MGMT 320 Assignment 2

Predictive Analytics Report

Aashish Agrawal
(SID- 44009240)

Task 1:

Introduction

This report looks into the Titanic data set and performs alternative predictive analytics for predicting survivals, selecting the better results. The report uses orange and rapidminer as statistical softwares in order to make inferences. It compares the predictive outputs from the two softwares and provides an opinion on which software is more preferable.

The first and foremost thing was to clean the data so that we can get more accurate results. In order to do so we used excel and found that variables like lifeboats, ticket number, name and cabin number was not useful for analysing our prediction, so we removed the variables. There were quite a few missing values for age of passengers so we used the median age of the passengers for the missing values. Since it is not good to test a model with the same data that we used to create the model, and in this particular case there was only one titanic, hence no other alternative data, we divide original data into training data and test data in proportion of 0.7 and 0.3 respectively.

Output Review

Reflecting on some of the outputs below, in specific if we look at the confusion matrix, we see that SVM analysis and Random Forest provide the best results in terms of predictive analytics solution for predicting survival for orange and rapidminer respectively. Out of the people who actually survived, 92.70% of the prediction was accurate when we used SVM analysis in orange and similarly 78.40% of the prediction was accurate when we used Random Forest in rapidminer.

SVM and Random Forest predictive analysis are best results produced by their respective softwares which is supported by the ROC Analysis graph. If we look at the one produced by orange, we see the light green line which is SVM touches the tangent line on few occasions, certainly the reason producing the best result. For instance, the optimal point we would like to be at could be where true positive rate is approximately 0.75 with a 0.18 false positive rate. Similarly for Random Forest, ROC analysis shows the optimal point at a true positive rate 0.82 with a false positive rate of 0.27. These points could be different depending on what we are trying to figure out.

We would ideally like the recall measure to be higher for SVM analysis in orange and for Random Forest in rapidminer, which are 0.706 and 0.715 respectively. Both results are lower than expected when compared to other predictive analytics used within the softwares namely decision tree, naive bayes and logistic regression.

Test & Score

Orange

Method	AUC	CA	F1	Precision	Recall
Tree	0.733	0.765	0.756	0.777	0.765
SVM	0.684	0.706	0.670	0.778	0.706
Random Forest	0.853	0.780	0.778	0.780	0.780
Naive Bayes	0.828	0.752	0.750	0.752	0.752
Logistic Regression	0.867	0.786	0.782	0.790	0.786

Rapidminer

Method	AUC	CA	F1	Precision	Recall
Tree	0.758	0.779	0.715	0.741	0.696
SVM	0.817	0.784	0.727	0.738	0.723
Random Forest	0.847	0.800	0.739	0.771	0.715
Naive Bayes	0.832	0.786	0.711	0.760	0.670
Logistic Regression	0.847	0.781	0.737	0.706	0.771

Confusion Matrix

Orange

SVM Analysis

		Predicted		
		No	Yes	
Actual	No	66.20%	7.30%	184
	Yes	33.80%	92.70%	143
		272	55	327

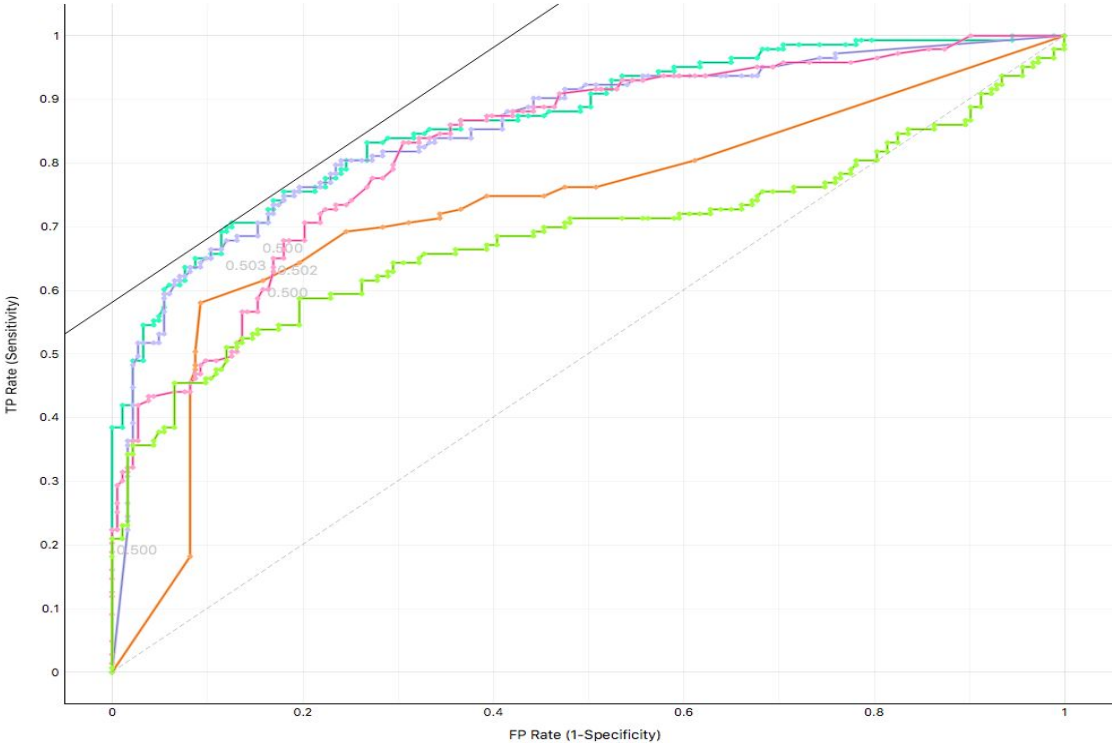
Rapidminer

Random Forest

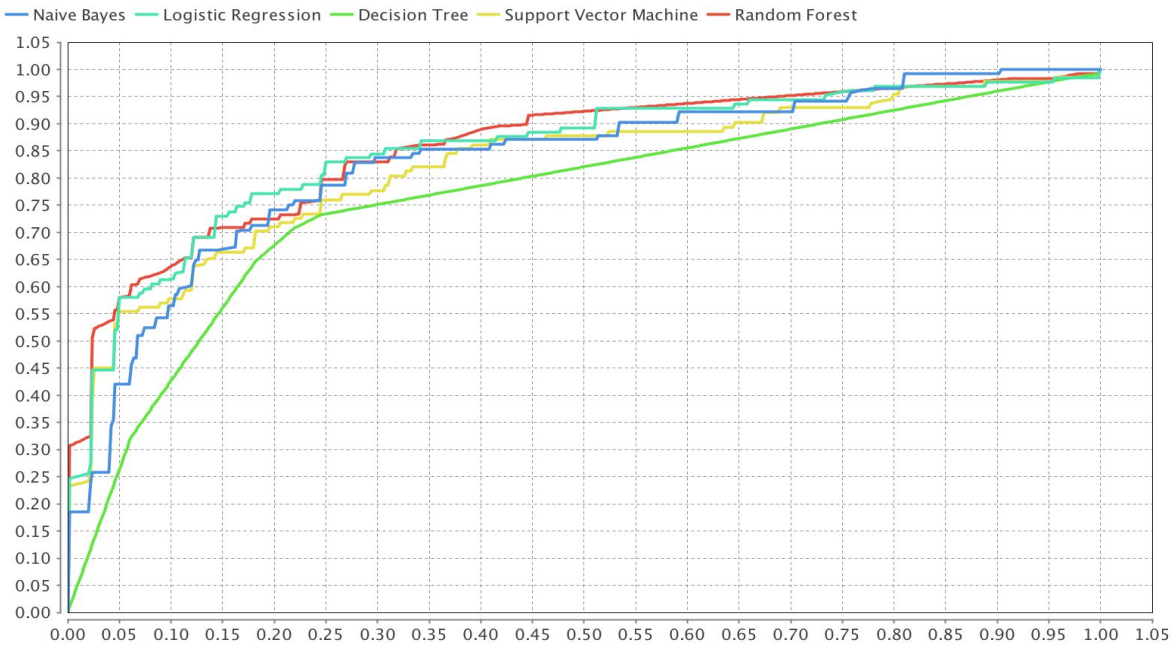
		Predicted		
		No	Yes	
Actual	No	77.70%	21.60%	184
	Yes	22.30%	78.40%	143
		202	125	327

ROC Analysis

Orange



Rapidminer



Task 2:

Comparison between software used for analysis

	Orange	Rapidminer
Ease of use	User friendly interface. Drag and chose different methods, visualization and test tools.	Has a GUI to design and execute analytics workflow. In particular the four click option by automodel is so convenient and efficient, the quickest way to get results without losing the effectiveness of data..
Flexibility	Allows users to add additional predictive analytic tools other than the ones already available.	Users can use only the built in tools.
Integration with outside data sources	Comparatively integration with outside data less friendly when compared to rapidminer. Would result in more accurate analytical predictions if data is clean and variables easy to interpret.	Strong built in algorithms makes integration with outside data easy and even if data set is not that clean, rapidminer takes those factors into considerations and provide really good results.
Interpretation of output	Interpretation of output produced by orange is simple till the time you have basic understanding of statistical terminology. Output are produced efficiently by simply dragging tools and connecting each tool to one another.	In terms of interpretation of output, the outputs are interpreted and produced in quite a similar manner as they are in orange. But it is comparatively more convenient to interpret them in orange than in rapidminer since while doing predictive analysis, the interpretation of confusion matrix seemed to be confusing in rapidminer but quite straightforward on orange. But, also to note is the fact that rapidminer provides more graphical displays for small little details in

		comparison to orange.
Price/License	Free Software (GPL)	Freeware Various fee-based versions
Programming language	Software Core: C++ Extensions and query language: Python	Java

Recommendation

I would recommend the client to use orange for future predictive analytics tasks. That's not because rapidminer is a bad software but because orange helps provide better analytical results in comparison to rapidminer as seen in this data set. Also, it is easier to interpret the output produced by orange, since we can see above when we analysed the ROC there was no tangent line produced my rapidminer which makes it hard to figure out the optimal point and not only that but the confusion matrix is hard to interpret as well, the one produced by rapidminer. For the above scenario, I had to make some changes to the confusion matrix produced my rapidminer in order to make it consistent with the output produced by orange, so that comparisons between the two softwares can be made.

Orange seemed to be more user friendly than rapidminer which required more initial assistant specially if someone from a non-statistical background was working on it. Additionally, orange allows you to play around with additional analytical tools other than the ones already available and make different inferences about the results while rapidminer does not have such a feature.

The graphs, confusion matrix table was much easy to compare in orange since they displayed the values of all predictive analysis on the same page. For instance we got the f measure, recall measure, precision measure for decision tree, random forest, naive bayes, logistic regression and SVM all in the same page but in rapidminer they just gave f measure for all the above test and then again we had to choose recall measure which would then again give values for all test, which made the interpretation and comparison harder.

To conclude, orange would be more suitable for future predictive analysis and would be my choice to use for interpreting the titanic data set.