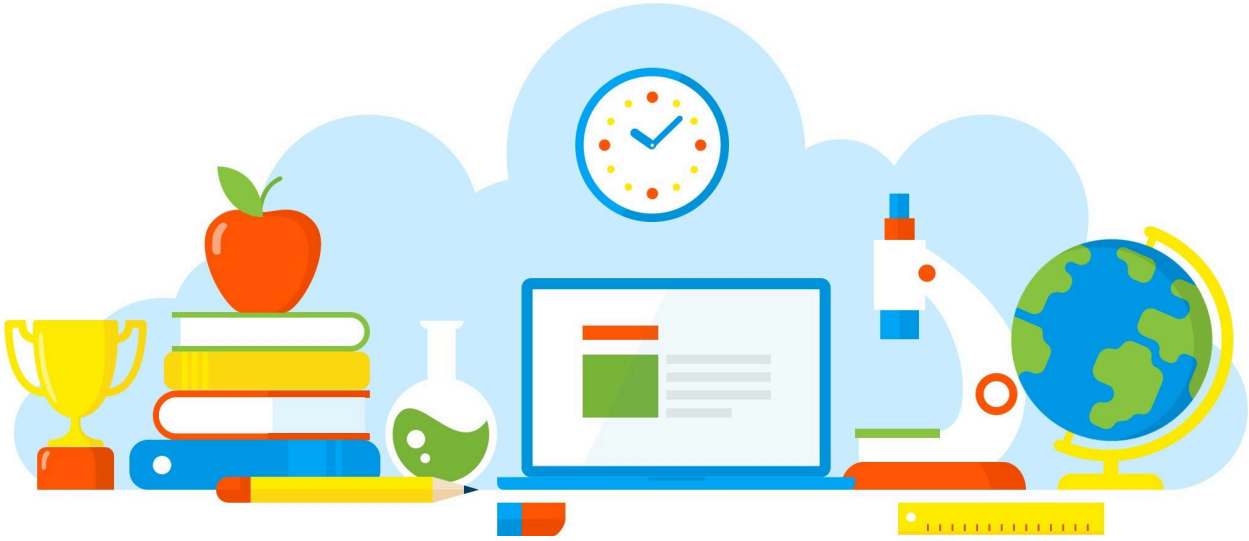


Next Vessel Destination Prediction

Amine DADOUN
15/08/2024

Abstract

Maritime shipping plays a crucial role in global trade, with vessel destination prediction being a key factor in optimizing port operations and supply chain efficiency. This paper presents an approach to predicting a vessel's next destination using machine learning techniques. We address the challenge of handling high-cardinality categorical data in maritime datasets by first narrowing down potential destination countries based on historical voyage patterns. Our method then employs a neural network model with embedding layers to effectively capture complex relationships between various features such as vessel characteristics, cargo details, and historical port visits. By utilizing embeddings, we overcome the limitations of traditional one-hot encoding for high-cardinality variables, enabling the model to learn rich, low-dimensional representations of categorical data. Our results demonstrate significant improvements in prediction accuracy, with the model achieving **56%** top-1 accuracy and **73%** top-3 accuracy in destination prediction. This approach not only enhances predictive performance but also provides a flexible foundation for future model iterations and expansions. We also provide github code of the prototype with a readable code structure, facilitating collaborative development and ongoing refinement of the prediction system.

Introduction

The maritime shipping industry plays a pivotal role in global trade, facilitating the movement of goods across oceans and continents. In recent years, the ability to predict vessel destinations has become increasingly important for various stakeholders in the shipping and commodities trading sectors. Accurate destination forecasting offers numerous benefits:

1. Supply Chain Optimization: Knowing where vessels are headed allows for better planning and coordination of port operations, reducing congestion and improving efficiency.
2. Environmental Impact: Accurate predictions can help optimize routes, potentially reducing fuel consumption and emissions.
3. Market Intelligence: For commodities traders, understanding vessel movements provides valuable insights into supply and demand dynamics.
4. Security and Safety: Predicting vessel destinations aids in maritime security efforts and enhances safety measures.
5. Resource Allocation: Ports and shipping companies can better allocate resources based on anticipated vessel arrivals.
6. Risk Management: Improved forecasting helps in assessing and mitigating risks associated with delays or changes in vessel routes.

Given these significant advantages, there has been growing interest in developing sophisticated methods for vessel destination prediction. This field has seen substantial research efforts, leveraging advances in machine learning and data analytics to improve prediction accuracy.

Literature Review

The challenge of vessel destination prediction has been approached from various angles in the literature. Early works often relied on historical data and simple statistical

models. However, with the advent of more powerful computing resources and advanced machine learning techniques, the field has seen significant advancements.

One of the seminal works in this area was conducted by Calabrese et al. (2018), who proposed a grid-based system for mapping GPS coordinates to waypoints, combined with Markov chains to estimate the most likely destination ports for ships in the Mediterranean Sea. Their approach took into account vessel characteristics such as draft and speed, demonstrating the importance of incorporating multiple features for accurate predictions.

Building on this, Zhang et al. (2020) introduced a more comprehensive approach using AIS (Automatic Identification System) data. Their method involved three main steps: building a historical trajectory database using DBSCAN clustering, measuring similarities between trajectories using Random Forest, and employing a port frequency-based decision strategy for final predictions. This work highlighted the potential of combining multiple machine learning techniques for improved accuracy.

Magnussen et al. (2021) took a different approach, utilizing graph-based representations of global tanker maritime traffic. They discretized port-to-port trajectories into sequences and used these as training data for a recurrent neural network (RNN) model. Their work demonstrated the potential of deep learning techniques in this domain, particularly in capturing the sequential nature of vessel movements.

These studies, among others, have laid a strong foundation for vessel destination prediction. However, they often face challenges when dealing with high cardinality categorical features, such as vessel IDs or port names, which are common in maritime datasets.

Our Approach and Contribution

In this work, we propose a novel approach to vessel destination prediction that addresses the challenge of high cardinality categorical features through the use of embeddings. Our method draws inspiration from recommendation systems, which face similar challenges in dealing with large numbers of users and items.

In our context, vessels can be thought of as "users" and destinations as "items". Just as recommendation systems aim to predict which items a user might prefer based on their history and characteristics, our system predicts which destinations a vessel is likely to visit next, given its past voyages and current attributes.

The key contribution of this work lies in the application of embedding techniques to handle high cardinality categorical features in the maritime domain. Embeddings allow us to represent categorical variables in a dense, low-dimensional space, capturing intrinsic relationships between different categories. This approach offers several advantages:

1. Dimensionality Reduction: Embeddings effectively reduce the dimensionality of high cardinality features, making the model more computationally efficient.
2. Capturing Semantic Relationships: Embeddings can capture semantic similarities between different categories, potentially improving the model's ability to generalize.
3. Handling Unseen Categories: Embeddings provide a way to handle new or rare categories at test time, which is particularly useful in the dynamic maritime environment.
4. Improved Model Performance: By providing a more nuanced representation of categorical features, embeddings can lead to improved prediction accuracy.

Our approach involves creating embeddings for vessel IDs, origin ports, and destination ports. These embeddings are learned jointly with the main prediction task, allowing the model to capture relevant patterns in the data. We then use these embeddings as input to a neural network model that predicts the probability distribution over possible destinations.

This method draws parallels with collaborative filtering techniques used in recommendation systems. Just as recommendation systems learn user and item embeddings to predict user-item interactions, our system learns vessel and port embeddings to predict vessel-destination relationships.

By framing the problem in this way, we can leverage advances from the field of recommendation systems to improve vessel destination prediction. This includes techniques for handling cold-start problems (new vessels or ports), incorporating side

information (such as vessel characteristics or port attributes), and dealing with the inherent sparsity of maritime trajectory data.

In the following sections, we will detail our methodology, including data preprocessing, model architecture, and training procedure. We will then present experimental results demonstrating the effectiveness of our approach, comparing it to baseline methods and discussing its strengths and limitations. Finally, we will conclude with a discussion of the implications of this work and potential directions for future research.

Through this work, we aim to contribute to the growing field of maritime analytics, offering a novel approach to vessel destination prediction that can enhance decision-making processes for various stakeholders in the shipping and trading industries.

Dataset

Port Calls and Vessels

The dataset encompasses maritime traffic data for the year 2023, providing comprehensive information on vessel movements, characteristics, and port activities. This section presents an analysis of the key features within the dataset, focusing on port calls and vessel attributes. For the ease of the analysis, the port calls and vessels dataset have been joined.

Vessel Type Distribution:

The dataset primarily consists of tanker vessels, with four main categories dominating the fleet composition:

Vessel Type	Percentage
Crude Oil Products Tanker	42.2%
Chemical/Oil Products Tanker	26.2%
Products Tanker	17.3%
Crude Oil Tanker	9.4%
Other types	4.9%

These four categories account for 95.1% of all vessels in the dataset, highlighting the focus on oil and chemical shipping sectors.

Port Call Frequency:

Analysis of port calls reveals significant variations in both vessel activity and port utilization. The distribution of port calls per vessel is right-skewed, with the following statistics:

Statistic	Value
Mean port calls	39.02
Median port calls	30.00
Minimum port calls	1
Maximum port calls	323

Similarly, port calls per destination exhibit a right-skewed distribution:

Statistic	Value
Mean port calls	57.52
Mean calls per destination	28.00
Median calls per dest	1
Minimum calls per dest.	317

Singapore, Rotterdam, and Ulsan emerged as the most frequented ports, indicating their significance in global maritime trade networks.

Visit Duration and Cargo Operations:

Port visit durations predominantly range from 10 to 40 hours, with a median of approximately 25 hours. The distribution is right-skewed, featuring a concentration of brief visits under 10 hours and an extended tail representing prolonged stays. A slight positive correlation was observed between visit duration and cargo volume, particularly for Crude/Oil Products and Crude Oil Tankers, which generally handle larger cargo volumes.

Vessel Age Distribution:

The age distribution varies across vessel types. Crude/Oil Products Tankers and Chemical/Oil Products Tankers exhibit similar age profiles, with median ages around 15 years. Product Tankers show a lower median age, suggesting a relatively younger fleet in this category. Crude Oil Tankers display the widest interquartile range, indicating more variability in vessel ages. Notably, some vessels across all categories exceed 40 years of age.

Anomaly Detection:

Using z-score analysis ($|z| > 3$), we identified anomalies in cargo volume and visit duration:

- 7,030 anomalous cargo volume records

- 1,926 anomalous visit duration records

Crude Oil Tankers were predominantly associated with cargo volume anomalies, while visit duration anomalies spanned various vessel types, with some durations exceeding 700 hours.

Trades

The dataset represents maritime trade information, combining port call and vessel data. It includes details on individual trades, such as origin and destination ports, traded volumes, distances, and timestamps. This dataset results from joining the previously discussed port calls dataset with vessel information, providing a holistic view of maritime shipping activities where each record represents a specific trade journey, linking port calls with the associated vessel and cargo details.

The analysis of the maritime trade dataset provides comprehensive insights into vessel operations, cargo distribution patterns, and trade dynamics. In the notebook exploring trade data, one figure illustrates the relationship between voyage distance and duration for the top four vessel types, with outliers removed. Chemical/Oil Products Tankers and Crude Oil Tankers demonstrate the widest range of both distance and duration, suggesting their involvement in diverse global trade routes. Crude Oil Tankers, in particular, show a distinct cluster of long-distance voyages (15,000-25,000 km) with durations of 30-60 days, indicative of intercontinental oil transportation, likely between major oil-producing regions and consumption centers.

Another figure depicts the monthly traded volume over the year 2023, revealing a relatively stable trend from January to September, fluctuating around 400 million units (presumably barrels or tons). However, a significant decline was observed in the last quarter, with volumes dropping to approximately 200 million units by December. This sharp decrease could be attributed to seasonal variations, geopolitical events, or global economic factors affecting maritime trade volumes.

In the notebook, we have 2 box plots of voyage distance and duration, respectively, for the top four vessel types. Crude Oil Tankers exhibit the highest median values (about 7,000 km and 18 days) and widest interquartile ranges for both metrics, consistent with their global operational scope. Chemical/Oil Products Tankers show a wide range but lower median values, suggesting a mix of short and long-haul operations. In contrast, Products Tankers show the lowest median values (around 1,000 km and 4 days) and narrowest ranges, indicating more localized or regional operations, possibly serving coastal or short-sea shipping routes.

The distribution of product families, as shown in the notebook figure, it indicates that clean petroleum products dominate the traded goods at approximately 41%, followed by

crude oil/bio at around 28%. Dirty petroleum products and chemical/bio products each constitute about 15% of the trade volume. This distribution reflects the significant role of refined petroleum products in global maritime trade, while also highlighting the diversification of cargo types. The presence of minor bulks and other categories, albeit in smaller percentages, underscores the variety of commodities transported by sea.

Statistical analysis further supports these observations. For instance, Chemical/Oil Products Tankers have the highest count of voyages (30,657) but a lower mean distance (1,709 km) compared to Crude Oil Tankers, which have fewer voyages (10,811) but a much higher mean distance (7,602 km). This suggests that Chemical/Oil Products Tankers are more frequently employed for shorter routes, while Crude Oil Tankers are primarily used for long-haul transportation.

These findings provide valuable insights into the operational characteristics of different vessel types and the composition of maritime trade. The data reveals clear distinctions in voyage patterns among vessel types, likely influenced by the nature of their cargo and global trade demands. The seasonal variation in trade volumes and the predominance of petroleum products in the cargo mix highlight the dynamic nature of maritime trade and its close ties to global energy markets. Such insights can inform strategic decision-making in the shipping industry, port operations, and global trade policies, particularly in areas of fleet management, route optimization, and infrastructure development.

Data Filtering and Augmentation

In this section, we give details on some data transformation that we made in order to clean a bit of the data and prepare for the model training.

Multiple products trades:

We first notice that there are some trades records for which the vessel makes the same trip with all variables that are equal except the product variable that has a different value, that means transported that the vessel can transport several products and this is represented in different records. Thus we decide to merge these records, and create an aggregate column namely products that is replacing the original product column.

By Performing this first data transformation we reduce the port calls dataset from 240.038 records to 221.591 records.

Voyage Definition:

In our data transformation, we define a voyage as a sequence of port calls made by a vessel, starting from a unique origin. This definition is operationalized using a dense ranking system based on the vessel's ID and the origin port call ID. Each time a vessel departs from a new origin port, it marks the beginning of a new voyage, regardless of the number of intermediate stops. This approach allows us to track the complete journey of a vessel from its initial departure point through multiple destinations, providing a comprehensive view of maritime trade.

Finally in order to enhance data quality and relevance for our analysis and port-to-port destination modeling, we perform the following steps:

1. **Duplicate Visit Elimination:** We removed instances where a vessel visited the same destination multiple times from a single origin within the same voyage. Only the first visit was retained. This step helps in focusing on unique port-to-port movements and prevents overrepresentation of repeated short-distance shuttles.
2. **Origin Consistency:** In cases where a vessel arrived at a destination from multiple different origins, we kept only the most recent origin-destination pair. This transformation ensures that each destination in our dataset is associated with a single, most relevant origin, simplifying route analysis and reducing noise from potential data inconsistencies.

3. **Self-Loop Removal:** We eliminated records where the origin port was identical to the destination port. This step removes potential data errors or specialized operations (like offshore loading) that do not represent typical port-to-port voyages, ensuring our analysis focuses on actual maritime routes.

These transformations serve to reduce data redundancy, enhance the clarity of vessel movement patterns, focus the analysis and prototyping on meaningful port-to-port routes and minimize the impact of potential data recording errors or specialized operations.

By implementing these data cleaning steps, we created a refined dataset that more accurately represents the global maritime trade network, allowing for more reliable analysis of trade routes, vessel behaviors, and port interactions. More concretely, we reduced the size of the dataset from 197.301 records that are obtained after the joint operation between port calls and trades datasets to 140.196 records.

Data Augmentation

To address the challenge of high cardinality in port destination prediction, we augmented our maritime trade dataset using an open-source geographical database. This database, containing detailed country boundary information including ISO codes, continent, and region classifications, was integrated with our preprocessed trade data through a geospatial joining procedure.

The primary motivation for this data augmentation was to reduce the complexity of our prediction task. With 1,578 unique port destinations, we faced an extreme multi-class classification problem. By enriching our data with country, region, and continent codes, we aimed to implement a hierarchical prediction approach: first predicting the most probable visited countries, then narrowing down to specific ports within those countries.

Our geospatial joining procedure involved two steps:

1. Exact matching: Identifying the country whose boundaries contain the port coordinates.

2. Proximity matching: For ports not exactly within country boundaries, finding the closest country based on geographical distance.

This approach yielded comprehensive geographical context for all origin and destination ports in our dataset. The results of this augmentation were significant:

- 42% of ports were exactly matched to their respective countries.
- 58% were assigned to the closest country, with an average distance of 6.5 km to the nearest border.

This enriched dataset provides a robust foundation for our hierarchical prediction model, potentially improving the accuracy and efficiency of port destination forecasting in maritime trade analysis.

If we consider the history of vessels, we have on average ~5 visited countries per vessel (with a maximum of 15 countries). Hence, if we compute the potential ports that could be predicted based on the historical visited countries, the volume drops from 1,578 potential port destinations to 150 possible port destinations that belong to the past visited countries. Thus we divided the number of possible predicted destinations by 10 on average. This is not only beneficial for prediction accuracy but also for prediction time.

Processed Trades Data

NB: Please relate to EDA_ProcessedTrades notebook

1. Analysis of Historical Routes and Patterns:

Figure 1 illustrates the relationship between distance, traded volume, and selected major ports (Ningbo, Singapore, Mumbai, Antwerp, and New York). This visualization reveals several key insights:

- Singapore stands out with the highest number of trips (3,380) and the largest average traded volume (302,076), suggesting its role as a major global hub for maritime trade.

- Antwerp shows a wide range of distances (up to 31,854,322 km) with relatively consistent volumes, indicating its importance in both short and long-distance European trade routes.
- Mumbai has the lowest trip count (925) among these ports, but with significant volumes, possibly indicating its role in specific, high-volume trade routes.
- Ningbo demonstrates a pattern of shorter distance, high-volume trades, reflecting its position in regional Asian trade networks.
- New York shows a unique pattern of very short-distance, lower-volume trades, which could indicate its focus on coastal or short-sea shipping within North America.

The varying correlations between distance and traded volume for each port (ranging from -0.0067 for New York to 0.5413 for Ningbo) suggest that different ports have distinct roles in global trade networks, influenced by their geographical location and economic hinterlands.

2. Impact of Vessel Type on Destination (Figure 5):

The treemap in Figure 5 provides a clear visualization of the most common destinations by vessel type:

- Crude/Oil Products Tankers dominate the chart, with Singapore, Antwerp, and Rotterdam as primary destinations. This suggests these ports have significant oil refining or storage capabilities.
- Chemical/Oil Products Tankers show a preference for Ulsan and Ningbo, indicating these ports' specialization in chemical processing or distribution.
- Product Tankers have a more diverse set of destinations, with notable presence in Asian ports like Singapore and Kaohsiung.
- The smaller representation of other vessel types implies a concentration of the dataset on oil and chemical transportation.

This distribution reflects the specialization of ports and regions in handling specific types of cargo, as well as the global patterns of energy and chemical product trade.

3. Seasonal Patterns in Vessel Movements (Figure 3):

Figure 3 presents a heatmap of monthly patterns in vessel destinations for the top 10 ports:

- Singapore consistently shows high activity throughout the year, with slight increases in the middle and end of the year.
- Rotterdam and Antwerp display similar patterns, with higher activity in the summer months (June-August). This could be due to increased energy demand in Europe during this period.
- Ports like Ningbo and Incheon Port show more variability, with distinct peak seasons.
- New York and Mumbai demonstrate relatively stable patterns throughout the year, suggesting less seasonal influence on their trade.

These patterns could be influenced by factors such as regional weather conditions, seasonal energy demands, or cyclical economic activities in different parts of the world.

4. Influence of Cargo Type on Destination (Figure 2):

Figure 2 illustrates the relationship between cargo type, volume, and trip frequency:

- Crude oil/condensate shows the highest average volumes per trip, with some destinations handling extremely large shipments (up to 298,419 units).
- Clean petroleum products have the highest number of total trips (57,165), indicating frequent, smaller shipments.
- Chem/bio products show a wide range of volumes but fewer trips, suggesting more specialized trade routes.
- Dirty petroleum products have the lowest average volume but a significant number of trips, possibly indicating regular short-distance transportation.

This distribution reflects the global energy trade structure, with crude oil moving in large volumes to refining centers, and refined products being more widely distributed in smaller quantities.

5. Impact of Distance on Destination Choice (Figure 1):

Revisiting Figure 1 with a focus on distance:

- Singapore shows trades across a wide range of distances, confirming its role as a global hub.
- Antwerp's long-distance trades (up to 31,854,322 km) suggest its importance in intercontinental routes, particularly for European imports/exports.
- Mumbai's pattern indicates it primarily serves long-distance trades, possibly connecting Asia with Europe or the Americas.
- Ningbo's concentration of shorter-distance, high-volume trades reflects its role in Asian regional trade.
- New York's short-distance pattern suggests a focus on coastal trade or servicing the immediate North American market.

The varying distance patterns for each port highlight how geographical location, port facilities, and economic relationships shape global maritime trade routes.

7. Vessel Flag and Age Analysis:

Figure 5 shows the top destinations by vessel flag, while Figure 4 illustrates the influence of vessel age on destinations.

Vessel Flag Analysis (Figure 5):

- The pie chart indicates a diverse range of flags represented in the dataset.
- Major destinations like Singapore, Ningbo, and Rotterdam appear multiple times, suggesting they attract vessels from various flag states.
- This diversity could indicate the global nature of maritime trade and possibly the use of flags of convenience.

Vessel Age Analysis (Figure 4):

- There's a concentration of vessels between 5-20 years old across all major destinations.
- Singapore, Incheon Port, and Rotterdam show the highest number of trips for vessels around 15 years old.
- Fewer trips are observed for very new (0-5 years) and older (20+ years) vessels.
- This pattern might reflect the optimal operational age for vessels, balancing efficiency and operational costs.

8. Port Restriction/Environment Factors Analysis:

Key Observations:

- Most destinations show a concentration of draught changes between -5 and 0 meters.
- Mumbai and New York show a narrower range of draught changes compared to ports like Singapore or Rotterdam.
- Chemical/Oil Products Tankers show the widest range of draught changes across most ports.

The variation in draught changes across ports could indicate differences in port infrastructure, tidal conditions, or loading/unloading practices. Ports with a wider range of draught changes (e.g., Singapore, Rotterdam) might have more flexible infrastructure or handling capabilities.

Next Port Destination Prediction: Key Insights from Data Analysis

1. Regional Patterns and Intra-Regional Preferences:

The Region-to-Region Trade Heatmap reveals a strong tendency for vessels to remain within the same region for their next destination. This intra-regional preference is crucial for our prediction model:

- High probability of short-range predictions: Our model should give higher weights to destinations within the same region as the origin port.
- Regional specialization: We can develop region-specific sub-models that capture unique patterns within each geographical area.

2. Vessel Type and Destination Relationship:

Figure 1 shows the distribution of vessel types across destination regions:

- Eastern Asia and South-Eastern Asia attract a diverse range of vessel types, suggesting these regions should have more complex prediction patterns.
- Some regions show a preference for specific vessel types, which can be a strong predictor for next port destination.

- Incorporating vessel type as a feature in our model will likely improve prediction accuracy, especially for regions with clear vessel type preferences.

3. Global Trade Routes and Major Hubs:

The Top 20 Country-to-Country Trade Routes (Figure 4) and the list of top origin and destination countries (Image 3) provide valuable insights:

- China, the United States, and Japan appear as major destinations across multiple routes. Our model should give higher probabilities to these countries as potential next destinations, especially for vessels originating from their top trading partners.
- The strong bilateral trade between certain country pairs (e.g., Indonesia-China, Russia-China) suggests that for vessels originating from one country in these pairs, the other country has a high probability of being the next destination.

4. Product Family Influence:

The observation about product family distribution varying by region (e.g., clean petroleum products being more common in African regions) is crucial for our prediction model:

- Product type should be a key feature in our model, as it strongly influences the likely next destination.
- We can create product-region association matrices to boost prediction probabilities for destinations that commonly receive specific product types.

5. Diversity of Destinations:

With 176 unique destination countries and 22 destination regions (Image 3), our model needs to handle a large number of possible outcomes:

- This suggests we might need a hierarchical prediction approach: first predicting the destination region, then the specific country or port within that region.
- Alternatively, we could use a multi-class classification model with techniques to handle a large number of classes, such as softmax with temperature or hierarchical softmax.

6. Seasonal and Temporal Patterns:

While not explicitly shown in these images, we should consider incorporating temporal features:

- Monthly or seasonal patterns in trade volumes or routes could be valuable predictors of next destinations.
- Historical data on a vessel's previous routes could inform predictions of its future destinations.

7. Port Infrastructure and Vessel Characteristics:

The relationship between draught changes and destinations (from previous analyses) suggests:

- Port infrastructure limitations could be a key factor in predicting next destinations. Vessels may be more likely to visit ports that can accommodate their size and draught.
- Incorporating vessel characteristics (size, type, age) and port capabilities in our model could significantly improve prediction accuracy.

In conclusion, our next port destination prediction model should leverage these insights by:

1. Prioritizing intra-regional predictions while accounting for major global trade routes.
2. Incorporating vessel type, product family, and origin-destination pair history as key features.
3. Considering a hierarchical or multi-stage prediction approach to handle the large number of possible destinations.
4. Including temporal features to capture seasonal or cyclical patterns in trade.
5. Factoring in vessel characteristics and port infrastructure capabilities.

Data Preparation and Preliminary Analysis

To evaluate our model's predictive capabilities, we implemented a temporal split in our dataset. This approach involves isolating the most recent trip for each vessel as our test set, while using all previous trips as our training data. This temporal separation ensures

that our model is tested on future trips, closely mimicking real-world prediction scenarios.

Before developing our predictive model, we conducted a preliminary analysis to assess the potential effectiveness of a hierarchical prediction method. Using the training set, we examined the historical patterns of each vessel's voyages. Our findings revealed a significant insight: when considering the list of ports within countries previously visited by a vessel, we found a 78% likelihood that the actual destination would be among these ports. In contrast, when we narrowed our focus to only the specific ports previously visited by a vessel, the likelihood of the actual destination being in this set dropped to 48%.

This substantial difference in predictive power (78% vs. 48%) strongly suggests the value of a two-tiered, hierarchical approach to destination prediction. By first predicting the destination country and then narrowing down to specific ports within that country, we can potentially leverage the higher accuracy at the country level to improve our overall port predictions. This insight forms a crucial foundation for our model design, indicating that incorporating historical country-level data could significantly enhance our prediction accuracy.

Approach

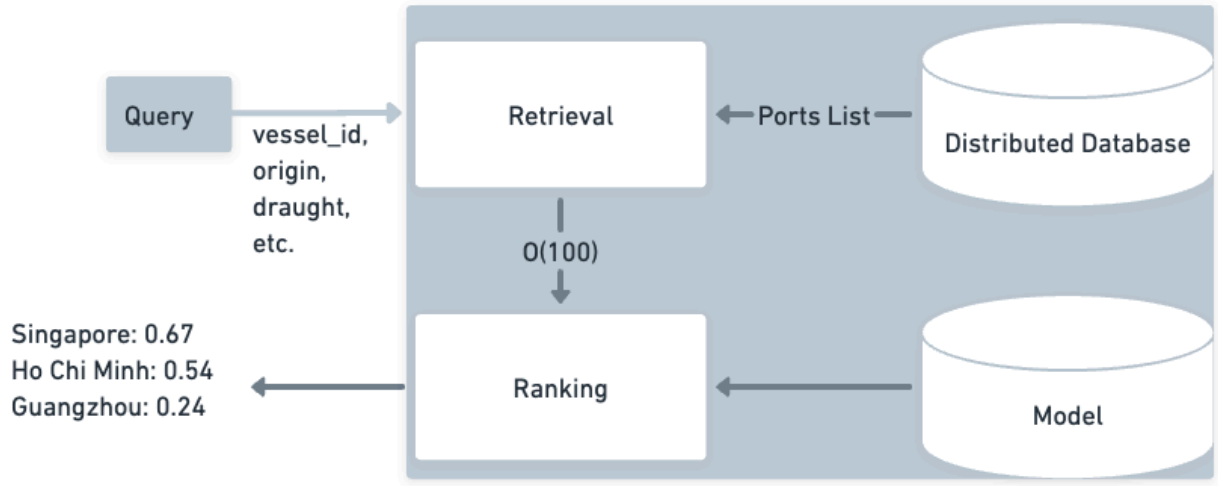
An overview of the next port destination prediction system is shown in the figure below. A query, which includes vessel identifier and contextual features such as origin, draught change, and other relevant information, is generated when a prediction for the next destination is needed. The system returns a ranked list of potential destination ports with their associated probabilities.

Since there are over 1,600 possible port destinations in the database, it is computationally intensive to exhaustively score every port for every query within a required serving latency when we are in a production set up. Therefore, the first step upon receiving a query is retrieval. The retrieval system accesses a distributed database containing historical port data and returns a shortlist of approximately 100 ports

(O(100)) that best match the query using various signals, usually a combination of machine-learned models and historical patterns.

After reducing the candidate pool, the ranking system ranks all retrieved ports by their scores. The scores are typically P(y|x), the probability of a port being the next destination y given the features x, including origin port, vessel features (e.g., vessel type, size), contextual features (e.g., current location, time of year), and traded information (e.g., product family, traded volume).

The final output is a ranked list of the most probable next destinations. In the example shown, Singapore has the highest probability at 0.67, followed by Ho Chi Minh at 0.54, and Guangzhou at 0.24.



Location Embeddings For Destination Prediction

Our proposed model architecture consists of three primary components: (1) an element-wise product module inspired by collaborative filtering, (2) a feed-forward neural network for feature integration, and (3) a final step to merge these two components. This approach allows for effective handling of high-cardinality categorical variables while capturing complex relationships between different entities in the maritime domain.

Element-wise Product Module

The element-wise product module is designed to capture latent relationships between key entities, similar to matrix factorization techniques in collaborative filtering. Let $e_v \in \mathbb{R}^d$, $e_o \in \mathbb{R}^d$, and $e_d \in \mathbb{R}^d$ represent the embeddings of vessel, origin port, and destination port respectively, where d is the embedding dimension. We compute two element-wise products:

1. Vessel-Destination interaction: Dot product of the two embeddings
2. Origin-Destination interaction: Dot product of the two embeddings

where $\odot$ denotes the Hadamard product. These operations enable the model to learn intrinsic affinities between vessels and destinations, as well as transition probabilities between port pairs.

Feed-Forward Neural Network

The feed-forward component of our model integrates various feature types:

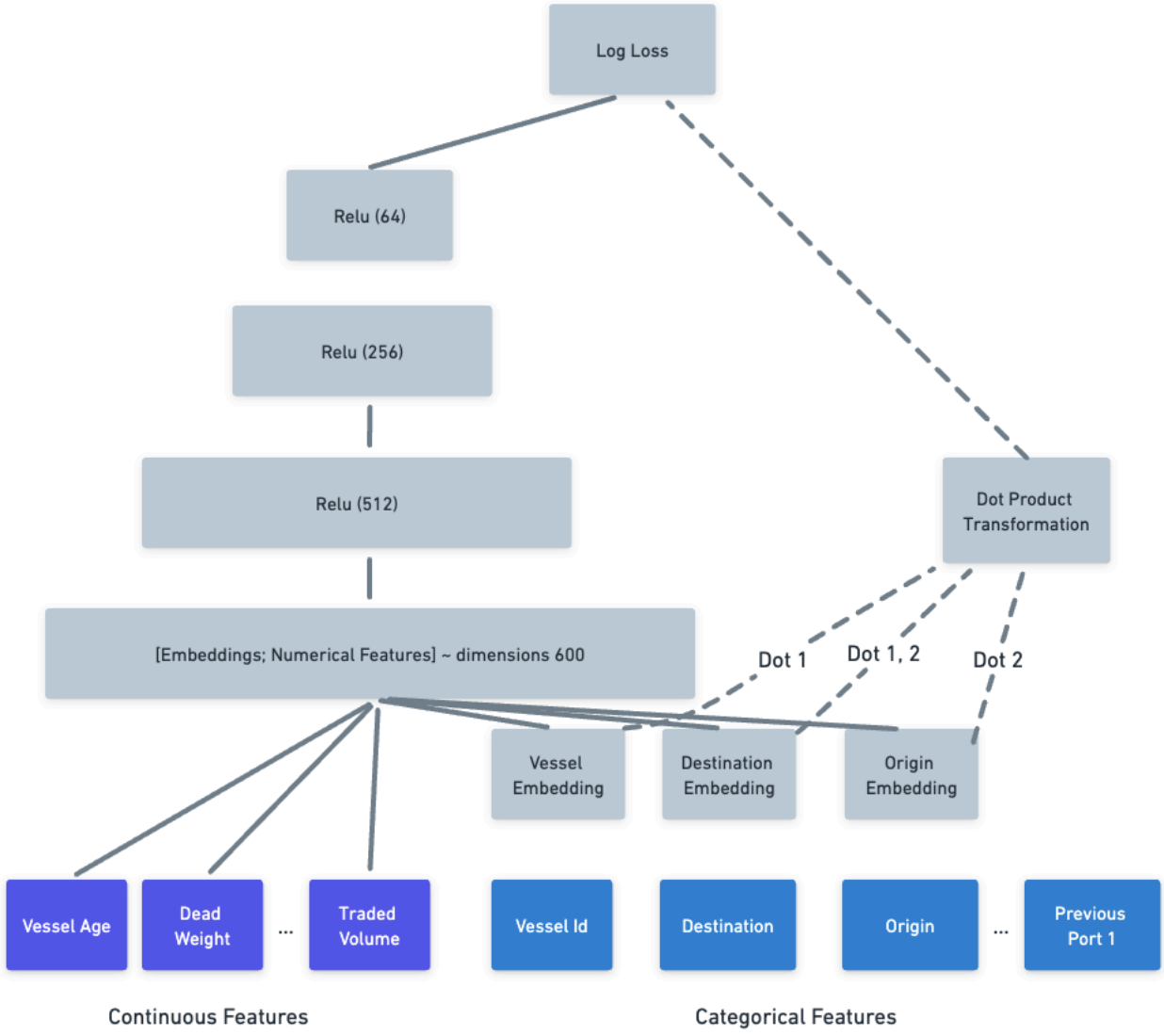
1. Embedding layers transform high-cardinality categorical variables into dense vectors
2. Numerical features are normalized and concatenated with the embedded categorical features:
3. The concatenated features are processed through multiple dense layers with ReLU activation

Final Architecture Layer

The outputs from the element-wise product module and the feed-forward network are merged as follows:

1. Concatenation of all components: output of the final hidden layer of the feed-forward network.
2. Final dense layer with sigmoid activation.

The model is trained end-to-end using binary cross-entropy loss



Feature Engineering

Temporal Feature Extraction

To capture cyclical patterns in maritime traffic, we extracted temporal features from the parsed start date and time of each voyage. Specifically, we derived:

1. Day of Week: This feature enables the model to discern weekly patterns in maritime activities.
2. Month: Extraction of the month facilitates the capture of seasonal trends in shipping routes and frequencies.

These temporal features allow our model to account for time-dependent variations in maritime behavior, potentially improving its predictive accuracy.

Voyage Sequencing

To establish a chronological context for each maritime journey, we implemented a series of sequencing features:

1. Voyage Number: A dense rank of trips for each vessel, providing a unique identifier for each voyage.
2. Leg Index: This feature represents the sequence of stops within a specific voyage.
3. Stop Index: An overall sequence of stops for a vessel, irrespective of voyage boundaries.

These sequencing features enable our model to comprehend the temporal ordering of maritime activities, potentially uncovering patterns in route selection based on voyage history.

Historical Port Visits

Leveraging historical data is crucial for predicting future behavior. We engineered several features to capture a vessel's port visit history:

1. Previous Visited Ports: The last three distinct ports visited by each vessel.
2. Previous Ports List: An array of all previous ports in the current voyage.
3. Previous Visited Ports List: An comprehensive array of all previously visited ports for each vessel.
4. Previous Different Visited Ports List: A list of unique previously visited ports, eliminating consecutive duplicates.

These historical features provide our model with valuable context about a vessel's past behavior, potentially informing predictions about future destinations.

Distance Features

To incorporate geographical context into our model, we compute distance from Previous Port: Calculates the distance traveled from the last port of call.

These features allow our model to consider spatial relationships and travel distances in its predictions.

Cargo and Vessel Features

Characteristics of the cargo and vessel play a significant role in determining shipping routes. We included the following features:

1. Draught Change: Captured for both origin and destination ports, indicating loading/unloading activities.
2. Product Family and Products: Categorization of the cargo being transported.
3. Vessel Characteristics: Including dead weight, flag name, build year and vessel age.

These features enable our model to account for the physical and operational constraints of different vessel types and cargo categories.

Port and Country Features

To capture patterns in international trade, we incorporated features related to ports and countries:

1. Origin and Destination Country Codes: Enabling analysis of international trade patterns.
2. Cargo Volume: Captured for both origin and destination ports.

Rolling Window Features

Finally, we introduced a set of rolling window features based on historical data. These features capture temporal dynamics and evolving patterns in our model with valuable context for more accurate predictions. The rolling window approach allows us to compute probabilities and statistics that adapt to recent trends while still considering longer-term patterns.

1. Rolling Window Probabilities:

We compute four different levels of rolling window probabilities, each capturing increasingly specific patterns:

- Probability of a destination given an origin:
  - Calculated weekly for each origin-destination pair
  - Provides a baseline probability based on general trade routes

- Probability of a destination given an origin and vessel ID
  - Computed weekly for each unique combination of origin, vessel ID, and destination
  - Captures vessel-specific patterns and preferences

- Probability of a destination given an origin, vessel ID, and product family
    - Calculated weekly for each unique combination of origin, vessel ID, product family, and destination
    - Incorporates cargo type into the probability calculation

- Probability of a destination given vessel ID, product family, and previous visited port
    - Computed weekly for each unique combination of vessel ID, product family, previous visited port, and destination
    - Captures the influence of the immediate previous port on the next destination

These rolling window probabilities are calculated using cumulative counts over time, allowing the model to capture both long-term trends and recent changes in maritime trade patterns.

2. Origin-Destination Distance:

We included the average distance between each origin-destination pair as a feature. This information helps the model understand typical voyage lengths and potential constraints on vessel movements.

3. H3 Geospatial Index:

To handle cases where direct origin-destination distance data might be missing, we implemented a fallback using the H3 geospatial indexing system:

- Origin and destination coordinates are converted to H3 indices at resolution 2

- Average distances between H3 index pairs are calculated

- This provides an approximate distance when exact port-to-port distances are unavailable

The combination of these rolling window features with our existing static features creates a comprehensive input set that balances historical trends with current context, potentially leading to more accurate and nuanced predictions of vessel destinations.

Experiments

Evaluation Metrics

To comprehensively assess the performance of our vessel destination prediction models, we employ a set of metrics that capture different aspects of predictive accuracy. These metrics are particularly well-suited for evaluating ranking problems in (e.g. recommendation systems) and information retrieval tasks. The chosen metrics are:

1. Accuracy
    - Definition: The proportion of cases where the model correctly predicts the actual destination as the top-ranked choice.
    - Interpretation: This metric provides a straightforward measure of how often our model is exactly correct in its primary prediction.
    - Calculation: (Number of correct top predictions) / (Total number of predictions)
2. Top-3 Accuracy
    - Definition: The proportion of cases where the actual destination appears within the top 3 ranked predictions.
    - Interpretation: This metric allows for some flexibility, recognizing that having the correct destination among the top few predictions can still be valuable in practical applications.
    - Calculation: (Number of times actual destination is in top 3) / (Total number of predictions)
3. Top-10 Accuracy
    - Definition: The proportion of cases where the actual destination appears within the top 10 ranked predictions.

    - Interpretation: This metric further relaxes the strictness of evaluation, accounting for scenarios where a broader set of likely destinations is useful.
    - Calculation: (Number of times actual destination is in top 10) / (Total number of predictions)
4. Mean Reciprocal Rank (MRR)
    - Definition: The average of the reciprocal ranks of the actual destinations in the predicted rankings.
    - Interpretation: MRR provides a nuanced view of model performance, rewarding higher placements of the correct destination while still giving credit for correct predictions lower in the ranking.
    - Calculation: MRR = (1/N) * Σ(1/rank_i), where N is the number of queries and rank_i is the position of the correct destination for the i-th query.
    - Range: 0 to 1, where 1 is the best possible score (all correct destinations ranked first).

These metrics offer a balanced evaluation approach:

- Accuracy focuses on perfect predictions.
- Top-3 and Top-10 Accuracy allow for near-misses, which can be practically useful in decision-support scenarios.
- MRR provides a single scalar value that captures overall ranking performance, weighing higher ranks more heavily.

Baseline Models

To establish a comprehensive benchmark for our vessel destination prediction task, we implemented three distinct baseline models. Each model represents a different approach to the problem, ranging from simple statistical methods to more complex machine learning techniques. These baselines provide a robust foundation for comparing the performance of our proposed neural network model.

1. Statistical Model (Probability-Based Baseline)
    - Approach: This model leverages the rolling window probabilities (P1, P2, P3, P4) computed from historical data.

- Implementation:
  - The model uses a hierarchical approach, starting with the most specific probability (P4) and falling back to more general probabilities when necessary.
  - For each prediction, it first checks P4 (vessel ID, product family, previous port). If not available, it falls back to P3, then P2, and finally P1.
  - The destination with the highest probability is selected as the prediction.
- Advantages:
  - Simple and interpretable
  - Captures historical patterns and trends
  - Computationally efficient
- Limitations:
  - May not capture complex, non-linear relationships in the data
  - Limited ability to generalize to new scenarios

2. Decision Tree Model
   - Approach: A single decision tree model trained on the feature set, with categorical variables encoded using target encoding.
   - Implementation:
     - Features include all engineered features described in the Feature Engineering section.
     - Categorical variables are encoded using target encoding, which replaces categories with their corresponding target mean values.
     - The model is trained using standard decision tree algorithms (e.g., CART).
   - Advantages:
     - Highly interpretable model structure
     - Can capture non-linear relationships
     - Handles mixed data types well
   - Limitations:
     - Prone to overfitting, especially with deep trees
     - May not capture some complex patterns in the data
3. Gradient Boosting Decision Tree (LightGBM)

    - Approach: An ensemble of decision trees using gradient boosting, implemented with the LightGBM framework.
    - Implementation:
        - Uses the same feature set as the single decision tree model.
        - LightGBM is chosen for its efficiency and effectiveness in handling large datasets with high-dimensional feature spaces.
        - Hyperparameters are tuned using cross-validation to optimize performance.
    - Advantages:
        - High predictive power, often achieving state-of-the-art performance on tabular data
        - Can capture complex, non-linear relationships
        - Efficient training and inference, even on large datasets
    - Limitations:
        - Less interpretable than single decision trees
        - Risk of overfitting if not properly regularized
        - Requires careful hyperparameter tuning for optimal performance

These baseline models provide a spectrum of approaches against which we can compare our proposed neural network model:

- The statistical model serves as a simple, interpretable baseline that leverages domain knowledge.
- The decision tree model offers a balance between interpretability and the ability to capture non-linear relationships.
- The LightGBM model represents a powerful, state-of-the-art approach for tabular data, setting a high bar for performance.

By comparing our proposed model against these baselines, we can assess its relative strengths and weaknesses across different evaluation metrics. This comparison will help us understand whether the added complexity of our neural network approach translates to meaningful improvements in predictive accuracy and generalization capability for the vessel destination prediction task.

Neural Network VS other baselines:

To evaluate the efficacy of our proposed neural network model, we conducted a comparative analysis against two baseline approaches: a statistical model and a Light Gradient Boosting Machine (LGBM) model. The performance of each model was assessed using four key metrics: Mean Reciprocal Rank (MRR), Top 1 Accuracy, Top 3 Accuracy, and Top 10 Accuracy. The results of this comparison are presented in the table below.

Model	MRR	Top 1 Accuracy	Top 3 Accuracy	Top 10 Accuracy
Statistical Model	0.32	0.21	0.38	0.56
LGBM	0.50	0.40	0.57	0.68
Neural Network	0.65	0.56	0.73	0.78

The neural network model consistently outperforms both baseline models across all evaluation metrics. With an MRR of 0.65, it demonstrates a substantial improvement over the LGBM (0.50) and statistical (0.32) models. This indicates that the neural network more frequently ranks the correct destination higher in its predictions.

In terms of Top 1 Accuracy, the neural network achieves 56% accuracy, significantly surpassing the LGBM (40%) and statistical (21%) models. This suggests that the neural network is more adept at precisely identifying the correct destination as its primary prediction.

The performance disparity is further evident in the Top 3 and Top 10 Accuracy metrics. The neural network's Top 3 Accuracy of 73% and Top 10 Accuracy of 78% are notably higher than those of the LGBM (57% and 68%, respectively) and statistical (38% and 56%, respectively) models. This indicates that even when the neural network's top prediction is incorrect, it is more likely to include the correct destination within its top few predictions.

The statistical model's relatively poor performance across all metrics suggests that simple probability-based approaches are insufficient for capturing the complex patterns inherent in maritime trade routes. While the LGBM model shows a marked improvement over the statistical model, it still falls short of the neural network's performance.

The neural network's superior performance can be attributed to its ability to learn complex, non-linear relationships from the data, as well as its effective use of embeddings to handle high-cardinality categorical variables. This allows the model to capture intricate patterns in vessel behavior, cargo types, and port relationships that may be overlooked by simpler models.

It is worth noting that while the neural network demonstrates clear advantages, there is still room for improvement, particularly in Top 1 Accuracy.

Evaluation of Neural Network Performance

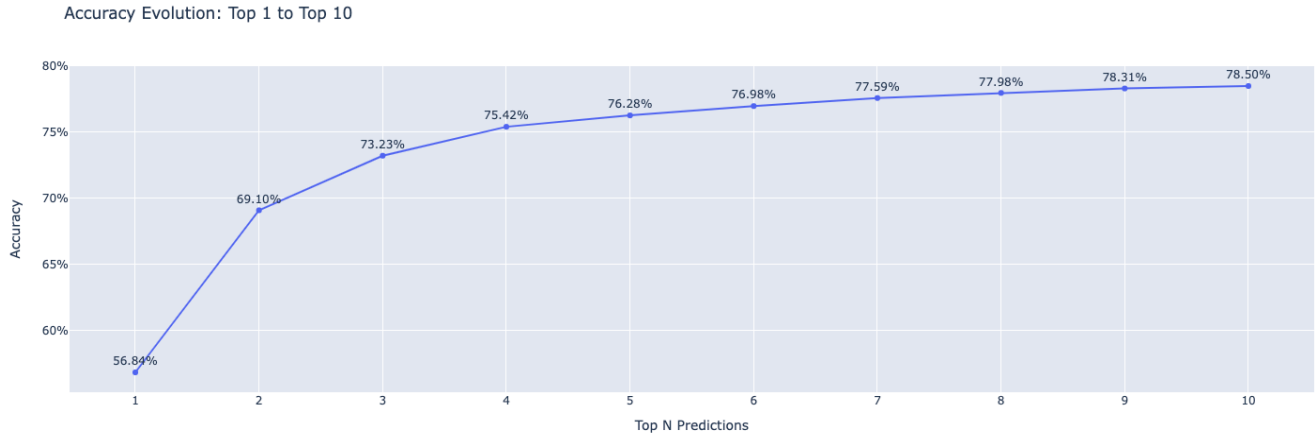
The neural network model's performance for vessel destination prediction was evaluated using multiple metrics and analyses. The results demonstrate the model's efficacy in capturing complex maritime trade patterns and its ability to generalize across various vessel types and cargo categories.

Accuracy Evolution Analysis

The figure below illustrates the model's accuracy as a function of the number of top predictions considered. The model achieves a top-1 accuracy of 56.84%, indicating that it correctly predicts the exact next destination in approximately 57% of cases. A substantial improvement is observed when considering the top-3 predictions, with accuracy increasing to 73.23%. This significant enhancement suggests that even when

the model's primary prediction is incorrect, the actual destination is frequently among its next two highest-probability predictions.

The accuracy continues to improve as more top predictions are considered, albeit at a diminishing rate, reaching 78.50% for the top-10 predictions. This asymptotic behavior indicates that the model effectively narrows down the potential destinations to a small subset in the majority of cases, which is particularly noteworthy given the complexity of maritime route prediction and the extensive number of possible destinations.

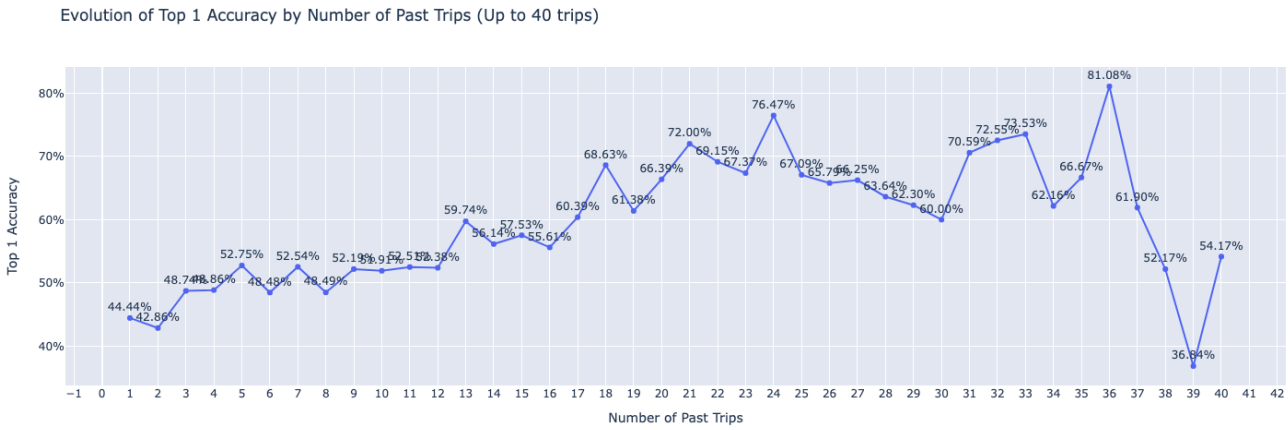


Historical Data Impact

The figure below depicts the relationship between the model's top-1 accuracy and the number of historical trips for each vessel. The graph reveals a general upward trend, with accuracy improving as the number of past trips increases. For vessels with 0-5 previous trips, the model achieves an accuracy of approximately 45-50%, demonstrating its capability to make reasonable predictions even with limited historical data.

The accuracy shows notable improvements for vessels with 20-30 past trips, often exceeding 70%. However, significant fluctuations are observed, particularly for higher numbers of past trips. These variations may be attributed to smaller sample sizes for vessels with extensive trip histories, leading to increased volatility in accuracy

measurements. The peak accuracy of 81.08% is achieved for vessels with 36 past trips, though this may be an outlier due to a potentially small sample size.



Vessel Type and Product Family Analysis

The two figures below present a breakdown of the model's performance across different vessel types and product families. The analysis reveals varying levels of predictive accuracy:

Vessel Type Performance:

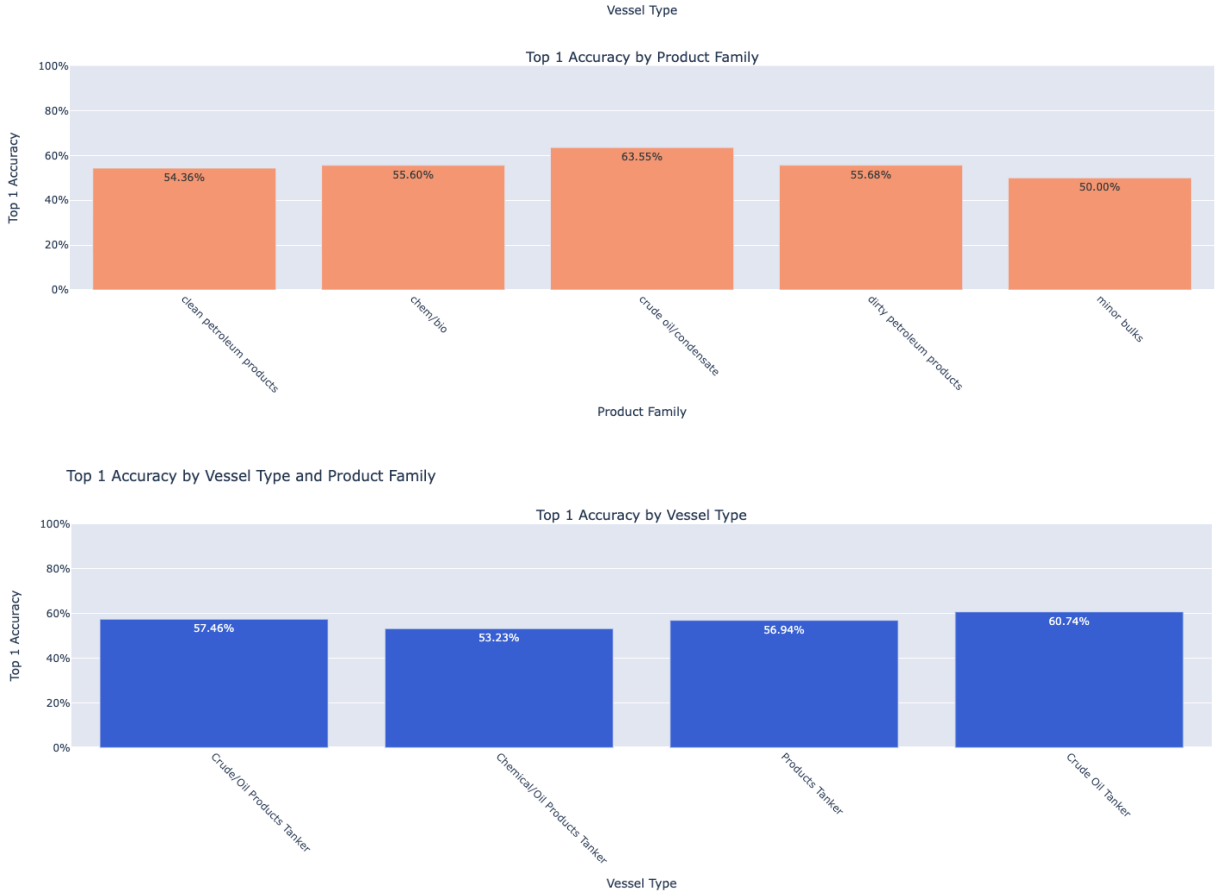
1. Crude Oil Tanker: 60.74%
2. Crude/Oil Products Tanker: 57.46%
3. Products Tanker: 56.94%
4. Chemical/Oil Products Tanker: 53.23%

Product Family Performance:

1. Crude oil/condensate: 63.55%
2. Chem/bio: 55.60%
3. Dirty petroleum products: 55.68%
4. Clean petroleum products: 54.36%
5. Minor bulks: 50.00%

The model exhibits superior performance for Crude Oil Tankers and crude oil/condensate products, potentially due to more predictable routes associated with these vessel and cargo types. Conversely, Chemical/Oil Products Tankers and minor

bulks present lower accuracy rates, possibly reflecting more complex and varied shipping patterns in these categories.



Conclusion

The neural network model demonstrates robust performance in predicting vessel destinations, with particular efficacy in crude oil-related shipping. The model's ability to improve predictions with increased historical data highlights its capacity to learn from past voyages effectively. The significant accuracy gain when considering top-3 predictions (from 56.84% to 73.23%) underscores the model's value in narrowing down potential destinations, even when the top prediction is not correct.

The varying performance across vessel types and product families indicates that the model has successfully captured some of the inherent complexities and specializations within maritime trade patterns. However, the lower accuracy rates for certain categories,

such as Chemical/Oil Products Tankers and minor bulks, suggest potential areas for model refinement.

In summary, the neural network approach shows promising results in the complex domain of maritime destination prediction. Its ability to provide accurate predictions within its top few choices, even with limited historical data, makes it a valuable tool for maritime logistics and planning. Future work may focus on enhancing the model's performance for currently underperforming categories and investigating additional features or architectural modifications to further improve overall accuracy.

Conclusion and Future Directions

The comparative analysis of our proposed neural network model against baseline approaches demonstrates its superior performance in predicting vessel destinations. However, the maritime domain presents unique challenges due to the sequential nature of vessel movements and the high cardinality of categorical variables such as ports and vessel identifiers. To address these challenges and potentially enhance predictive accuracy, we propose exploring the application of Recurrent Neural Networks (RNNs) in future research.

RNNs, particularly in their many-to-one configuration, offer several advantages that align well with the vessel destination prediction task:

1. Sequence Modeling: RNNs are inherently designed to process sequential data, making them well-suited to capture the temporal dependencies in a vessel's journey history. By maintaining an internal state that updates with each time step, RNNs can effectively model the influence of past port visits on future destinations.

2. Variable-Length Input: Unlike our current model, which relies on a fixed number of historical trips, RNNs can handle input sequences of varying lengths. This flexibility allows for more efficient use of available historical data, potentially improving predictions for vessels with both short and long operational histories.

3. Long-Term Dependencies: Advanced RNN architectures such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) are capable of capturing long-term dependencies in the data. This property could be particularly beneficial in identifying cyclical patterns or long-term trends in vessel movements that may span multiple voyages.

4. Implicit Feature Engineering: RNNs can learn to extract relevant features from the sequence of past ports and other time-dependent variables, potentially reducing the need for explicit feature engineering of historical data.

5. Handling High Cardinality: By processing port visits sequentially, RNNs may be able to learn more nuanced representations of ports and their relationships, potentially mitigating some of the challenges associated with the high cardinality of these categorical variables.

The implementation of an RNN-based model for vessel destination prediction would involve restructuring our data to represent each vessel's history as a sequence of port visits and associated features. The model would then be trained to predict the next destination based on this sequence.

While the potential benefits of RNNs are promising, it is important to note that they also present challenges, such as increased computational complexity and the potential for vanishing or exploding gradients during training. These challenges would need to be carefully addressed in the implementation and training process.

In conclusion, while our current neural network model demonstrates strong performance in vessel destination prediction, the exploration of RNN-based approaches represents a promising avenue for future research. By leveraging the sequential nature of maritime data, we may be able to capture more complex patterns and dependencies, potentially leading to further improvements in predictive accuracy and model generalization.

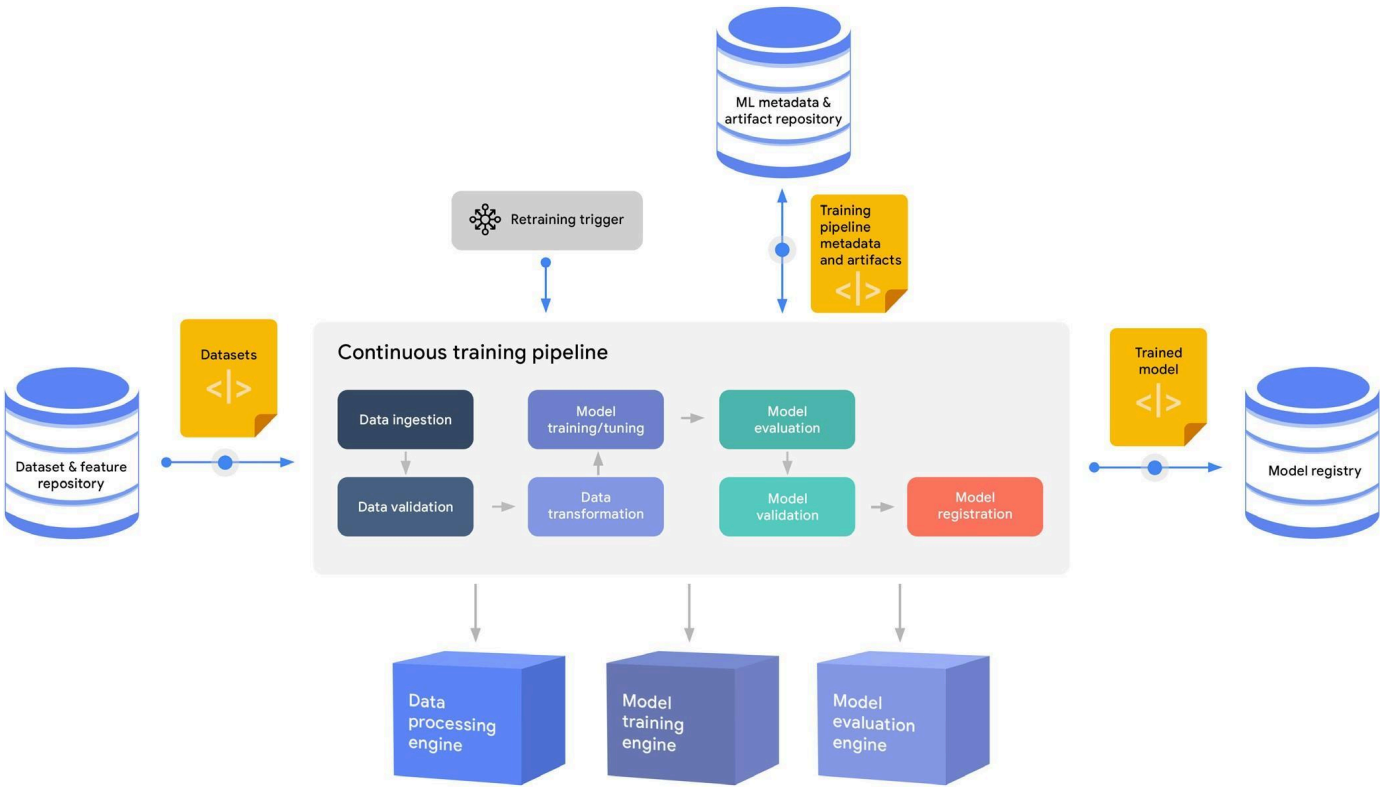
Section on Model Deployment.

NB: (To be discussed during interview - plots are taken from Google)

(Re)Training Pipeline

A comprehensive ML training pipeline should incorporate these key elements:

1. Data Collection: Extract relevant training information from databases and feature stores using specific criteria, such as the most recent update timestamp.
2. Data Verification: Examine the gathered training data to ensure it's not biased or corrupted, maintaining data quality for model training.
3. Data Preprocessing: Divide the dataset into training, validation, and testing subsets. Apply necessary transformations and engineer features to meet model requirements.
4. Model Development and Optimization: Train the ML model and fine-tune its parameters using the prepared data subsets to achieve optimal performance.
5. Model Assessment: Evaluate the model's performance using the test dataset, applying various metrics across different data segments.
6. Model Quality Check: Verify that the model's evaluation results meet predetermined performance standards.
7. Model Storage: Save the validated model in a centralized repository along with its associated metadata.



The ongoing training process is initiated by a retraining trigger. Upon activation, the pipeline fetches fresh data from relevant sources, executes the ML workflow steps, and submits the newly trained model to the storage system. Each run's information and outputs are recorded in a metadata and artifact repository.

While an automated training pipeline mirrors the typical data science workflow, it differs in key aspects:

- Data and model validation play crucial roles as gatekeepers in the automated process.
- Automated runs occur without direct supervision, so data processing and training procedures may become suboptimal or invalid as training data evolves over time.

The data validation step can identify anomalies such as new or missing features, changes in feature domains, or significant shifts in feature distributions. It does this by comparing new data against expected schemas and baseline statistics.

Model validation detects issues like stagnation or decline in new model performance. This step can involve complex validation logic, including multiple evaluation metrics, input sensitivity analysis, calibration checks, and fairness indicators.

A critical aspect of continuous training is comprehensive tracking. Pipeline runs must record metadata and artifacts to enable debugging, reproducibility, and lineage analysis. This tracking allows users to:

- Access training hyperparameters
- Review all pipeline evaluations
- Retrieve processed data snapshots (when feasible)
- Access data summaries including statistics, schemas, and feature distributions

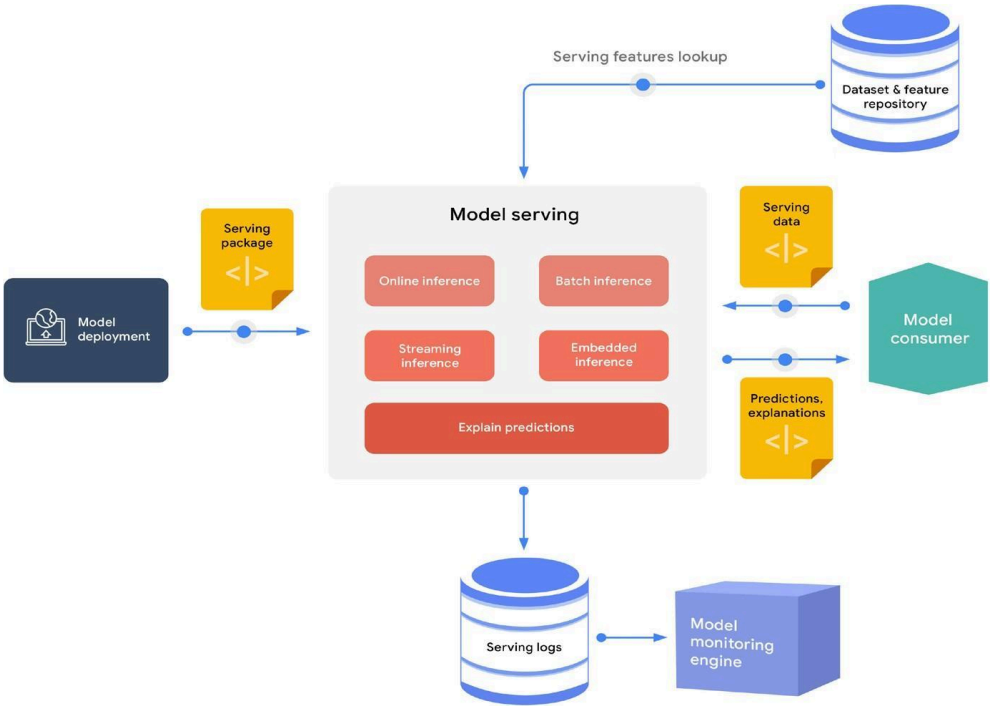
Model Deployment

After successful validation, the pipeline can register the new model candidate in the central repository.

Once a model is trained, validated, and registered, it's ready for deployment. This process involves packaging, testing, and deploying the model to its target environment, often including multiple testing stages and environments.

For users of simplified development platforms, the deployment process is often streamlined. Typically, users specify the model's location in the registry, and the system handles deployment automatically using stored metadata and artifacts.

When you use no-code or low-code solutions, the model deployment process is streamlined and abstracted from the perspective of the data scientists and ML engineers. Typically, you point to the entry for the model in the model registry, and the model is deployed automatically using the metadata and the artifacts that are stored for that model.



Prediction Serving

After deployment, the model begins accepting prediction requests and returning responses. The serving system can provide predictions in several ways:

1. Real-time inference for high-frequency individual or small batch requests, often via REST APIs.
2. Near real-time streaming inference through event processing systems.
3. Batch inference for large-scale data scoring, typically integrated with data pipeline processes.
4. On-device inference for embedded systems or edge devices.

In some cases, the serving system may need to retrieve additional feature values related to a request. For example, a product recommendation model might receive only customer and product IDs, requiring the system to fetch relevant features from a repository before making a prediction. An important aspect of ML system reliability is the ability to interpret models and explain their predictions. These explanations should provide insight into the reasoning behind predictions, such as

by identifying which features contributed most significantly to a particular outcome.