

|  |
| --- |
| **ArcBest** |
| **Dock Bills per Hour** |
| **3rd June 2022** |

Contents

[1. Project Objective 3](#_Toc105167358)

[2. Approach Summary 3](#_Toc105167359)

[2.1 Defining Datasets: 3](#_Toc105167360)

[2.2 Datamart Creation: 5](#_Toc105167361)

[2.3 Variable Selection Method: 5](#_Toc105167362)

[2.3.1 Regression: 5](#_Toc105167363)

[2.3.2 Random Forest: 5](#_Toc105167364)

[2.3.3 Variance Inflation Factor (VIF): 6](#_Toc105167365)

[2.4 Decision on Training/ Scoring: 6](#_Toc105167366)

[2.5 Training: 6](#_Toc105167367)

[2.6 Scoring: 7](#_Toc105167368)

[2.7 Output Transfer: 7](#_Toc105167369)

**Dock Bills Per Hour**

# Project Objective

The objective of this phase is to utilize predictive analytics to provide a control limit for dock bills per hour for each station. The project will help the station manager keep track if their dock bills per hour fall in an acceptable range or not.

# Approach Summary

## Defining Datasets:

Below are the datasets that were used to predict dock\_bills\_per\_hour

1. Dockhist: Table contains manifest information



1. Actdates: Table contains Dock\_bills\_per\_hour information



1. H2H Predictions: Table contains information about probability of shipments that can be tagged as Hard to Handle



1. Ldrinfo: Contains Loader information



1. Laborunits: Provides information about total labor units for the terminal



1. Ldrtenure: Provides tenure information for Loaders



1. Proinfo: Provides information about pronumber



1. Stainfo: Provides Station/Terminal information



1. Workdays\_data: Provides Workday Information



## Datamart Creation:

1. Created H2HStrip dataset by joining Ldrtenure, ldrinfo, h2hprediction, proinfo & dockhist tables.
2. Then created H2HTermMonth by aggregating the above data at Station, MFST\_DW\_DATE level, and joining Bills\_Dock\_Hr & laborunits dataset on MFST\_DW\_DATE & station.
3. Finally, created multiple dummy variables of station(9) & months(11) columns using proc glmselect.

## Variable Selection Method:

### 2.3.1 Regression:

* Defined train dataset: MFST\_DW\_DATE<=01dec2018, test dataset: 01Jan2019 – 01Jan2021
* Defined continuous variable & dummy variables in different macros
* Applied linear regression technique on different values of sl\_stay & sl\_entry (Forward & Backward selection methods of stepwise regression, values between 0.01-0.19) and obtained model estimates.
* Then scored the test dataset using the model estimates obtained from above.
* Finally, selected the model with least MAPE value.
* This method was used for selecting the best set of sl\_stay & sl\_entry.

### 2.3.2 Random Forest:

* Defined train dataset: MFST\_DW\_DATE<=01dec2018, test dataset: 01Jan2019 – 01Jan2021
* Defined continuous variable & dummy variables in different macros
* Did hyperparameter tuning by running proc hpforest on different combinations of its hyperparameters, eg.,

Maxtrees = 50 to 100, varTry = 7 to 12, maxDepth = 30 to 35, leafsize = 4 to 10.

* Selected the model with the least average value of PredOob
* Variable importance file is one of the outputs of the above step, hence the variables corresponding to the selected model is finalized.

### 2.3.3 Variance Inflation Factor (VIF):

* This code runs on continuous variables.
* Defined train dataset: MFST\_DW\_DATE<=01dec2018, test dataset: 01Jan2019 – 01Jan2021
* Defined continuous variable & dummy variables in different macros
* Using VIF loop , selected those continuous variables which had VIF<=5

## Decision on Training/ Scoring:

Once the costing run is completed on let’s say 12th May 2022, all the data from last 4 years till today is scored using the Training Code. And as we move on to the next day i.e., 13th May 2022, all the data from 1st May 2022 till 13th May 2022 are scored. This process continues till the next costing run in next month.

Suppose the next costing run is scheduled on 13th June 2022, then on 12th June 2022, all the data from 01May2022 – 12June2022 will be scored, and on 13th June 2022 training code will score all the data from past 4 years till today.

Steps that we followed were:

1. Found maximum date from procost\_train dataset & stored it in macro max\_dt.
2. Found number of months difference between max\_dt & sys\_date and store in macro month\_diff.
3. If month\_diff=1 then we followed below steps:
   1. Found Month number of last Training date & stored in lsttrndate\_mo
   2. Find month of sys\_date and store it in sys\_date\_mo
   3. If sys\_date\_mo NE lsttrndate then we run the training code, else we run the Scoring code.
4. If month\_diff=2 then we run the scoring code.

## Training:

1. Frequency of run: Training code gets triggered once each month after the costing run is over.
2. Defined sys\_date as Today’s date. Eg. 30May2022
3. Training time period: Data of last 4 years till last month, i.e., 01May2018-01Apr2022
4. Found model estimates using regression, & hence MAPE.
5. For scoring the historical data we followed the below steps:
   * Test timeframe: All the data till sys\_date i.e., till 30May2022
   * Dropped dependent variable i.e., Bills\_Dock\_Hr
   * Transposed and renamed columns in model estimates obtained from step- c
   * Using proc score, obtained Upper & Lower control limits along with the predictions.

## Scoring:

1. Frequency of run: Scoring code runs everyday.
2. Scoring time period: Data from 1st day of last training date till today, i.e., 01May2022-30May2022
3. Dropped dependent variable i.e., Bills\_Dock\_Hr
4. Scored the above data using the model estimates obtained from Training code to get the Upper & Lower control limits along with the predictions. So, on the day of the costing run est\_model file is generated from Training code which stays the same till the next costing run i.e., the day on which the Training code runs.



## Output Transfer:

1. After the costing run is over, Training code gets triggered & it scores all the data from last 4 years till today’s date. On the ABF DOCK\_BILLS\_PREDICTION dataset, all the previous data (data from 01June2017 onwards) are first deleted & this new predicted dataset is freshly uploaded into the same.
2. In all other days, scoring code is run. On the ABF DOCK\_BILLS\_PREDICTION dataset, all the data from 1st day of the month of the training date till today’s date-1 are first deleted and this new predicted dataset is uploaded.