ST3189 MACHINE LEARNING COURSEWORK REPORT

CONTAINER PRE-MARSHALLING PROBLEM &

BLOOD TRANSFUSION SERVICE CENTRE PROBLEM

Candidate Number: 200664383

Content Page

1. Introduction……...2

2. Problem Definition ..……2

3. Research Objectives……3

4. Machine Learning Methodologies……3

5. Data Processing……4

5.1 Container Pre-marshalling Problem……4

5.2 Blood Transfusion Service Centre Problem ……7

6. Conclusion……..8

7. Proposed Further Research ……..…9

1 Introduction

The container pre-marshalling problem (CPMP) is one of the important and hard to make decisions in modern container terminals that has a direct impact on vessel berthing times and hence on the performance of the container terminal (Gheith, 2012). The CPMP leads to larger problems like in maritime transport. Maritime transport is important for international trading, and it is responsible for 90% of goods in the world, with 60% of it being transported by containers (Bhonsle, 2022).

Blood donation is very essential in our lives as blood is needed every second of each day to save lives as it is very easy to lose blood due to accidents occurring, requiring fresh blood which there is a great demand for. Tracking donors will allow for easier identification in the future, and thus using the blood donation dataset, we are able to predict whether donors will donate again in the future.

2 Problem Definition

This real-world issue needs to be tackled as it requires the re-ordering of the containers during off-peak periods so that containers can be taken from the terminals efficiently during peak season when it gets busy. Thus, using the database we have on the CPMP, we will use different predictive modelling frameworks to analyse and to provide suggestions to improve the efficiency of container retrieval from the port, improving vessel berthing and port operations as a whole which may be applied to the various industries the CPMP can be found in.

The real-world problem of blood donation is also of importance because it is an essential in medical institutions like hospitals worldwide. In a world so unpredictable like ours today where virus outbreaks may occur any moment, blood donors and blood banks are critical in maintaining a healthy blood bank level so that in times of need there will be blood stock ready to be used.

3 Research Objectives

Figuring out the important variables that lead to the real-world problems we face today is important and thus, there are some research objectives that can be identified for us to dive deeper to understand these real-world issues even better.

Firstly, we can find out the variables that are highly related to the cause of mis-overlays in container terminals and what causes them to occur. Some possible variables identified to affect vessel berthing times in the dataset is the percentage of over stowing of the containers, number of stacks, tiers and number of containers there is in the left hand side of the container stack in the bay. Finding out the level of correlation of each variable and how relevant they are to the main target variable– amount of time taken to load the vessels, will allow more suggestions and solutions which are more targeted and efficient.

Secondly, we can also find out the variables highly related to whether blood donors will donate blood again, which is critical for medical institutions like hospitals and clinics as blood is always needed to save lives.

4 Machine Learning Methodologies

In this research report, unsupervised learning techniques– Principal Component Analysis(PCA) and K Means Clustering(KMC) are observed for the analysis of respective real world problems. PCA and KMC help to reduce the complexity of the variables present in the data, which aids the following regression and classification prediction models used to conduct prediction and analysis to help us understand on a deeper level why and how to solve or make improvements to the current problems we face which may impact our daily lives. As all data collected on real-world problems vary from data to data, the results of the prediction models used differ greatly and it requires numerous trial and error for the best results and analysis to the research questions identified.

Classification and Regression Trees (CART), randomForest and linear regression is used on the CPMP dataset to predict the time taken for loading and unloading of containers, while using CART and logistic regression to predict the categorical feature of whether the donor will donate blood in the future, comparing the results from different models to see which fits and represents the data more accurately.

In the following paragraphs, there would be mentions of some terms used to indicate the performance of the models used to process the data:
- R-squared (RSQ) value measures the goodness of fit or the best-fit line of the model
- Mean squared error (MSE), represents the quality of a regression model (Kumar, 2022)
- Accuracy, used to evaluate classification models, is the fraction of predictions the model got right,
- Precision, is the proportion of positive identifications that is deduced correctly,
- Recall, is the proportion of true positives identified correctly (Google Developers, 2022)
- F1 score, measures a model's accuracy by computing number of times a model made a correct prediction across the entire dataset (Kundu, 2023).

5 Data processing

There are some questions for us to consider when wanting to solve or try to improve real world issues by finding out what factors contribute to the problems.

5.1 Container pre-marshalling problem

To improve the vessel berthing times– the time taken for vessels to reach dock and unload the containers:
Q1: Which variables are the most important in predicting the amount of time taken for the vessel to berth?
Q2: How else besides predicting the time taken(runtime) for vessel berthing based on different predictor variables, can we aim to improve the performance of container terminals?

Using CART results in a decision output tree which shows how significant each predictor variable is to the target variable. The predictor variables used to predict are stacks, tiers, number of overstowed containers in one stack(overstowing.stack.pct), percentage of empty stacks and the average density of the leftmost stack of the container terminal. The most significant variable, making the first split in the tree was the left stack density, however the RSQ value (0.219) was very low and MSE value (1965684) was high, indicating that the model did not fit the data very well. PCA–to decrease the complexity of the data for the model to perform better, was then run before running the data through CART again. The PCA results showed that the number of overstowed containers have a positive linear relationship with the runtime. The principal components (PCs)are shown below.

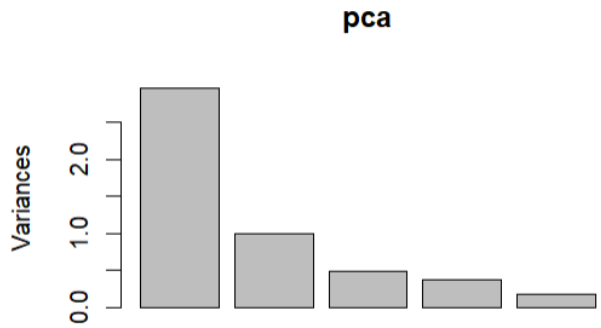

Figure 1: PC1 to PC5 from left to right

The figure above shows the decreasing variance of the 5 PCs, where the variable overstowing.stack.pct has the highest contribution to PC1, showing that it has the strongest relationship. However, despite it being the highest contributor at 0.534, the other variables had similar values of -0.5(stacks), 0.44(tiers), -0.49(empty.stack.pct) and left.density was the highest contributor for PC2 at -0.94. Thus, this shows that these 5 variables are important in predicting the amount of time taken for the vessel to berth.

However, after modelling the variables again after PCA with CART, the results were as shown below.

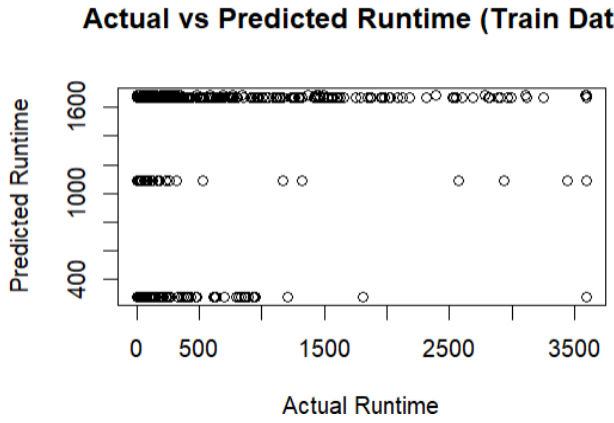

Figure 2: Actual vs Predicted runtime

Figure 2 shows how the actual runtime does not match the actual predicted runtime values, with MSE value (2134897) still as high and RSQ(0.1499) showing that the model may not be a good fit for this dataset.

However, using another prediction model, randomForest– showed some promising results, with a lowered MSE (228206) and high RSQ(0.914), indicating a strong correlation between the 5 variables and the time taken for a vessel to berth.

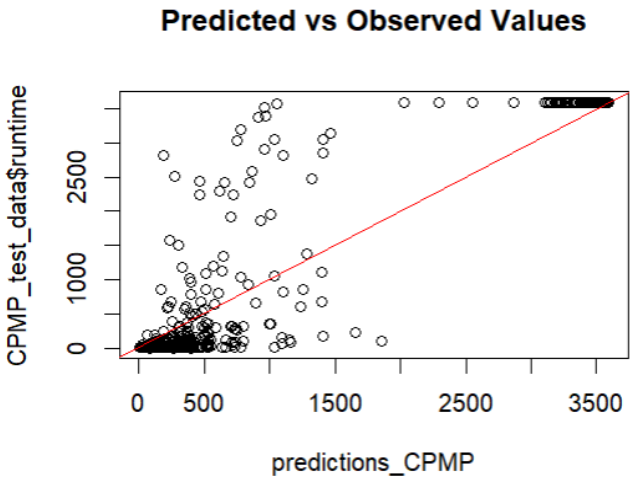

Figure 3:Predictions vs observed values after running randomForest model

Furthermore, another variable, overstowage.pct– the percentage of containers that are overstowed in the whole container terminal, was ran under randomForest predictive modelling which produced a relatively low MSE (0.000235) and high RSQ(0.970), showing that the fit of the variable against the vessel berthing time is good.

The same variable of overstowage.pct was modelled via linear regression, where the results are shown in the figure 4 below. As the performance of the terminals directly affect vessel berthing times (Gheith et al., 2012), the overstowage percentage is something to improve on if the overall performance of the container terminals thus time taken for vessels to be docked and unloaded needs to shorten.

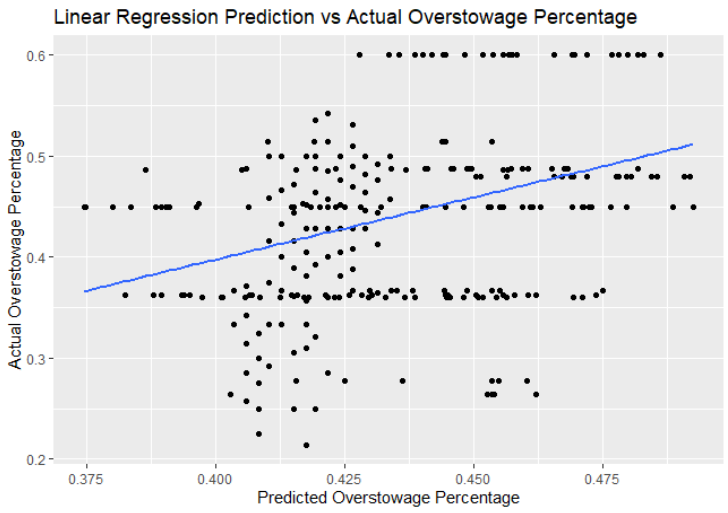

Figure 4: Linear Regression on overstowage

5.2 Blood donation

To improve the rate of blood donors going back to donate:
Q3: Which variable is the returning rate of donation most dependent on?
Q4: How to prevent predictive models from predicting false negatives (predict a false donation)?

Using CART, we were able to observe which variable is more significant due to it being the splitting factor of the first node in the decision tree from running CART, as seen in the below figure, which is the recency of the previous blood donation.

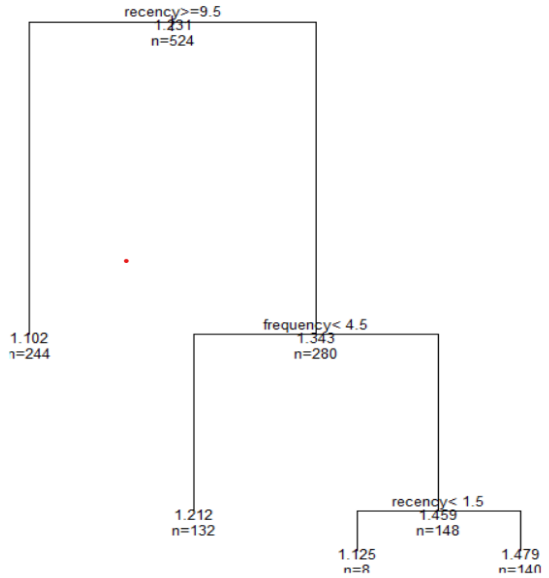

Figure 5: Decision tree from CART

However, it is seen that in this data set, there is class imbalance as there are more positive classes– those who did not go back for donation again in that period, than negative classes– those that did donate blood again. With accuracy (0), precision (0.814), recall (0.797) and F1 score (0.805), the F1 score is a better measure compared to accuracy due to the uneven distribution of classes in the data(Huilgol, 2019). This is because it is worse to predict false negatives (Zach, 2021)– where it is predicted that a blood donor will donate their blood again, but the prediction is wrong, rather than false positives– where a blood donor is predicted to not donate blood in the future but does. False negatives may lead to a false security that there would be an extra unit of blood for hospitals to use, but they in fact do not have it, which may put the patient in need in danger of losing their life. Thus, using different metrics according to the nature of the dataset is how they can prevent predicting false negatives.

To compare other prediction models, logistic regression is also carried out to see if the results will give more meaningful insights.

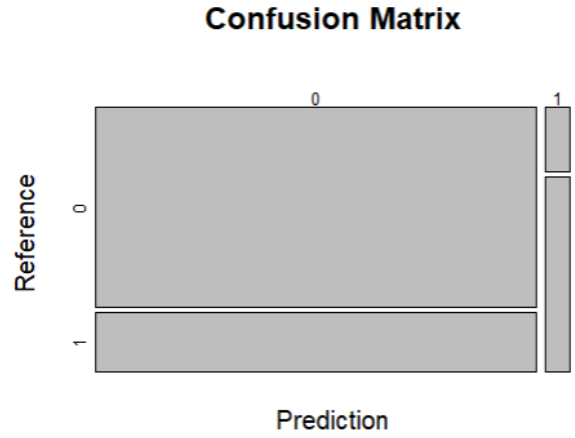

Figure 6: Confusion matrix of relationship between actual and predicted classes.

The confusion matrix provides a summary of the logistic regression model performance, where the top left is the true negativity measured by specificity (0.158), bottom right is true positivity measured by sensitivity (0.982) shows that the model incorrectly identifies many future donors as non-donors. Thus, there is a balanced accuracy– the average of sensitivity and specificity, of only 0.570.

6 Conclusion

In conclusion, the CPMP is a difficult task to perfect in vessel berthing prediction, where the goal is to make optimal use of the container terminal space as well as the container space on vessels, so that the containers in terminals are arranged in order and in a way to prevent overstowing to occur. This would allow docking and unloading to run smoothly, improving the speed and overall prevent a delay in operations.

Blood transfusion is a critical and important aspect of hospitals and should be taken very seriously as it involves peoples' lives. The goal of understanding the variables involved in blood donation is to reduce the rate of false negatives in this case, where it could cost lives if there is an over-estimation of the blood units available.

7 Proposed Further Research

In my opinion, there are many factors involved in trying to improve the performance of container terminals overall, while there are only data on the container terminal after the vessel has been docked and loaded. However, there are many factors that can cause delays in vessel berthing which happens on the vessels, and not the terminal yet. Thus I would suggest collecting data on the containers that are being transported to container terminals and this would allow a more direct and accurate solution to this complex problem.

I would also suggest more data on the blood transfusion service centre to be collected so that the class distribution can be more even which will lessen the complexity of the interpreting the results when training and testing any models. This can improve the predictive accuracy and allow hospitals and healthcare institutes to be better prepared considering any possibility of outbreaks of diseases or viruses in the future just like how Covid-19 escalated so quickly in 2020 for the world.

8 References

Bhonsle, J. (2022 , January 24). *Causes and consequences of vessel delays in container shipping*. Marine Insight. Retrieved March 30, 2023, from https://www.marineinsight.com/maritime-law/causes-and-consequences-of-vessel-delays-in-container-shipping/

Gheith, M. (2012, January). Container pre-marshalling problem: A review and Solution Framework. Container Pre-marshalling Problem: a Review and Solution Framework. Retrieved April 2, 2023, from https://www.researchgate.net/publication/258118547_Container_Pre-marshalling_Problem_a_Review_and_Solution_Framework

Gheith, M. S., El-Twail, A. B., & Harraz, N. A. (2012, January). *Container pre-marshalling problem: A review and Solution Framework*. ResearchGate. Retrieved April 1, 2023, from https://www.researchgate.net/publication/258118547_Container_Pre-marshalling_Problem_a_Review_and_Solution_Framework

Google. (2022, July 19). Classification: Precision and recall | machine learning | google developers. Google. Retrieved April 2, 2023, from https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Huilgol, P. (2019, August 24). Accuracy vs. F1-score. Medium. Retrieved April 2, 2023, from https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

Kumar, A. (2022, May 18). *Mean squared error or r-squared - which one to use?* Data Analytics. Retrieved April 1, 2023, from https://vitalflux.com/mean-square-error-r-squared-which-one-to-use/#:~:text=MSE%20represents%20the%20residual%20error,captured%20by%20the%20regression%20model.

Kundu, R. (2023, March 2). F1 score in Machine Learning: Intro & Calculation. V7. Retrieved April 2, 2023, from https://www.v7labs.com/blog/f1-score-guide#:~:text=for%20Machine%20Learning-,What%20is%20F1%20score%3F,prediction%20across%20the%20entire%20dataset.

Zach. (2021, September 8). F1 score vs. accuracy: Which should you use? Statology. Retrieved April 2, 2023, from https://www.statology.org/f1-score-vs-accuracy/