



MACQUARIE
University
SYDNEY • AUSTRALIA

BUSA3020 ADVANCED ANALYTICS TECHNIQUES

ASSIGNMENT #2: PREDICTIVE ANALYTICS TITANIC ANALYSIS & SOFTWARE RECOMMENDATION

Name: Chris Jerylle Vargas Oidem

Student ID: 45476624

Due Date: Friday, 17 April 2020, 11:55pm

TABLE OF CONTENTS

1. Predictive Analysis of Titanic Data Set.....	1
1.1 R Statistical Programming Language.....	4
1.2 Orange Data Miner.....	7
2. Comparison of the Two Programs.....	7
2.1 Ease of Use.....	8
2.2 Flexibility.....	8
2.3 Learning Curve.....	8
2.4 Features.....	8
3. Conclusion.....	8

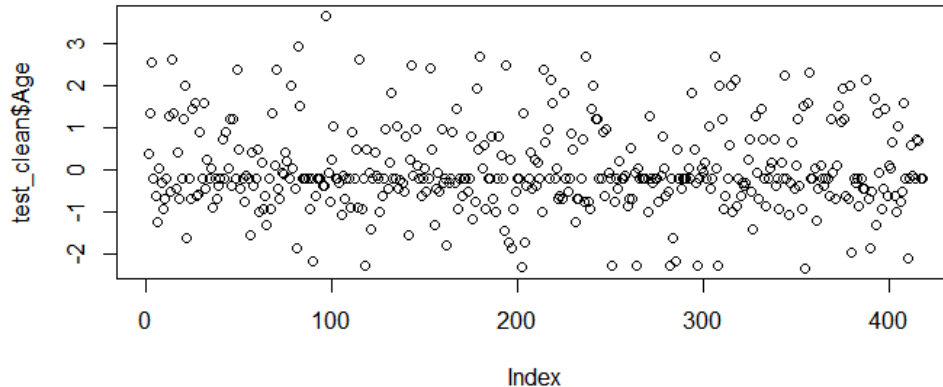
1. Predictive Analysis of Titanic Data Set

The Titanic data analysis was conducted using Logistical Regression, Decision Tree, and Support Vector Machine (SVM).

Before data manipulation:

Variables	No. of Missing Values in Dataset
Survived	0
Passenger Class	0
Name	0
Sex	0
Age	263
No of Siblings or Spouses on Board	0
No of Parents or Children on Board	0
Ticket Number	0
Passenger Fare	1
Cabin	1014
Port of Embarkation	2
Life Boat	823

Missing values for the Age variable were replaced with a median value of 27. The rest of the variables with missing values were replaced with "NA".



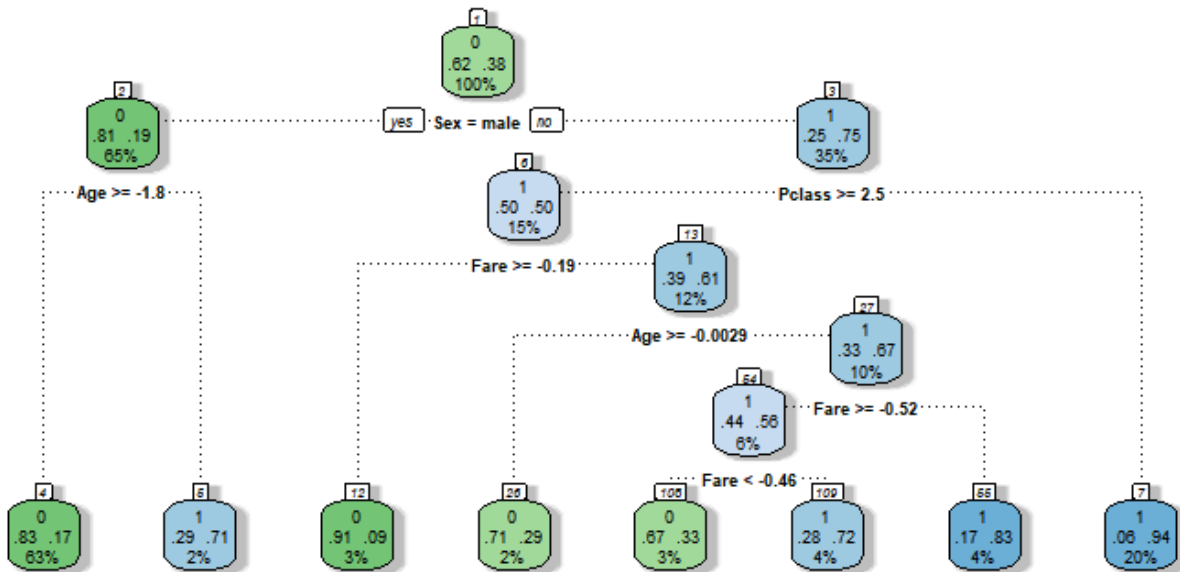
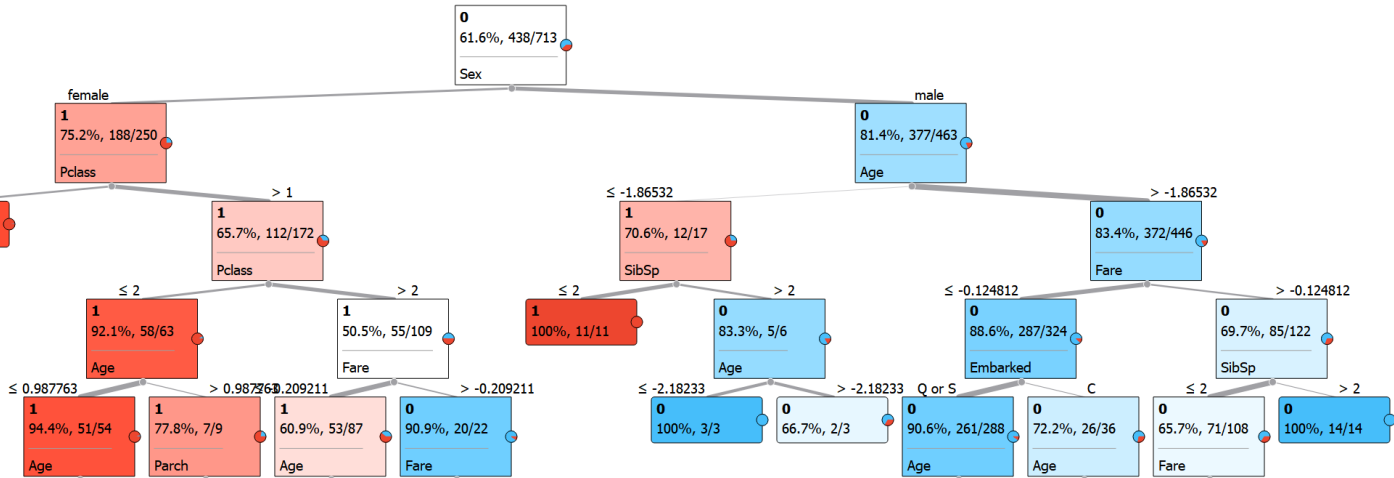
After dropping variables that were not beneficial to the model, the following remains for the analysis:

- Survived
- Passenger Class
- Sex
- Age
- No of Siblings or Spouses on Board
- No of Parents or Children on Board
- Passenger Fare
- Port of Embarkation

In comparing the performance of the various models, data have been split into datasets consisting of 80% training and 20% testing.

The below tables are a summary of the bar graphs, box plot, and decision tree that were run in both RStudio and Orange.

	R Statistical Programming Language	Orange Data Miner
Survival Rate by Gender		
Survival Rate by Passenger Class		
Survival Rate by Passenger Age		
Survival Rate by Age and Sex		

Software	Decision Tree Output
R Statistical Programming Language	 <p>Rattle 2020-Apr-05 13:45:43 Chris Jerylle</p>
Orange Data Miner	

1.1 R Statistical Programming Language

Further analysis had been run in RStudio, and findings were presented below:



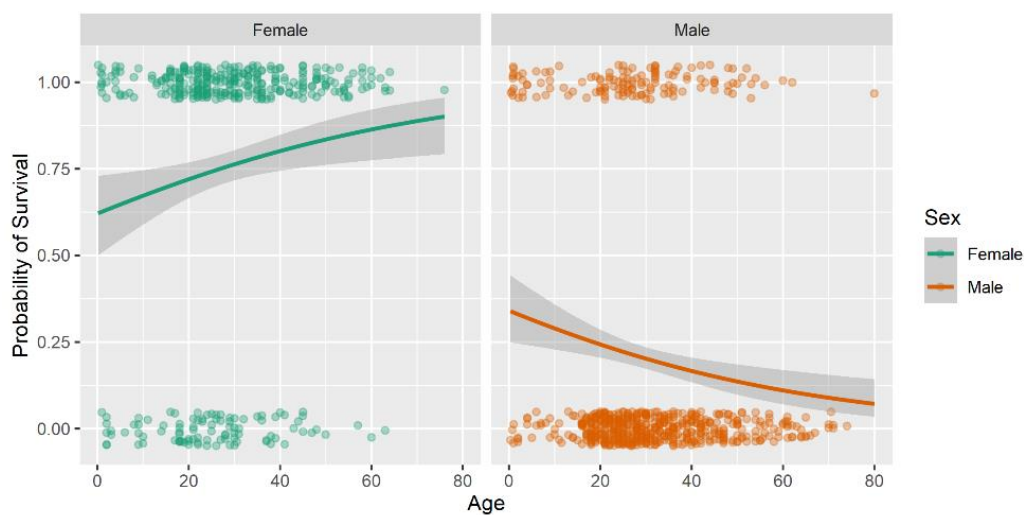
Decision Tree

	DecisionTree_Prediction	
	0	1
0	95	15
1	25	43

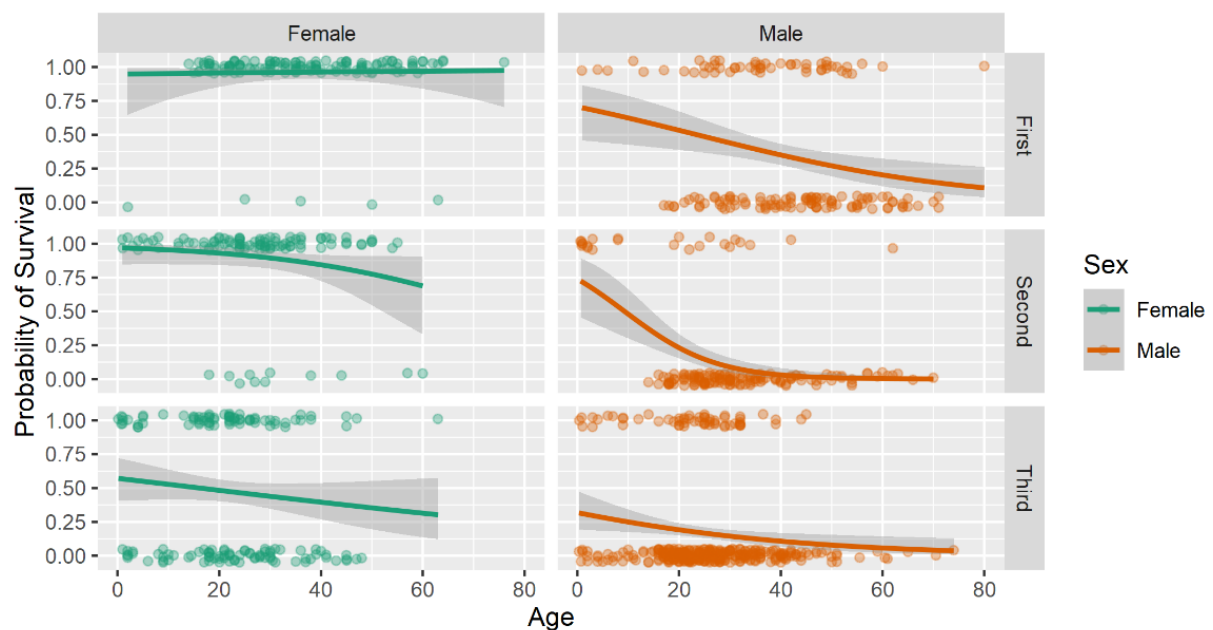
The prediction success rate for the decision tree is 77.53% (63.24% survival and 83.36% non-survival).

Logistic Regression

Probability of survival as a function of age for female and male passengers:



Extended logistic regression model to consider how sex, age, and class affects survival on the Titanic:



The output is as follows:

```
Call: glm(formula = Survived ~ ., family = binomial, data = training_set_clean)

Coefficients:
(Intercept)      Pclass      Sexmale      Age      SibSp
  4.20722      -1.07942      -2.73679      -0.51166      -0.27968
  Parch      Fare      EmbarkedQ      EmbarkedS
 -0.04511      -0.03024      -0.02183      -0.57093

Degrees of Freedom: 710 Total (i.e. Null); 702 Residual
(2 observations deleted due to missingness)
Null Deviance: 946.1
Residual Deviance: 628.9      AIC: 646.9
```

	LogisticRegression_Prediction	
	0	1
0	90	20
1	21	47

The prediction success rate for the logistic regression is 76.97% (69.12% survival and 81.82% non-survival).

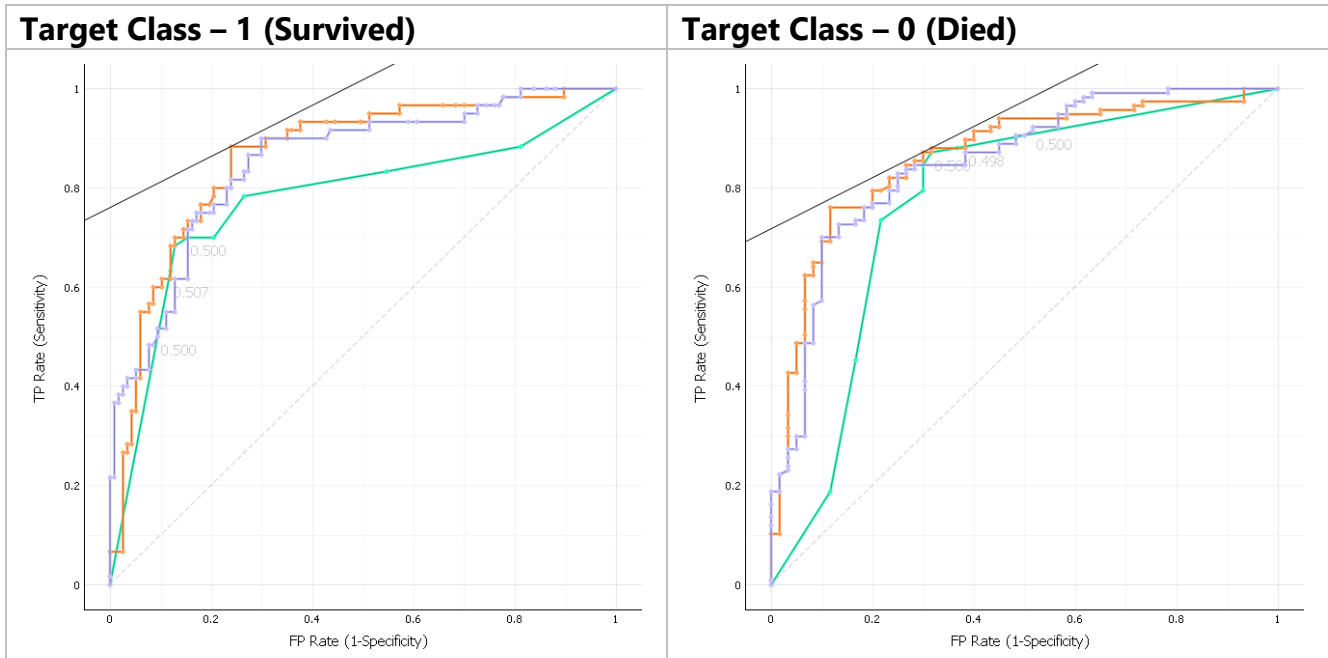
Support Vector Machine

	SVM_Prediction	
	0	1
0	91	19
1	23	45

The prediction success rate for the SVM is 76.40% (66.17% survival and 82.73% non-survival).

1.2 Orange Data Miner

Receiver Operating Characteristic (ROC) curve of three predictors – Tree, Logistic Regression, and SVM – of survival in Titanic:



Precision Accuracy Comparison

	R Statistical Programming Language	Orange Data Miner
Decision Tree	77.53%	80.60%
Logistic Regression	76.97%	79.30%
Support Vector Machine	76.40%	76.40%

The precision accuracy percentage is conclusive of a consistent result from both R and Orange. The Decision Tree yields the highest percentage, while the SVM has a similar result of 76.40% in both programs.

2. Comparison of the Two Programs

2.1 Ease of Use

Orange is easier to use, visualize concepts, and process data compared to RStudio. Orange uses a drag and drop functionality to conduct analysis, and workflows are created by linking predefined components called widgets, requiring minimal input from the user.

2.2 Flexibility

Data manipulation – such as creating data subsets from existing data, calculating columns, and performing statistical modelling – is more straightforward in RStudio. Once the script has the analysis written, it becomes simpler to maintain, edit, and run repeatedly.

2.3 Learning Curve

RStudio has a steeper learning curve compared to Orange. RStudio requires using its programming language to operate the program. The user also needs to dedicate additional time to learn about the various packages available.

2.4 Features

Both programs have an open-source environment with a great community. RStudio has more features and is equipped with a set of packages to perform data mining. Other packages such as '*dplyr*', '*lattice*', and '*ggplot2*' are available for download and installation.

3. Conclusion

Software	Pros	Cons
R Statistical Programming Language	<ul style="list-style-type: none">▪ R User Community▪ Open source▪ The interface is clear and customizable▪ Flexibility to generate reports▪ Readily available packages▪ Able to handle copious amounts of data	<ul style="list-style-type: none">▪ Finding packages can be time-consuming▪ Mandatory to include libraries to achieve proper output▪ Steeper learning curve
Orange Data Miner	<ul style="list-style-type: none">▪ Open source▪ User/Beginner-friendly▪ Drag & drop functionality▪ Lesser training time▪ Interactive data visualization	<ul style="list-style-type: none">▪ Debugging is more challenging▪ Lack of documentation, tutorials, and community forums

The summary above shows RStudio being a better program to adopt for future predictive analysis tasks. It is a free and open-source software for statistical analysis, machine learning, and data visualization supported by a strong online community.