REPORT: A COMPARISON BETWEEN PREDICTIVE ANALYTICS SOFTWARE USING TITANIC DATA

Cyrus Kwan

ID: 45200165 Date: 4/04/2021

CONTENTS

NOTE: Data preparation and analysis sections were performed in the R statistical language. Performing similar analyses of data that has not been passed through an algorithm is assumed to yield the same distribution of attributes and was thus not also performed in Orange data miner.

NOTE: All project files can be found at https://github.com/MQCyrusKwan/BUSA3020-Assessment-2---Predictive-Analysis.git

DATA PREPARATION:

STRUCTURE OF TITANIC DATA:

DATA DICTIONARY:

Variable	Definition	Variable Type	Key
Survived	Survival	Character	Yes, No
Passenger Class	Ticket Class	Character	First, Second, Third
Name	Passenger name	Character	
Sex	Sex	Character	Female, Male
Age	Passenger Age	Number	
No of Siblings or Spouses on Board	no. of siblings / spouses aboard the Titanic	Integer	
No of Parents or Children on Board	no. of parents / children aboard the Titanic	Integer	
Ticket Number	Ticket Number	Character	
Passenger Fare	Passenger fare in £ sterling	Number	
Cabin	Cabin number	Character	
Port of Embarkation	Where passengers boarded	Character	Cherbourg, Queenstown, Southampton
Life Boat	Lifeboat boarded	Character	

SUMMARY OF TITANIC DATA:

```
> list_NA(titanic_data)
[1] "Age"                "Passenger.Fare"     "Cabin"
[4] "Port.of.Embarkation" "Life.Boat"
```

Output 1: list_NA() returns the objects in the data set that contain NA values. Some attributes like 'Life.Boat' are automatically assumed to not have any impact on the outcome of passenger survival.

```
> ncol(titanic_data)
[1] 12
> nrow(titanic_data)
[1] 1309
```

Output 2: Both function calls reveal the dimensions of the data set; 12 columns, and 1309 rows.

DATA CLEANING

IMPUTATION

Used KNN imputation because it is more robust than filling out the NAs with the mean or median which could potentially further skew the distribution of the dataset and reduce the accuracy of models. Note that entire rows were not removed from the data set as there are only 1309. Imputation balances the data set for further use.

ADDING FEATURES

Extracted and combined unique titles of passengers, allowing too many categories could result in minor categories being lost in a single set upon splitting the data. Family size was added as the addition between number of parents, children, spouses, or siblings on board. Finally, dummy variables were created for character variable type data (refer to data dictionary in Structure section) used to parse into models that only accept numeric data. ‘nCabins’ shows the distribution of the number of cabins each passenger had access to.

DROPPING ATTRIBUTES

- Removed ‘Life Boat’ as it is assumed that the passenger survived if they were on a life boat
- 'Life Boat', 'Ticket Number', and 'Cabin' don't seem to have any correlation to survival
- 'Number of Siblings or Spouses on Board', and 'Number of Parents or Children on Board' can also be removed since they appear to have the same correlation as each other which we represented with 'Family Size'
- ‘Names vary’ so much that they cannot be categorized
- Splitting rare titles can lead to model errors and ‘Normal Title’ achieves a similar outcome

COMPARING ANALYSIS USING R STATISTICAL LANGUAGE & ORANGE DATA MINER:

TITANIC ATTRIBUTE DISTRIBUTION:

AGE

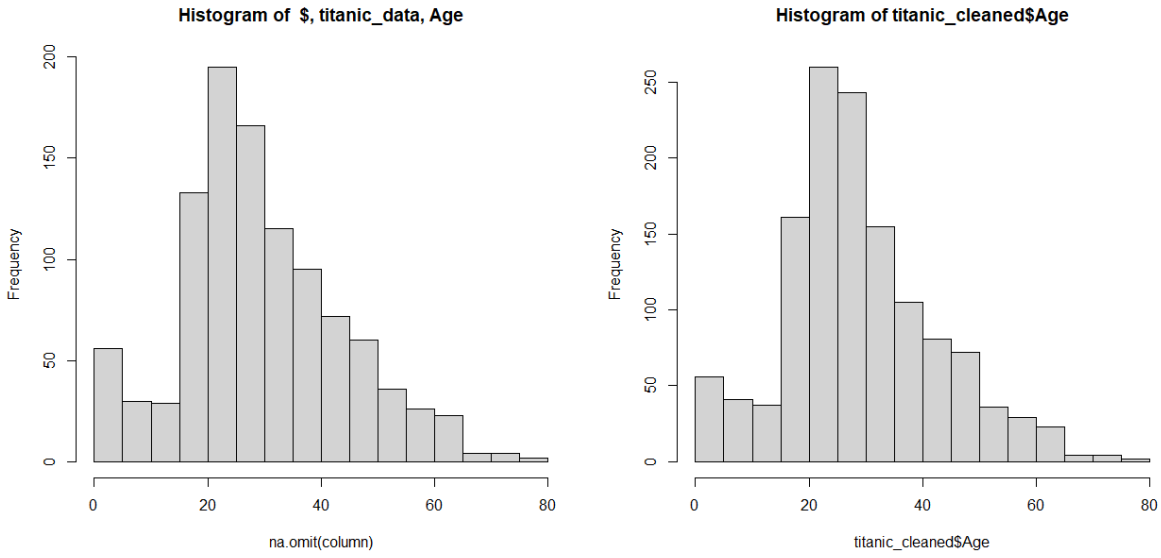

Figure 1: Comparison of 'Age' attribute before(left) and after(right) kNN imputation. Form of distribution stays mostly similar before imputation.

PASSENGER FARE

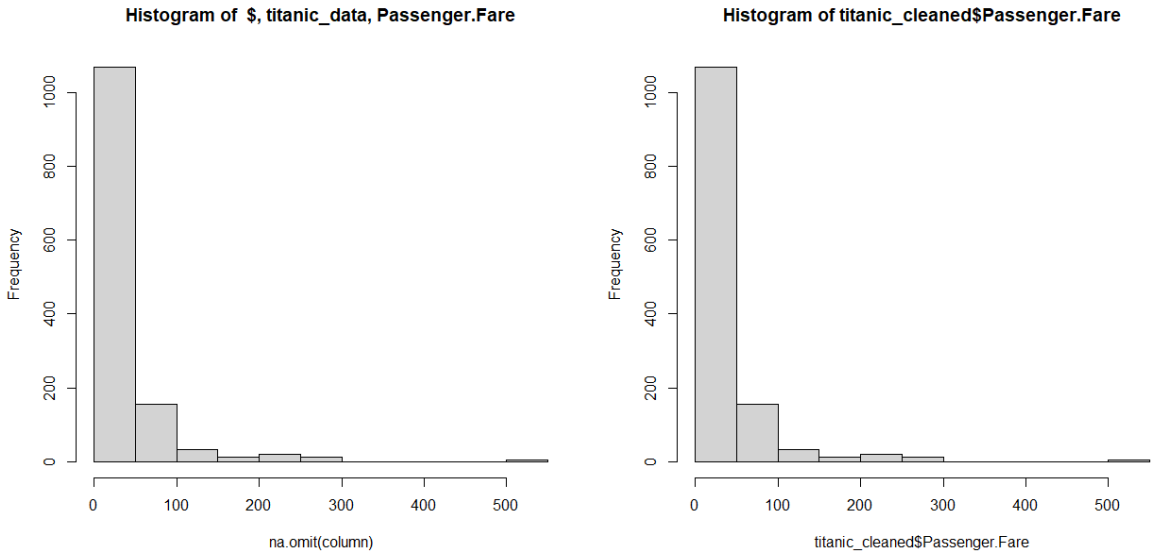

Figure 2: Passenger fare maintains a similar distribution both before(left) and after(right) imputation. Plot shows that most passengers paid less than or equal to £50 to board the Titanic.

PORT OF EMBARKATION

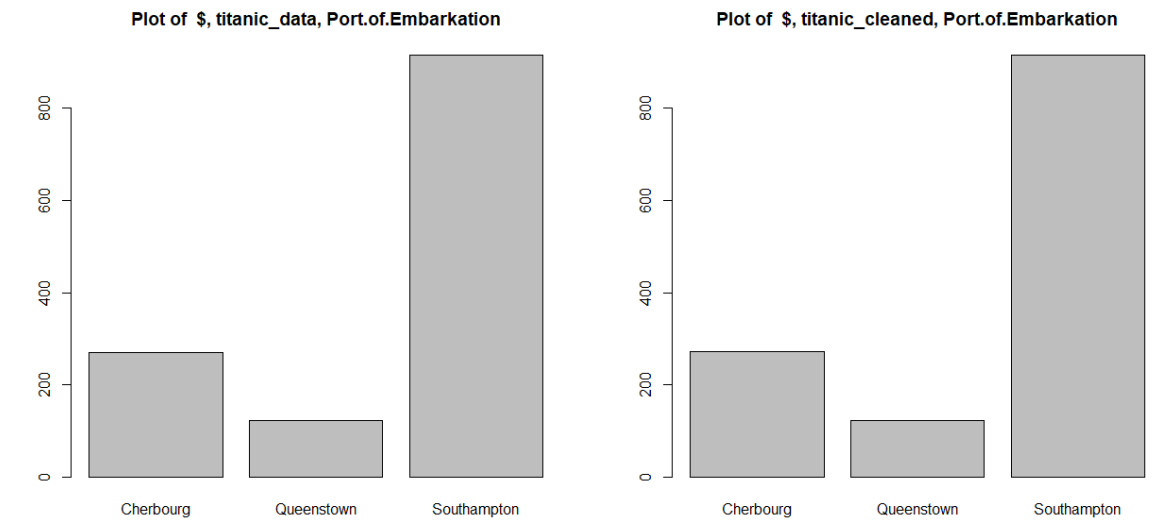

Figure 3: Distribution of port of embarkation remains similar. Most passenger boarded from Southampton, further insights could be produced based on historic and demographic data.

NUMBER OF CABINS

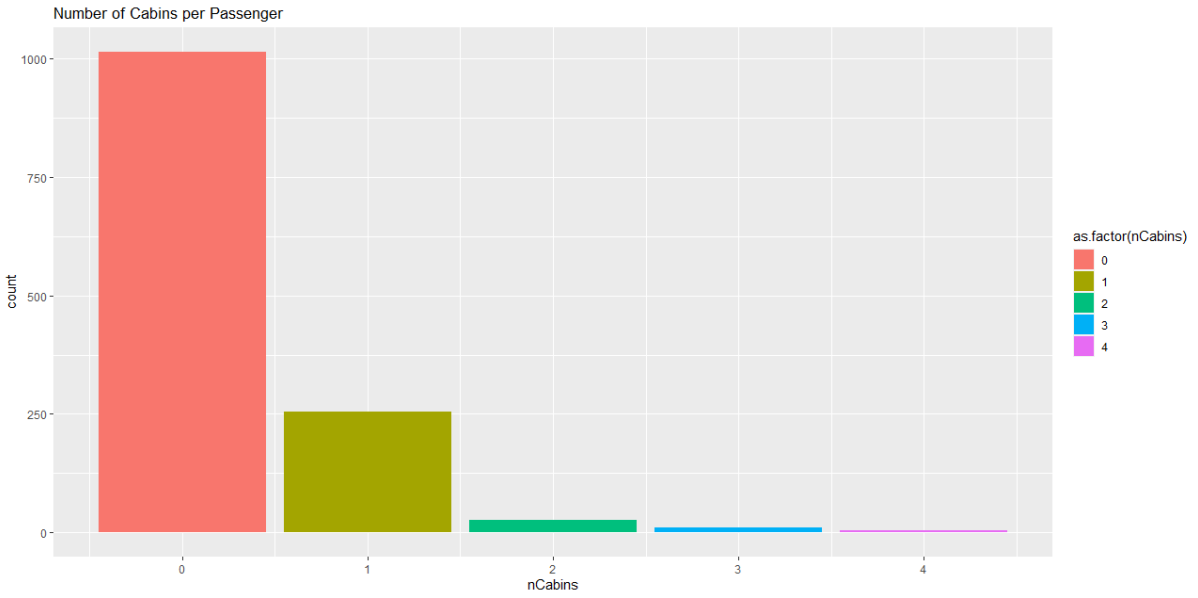

Figure 4: Plot reveals that only the minority of passengers had access to a cabin.

EXPLORATORY ANALYSIS:

AGE GROUP

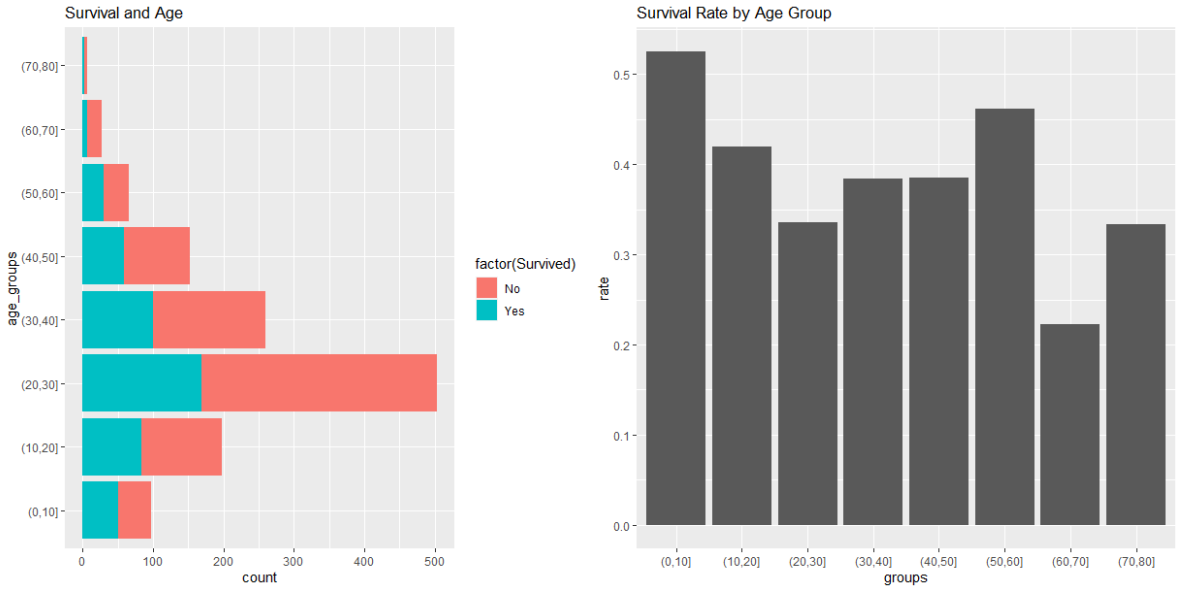

Figure 5: Survival and Age plot shows that passenger age was skewed towards 20-30 year old's. Survival Rate by Age plot shows that children were the most likely to survive.

SEX

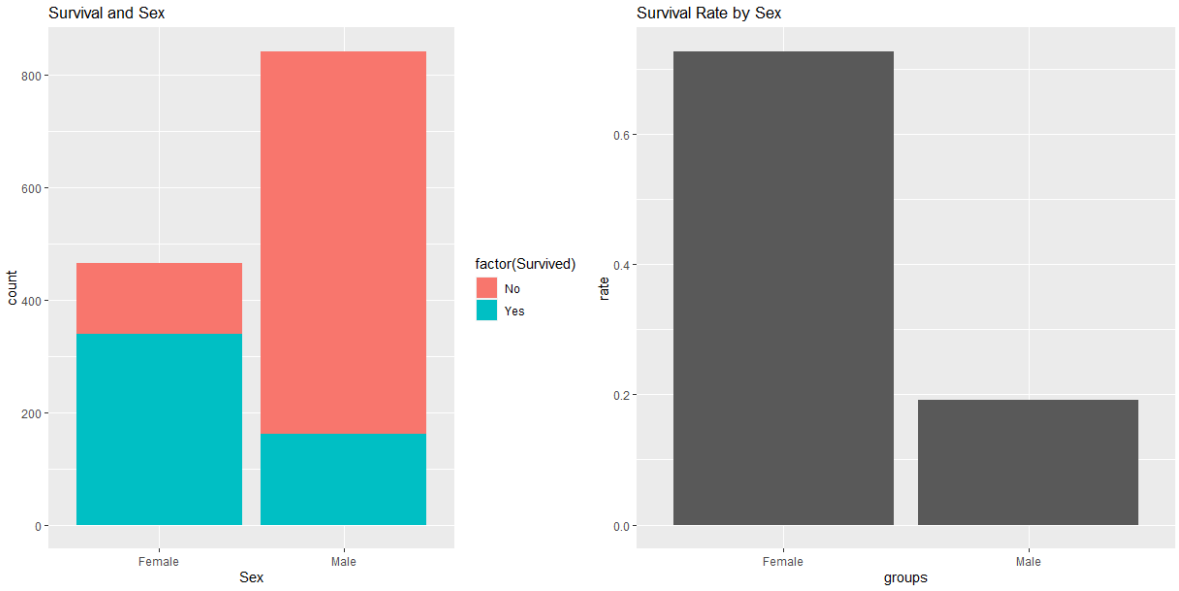

Figure 6: Survival plot of sex shows that women were significantly more likely to survive. Survival rate plot of sex shows the same story with an almost 75% rate of survival while men had a less than 20% chance of survival.

CLASS

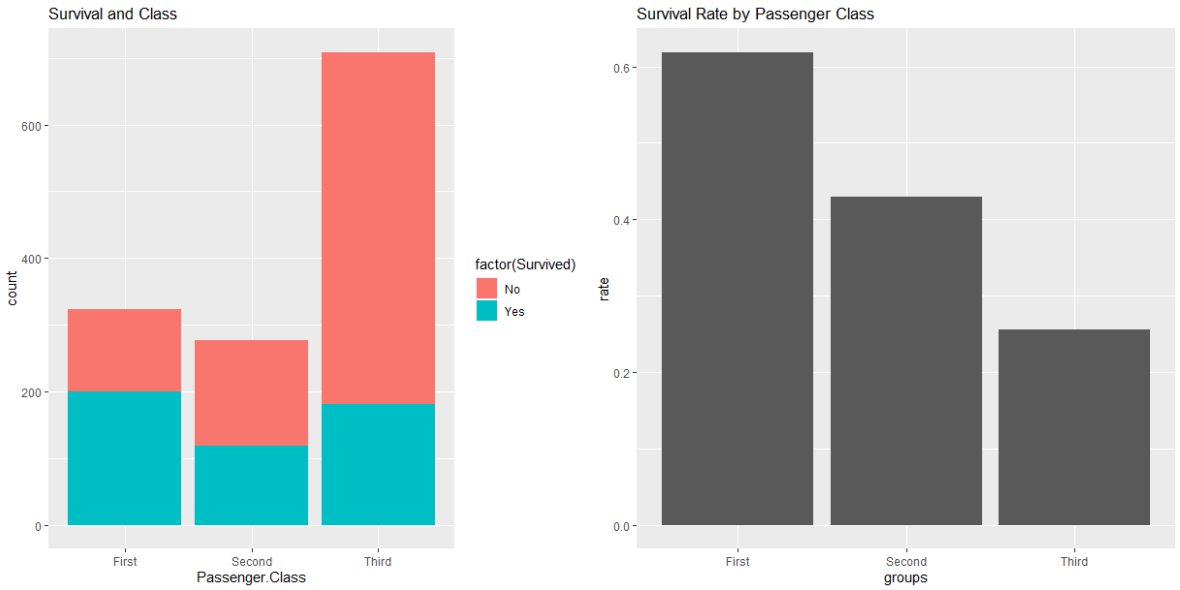

Figure 7: Survival plot shows that most passengers were third class followed by first class. An inference could made ticket prices were low enough that more passengers opted for first class (see figure 2). First class passengers had the highest survival rate and may have had their safety prioritised.

TITLE

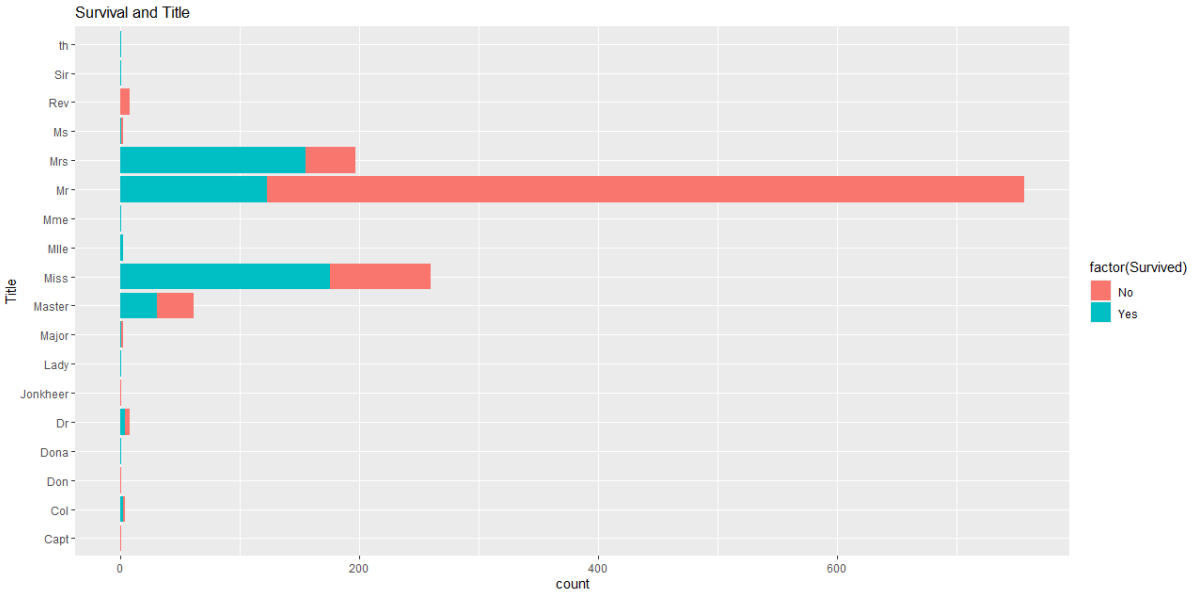

Figure 8: Title distribution shows that over half the passengers on the Titanic were male.

NORMALIZED TITLES

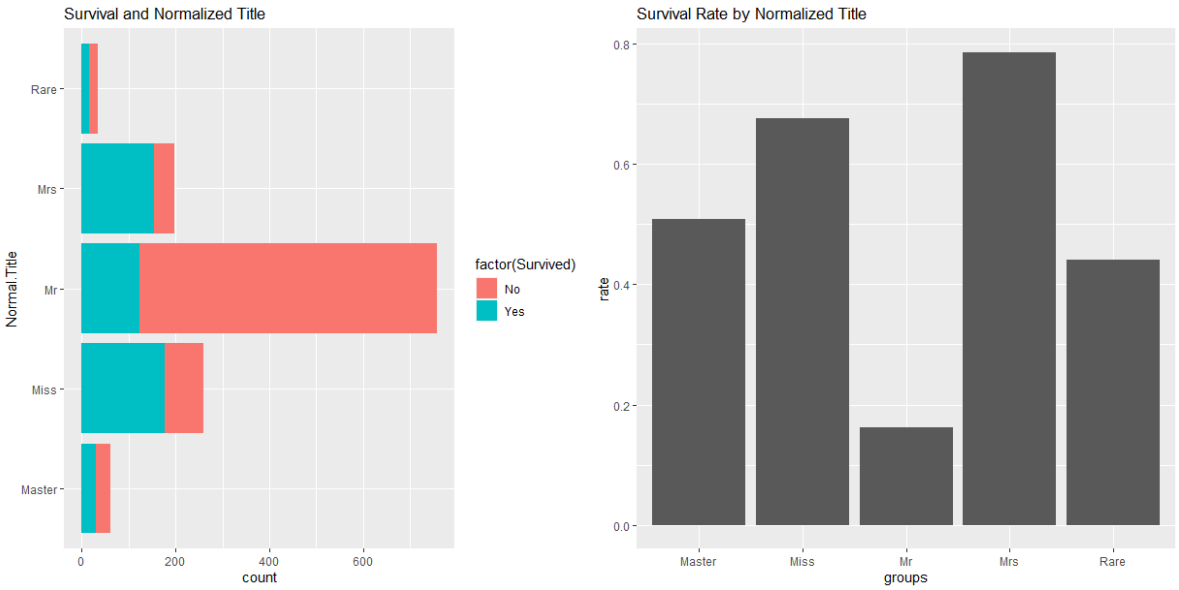

Figure 9: Survival rate shows that the title of 'Master' would significantly increase the likelihood of survival (assuming all masters are male).

FAMILY SIZE

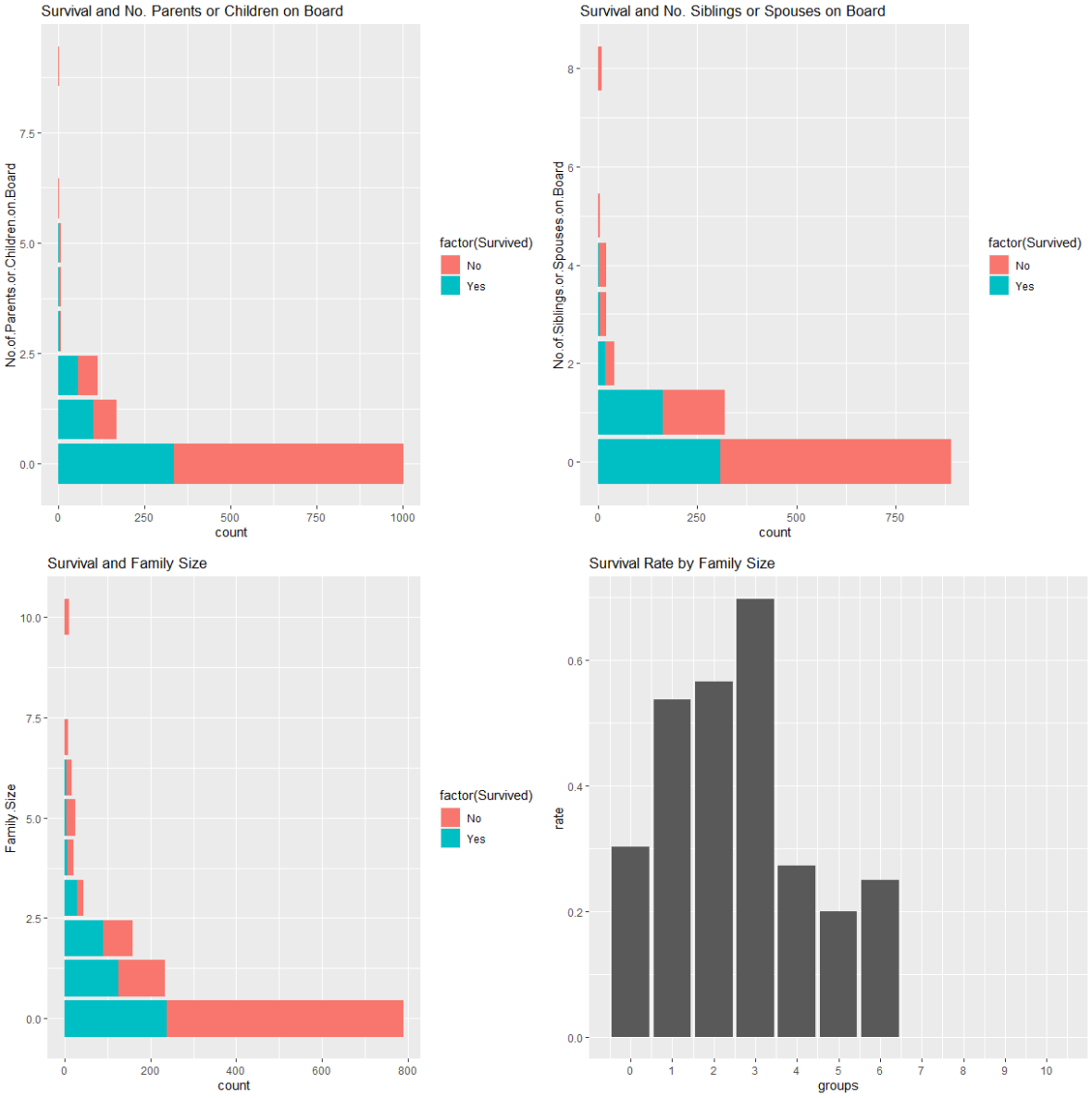

Figure 10: Plots of parents & children, and siblings & spouses appear to show the same trend. Survival plot also shows that family size was skewed towards 0. Survival rate appears somewhat normally distributed about 2-3 family members. An inference could be made that people with fewer than 2-3 family members had less help from others, and those with more had to help more people.

NUMBER OF CABINS

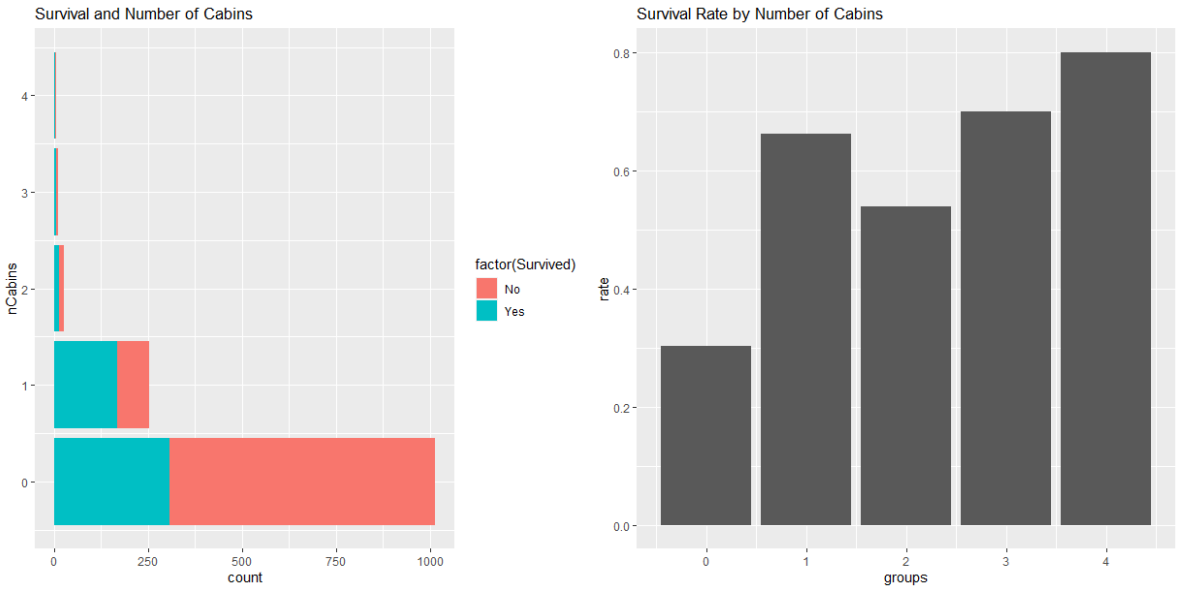

Figure 11: Distribution of cabins per passenger is skewed towards 0. Passengers with more cabins were likely wealthier and had priority to life boats.

MODELLING & EVALUATION:

NOTE: Performance measures will be limited to area under the curve (AUC) and class accuracy which is more appropriate for a balanced data set (see Imputation in Data Preparation > Data Cleaning).

NOTE: Models were chosen from a pool of classification models to predict more appropriately 'Survived'.

RANDOM FOREST

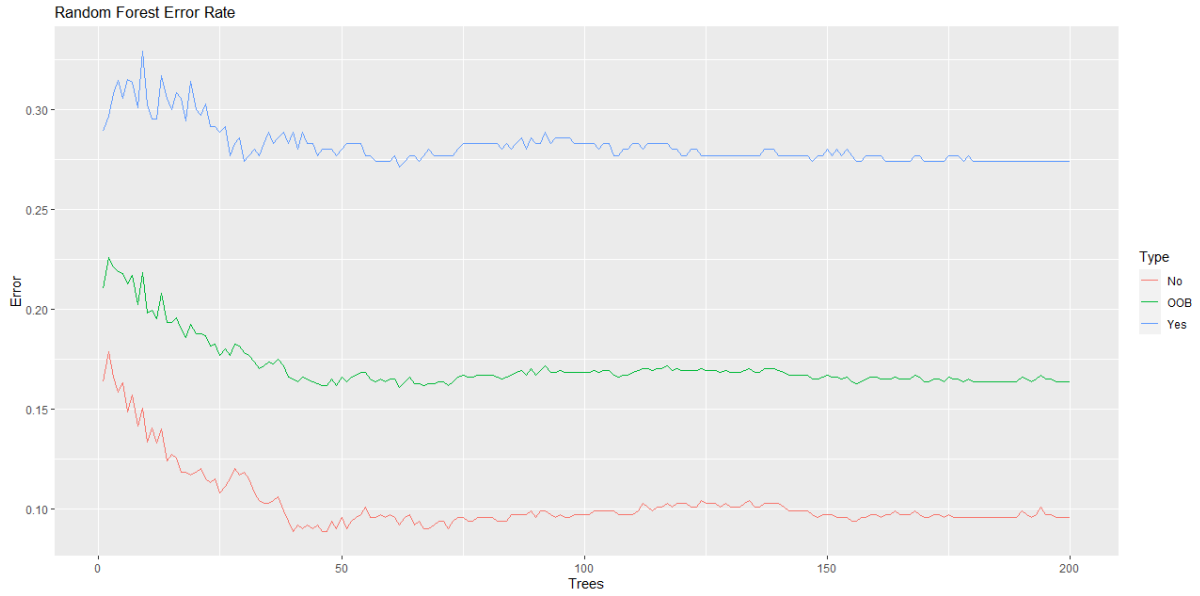

Figure 12: Error plot of random forest model. Accuracy appears to plateau past approximately 100 iterations.

```
     Trees            Type              Error
 Min.   :  1.00   Length:600         Min.   :0.0885
 1st Qu.: 50.75   Class :character   1st Qu.:0.1027
 Median :100.50   Mode  :character   Median :0.1672
 Mean   :100.50                      Mean   :0.1852
 3rd Qu.:150.25                      3rd Qu.:0.2771
 Max.   :200.00                      Max.   :0.3294
```

Output 3: Summary output of errors. Median, 1st , and 3rd quartiles show that the distribution of errors is somewhat skewed likely as the model becomes more accurate over consecutive iterations.

R Random Forest: Confusion Matrix			
	Actual		
Predicted		No	Yes
	No	219	52
	Yes	24	97

R Random Forest: Model Evaluation	
Accuracy	Area Under the Curve
0.806	0.854

Orange Random Forest: Confusion Matrix			
	Actual		
Predicted		No	Yes
	No	215	29
	Yes	49	100

Orange Random Forest: Model Evaluation	
Accuracy	Area Under the Curve
0.802	0.844

MULTINOMIAL LOGISTIC REGRESSION

R Multinomial Logistic Regression: Confusion Matrix			
	Actual		
Predicted		No	Yes
	No	167	77
	Yes	76	72

R Multinomial Logistic Regression: Model Evaluation	
Accuracy	Area Under the Curve
0.610	0.641

Orange Multinomial Logistic Regression: Confusion Matrix			
	Actual		
Predicted		No	Yes
	No	209	35
	Yes	43	106

Orange Multinomial Logistic Regression: Model Evaluation	
Accuracy	Area Under the Curve
0.802	0.829

NEURAL NETWORK

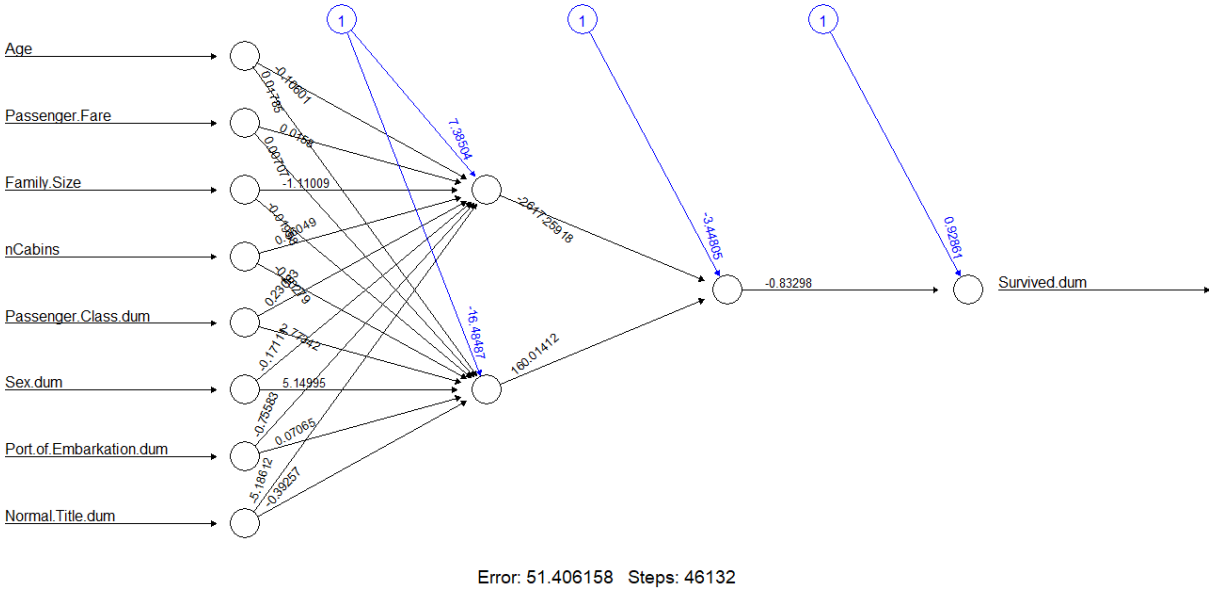

Figure 13: Neural Network plot of titanic data. A layer of 2 and 1 hidden neurons are present where more layers can be added to increase the depth of the network.

R Neural Network: Confusion Matrix			
	Actual		
Predicted		No	Yes
	No	215	56
	Yes	28	93

R Neural Network: Model Evaluation	
Accuracy	**Area Under the Curve**
0.788	0.818

Orange Neural Network: Confusion Matrix			
	Actual		
Predicted		No	Yes
	No	214	30
	Yes	57	92

Orange Neural Network: Model Evaluation	
Accuracy	**Area Under the Curve**
0.779	0.829

ROC COMPARISON

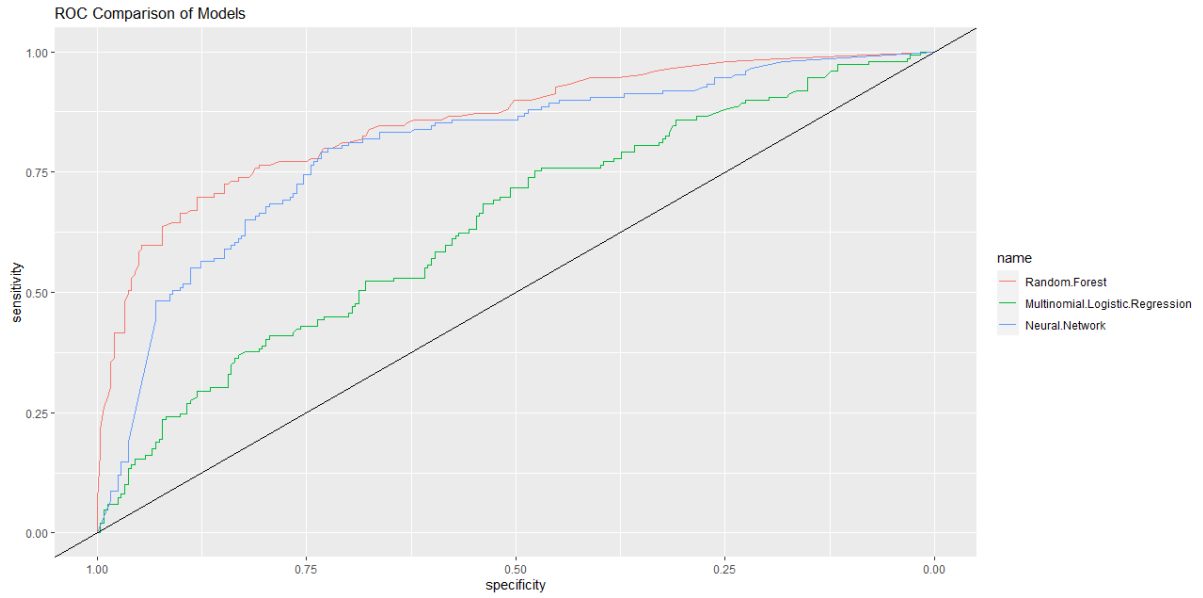

Figure 14: ROC curve produced by models R statistical language. Logistic regression is the worst performing model in this graph with random forest being the best performing model.

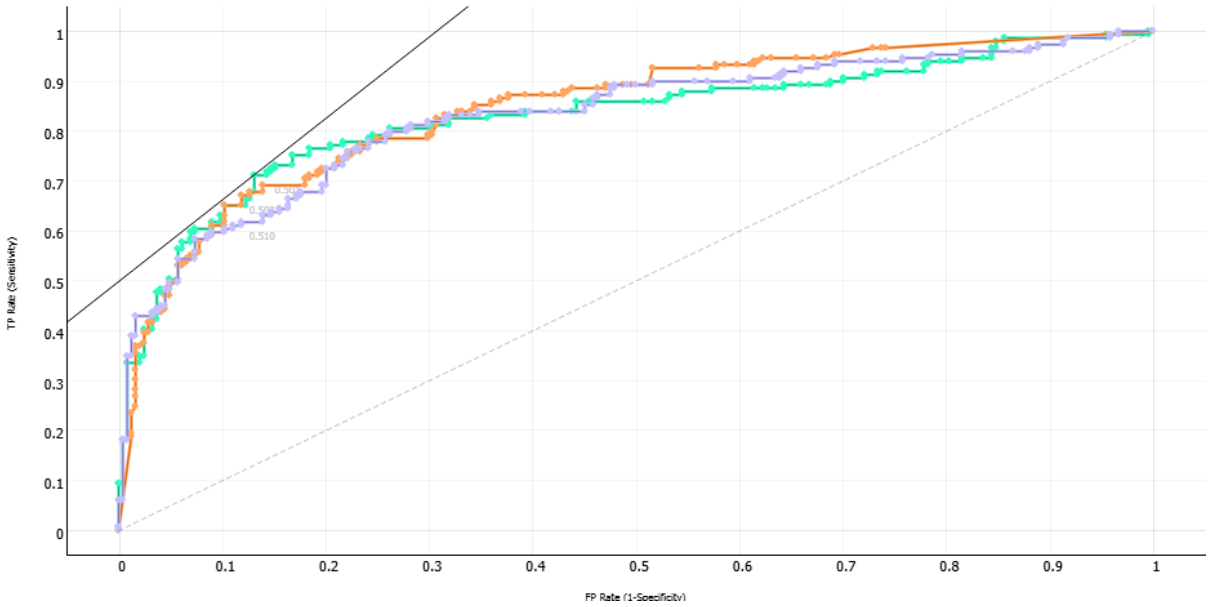

Figure 15: Logistic Regression = Green Random Forest = Orange Neural Network = Blue Logistic regression and neural networks perform significantly better compared to R.

RECOMMENDATION:

EVALUATION:

Rating table of R Statistical Language and Orange Data Miner out of 10								
	Compatibility	**Documentation**	**Ease of Use**	**Flexibility**	**Interpretation of Output**	**Learning Curve**	**Speed/ Workflow**	**Utility/Use Cases**
R	9	9	5	8	7	3	3	7
Orange	4	5	8	6	8	9	9	5

Total Rating: *R = 51* *Orange = 54*

R STATISTICAL LANGUAGE:

The R language has very concise and quickly accessible documentation. Additionally, it has a very active user base of which can be easily consulted for isolated problems. As a programming language, R naturally has more utility for use cases outside of just predictive modelling; learning the R language would be an asset to any organisation. In contrast, R has a very steep learning curve compared to other predictive modelling software. As such, workflow is considerable slower compared to its competitors and can be difficult to use for unpractised users.

ORANGE DATA MINER:

Orange is a very user-friendly predictive analytics software. It has consistency across its models and is easily interpretable. Orange benefits from a quick workflow and can be learned very quickly by untrained users. Unfortunately, it is hard to transfer output to reports outside of Orange and many forms of output are limited by the software i.e. errors, plots of certain models. In summary, Orange performs the job of predictive analytics modelling consistently well and is easy to pick up.

CONCLUSION:

Management should consider the scope of the project before implementing either software package into their organisation. If the task covers more than just predictive modelling and requires a lot of flexibility for highly capable staff, then it would be more appropriate to use R. Otherwise, for organisations whose only objective is to accurately use predictive modelling on simpler data then Orange should be implemented. Alternatively, organisations could use R to finely manipulate data sets and use Orange for its more accurate models.