Onsite MSBA Applied Project, Spring 2017

**W.P. Carey, ASU**

**Final Project Report**

|  |  |
| --- | --- |
| Team B-04 | *Aditi Sharma, Fan Chen, Hsing Yuan, Sreerag Raghunathan* |
| Topic | *Container Availability Prediction* |
| Client | *Ports America* |

MSBA Applied Project Report Template

Spring 2017

W.P. Carey, ASU

# Executive Summary

Import containers constitute a large part of Port America’s business, the largest terminal operator in the US. The ability of the company to accurately identify the availability of containers translates to a better control of variation, budget, waste, and transportation within the operational cycle, which in turn translates to competitive edge in customer service and an increased opportunity for revenue generation.

Thus, the aim of this project is to build a predictive model which could accurately estimate import container availability. This would enable Ports America to manage the arrival, planning, and delivery of containers and drive down operating costs by identifying factors and outliers affecting the operations through analysis of its historical data covering a 1-year time horizon.

The team was provided with a combination of datasets from Ports America’s Terminal Operating System (TOS) and Vessel Productivity System (VPS), of its WCBT terminal. Essential contributing factors were narrowed down, and selected to finally build a regression model using Python, which can estimate the availability with a margin of 3 days. A boosted decision tree algorithm was used after comparing other machine learning algorithms, for improved accuracy. The prediction of the model was validated internally by splitting the model into training and test dataset initially, and using the training dataset for the algorithm’s learning. The predictions were then made on the test dataset, which are the final results.

The initial assumptions made by the model, including the factors affecting variability were validated by the model results.

Further steps can be taken such as incorporation of empirical vessel efficiency data, and using an ensemble of multiple machine learning algorithms, to improve the model results.

# Background

The United States of America is the world’s largest exporter and importer of goods and services. The US economy is heavily reliant on its coastal ports and inland waterways for trade. Ports America is the largest American ports and terminals operating company in the US. They handle approximately twelve million TEU annually, have a network of thirty-six terminals handling containers in twenty-four U.S. ports, and hold approximately one third of the container market share in North America. Needless to say, importing containers constitutes an essential and major part of Ports America’s operations and revenue.

**Operations:**

We concentrate on the containers received by the Ports America terminals, for our analysis. The import containers are received into the marine terminal via vessel discharge and decked to an initial yard location. The containers not designated for the particular port are put on hold within the vessel, and eventually loaded back. Containers can also be discharged and loaded back onto the vessel for shifting onto other sections of the vessel.

The containers movement operation is achieved from a combination of various modes of transportation within the terminal. The containers are discharged by cranes working against the vessel. Straddle Carriers are used to transport containers to or from ground. Utility Tractors (UTR’s) use bomb carts to transport containers if they are to be decked, or Chassis if the containers are to be wheeled. Top Handler is used to lift containers to or from a UTR and ground.

The efficient transportation for intra yard movements of the containers is essential for availability of the containers on time. The team was given containers, cranes, locations, and vessel productivity data for Ports America’s WCBT terminal to build a predictive model that could accurately identify when a particular container would be available for pickup, and also identify variation in the availability for optimized operations.

# Problem Statement

The ability to predict the imported containers’ availability for pick up by a trucking company working for a shipper or a consignee is essential for optimizing the ports operations, and translates directly to revenue. The main aim of this project was to build a predictive model to enable Ports America to establish a baseline of availability of containers for pickup and reduce the variance of service availability, by analyzing the given historical data associated with import containers arriving at Ports America’s container terminal. The result will be a reduced inventory turnover rate, improvement in overall service quality and thus an increase in profit. The team was given a historical dataset extracted from Ports America’s Terminal Operating System (TOS) and Vessel Productivity System (VPS), spanning approximately a year for the WCBT terminal wherein a non-linear relationship was assumed among the attributes for practical purposes, and a non-linear regression model was built to provide an estimate of the availability of containers, with root mean square error taken as a measure of the model’s prediction error.

# Methods

The following steps were taken to build the predictive model:

1. The team was provided with a dataset comprising of multiple tables namely container availability, attributes, locations, vessel productivity, and vessel sequence. Combination and inner join operations were performed on these tables using Python, for better analysis. Relevant data was identified using exploratory data analysis techniques, followed by data cleaning and treatment of missing values.
2. The features essential for prediction were selected, and extracted. The features finally used in the model prediction were:
   * Vessel positions - import container's bay, row, and tier position on the vessel
   * Yard location updates - number of changes in location from first yard location update to last yard location update
   * Volume - calculated for each vessel type (group by vessel type) from given number of container lifts (both discharge and loadout) completed by each crane working against the vessel
   * Restows - total number of container moves that had to be shifted on vessel in order to accommodate the discharge and loadback of the vessel over the course of its visit to the terminal
   * Standby hours - number of non-productive hours or hours where vessel operations ceased over the course of the vessel's visit to the terminal
3. The output column Availability (gives the duration of the container on the terminal in days) was derived from the in-time (date and time of container entering the terminal) and out-time (date and time of container exit from the terminal).
4. The output column was then analyzed, and the duration ranged from 0 days to 400 days (Fig 1). On plotting the values sans outliers using Python, it was observed that maximum data points lied within 0-12 days. This variation in the values was accounted to incorrect or late data entry into the terminal system, as per client discussion. Thus, a threshold of 15 was considered, and a list-wise deletion of records with availability values > 15 days was carried out, which was approx. 15% of the available records.

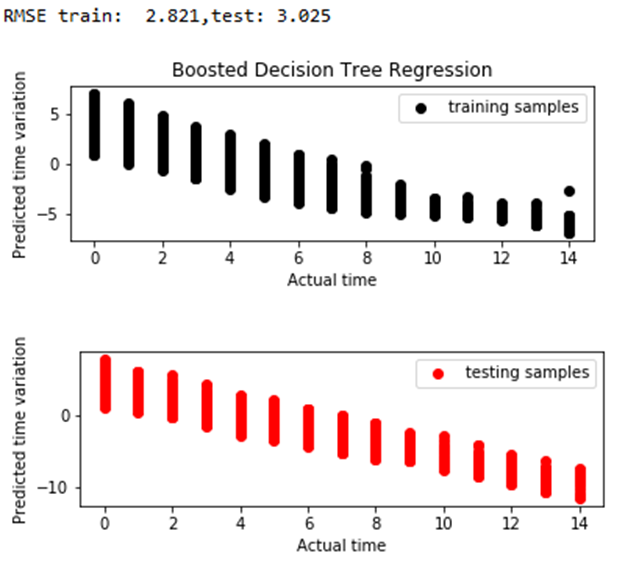
|  |  |
| --- | --- |
| Availability with outliers | Availability without outliers |

*Fig 1. Plot of Output (availability duration) with and without outliers*

1. The dataset was randomly split into a 60-40 ratio. Various machine learning algorithms - linear regression, random forest regression, Bayesian linear regression, and boosted decision tree regression, were tried and compared, and finally boosted decision tree regression was chosen for model prediction due to most accurate results. For the algorithm, the maximum depth parameter was set at 8 to limit the maximum depth of trees, Number of trees were set to 1000, to gain a good accuracy of prediction results and the learning rate was set at 0.2. Random shuffling was chosen using a random number generator for a practical output.
2. The model was trained on 60% of the data (training data), and its predictive power was validated against 40% of the data (test data). The model performance was judged based on the RMSE of test dataset, which is a measure of differences between the estimated and actual values.
3. Initial EDA and feature selection was accomplished using MS Excel and MS Azure ML. The predictive model and analysis was carried out using Python.

# Results and Conclusions

The results obtained from the predictive model are shown in the figure below (Fig 2). It shows a plot of predicted time variation vs actual time for the training as well as test data, and their respective RMSE values. The RMSE is an absolute measure of differences between the estimated and actual values. The variation in predicted vales with respect to the actual values is quite constant, thus validating the model’s initial assumptions and prediction.



*Fig 2. Results – Predicted vs actual plot*

The RMSE value for training dataset is 2.821, and test dataset is 3.025 which narrows the prediction accuracy of the model to a span of 3 days. This model can be used by Ports America to provide the estimate (with a margin of 3 days) of a containers availability to its customers. Changes can be made to certain input parameters to identify the source of variation in availability. From the model’s assumptions and final results, it can be concluded that the factors affecting a container’s availability for pickup are mainly vessel position, yard movements and vessel operational efficiency.

Also, as observed and stated earlier, there was a huge variation in output time which was accounted to invalid or incorrect data entered into the system. A recommendation was made to Ports America to further look into these outliers to eliminate unnecessary causes of variation in the dataset, and thus gather cleaner and more insightful data for best data analysis results.

Furthermore, currently a uniform probability distribution of vessel efficiency was assumed because of unavailability of actual data. The model results can be further improved by including empirical vessel efficiency probability distribution model w.r.t. container position within the vessel. An ensemble of other machine learning algorithms can be incorporated in the future to improve prediction accuracy, which the team was unable to take up due to time constraints.

# References

1. http://scikit-learn.org/stable/auto\_examples/ensemble/plot\_adaboost\_regression.html
2. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
3. https://www.portsamerica.com/services-containers.html
4. Ports America documentation: Import Container Process\_170303
5. http://www.iwr.usace.army.mil/Portals/70/docs/portswaterways/rpt/June\_20\_U.S.\_Port\_and\_Inland\_Waterways\_Preparing\_for\_Post\_Panamax\_Vessels.pdf
6. https://www.portsamerica.com/services-containers.html

# Appendix

## Reproduction of Results

1. Put all the dataset files in csv format in any specified directory
2. Update the directory location of the files in the code (‘TeamB-04\_FinalCode.py’, lines 21-24)
3. Run the Python code for model building.
4. Enter the path where predictions are to be stored in csv format (line 170)
5. Plot the results by running the plot code