Predicting Ship Arrival

Analysis of Ship Arrival Times to Optimize Operational Efficiency at Ports

OIT276: Regression Project

Prepared For

Team

Amee Kapadia

Farrukh Hamza

Lydia Deneen

James Park

STANFORD GRADUATE SCHOOL OF BUSINESS

STANFORD GRADUATE SCHOOL OF BUSINESS

Table of Contents

Executive Summary 3
- **1. The Problem** 3
- **2. The Approach** 3
- **3. Key Findings and Recommendations** 3

Problem Statement 4
Data 5
Analysis 6
Regression Modeling 7
- **Linear Regression with One Independent Variable** 7
- **Linear Regression with Multiple Independent Variables** 7
- **Regression Decision Tree** 7
- **Random Forest** 7
- **Feature Extraction** 8

Principal Findings 9
- **1. Days of the Week** 9
- **2. Number of Journeys** 9
- **3. Specific Ships** 10

Conclusions and Recommendations 11

STANFORD GRADUATE SCHOOL OF BUSINESS

Executive Summary

1. The Problem

The dataset for this study was provided by Sheikha Amneh Al Qasimi from the Investment and Development Office (IDO) in Ras Al Khaimah (RAK), United Arab Emirates. The strategic goal of the IDO is to transform RAK into a leading destination for travel, leisure, and business through a series of redevelopment initiatives. An important factor in achieving this transformation is to enhance the operational efficiency of Saqr Port, which serves as a crucial industrial hub in the Middle East region. One method to improve efficiency is to enhance the prediction accuracy of ship arrivals, which could potentially reduce waiting times, increase the throughput of goods, and aid the IDO's objectives by guaranteeing the timely supply of materials for various projects. Untimely operations at the port could result in substantial demurrage fees. Hence, the IDO has requested that our team utilize the Automatic Identification System (AIS) ship data to analyze and evaluate the accuracy of the port's arrival predictions, focusing on how this accuracy changes as ships approach the port.

2. The Approach

We first examined the data to identify variables that could facilitate meaningful analysis. We conducted four types of regression analyses: a linear regression with a single independent variable, a linear regression with multiple independent variables, a regression tree model, and a random forest model. For each ship in the dataset, we calculated a quality score, which we defined as the difference between the final and initial mean squared error (MSE) values. Additionally, we systematically recorded the average quality associated with each day of the week, the total number of journeys undertaken by each ship, and the specific characteristics of each ship to gain a nuanced understanding of the predictive accuracy.

3. Key Findings and Recommendations

Our analysis indicated that the MSE diminishes as more independent variables are integrated into the regression models, suggesting that a more extensive dataset can help in constructing a more precise and robust predictive model. Notably, the regression tree and random forest models emerged as more effective in accurately predicting ship arrivals at the port. Our findings highlighted that various factors influence the overall quality of the arrival time predictions. We observed a nontrivial variability in the predictive accuracy among different ships. To refine the predictive accuracy further, we advocate for the integration of additional datasets, such as weather pattern records, or the segmentation of ships based on their commerce seasonality, as these strategies may unveil more definitive predictive patterns. To leverage these insights for the tangible benefit of RAK, we recommend optimizing the deployment of operations personnel at Saqr Port based on these enhanced predictive models, and devising projection-based datasets that can adapt to and accommodate the anticipated rapid growth of RAK in the forthcoming years.

STANFORD GRADUATE SCHOOL OF BUSINESS

Problem Statement

A fundamental element in the transformation of Ras Al Khaimah (RAK) is enhancing the operational efficiency of Saqr Port, the premier maritime and industrial commerce hub in the Middle East. See **Exhibit 1** for an illustration of a ship's journey into Saqr Port. Given Saqr Port's crucial role in supply chains, particularly in supplying construction materials for the UAE's real estate ventures and bolstering the mining industry, establishing a highly efficient and predictive operational framework is imperative. Accurate predictions of ship arrivals would facilitate improved scheduling, diminish waiting times, and boost the port's overall throughput capacity. This improvement would further the Investment and Development Office's (IDO) broader objectives by ensuring a timely supply of materials for infrastructure projects, reducing logistics costs, and drawing more investment to the region. Moreover, a significant motivation for this project stems from the substantial demurrage fines of $20,000 per hour that the IDO incurs for any delays.

Figure 1 Location of Saqr Port in the UAE (source: Google Maps)

The problem statement has two parts:

1. *What is the quality of prediction from the AIS Data?*
2. *How does it change over time as the ship approaches?*

Predicting ship arrivals with high precision is complex. Although AIS is designed to broadcast information automatically, key data like the destination and estimated time of arrival (ETA) require manual input by the ship's crew. Furthermore, AIS data is transmitted through Very High Frequency (VHF) radio channels, the reliability of which can be affected by physical barriers, such as the congested environments typical of busy ports. Maritime logistics are further complicated by factors such as weather conditions, traffic congestion at sea, and varying speeds

of the ships, all of which add layers of complexity to the prediction models. Additionally, the integrity of the data is a crucial aspect; the presence of missing or poorly formatted records can significantly undermine the accuracy of predictions.

Data

Our dataset comprised 23,626 text files, amounting to 2.18 GB in total. Each text file was structured as a JSON dictionary, containing objects as AIS records. The format of these records varied depending on the "status" field, which could be one of five values: *Underway*, *DestinationChanged*, *NotMoving*, *UnableToPredict*, or *Arrived*. The fields of interest for our analysis were *recordTime*, indicating when the AIS record was made, and *eventTime*, predicting the ship's time of arrival at the destination.

Figure 2 Example of a record contained in each JSON file


```
{
    "mmsi": 447215000,                          # unique ID for ship
    "port": "AEMSA",                            # code for port destination
    "recordTime": "2023-06-08T16:02:58Z",       # when record was made
    "eventTime": "2023-06-11T01:52:45.700Z",    # predicted time of arrival
    "status": "Underway",                       # one of five statuses
    "warnings": [],                             # related to ship's journey
    "timeType": "Estimate",                     # only "Actual" when ship status is "Arrived"
    "travelDistance": 244.73218142548595,       # estimated distance to arrival
    "waypoints": [                              # predicted waypoints, if status is "Underway"
        {
            "coordinates": [
                52.07855337351741,
                27.166483145568307
            ],
            "type": "Point"
        }
    ],
    .....
    "destinationSource": {
        "type": "aisDestination"
    },
    "destination": {
        "port": "AEMSA",
        "source": "AISDestination",
        "aisDestination": "MINA SAQR"
    },
    "ship": {
        "imo": "9547972",
        "mmsi": "447215000",
        "callSign": "9KGO",
        "name": "HEISCO 23"
    },
    "portcallId": "5164ca13-3e42-36dd-9a7a-82bff40db9c2",
    "predictionRequestId": "6dd5f70b-6699-3c2e-9ec7-1a419ba184f1"
},
```


Each file name corresponded to the *recordTime* of each record. However, the organization of the records within the files was not straightforward. For example, the file named api_response_2023-06-08_17-01-17.json contained eighteen records, all with *recordTime*'s before 17:01:17 UTC on June 8, 2023. A significant issue arose with records for ships far from their destination, where the estimated time of arrival could be days or weeks ahead; subsequent updates for these records were not necessarily in the same file. Additionally, the dataset

contained duplicate records. Our initial challenge was to process and organize the data effectively to allow for meaningful analysis.

Analysis

To facilitate our data cleaning process, we created a second copy of the dataset. We executed a script, unique_ships.py, to count the unique ships within the data, identifying a total of 904 distinct vessels. For each ship, we generated a separate text file to organize its specific data. Our primary focus was on the *eventTime*, the estimated time of arrival for each ship, which guided our analysis and helped in assessing the prediction quality.

We sorted the records within each ship's file by *recordTime*, creating a chronological narrative of the ship's journey. Records with a *Status: Arrived* indicated that *recordTime* and *eventTime* were the same. Our analysis concentrated on records where the status was either *Underway* or *Arrived*, leading us to eliminate records with statuses such as *DestinationChanged*, *NotMoving*, or *UnableToPredict*. Consequently, each ship's text file portrayed a sequence of Status: *Underway* records, culminating in a *Status: Arrived* entry.

During our analysis, we noticed that some ships' files started with a *Status: Arrived* record. We excluded these from our dataset, as they lacked preceding estimated eventTime entries for comparative analysis. Ultimately, each ship's text file consisted solely of records representing valid journeys, with each file containing one or more journeys.

Figure 3 Each ship's journey is processed into one file

```
Record time: 2023-06-14T02:14:22Z
Event time: 2023-06-14T05:12:19.750Z
Status: Underway
Travel distance: 41.94924406047516
Num waypoints: 10
Waypoints: (55.50845809838617, 25.6242707117853), (55.50845809838617, 25.694282

Record time: 2023-06-14T03:35:33Z
Event time: 2023-06-14T04:50:15.500Z
Status: Underway
Travel distance: 18.86879049676026
Num waypoints: 5
Waypoints: (55.78845032082444, 25.8343057176196), (55.858448376433984, 25.90431

Record time: 2023-06-14T05:06:31Z
Event time: 2023-06-14T05:06:31Z
Status: Arrived
Travel distance: N/A
Num waypoints: 0
Waypoints:
```

Regression Modeling

We employed four regression analysis methods:

1. Regression with one independent variable
2. Regression with more independent variables
3. Regression tree
4. Random forest

The first model used travel distance as a singular independent variable. The second model expanded on this by incorporating multiple independent variables: travel distance, day of the week, ship MMSI, and the number of journeys undertaken by each ship. The third model, a decision tree regressor, provided a non-linear approach to regression analysis. The final model utilized a random forest regressor for its predictive analysis.

Linear Regression with One Independent Variable

This method conducts a straightforward linear regression analysis. For each input file, it processes the records to extract necessary features and segregate different journeys. Although the script includes a *perform_space_regression* function for predicting travel distance, this function was not utilized for this project. However, it remains in the source code as a potential avenue for further analysis. The model establishes a simple linear relationship between time and distance and calculates the MSE using test data.

Linear Regression with Multiple Independent Variables

Expanding on the simple regression model, this approach introduces additional independent variables to create a more complex model. This multiple linear regression aims to predict total travel time based on these variables, enhancing prediction accuracy by accounting for weekly patterns (day of the week), ship-specific characteristics (MMSI), and journey frequency (total number of journeys).

Regression Decision Tree

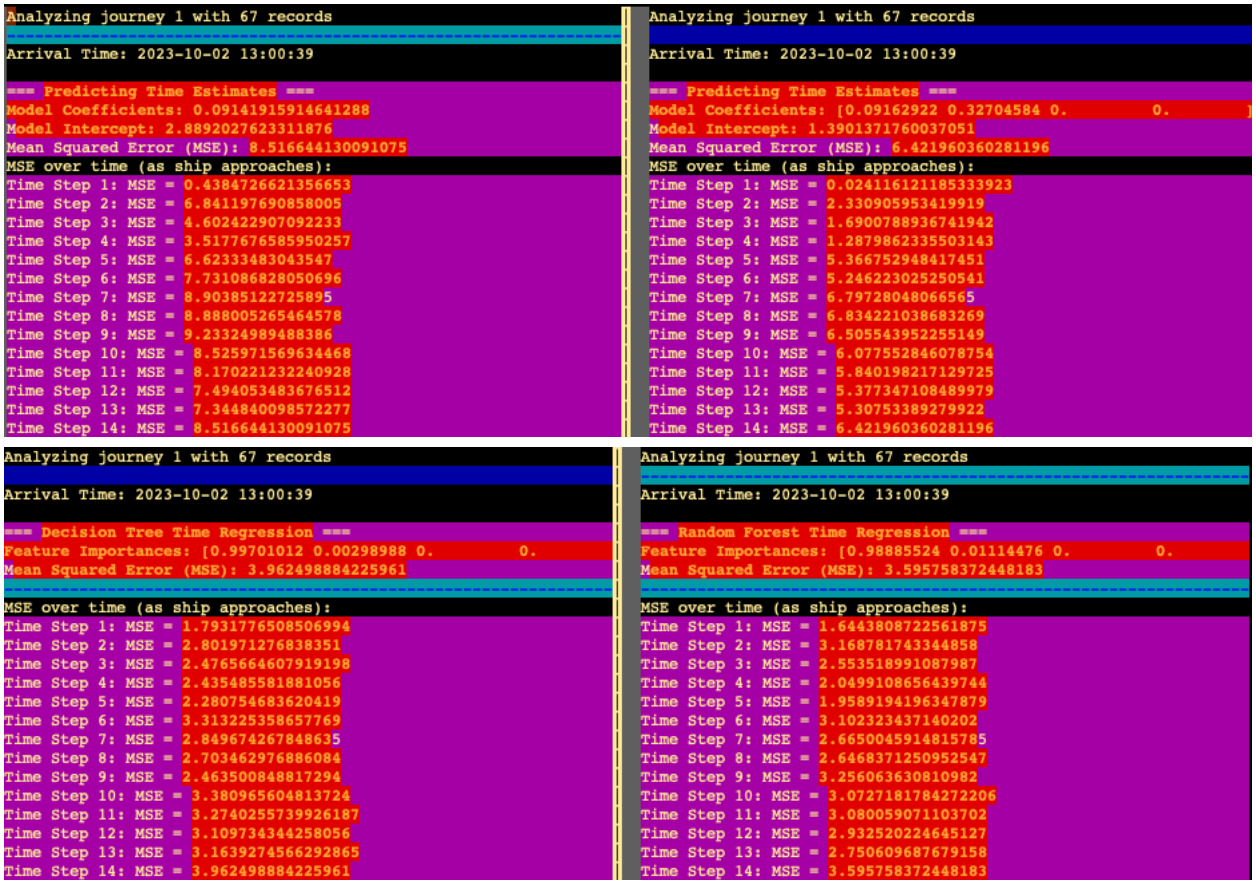
This model employs a Decision Tree Regressor to capture complex relationships between variables, offering a non-linear perspective that linear models might miss. By making successive decisions based on the input features, the decision tree predicts outcomes and assesses the importance of each feature in determining total travel time, providing valuable insights into the most significant predictors.

Random Forest

Building upon the decision tree model, the Random Forest Regressor utilizes an ensemble learning method. By generating multiple decision trees from different subsets of the data and

averaging their predictions, it enhances predictive accuracy and mitigates overfitting. This model is preferred for its robustness and improved accuracy over a singular decision tree, balancing variance reduction without significantly increasing bias.

Figure 4 Snapshot of the first journey's regression results for one particular ship



Feature Extraction

Each regression model directed its outputs to distinct folders named 1_results, 2_results, 3_results, and 4_results. Consistent with our data formatting approach, each folder contained 904 files. Post-regression, we executed an analysis script to decipher the results, focusing particularly on the influence of various independent variables and the interpretive power of the models. Our goal was to evaluate the quality of predictions. Since additional independent variables were incorporated starting from the second regression, our analysis primarily focused on the second (Linear Regression), third (Regression Tree), and fourth (Random Forest) regressions.

The analysis entailed calculating the prediction quality based on the MSE over time, determining the average MSE quality per day of the week, and assessing the average MSE quality per journey count. For each file in a given folder, the script reads the content, employing two functions: *extract_mse_over_time* to ascertain MSE fluctuations over time, and *extract_day_of_week* to

identify the weekdays of arrivals. It computes the **relative** prediction quality by subtracting the initial MSE from the final MSE, positing that a reduction in MSE signifies enhanced accuracy. This quality metric is then compiled based on ship ID, day of the week, and journey count, providing a comprehensive view of how prediction quality fluctuates across these categories.

Principal Findings

1. Days of the Week

The prediction quality for ship arrivals fluctuates significantly throughout the week, suggesting a strong correlation between operational activities or weekly patterns and ship arrival times. In Linear Regression analysis, the variability in prediction quality across different days is evident through the oscillating heights of the blue bars, indicative of the model's sensitivity to day-dependent operational factors. The Regression Tree model, represented by orange bars, shows considerable day-to-day prediction quality variation, likely due to its capacity to respond to specific daily conditions such as traffic or operational changes at the port. The Random Forest model, visualized with gray bars, demonstrates a similar variability, highlighting the impact of daily factors on ship arrival predictability. Both Linear Regression and Random Forest models offer comparable prediction quality assessments, with Random Forest generally outperforming Linear Regression, particularly on Sundays and Mondays, whereas Wednesdays exhibit the lowest prediction accuracy, likely due to midweek operational dynamics. See **Exhibit 2** for the quality of prediction quality scores across different models for each day.

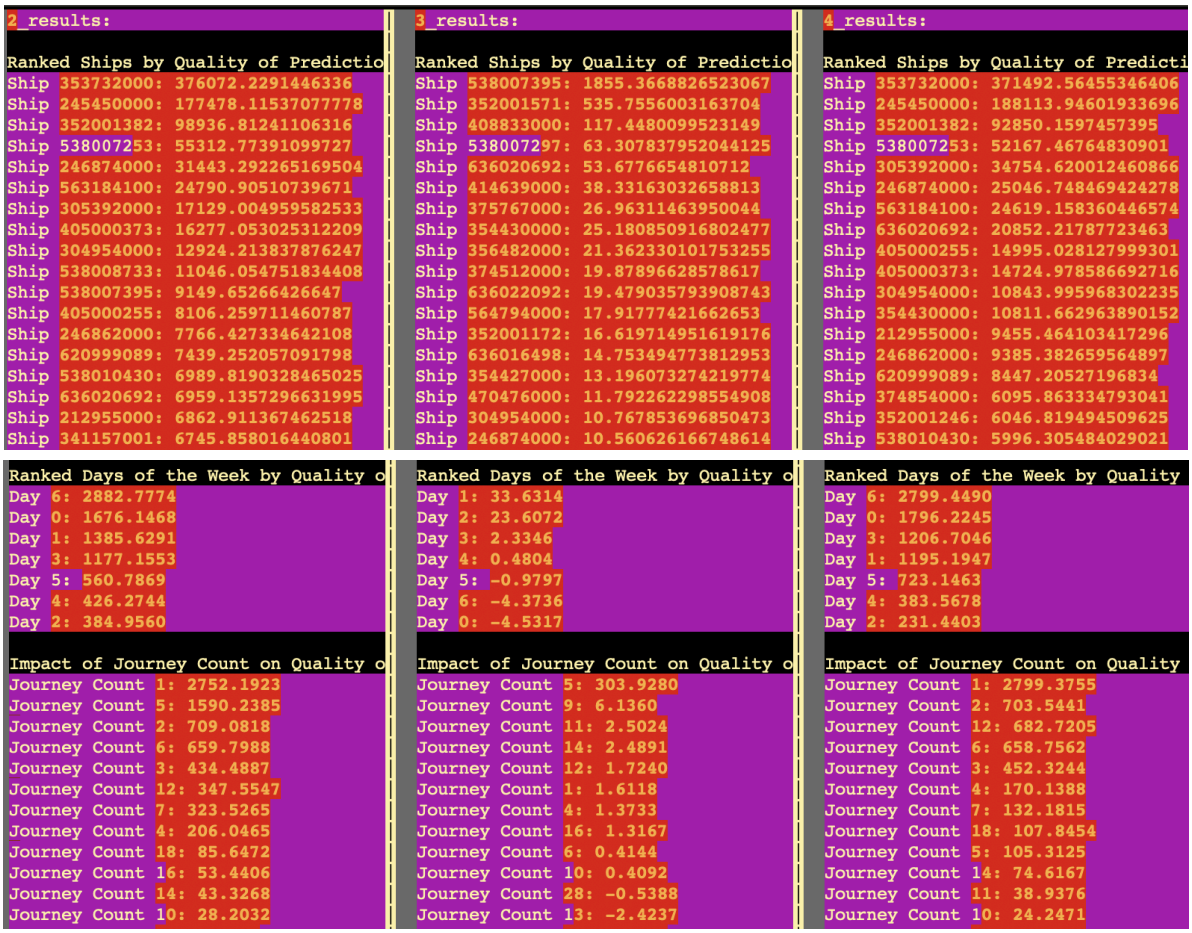
2. Number of Journeys

In the Linear Regression model, the prediction quality's nuanced response to the total number of journeys undertaken by a ship suggests that while increased journey data might enhance the model's learning, it also potentially introduces more variability or noise, impacting the predictions negatively. This is depicted through the fluctuating pattern of the blue bars across different journey counts. The Regression Tree model, with its significant variation in prediction quality across journey counts (indicated by orange bars), underscores the model's sensitivity to the amount and diversity of data, likely reflecting the varying operational patterns or environmental conditions experienced across different journeys. The Random Forest model, represented by gray bars, reveals a complex relationship between prediction quality and journey count, indicating its ability to balance the benefits of extensive data against the increased complexity and variability. This model's performance suggests that while additional data from more journeys can be advantageous, it also necessitates sophisticated handling of the resultant complexity to avoid diminishing the prediction accuracy. See **Exhibit 3** for the quality of prediction quality scores for the different models across various journey counts.

3. Specific Ships

Enriching Linear Regression with additional variables, including the Maritime Mobile Service Identity (MMSI), leads to improved prediction quality, illustrating the model's capability to encapsulate unique ship-specific patterns or operational behaviors, which are critical in determining the predictability of ship arrival times. The substantial variation in prediction quality among different ships, as observed in the Regression Tree model, highlights its efficacy in deciphering complex, ship-specific patterns, which might be obscured in more simplistic models. This indicates a high level of model responsiveness to individual ship characteristics or historical performance data. The Random Forest model consistently offers reliable and precise predictions across a variety of ships, attributing its success to the ensemble method which reduces the risk of overfitting while capturing a comprehensive range of patterns within the dataset. This consistency suggests that the Random Forest model is adept at integrating and analyzing the diverse operational and historical data pertinent to each ship, leading to a robust predictive performance. See **Exhibit 4** for the quality of prediction quality scores across different models for each MMSI.

Figure 5 Side by side comparison of quality assessments of Linear Regression, Regression Tree, and Random Forest



Conclusions and Recommendations

In our analysis, we aimed to address two critical questions about the AIS data quality and its temporal reliability for predicting ship arrivals at Saqr Port. Our findings reveal that integrating multiple independent variables does enhance the predictive model's accuracy, indicating that a broader dataset can establish a more precise and robust model. This is particularly evident in the Regression Tree and Random Forest models, which outperformed simpler linear models in predicting ship arrivals. The analysis underscores the influence of temporal factors, such as days of the week, on arrival time predictions, suggesting that operational dynamics and weekly patterns play a crucial role in forecasting accuracy. Furthermore, the variance in prediction quality across different vessels highlights that specific ships have better predictability than other ships. These insights address the primary concerns of AIS data quality and its variation as ships approach the port, suggesting that comprehensive and advanced modeling techniques can significantly improve arrival time predictions at Saqr Port.

To enhance model accuracy, we recommend integrating weather pattern data into the dataset to understand environmental impacts on ship arrivals. Segmenting ships based on time-related aspects, like seasonal commerce fluctuations or peak periods, could refine predictions by considering changing traffic patterns and operational conditions. Utilizing the predictive model to optimize the number of operations personnel for loading and unloading operations can lead to more efficient resource allocation and reduced operational costs. With RAK's ongoing development, it is vital to project future growth in the analysis, including anticipating increased ship traffic and identifying necessary port expansions or enhancements to accommodate this growth effectively. Implementing these recommendations can further refine the predictive model and improve the operational efficiency of Port Saqr, aiding RAK's broader goals of becoming a travel, leisure, and industry hub.

Exhibit 1 AIS predicted waypoints for a ship at various time intervals

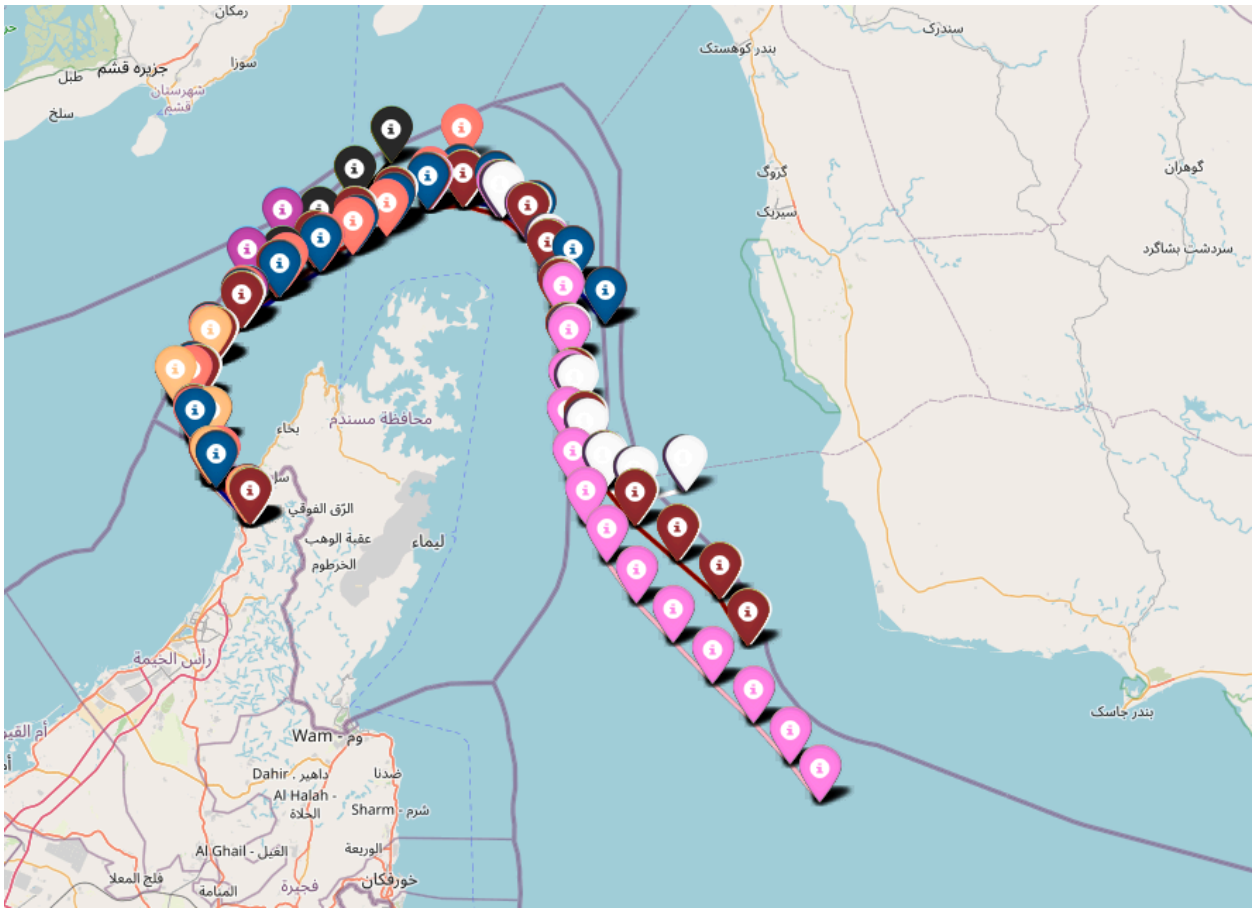


Exhibit 2 Quality of prediction scores for Linear Regression, Regression Tree, and Random Forest, for each day of the week

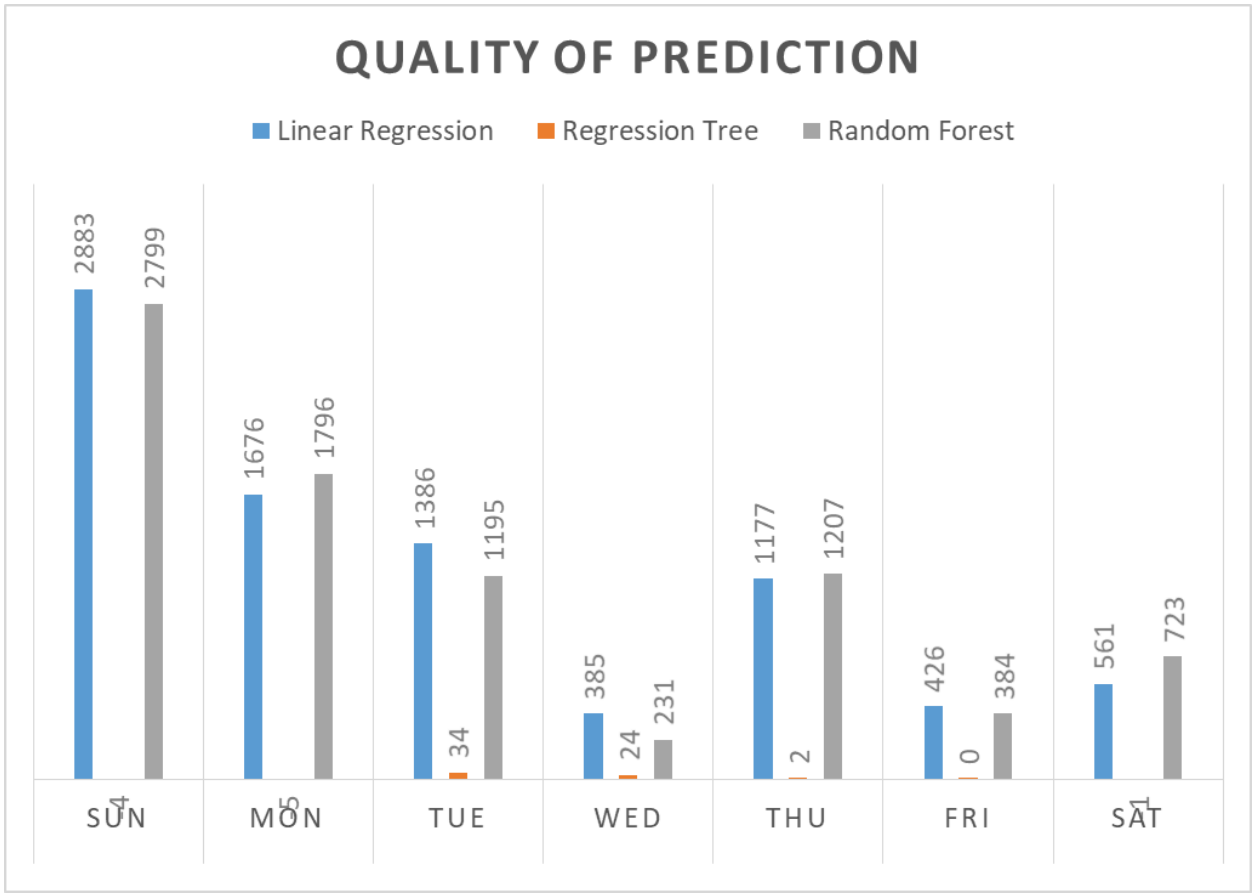


Exhibit 3 Quality of prediction scores for Linear Regression, Regression Tree, and Random Forest, for different journey counts

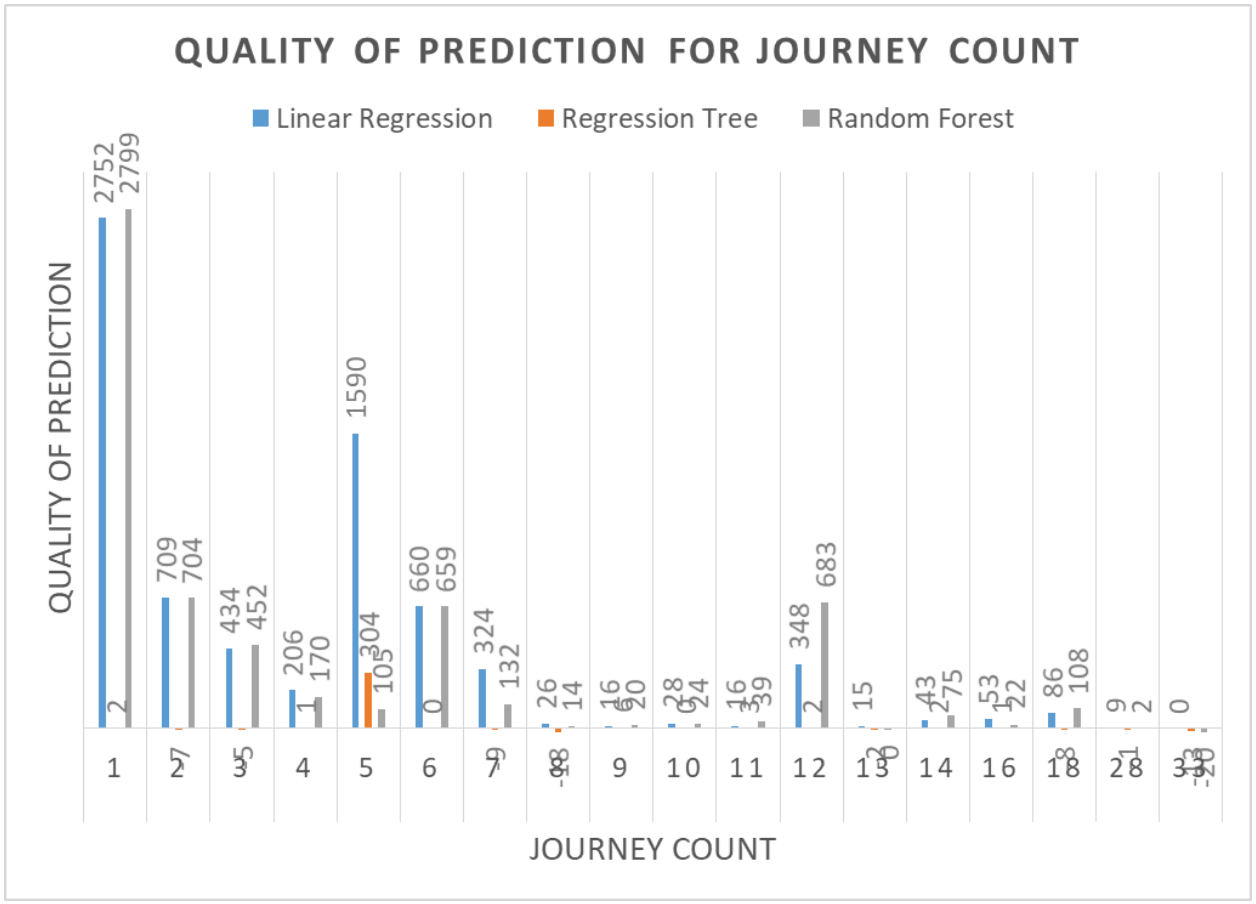


Exhibit 4 Quality of prediction scores for Linear Regression, Regression Tree, and Random Forest, for different ship's (MMSI's). "Top Five" and "Bottom Five," indicating the five ships with the highest and lowest prediction qualities, respectively

Regression		Regression Tree		Random Forest	
Ship mmsi	Quality of Prediction	Ship mmsi	Quality of Prediction	Ship mmsi	Quality of Prediction
Top Five					
353732000	376072	538007395	1855	353732000	371493
245450000	177478	352001571	536	245450000	188114
352001382	98937	408833000	117	352001382	92850
538007253	55313	538007297	63	538007253	52167
246874000	31443	636020692	54	305392000	34755
Bottom Five					
471174000	-15	355023000	-45	372122000	-46
470774000	-32	563092700	-55	626250000	-49
563142300	-176	357549000	-65	447215000	-51
308521000	-2373	447215000	-78	357549000	-69
374512000	-3216	563142300	-543	563142300	-196