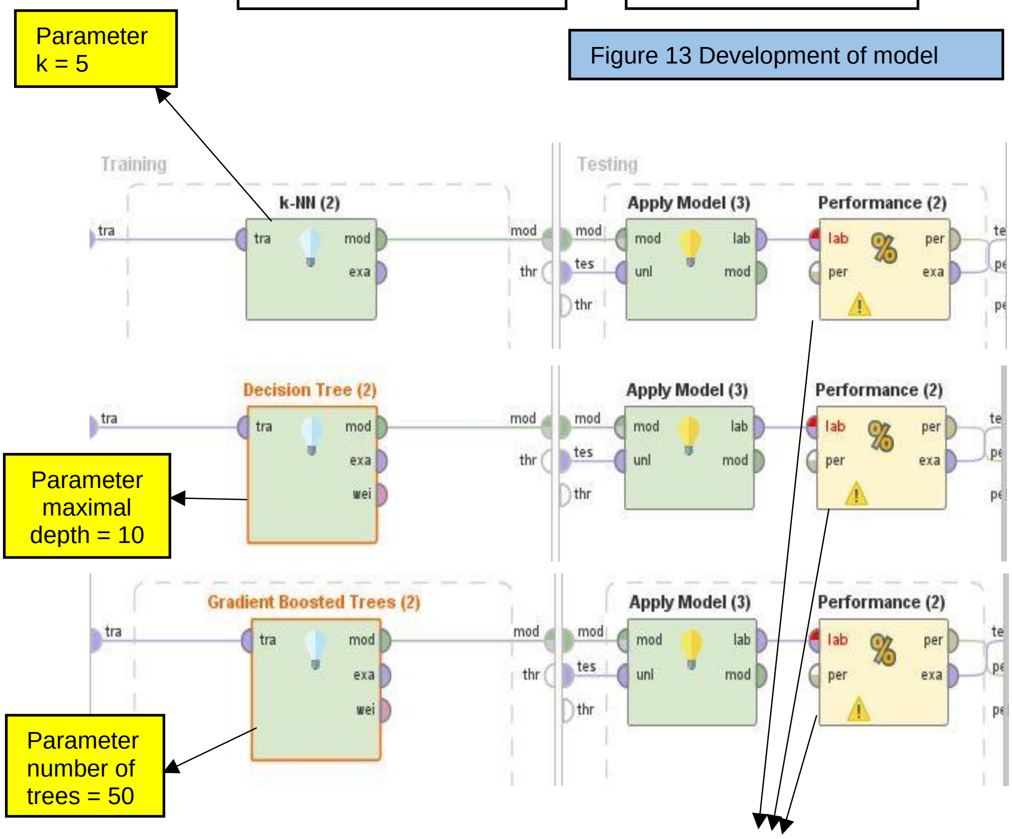


Figure 13 Development of model



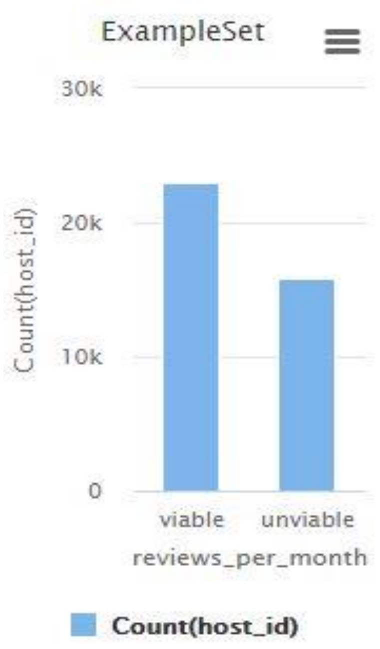


Figure 14 Development of model using k-NN, Decision tree, Gradient Boosted trees

Figure 15 Investigating for imbalanced data (therefore used SMOTE operator in model)

Model Evaluation (one page)

kappa: 0.607 +/- 0.019 (micro average: 0.607)

Kappa = 0.607

	true viable	true unviable	class precision
pred. viable	3903	870	81.77%
pred. unviable	1097	4130	79.01%
class recall	78.06%	82.60%	

Figure 16 Confusion Matrix for k-NN

kappa: 0.637 +/- 0.013 (micro average: 0.637)

Kappa = 0.637

	true viable	true unviable	class precision
pred. viable	3960	773	83.67%
pred. unviable	1040	4227	80.25%
class recall	79.20%	84.54%	

Figure 17 Confusion Matrix for Decision Tree

kappa: 0.646 +/- 0.018 (micro average: 0.646)

Kappa = 0.646

	true viable	true unviable	class precision
pred. viable	3854	623	86.08%
pred. unviable	1146	4377	79.25%
class recall	77.08%	87.54%	

Figure 18 Confusion Matrix for Gradient Boosted Trees

Statement: Since we are dealing with imbalanced data as shown in Figure 15 and accuracy measure is more sensitive to distribution of imbalanced labels, Kappa measure is chosen for comparing between all the three classifiers and Gradient Boosted Trees has the highest value of kappa which is 0.646. Kappa above 0.5 is a good model. Therefore, Gradient Boosted Trees is chosen to be the best classifier for developing model for this data. There is other performance measure such as Area Under the Curve for ROC chart which should be close to 1 for ideal model.

Best performance measure of kappa = 0.651 at gradient boosted number of tress = 80

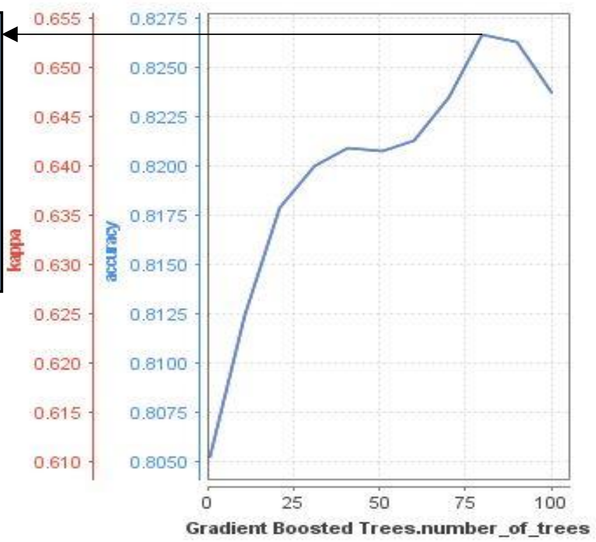


Figure 20 Using optimisation grid to determine set of parameters for best performance

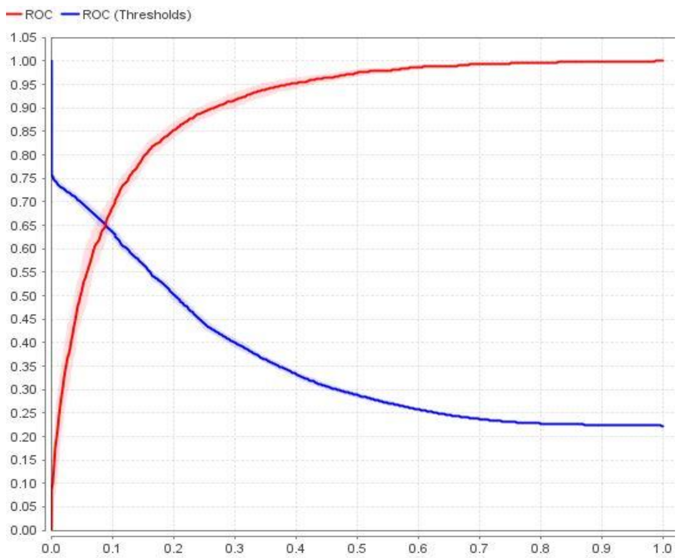


Figure 21 AUC/ROC Curve for final model with AUC=0.899+/-0.009

kappa: 0.651 +/- 0.018 (micro average: 0.651)

Kappa=0.651

	true viable	true unviable	class precision
pred. viable	3867	613	86.32%
pred. unviable	1133	4387	79.47%
class recall	77.34%	87.74%	

Figure 19 Confusion Matrix for final model

Cross Validation performance result