ETA PREDICTION OF CARGO VESSELS FINAL PROJECT REPORT

Group 10

Nate Hellmuth
(nah6qg)

Nicholas Phair
(np4ay)

Matt Walsh
(mw6es)

ABSTRACT

The global shipping industry is the backbone to transporting goods around the world. In fact, the International Chamber of Shipping describes it has a 14 trillion dollar industry that accounts for 80% of imports and exports in the European Union. It is clear that tracking a ship's time of arrival at a port is crucial to the success of the entire operation. However, these estimated times of arrivals (ETAs) are entirely self-reported by the vessel itself at departure time. They may not account for any delays occurred on the voyage and are often inaccurate. As such, these self-reported values cannot be trusted. To this end, we propose using automatic identification system (AIS) data and machine learning algorithms to better predict a ship's ETA. We evaluate two machine learning models, an SVR and LSTM, as well a simple kinematic model. We compare the predictions that these models make to the self-reported value to measure improvements to ETA predictions.

1 PROBLEM DEFINITION

The COVID-19 pandemic had a major effect on the maritime shipping industry and exposed many of the challenges the global shipping industry faces. One major challenge has been how to efficiently manage ports. Efficient scheduling allows for faster loading and off-loading of cargo in ports, but doing so is a challenge as current estimated time of arrival (ETA) in AIS messages are frequently inaccurate. The ETA is reported by the vessels and can vary greatly. The reported estimates can be calculated using software or can be entirely guessed. Often ETA is not provided at all. As machine learning continues to develop, we can use it to solve this real world problem.

There are many who benefit from improved ETA prediction. Primarily are ports and port operators. By improving estimated arrival port managers can schedule more accurately, which reduces time spent waiting for access to a dock. In addition to ports, the companies who manage and track cargo will be able to better predict when the things they need are expected to arrive. It can even lend itself to new applications of anomaly detection. If a vessel is not matching the ETA closely enough it may be an indication that something has gone wrong.

The input for this problem is AIS data that report vessel position and destination. Using additional fields included in those records we can develop a model to predict ETA and output the estimated time to arrival.

2 PROPOSED IDEAS

We propose two analytic model types for determining accurate prediction of ETA for maritime vessels. The two models we will be testing are support vector regression (SVR) and long short-term memory recurrent neural networks (LSTM).

These models were chosen based on a few factors. Firstly, previous work has suggested that both SVR and LSTM models are able to predict ETA accurately, but the two have not been directly compared. SVR models can be successful for a wide range of classification and regression problems based on the tuning of meta-parameters and are more resilient to over fitting. Alternatively, LSTM

models focus on sequences of data, which are highly applicable to routing data. Instead of just analyzing individual data points, LSTM models incorporate variable length sequences of data to more accurately extrapolate future data. This project will focus on the comparison of the two methods and their respective ability to reliably predict arrival times.

These models will be trained on features such as latitude, longitude, speed, course, heading, and other AIS data. The resulting prediction will be compared with current kinematic ETA prediction algorithms to determine if the proposed models provide an improvement to ETA prediction techniques. We will study select vessel types on a single major shipping route.

To determine the accuracy of our model, we will compare the predicted time of arrival with the true arrival time throughout the entire journey.

3 Related Work

There has been several papers published using machine learning methodologies to improve the state-of-the-art in ETA prediction. In 2017, Parolas proposed an SVR model trained on a small set of features extracted from AIS data. It was recommended as the best tool to improve uncertainty regarding expected vessel arrival times at the Port of Rotterdam. Other works, such as Predicting Ships Estimated Time of Arrival by El Mekkaoui et al., investigate a different set of models, including LSTMs and RNNs, to improve ETA predictions. The authors in this work recommend LSTMs as their model of choice. However, they did not consider an SVR model, as Meir recommended, in their analysis. In our research, we will pit both these recommendations against each other and evaluate how they compare. Further, both works only consider a small subset of AIS features in their encoding space. El Mekkaoui et al. describe only having access to anonymized ship identity code, speed, longitude and latitude, course, heading, timestamp of the sent message, and the last port of call. Our dataset has a far richer feature set. We will leverage these additional features to better represent the shipping vessels. Lastly, while machine learning models are a popular choice, simpler kinematic models, are also being used in industry. At GA-CCRi in Charlottesville Virgina, Restrepo describes improving ETA prediction by building a speed profile for each vessel based on its history. While kinematic models are simper to implement, they cannot express all of the parameter as well as some machine learning models. In our work, we will compare the results against a simple kinematic model and decide if the added complexity of a machine learning model noticeably improved accuracy.

4 Experiments

4.1 Experiment Setup

The dataset is a collection of AIS records received by payloads hosted on the Iridium Next satellite constellation. These records are broadcasted regularly by vessels and include information relevant to a the ship and its voyage. All of the data lies along the route between Shanghai to Los Angeles, from June 2019 to October 2020. In total, the dataset contains 548,000 records.

This data required minimal pre-processing for our experiments. The biggest effort is in formatting the data in an acceptable way to train the models. The final format of the pre-processed data is described in Table 1.

In addition, we remove fields that would have limited to no effect on the training of the data. Specifically, vessel_type has a shared value of Cargo, therefore the inclusion of that field has no effect on training data. In addition country_flag, mmsi, start_port and end_port are all identifying fields that are already well known or irrelevant to the prediction of the arrival. The final format of the starting data is presented in Table 2.

We used two models to find the best predictor SVR and LSTM

4.1.1 SVR

We used SKLearn SVR and LinearSVR to find the best predictor for arrival. We use the mean absolute error to determine accuracy to reduce the impact of outliers and malformed data on the

Pre-processed Fields
latitude
longitude
timestamp
sog (speed over ground)
cog (course over ground)
width
length
vessel_type
country_flag
time remaining to destination (seconds)

Table 1: Pre-processed Fields

Processed Fields
latitude
longitude
timestamp
sog (speed over ground)
cog (course over ground)
width
length
time remaining to destination (seconds)

Table 2: Processed Fields

error. Due to the size of the data the we start with the LinearSVR for speed of training and later move to the SVR model with RBF and Sigmoid kernels.

First, we begin by splitting the data into 80% train and 20% test. The default LinearSVR regression was first run to get a baseline error value. The regression parameters C and the loss type were optimized using the GridSearchCV function. After tuning the parameters we tested various combinations of columns of the data where lat and lon were always present.

Next, the traditional SVR was trained using a subset of the original data to reduce initial run time. Like with LinearSVR the default parameters were used. Once there was a baseline, the combinations of parameters were used to try and reduce the baseline error further. Finally, we used GridSearchCV to tune the kernel, $gamma$ and $epsilon$ parameters.

4.1.2 LSTM

The second model we analyzed was a Long Short-Term Memory (LSTM) Recurrent Neural Network. These models are trained on sequences of data to predict future values in time-series data. The inputs to this model are sequences of latitude, longitude, and speed over ground. The output is the predicted time to arrival at the end of each sequence. The data from each vessel is split into sequences of length n and concatenated into a matrix of all sequences sorted by time. The average time step in the data is 817 seconds, so $n = 10$ would correlate to a 2.25 hour sequence of data.

The model used consists of one LSTM layer and one dense output layer. To optimize the performance of this model, we tested the accuracy of many different parameter configurations. We tested time steps of $n = 3$, 5, 10, and 20, along with 50, 100, and 300 nodes in the LSTM layer. We used a ReLu activation function in the LSTM layer along with Adam optimization. Tuning these parameters led to increased accuracy, though more complex models that yielded better results came with higher training costs. The highest accuracy was achieved with $n = 10$ and 300 nodes in the LSTM layer.

4.2 Experiment Results

The results of the model training and mean absolute error (MAE) are provided in two sections: SVR and LSTM. The average duration for the route we selected was 352.62 hours, and results are reported as the mean average error between predicted time to arrival and actual time of arrival.

4.2.1 SVR

The LinearSVR model was the least accurate predictor with a baseline MAE of 426720.5 seconds or equivalent to 118.5 hours. The baseline uses all fields and the default LinearSVR parameters. Next, all combinations of fields were tried and the best results were found when only the lat, lon and sog fields were used resulting in a 189042.8 second MAE or 52.51 hours. This is significant improvement of about 66 hours over the complete set of fields. Finally, we tuned C in the LinearSVR model and found the default value of $C = 1.0$ to be the best.

For the SVR regression we calculated a baseline MAE of 270588.65 where the parameters were set to there default with the lat, lon and sog fields. Then using the suggested parameters from the work done by Parolas where $C = 100$ and $gamma = 0.001$, we improved the MAE to 74281.9 or 20.6 hours. Then after tuning the parameters further for our own dataset we found the optimal configuration of RBF kernel, $C = 1000000$ and $gamma = 0.001$ with the smallest MAE error of all SVR models of 47893.8 or 13.3 hours.

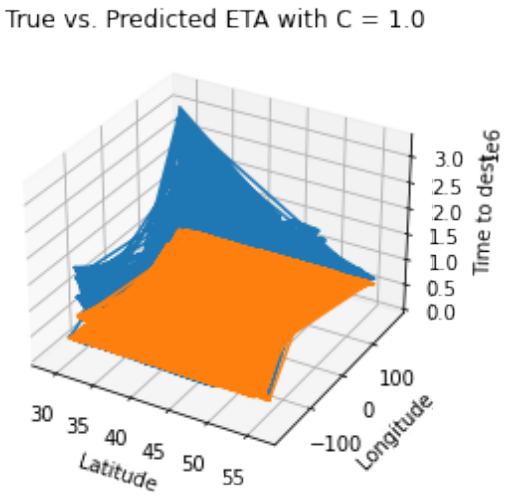

Figure 1: Predictions plotted on top of true values where C is default

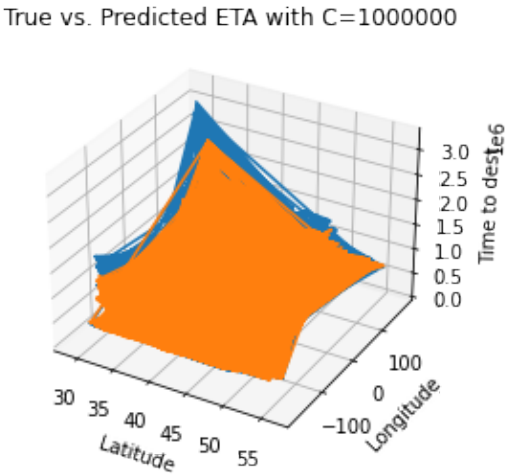

Figure 2: Predictions plotted on top of true values where C is optimal

4.2.2 LSTM

The baseline iteration of the LSTM model used a time step of $n = 10$ with 50 nodes in the LSTM yielded a MAE of 278,689 seconds or 77.41 hours. Reducing the feature set to just lat, lon, and sog reduced error to 64,766 seconds or 17.99 hours. Lowering the time step of sequences length $n = 5$ improved this error to 58,490 seconds or 16.25 hours. The optimal LSTM architecture, however, was determined to be $n = 10$ with 300 nodes in the LSTM layer. The MAE for the optimal model was 52,989 seconds or 14.72 hours. The optimal model therefore yields a percent error of 4.17% when compared to the total trip duration.

4.3 Experiment Analysis

After the experiment using the two models, it is clear that after some optimization the error significantly improved. Additionally, we find that the kinematic method of using the speed calculating time remaining using distance divided by speed is outperformed by both the SVR and LSTM model.

The kinematic model error varies greatly but sampling along the trip from Shanghai to Los Angeles finds that the average absolute error is 23.7 hours. The distribution can be seen in figure 3. The LSTM model improves the prediction error by 37.97% and the SVR model improves error by 43.88%. Which is a significant gain over the kinematic model and will improve prediction accuracy greatly, especially at the beginning of trips when there are greater unknowns.

Compared with each other the SVR and LSTM models have somewhat similar prediction accuracy, but in this case the SVR model was more accurate. The value of C had enormous effect as seen in figure 1 and figure 2. The predicted values are much closer to the true values when C is extremely large, which may indicate that the amount of data required to represent the truth is much greater than we had access to in this experiment. Despite this it had lower error by 9.5% than the LSTM model. This result is somewhat surprising as LSTM was designed to model time series data better. One possible explanation is that the was not enough data for the LSTM to outperform the SVR which usually has improved accuracy on smaller amounts of data.

5 Conclusions

In this paper we explore two separate ETA prediction methodologies. Namely, we pit a support vector regression model against a long short-term memory model. We train these models on AIS data collected by payloads hosted on the Iridium Next satellite constellation. To represent, real-world shipping routes, we focus our analysis on the popular journey from Los Angeles to Shanghai. The initial results are promising, predicting journeys within a half day of the actual arrival time.

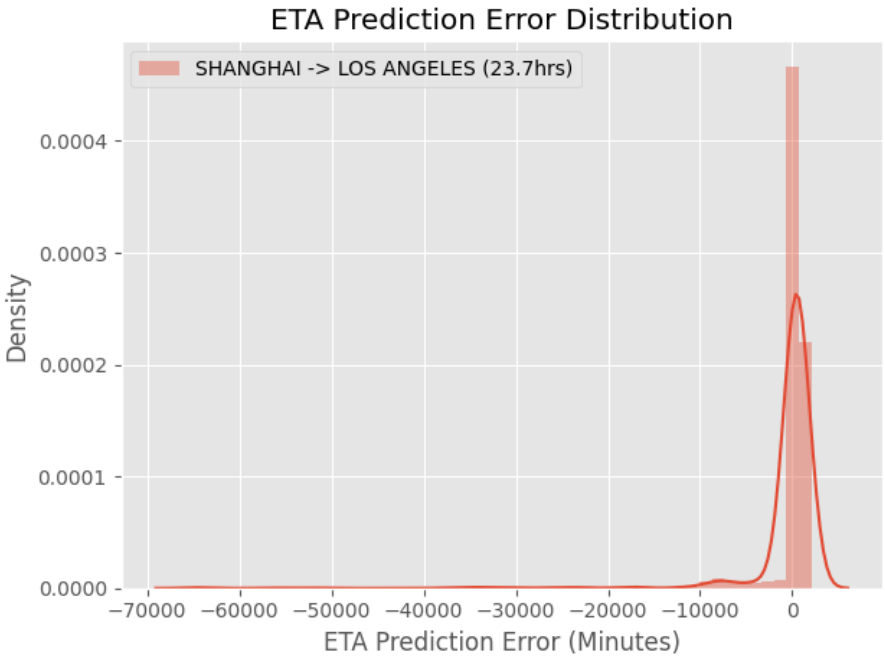

Figure 3: Kinematic Mean Absolute Error

These results are close to some of the complex kinematic models we analyzed. While the work does not outperform the kinematic models, we see opportunities in future work to expand our feature set and tune our algorithms to achieve better performance. Ultimately, the initial successes we have had suggest that machine learning approaches have a lot to offer and should continue to be explored.

ACKNOWLEDGMENTS

We thank exactEarth and GA-CCRi for providing the data for this project.

REFERENCES

Sara El Mekkaoui, Loubna Benabbou, and Abdelaziz Berrado. Predicting ships estimated time of arrival based on ais data. In *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, pp. 1–6, 2020.

ICS International Chamber of Shipping. Shipping and world trade: driving prosperity. URL `https://www.ics-shipping.org/shipping-fact/shipping-and-world-trade-driving-prosperity/`.

Ioannis Parolas. Eta prediction for containerships at the port of rotterdam using machine learning techniques. 2016.

Mark Restrepo. Analyzing and enhancing vessel schedule data, Aug 2021. URL `https://www.ga-ccri.com/analyzing-and-enhancing-vessel-schedule-data`.