Global Oil Export Destination Prediction: A Machine Learning Approach

Haiying Jia,[a] Roar Adland,[b] and Yuchen Wang[c]

ABSTRACT

We use classification methods from machine learning to predict the destination of global crude oil exports by utilising micro-level crude oil shipment data that incorporates attributes related to the contract, cargo specifications, vessel specifications and macroeconomic conditions. The results show that micro-level information about the oil shipment such as quality and cargo size dominates in the destination prediction. We contribute to the academic literature by providing the first machine learning application to oil shipment data, and by providing new knowledge on the determinants of global crude oil flows. The machine-learning models used to predict the importing country can reach an accuracy of above 71% for the major oil exporting countries based on out-of-sample tests and outperform both naïve models and discrete regression models.

Keywords: Random forests, Gradient boosted trees, Machine learning, Crude oil, Choice models

https://doi.org/10.5547/01956574.42.4.hjia

1. INTRODUCTION

Oil is one of the most important raw materials and is the lifeblood of the global economy. Indeed, oil and gas provide over 50% of primary energy supplies to the world (IEA, 2019) and is the primary power source for transportation. The fourteen member states of the Organization of the Petroleum Exporting Countries (OPEC) control over 80% of world crude oil reserves (OPEC, 2016), while consumption is mainly driven by the OECD (Organisation for Economic Co-operation and Development) countries, China and India. The geographical separation of oil consuming and producing nations means that oil needs to be transported at great distances, with forty percent of the annual global oil production transported via the oceans in specialised oil tankers (Clarksons, 2016; Adland et al., 2017). The global oil trade is greater than any other commodity in terms of value and it was the world's first trillion-dollar industry in terms of annual sales (Doyle, 1994). Major oil producing countries such as Saudi Arabia, Norway, Nigeria and Venezuela derive much of their national income from the production of oil. For many other countries, such as the US and China, the cost of importing oil is a major component of their foreign exchange balance. Thus, oil trade and the price of oil are crucial factors both for national and foreign policy.

a Department of Business and Management Science, Norwegian School of Economics; Center for Applied Research SNF at NHH, 5045 Bergen, Norway.

b Corresponding author. Department of Business and Management Science, Norwegian School of Economics, 5045 Bergen, Norway. E-mail: Roar.Adland@nhh.no.

c Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02142, USA

At the macro level, oil trade flow (i.e. the spatial supply and demand balance) is driven by factors such as population growth, per capita energy usage, and structural changes (e.g. innovation in energy efficiency and the emergence of alternative sources of energy). At the micro level, the ultimate destination of oil exports is the result of a complex and dynamic system including, for instance, contractual agreements (long-term bilateral agreements and short-term commercial contracts), political factors (sanctions or restrictions), new pipelines and refineries, the use of storage, and regional price fluctuations.

The objective of this paper is to predict the destination of oil exports at the micro level in a data-driven framework by utilizing actual oil shipment information and training machine learning algorithms based on supervised classification techniques. Based on crude oil shipment data for the period January 2013 through mid-March 2016, we investigate how destinations are determined based on four attribute clusters: cargo information (such as sellers' identity, cargo grade and cargo size), vessel information (such as vessel identity and its technical specifications), geographic information (load terminals and ports), and macroeconomic data (e.g. regional oil prices and crack spreads). We train the machine learning algorithm based on historical data and demonstrate the models' out-of-sample accuracy.

To our knowledge there is no comparable academic research in the oil trade domain. We contribute to the literature in at least three ways. Firstly, we contribute to the choice model methodology literature by applying cutting edge machine learning techniques in the prediction. Compared to traditional discrete choice models, our approach lessens the dependence on often unrealistic statistical assumptions (such as factor independence) and remain completely data-driven thanks to the increasing availability of maritime big data. Secondly, the unique dataset of micro-level oil shipment information, which is primarily derived from the Automated Identification System (AIS) for satellite tracking of vessels, provides a new and rich information of global oil trades. The high dimensionality in the attributes is key in training machine learning algorithms to predict trade patterns. Thirdly, the variety of machine learning models that are employed in this research provides a good combination of interpretability and accuracy.

This last contribution is key in real life applications and, thus, our methodology is potentially important as a building block in commercial applications that deal with oil and freight market analysis. For instance, the public destination information in ship tracking data is known to be of low quality and can be easily manipulated. Accordingly, analysts that want to track cargoes as a proxy for economic activity or to estimate short-term regional supply of crude oil need a tool to benchmark such information against the likely outcome predicted from past trading patterns and micro data. Importantly, our work suggests that micro data is substantially more valuable for predictive oil trade models than observable macroeconomic data such as crack spreads and oil prices.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature, which is followed by methodology in Section 3 and data description in Section 4. Section 5 describes the feature engineering process, Section 6 presents the results, and Section 7 concludes the paper.

2. LITERATURE REVIEW

The use of discrete choice models, both binary and multinomial, has been the dominating method in modelling destination choice in transportation, see for instance, Malchow and Kanafani (2004), Rich et al. (2009), Steven and Corsi (2012), Piendl et al. (2017), and Alizadeh et al. (2016). Discrete choice models in this context have dealt with destination choices for shopping trips (Tim-

mermans, 1996; Wang and Lo, 2007), car purchases (Train, 1986; Train and Winston, 2007) or the demolishing of ships (Alizadeh et al. 2016). Professor Daniel McFadden won the 2000 Nobel prize for his development of theory and methods for analyzing discrete choices (Manski and McFadden, 1981; McFadden, 1974, 1989; McFadden and Train, 2000). In these models, the choice made by entities (a person, firm or industry) is statistically related to the attributes of the choices. For example, the choice of which port a liner shipping company uses is statistically related to the port service level, vessel sizes, cargo information, and other attributes of each available alternative. The models estimate the probability that a particular alternative is chosen using econometric methods such as parametric models (see, for example, Allenby and Rossi, 1998; Andrews et al. 2002; Hensher and Greene, 2003) or nonparametric models (see, for example, McLachlan and Peel, 2005; Train, 2008). As an extension, the model is naturally used to predict how choices will change when the attributes of the alternatives change. However, the imposed statistical model cannot possibly include all the factors or information that lead to decisions as their determinants are only partially observed or imperfectly measured. Therefore, discrete choice models rely on statistical assumptions and specifications to account for, for example, individual taste differences (Vij and Krueger, 2017). Traditional statistical techniques were designed for relatively small datasets with standardized structures, i.e. similar type of variables. The underlying assumption is that the relationship is homogeneous, that is, the same relationship between variables hold across the entire measurement space. This leads to models where only a few parameters are necessary to trace the effects of the various factors involved (Breiman et al.1998).

As the result of increasing availability of information and the exponential growth in data in recent years, machine learning methods have been gaining popularity in various areas due to their ability to model large amounts of data without explicitly imposing a statistical model form. The term "machine learning" was coined by Samuel (1959), in which he suggests that computers can be programmed to "behave in a way which, if done by human beings or animals, would be described as involving the process of learning". Machine learning typically refers to the scientific study of algorithms that computer systems use to progressively improve their performance on a specific task (Bishop, 2006). Machine learning is today used in various research areas such as, for instance, image recognition for oil spills (Kubat et al. 1998), cancer prediction (Cruz and Wishart, 2006), information extraction (Freitag, 2000), and biology (Kampichler et al. 2010).

A large dataset not only involves a large number of observations for many variables, but also has high complexity in the data structure. This may include high dimensionality, a mixture of data types and nonstandard data structure (Breiman et al. 1998). High dimensionality in machine learning means there is a large number of attributes, which can be features required to represent data, or independent parameters. In this case, the number of observations may be less, but rich information for each observation leads to high dimensionality which demands better handling of the data. Mathematically, in a dataset with M dimensions, the number of parameters needed to specify distributions in M dimensions increases by the factor of M^2 for a normal distribution, unless one makes the very strong assumption that the variables are independent (the typical i.i.d. assumption in traditional statistical models). Indeed, thanks to the complex impact of high dimensionality on statistics, mathematicians have termed it "the curse of dimensionality" (Bellman, 1961). With accelerating computer capability, the analysis of complex high dimensional databases with mixed data types is increasingly feasible without imposing a model structure *a priori*. Micro-level oil shipment data represents exactly such a dataset, which motivates our choice of the machine learning methodology.

3. METHODOLOGY

3.1 Multinomial Logit Model

Discrete choice models that are based on utility maximization theory have gained popularity in transportation research, where the family of these models is typically used to predict individual choices in transport mode and routes. There is an extensive literature on the development of discrete choice models (see Cirillo and Xu, 2011, for a review). For our purpose, one of the most widely used models—the multinomial Logit model (MNL) - is a useful benchmark for the classification performance of machine learning techniques. MNL assumes that the probability of choosing one of the J alternatives y_j from the choice set Y (i.e. the multinomial output variable Y) is a function of a group attributes (i.e. input variable X) (Manski and McFadden, 1981; McFadden, 1987):

$$P\left(y_j \mid Y\right) = \frac{\exp(\beta \cdot X_i)}{\sum_{j=1}^{J} \exp(\beta \cdot X_j)} \tag{1}$$

where, y_j is a choice from a finite set of alternatives Y; X_i is a vector or attributes of alternative i, and β is a vector of parameters. McFadden (1987) suggested the size of alternatives Y should be limited to 100 categorical outputs. In practice, MNL has mainly been used to generalize small-scale problems with a modest number of alternatives and attributes.

3.2 Supervised Classification

The objective in our destination choice problem is to find a mapping function from the attributes (X), which are factors influencing seaborne oil trade, to the output variable—the export destinations (Y). Not only is the number of attributes large, which makes MNL model estimation difficult; most importantly, it is not plausible to impose any form of mathematical functions on this mapping. Therefore, we here introduce supervised machine learning techniques (as opposed to unsupervised, i.e. no observed output variable Y). The goal of supervised learning is to find algorithms (functions) that, given a sample of data, best approximates the relationship between X and Y.

There are several supervised machine learning techniques, such as Artificial Neural networks (ANN) (refer to Zhang, 2000, for an overview of ANNs), k-Nearest Neighbour (kNN), Bayesian network (BN), Support Vector Machines (SVMs), and Decision trees. ANN is inspired by the complexity in biological neuron networks and has a growing number of applications. However, our research question is not suitable for ANN due to the lack of "depth" in the data and the lack of interpretability of the resulting model. The SVM approach is a supervised machine learning technique (Kotsiantis, 2007; Vapnik, 1995) and the main idea is to separate the data by a hyperplane, creating the largest possible distance between the hyperplane and the instances on either side of it. However, as pointed out by Kotsiantis et al. (2006), SVMs may not be applicable in many real-world problems, as many cases involve non-separable data for which no hyperplane exists. Similarly, kNN has limitations in handling imbalanced data (an issue in our trade data set) and missing values. For the purpose of this research, we therefore focus on the following supervised classification techniques: Bayesian Network, Decision Trees, Random Forest and Gradient Boosted Trees.

3.2.1 Naïve Bayes Classification (NB)

NBs are very simple Bayesian networks with one strong assumption, which is the independence among the attributes (X). In other words, in order to calculate the probability of each event, it is assumed that the probabilities of each event are conditionally independent given the target value.

Thus, the probability of choosing output y_i based on the whole attribute set X, which is given as $P(y_i|X)$, is the unconditional probability of y_i occurring in each sub event. We then compare the probability of choosing y_i with the probability of choosing y_j:

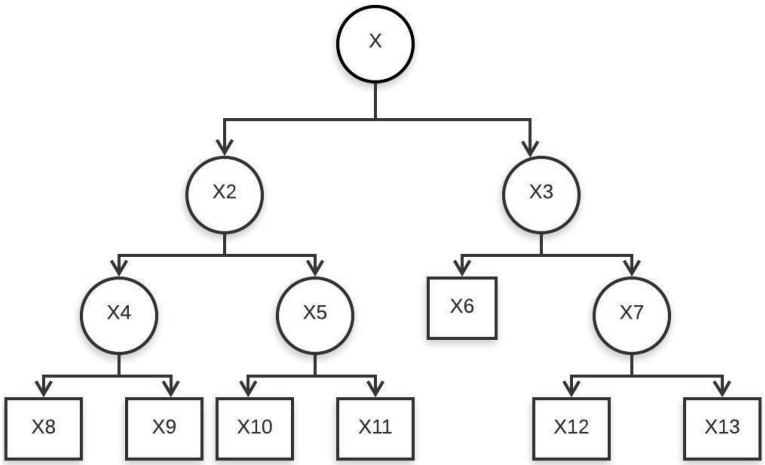
$$\frac{P(y_i \mid X)}{P(y_j \mid X)} = \frac{P(y_i)P(X \mid i)}{P(y_j)P(X \mid j)} = \frac{P(y_i)\Pi P(X_r \mid i)}{P(y_j)\Pi P(X_r \mid j)} \tag{2}$$

If $P(y_i \mid X) > P(y_j \mid X)$, the classification label value would be more likely to be y_i. NBs are fast and easy to implement but the strong assumption in the network is the biggest weakness as most real-life cases would have inter-dependent predictors or attributes X.

3.2.2 Decision Trees

Decision trees utilize a "divide-and-conquer" approach to the problem of supervised learning (Witten and Frank, 2005). Figure 1 illustrates a simplified binary decision tree, which is usually drawn upside down with its root at the top. In each node, a split into two descendant subsets is made based on features and set of conditions. The crucial point is how to determine the splits, the terminal nodes, and their assignments, in other words finding good splits and knowing when to stop. The fundamental idea is to select each split of a subset so that the data in each of the descendant subsets are "purer" (more homogeneous) than the data in the parent subset.

Figure 1: Decision Tree Diagram



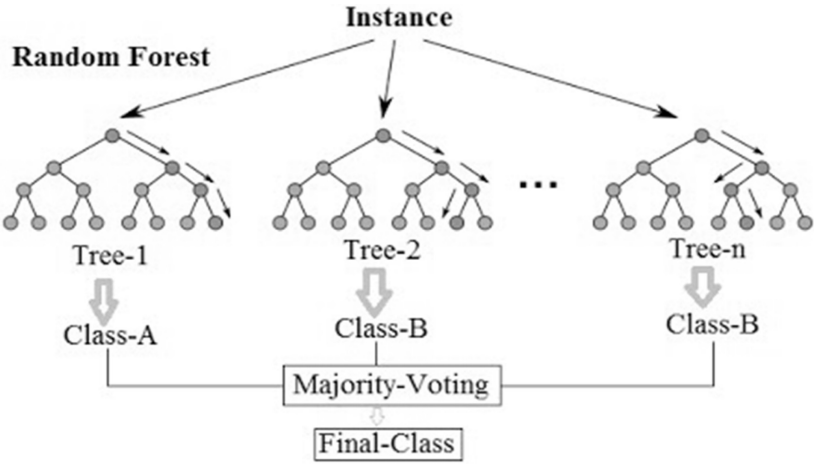
There are many algorithms to build decision trees for classification (see Wu et al. 2008 for a review). The Classification and Regression Trees (CART) of Breiman (1984) are among the most widely used and popular algorithms. CART employs a measure of node impurity, a Gini index, based on the distribution of the observed y values in the node, and splits a node by exhaustively searching over all x and s for the split $\{x \in s\}$ that minimizes the total impurity of its two child nodes. If all datapoints at one node belong to the same class then this node is considered "pure". Therefore, by minimizing the Gini index, the decision tree finds the features that separate the data best. The process is applied recursively on the data in each child node. The splitting stops if the relative decrease in impurity is below a pre-specified threshold. The disadvantages that have been identified by applications in

the literature include lower accuracy and instability comparing to other more complicated machine learning methods like Random Forest and Boosted Trees.

3.2.3 Random Forest

The Random Forest (Breiman, 2001) model, as implied by the name, builds multiple decision trees to form a forest by randomly selecting observations and features. It classifies a new object from an input vector by submitting the input vector to each of the trees in the forest with each tree producing a prediction. In the end, it merges the results from each tree to get a more accurate prediction by choosing the category which has the most votes over all the trees in the forest. The Random Forest algorithm in Figure 2 illustrates a simplified Random Forest structure with *n* decision trees.

Figure 2: Random Forest Diagram



Random Forest has certain advantages as it: i) handles categorical predictors naturally, ii) is quick to fit even for large problems, iii) makes no formal distributional assumptions and iv) automatically fits highly non-linear interactions. The main limitation of the Random Forest model is that it is easy to overfit and has a relative longer run-time as more accurate predictions demand more trees. Nevertheless, in most real-world applications, the Random Forest algorithm is fast enough.

3.2.4 Gradient Boosted Trees

Similar to Random Forest, Gradient Boosted trees (GBT), see Friedman (1999a, b, 2001, 2002), also build multiple trees and combine the outputs from individual trees to improve predictive accuracy. GBT differs in the way that trees are built one at a time, while each new tree helps to correct errors made by a previously trained tree. Boosting is an ensemble technique in which the predictors are not made independently, but sequentially (Grover, 2017). The boosting algorithms combine weak learners to form a strong rule for classification, essentially the algorithm converts relatively poor hypotheses (weak learners) into very good hypotheses (strong learners) (Kearns, 1988).

In GBT, the algorithm trains many models sequentially. Each new model gradually minimizes the loss function of the whole system using the Gradient Descent method, i.e. to find local minimum of the loss function by taking steps proportional to the negative of the gradient of the function at the current point. The learning procedure consecutively fit new models to provide a more accurate estimate of response variable (Pedregosa et al. 2011, Chen and Guestrin, 2016).

Note that GBT always uses regression trees even for classification problems. The learning rate can be controlled to reduce the risk of overfitting. The GBT algorithm performs well in applications with unbalanced data, which in this context means an unequal number of instances for different classes. Though unbalanced data is common, most machine learning classification algorithms are sensitive to imbalance in the predictor classes. For example, in a hypothetical extreme case, if a dataset consists of 10% class A and 90% class B, a machine learning model that has been trained and tested on such a dataset would predict class B for all samples and still gain a very high accuracy. Effectively, an unbalanced dataset will bias the prediction model towards the more common class.

4. DATA

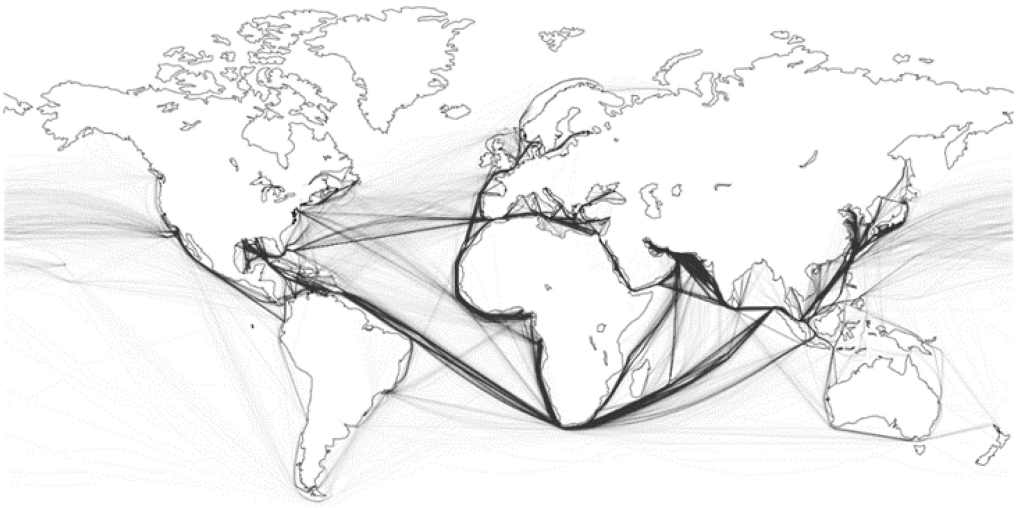
4.1 Seaborne Oil Trade Shipment Data

We utilize a unique dataset comprised of 73,312 oil shipments loading from 212 ports in 76 countries and exporting to 95 countries between 1 January 2013 and 15 March 2016. The dataset is a multi-dimensional database of oil shipment information including:

1. Cargo information: oil grade (e.g. ARAB Crude, Basrah light etc.); producer country/region; API gravity (e.g. light, medium or heavy), sulfur content (e.g. sweet or sour), cargo size (bbls) and seller identity.
2. Geographic trade information: load port/country/region, load date, offtake port/country/region, offtake date.
3. Vessel information: Details about the vessel undertaking each shipment (e.g. vessel name, IMO, flag, class, year-of-build, deadweight)

The cargo data is provided by ClipperData™ and enriched with technical vessel data from the World Fleet Register of Clarkson Research (2016). For the illustration of trading patterns in Figure 3, the cargo data has also been merged with ship positioning data derived from the Automatic Identification System (AIS).

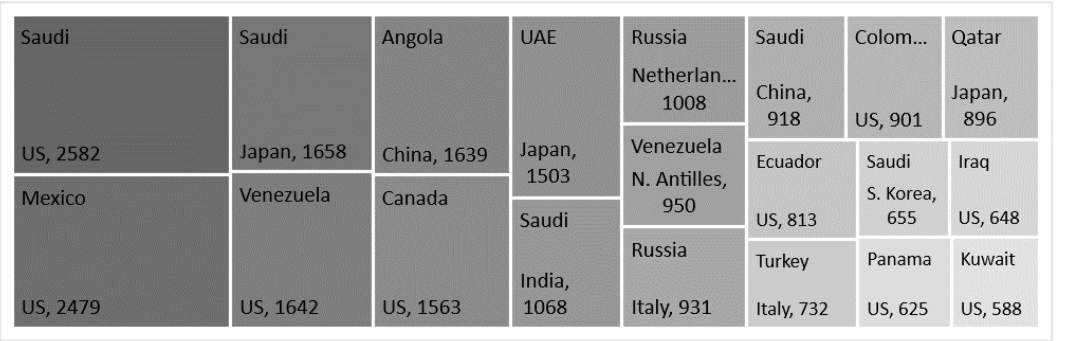
Figure 3: Global Seaborne Oil Trade Flows



Source: Sample data in combination with AIS paths

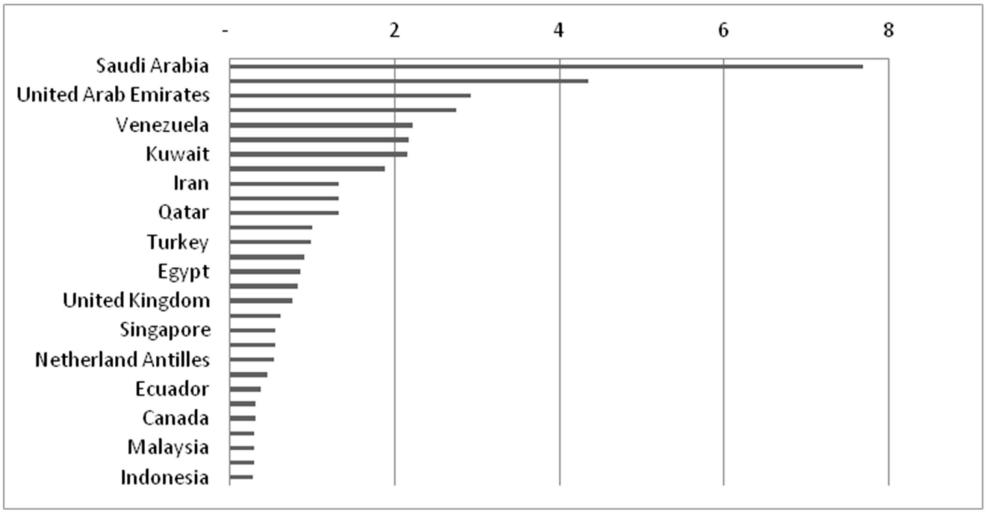
As an example, Figure 4 illustrates the major global seaborne oil flows during the sample period for the top 20 country pairs measured by the total number of shipments. The offtake/load (import/export) country matrix illustrates the frequency of crude oil trade between two countries. For instance, the most common trade by number of shipments is between the United States and Saudi Arabia (2,582 shipments), followed by US-Mexico (2,479 shipments) and Saudi Arabia-Japan (1,658 shipments). For Chinese oil imports, Angola and Saudi Arabia are the top two seaborne sources of oil in terms of the number of shipments.

Figure 4: Top-20 Crude Export (top) - Import (bottom) Country Pairs by No. of Shipments



Source: Derived from Clipper Data

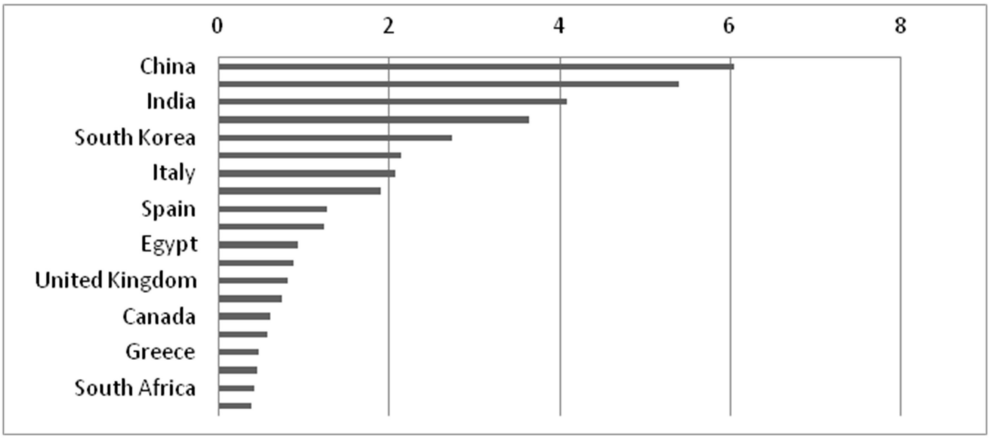
Figure 5: Top-20 Crude Oil Exporting Countries by Volume (billion barrels)



Source: Derived from ClipperData

Figure 5 ranks the exporting countries over the sample period in terms of total seaborne shipment volume (billion barrels). Saudi Arabia tops the list by exporting nearly 8 billion barrels by sea over the time period, followed by UAE, Venezuela, Kuwait and Iran. Similarly, Figure 6 ranks the top-20 oil importing countries in global seaborne oil trade. It shows that countries including China, India, and OECD countries are among the top destinations.

Figure 6: Top-20 Crude Oil Importing Countries by Volume (billion barrels)



Source: Derived from ClipperData

4.2 Oil Market Data and Other Economic Data

Most crude oil is sold on Free-on-board (FOB) basis, which means the buyer takes ownership at the point of loading and arranges and pays for seaborne transportation. Even though long-term offtake agreements dominate in the commercial relationships between oil buyers and sellers, ensuring some rigidity in global crude oil trade, the observable trading patterns are a result of complex commercial decisions. Firstly, as integrated oil companies and refining companies (the oil buyers) often have operations across multiple countries, cargoes may be directed to different ports depending on local tank storage levels, local refinery economics (crack spreads, foreign exchange rates) and logistical bottlenecks (e.g. port congestion). Secondly, third-party trading companies such as Glencore, Vitol, Trafigura and Gunvor are large players in the global crude oil trade and will trade crude oil in the spot market to take advantage of perceived arbitrages, such as geographical price spreads or the cost of physical storage vs futures prices. Thirdly, countries such as the United States or China will occasionally make strategic purchases of oil to build their national reserves. Fourthly, oil flows can be disrupted due to trade embargoes and company sanctions, political unrest and other exogenous factors such as natural disasters and acts of terror.

To proxy some of the macro-level attributes that affect the destination of crude oil exports we chose six categories of time series: (1) crude oil prices (incl. spot and futures prices in Europe and the US), (2) natural gas prices (spot and futures at different locations), (3) oil products prices (gasoline, jetfuel/kerosene in different locations, quoted on FOB or Cost-Insurance-Freight CIF basis), (4) crack spreads, (5) inventory levels (LPG), and (6) foreign exchange rates (USD, JPY, CNY, EUR, BRL spot and futures), in total 38 time series. The crack spread represents the differential between the price of the input in the refining process (crude oil) and the value of the output (petroleum products).

5. FEATURE ENGINEERING

Feature engineering refers to the process of data cleaning and deciding what information to include in the machine learning models. In total, we include 31 predictors reflecting the cargo and shipment information, as well as 38 time series predictors reflecting the market conditions. In this section we describe the steps taken to process the data prior to training the machine learning models.

Firstly, as the machine learning techniques applied in this research are not capable of directly handling categorical variables, such as geographical information, they are converted to numerical values. We experimented with different encoding methods, including one-hot encoding and label encoding. The trade-off is to balance model performance and the number of variables and, thus, the complexity of the model. One-hot encoding expands the number of variables drastically by creating a vector to denote the presence or absence of a categorical variable. The dimension of the vectors depends on the number of categories for an attribute. In our research problem, there are 30 categorical variables, and, as an example, one of the categorical variables (load port) alone has 212 categories (unique port names). As one-hot encoding method slows down the learning significantly, we apply label encoding by assigning a value from 1 to N_x, where N_x is the number of categories for attribute x.

Secondly, random missing data values are present for some of the categorical variables with a missing ratio of 1.8%. For all the variables except one (Sulphur), less than 0.5% is missing. The variable Sulphur represents the crude oil cargo's sulphur content as a percentage of volume, ranging from 0.1% to 6.8%, and is missing in 20% of the cases. There are various techniques to treat missing values: such as deleting, random filling, estimating based on another predictive model, finding the k-nearest neighbors, or creating a separate category as unknown. We choose to impute the missing values with the mean for the particular feature. This is a basic imputation method that serves our purpose well, though it has been criticised as reducing variance in the dataset. We also note that the variable API gravity (density of the crude oil cargo) provides supplementary information on cargo quality.

Thirdly, the dataset is also inherently imbalanced, in the sense that certain trading partners dominate in the sample either as exporters or importers. This is sometimes explained by geographical proximity but also long-term commercial offtake agreements for crude oil, bilateral free-trade agreements between countries, sanctions or security concerns etc. Because of this imbalance, we choose to split the dataset and apply the machine learning algorithms at the export country level. This decision is also driven by the empirically poor prediction performance when utilizing the full dataset. In an initial experiment, the prediction accuracy of the Decision Tree algorithm was only 28.8%. To improve performance, the dataset is divided by origin countries into 76 sub-samples: $DF = \{DF_1, \ldots, DF_n, \ldots, DF_N\}, (N = 76)$.

As our main cross-validation approach, each dataset DF_n is then divided into: i) a training set—the sample of data used to fit the models, ii) a validation set—the sample of data used to provide an unbiased evaluation of the fit of the model from the training dataset while tuning model hyper-parameters, and iii) a testing set—the sample of data used to provide an unbiased evaluation of a final model fit (out-of-sample performance). The size of the three periods are set simply by randomly allocating 60% of observations to the training set, 20% for validation and 20% for testing. In each case we train the three models, namely Decision tree, Random Forests and Boosted tree, respectively, on each subset.

6. RESULTS

6.1 Performance Measures

Once the best fitting model has been identified through training and validation, the model is then used to predict the oil export destination countries for each exporting country. By comparing the predicted values to the actual value in the test sample, the accuracy of the classification output

can be classified according to Table 1. For instance, True Positive (TP) means to correctly predict y_i when the actual value is in fact y_i.

Table 1: Parameters measuring prediction performance

		Predicted Value	
		Yes	No
Actual Value	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

Based on these four parameters (TP, TN, FP, FN), four prediction performance measures are calculated: accuracy, precision, recall and F1 score. Accuracy measures the ratio of correctly predicted values to the total number of out-of-sample values, i.e. Accuracy = (TP+TN)/(TP+TN+FP+FN). Precision measures the ratio of correctly predicted positive values to the total predicted positive values, i.e. Precision = TP/(TP+FP). Recall, which is also called sensitivity, measures the ratio of correctly predicted positive values to all the actual positive values, i.e. Recall = TP/(TP+FN). F1 scores is the weighted average of Precision and Recall, by taking into account both false positives and false negatives. When sample data is imbalanced, i.e. uneven class distribution, F1 scores gives a more realistic performance measure.

6.2 Overall Performance

Our aim is to predict the destinations of export shipments in the global seaborne oil trade. Thanks to the richness of the dataset, the geographical levels of prediction can be scaled down from regions to individual countries, ports or even terminals. In total, there are 76 exporting countries and 95 destination countries in the sample.

Table 2 reports the statistics of the average prediction accuracy from the three machine learning models in the training and test sample, measured as the average performance *within* a percentile when exporting countries DF_n are ranked by the total number of shipments. As an example "Mean 95 percentile of obs" refers to the average prediction accuracy for the countries with number of shipment below 3,666, which is the 95-percentile in the sample. Similarly, "Mean 10 percentile of obs" refers to the average prediction accuracy for the countries with number of shipments below 18, which is the lower 10-percentile in the sample. Overall, the test accuracy does not differ materially between the training sample and test sample for the majority of the subsamples. As the samples are of different sizes we have to be careful of drawing strong conclusions here, as the level of homogeneity could be different, but this is nevertheless encouraging. The relatively high accuracy also presumably reflects the fact that the trading patterns for crude oil are somewhat stable.

We note the tendency for the predictive power of the machine learning algorithms to drop when the number of observations in the subsample decreases from thousands to a few dozens. Countries in the lower quartile include, for instance, Poland (12 shipments), Philippines (40), South Africa (21), Lavia (21), Chile (17), Mauritania (13), St. Croix (12), Ireland (6) and Belgium (1). There are several points to note here. Firstly, when there is limited data to train the model, we face the difficulties of over-fitting and noise. Secondly, countries with such low volumes—like the ones listed above—tend not to be oil exporters at all and are present in the dataset only due to rare occurrences of transhipments or re-exports of crude oil cargoes. The random nature of such trades means they are by definition hard to predict using historical data. Thirdly, we note that the number of shipments (or number of destination countries for that matter) need not be a good proxy for the difficulty of the

classification problem. Indeed, a country with few shipments can have only a single trading partner (New Zealand being an example of this) and so has a 100% predictable outcome. Conversely, very large crude oil exporting countries will typically have a large and diverse group of trading partners, with flows that also change over time, making the destination somewhat less predictable. We see this reflected for the largest exporters in Table 3.

Table 2: Prediction accuracy statistics

	Decision Tree	Random Forest	Boosted Tree
Training sample			
Mean all	59.7%	63.5%	61.6%
Median	61.9%	66.7%	67.8%
Mean 95 percentile of obs	63.9%	66.1%	70.7%
Mean 75 percentile of obs	63.6%	67.8%	71.2%
Mean 25 percentile of obs	54.99%	55.65%	45.67%
Mean 10 percentile of obs	60.42%	60.42%	41.67%
Test sample			
Mean all	54.6%	58.1%	58.4%
Median	58.8%	62.4%	62.5%
Mean 95 percentile of obs	66.6%	68.2%	71.4%
Mean 75 percentile of obs	65.2%	69.4%	72.8%
Mean 25 percentile of obs	33.3%	33.4%	29.9%
Mean 10 percentile of obs	**26.0%**	16.7%	22.9%
Mean all	59.7%	63.5%	61.6%
Median	61.9%	66.7%	67.8%

As machine learning is a data-driven technique, results can be highly dependent on the data structure of a subsample and it may therefore be difficult draw strong generalized conclusions. In order to have a better understanding of prediction performance, we therefore focus on the F1 scores by load country subsample in the prediction models, together with the percentage of the most frequent offtake country class. Table 3 reports the accuracy performance, measured by F1 score, in predicting the offtake (importing) countries for the top-15 oil exporting countries, which together account for 76% of total seaborne shipment volume in our database. The table is ranked by the Boosted tree performance results for the test sample in descending order. We also include the subsample size (i.e. number of shipments) and the number of offtake countries.

There are a couple of important takeaways from Table 3. Firstly, without exception, all the machine learning models outperform the naïve benchmark models, i.e. simply guessing that the destination is the biggest importer ("most frequent class"), the classical multinomial Logit regression model, and the Naïve Bayesan prediction. This suggests that the machine learning approach—with its ability to utilize complex relationships between variables—is in fact a valuable addition to predictive modelling of crude oil flows.

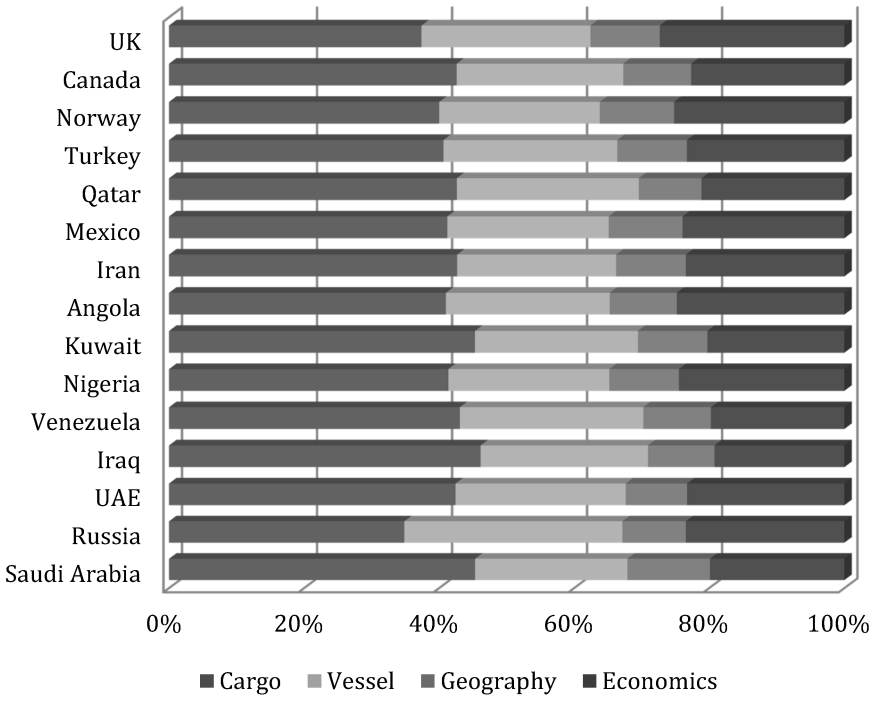
Secondly, we note that the degree of homogeneity in the exporting profile for each load country affects the prediction accuracy. Specifically, countries with highly concentrated seaborne oil exports tend to have higher prediction performance than other countries. For instance, 93% of Canada's oil export shipments are destined for the United States in the sample, resulting in the highest out-of-sample accuracy of 97.36%. Similarly, Mexico exports 68% of its oil shipments to the United States, making it the second most homogeneous country in terms of destination countries, with a correspondingly high 90.8% prediction accuracy. On the other hand, Russia's seaborne oil exporting profile is rather fragmented with only 14% of shipments going to its largest export destination, the Netherlands, with prediction F1-score below 40% for the test sample.

Table 3: Prediction performance of the models by load countries, F1 score for classifications

Load country	Num. offtake countries	Num. shipments	Most frequent class	Multinomial Logit	Naïve Bayes	Decision tree	Random forests	Boosted tree
Canada	18	1,683	92.9%	92.2%	84.5%	94.4%	97.2%	97.4%
Mexico	26	3,630	68.3%	81.8%	62.6%	86.2%	88.8%	90.8%
Venezuela	40	3,719	44.2%	46.0%	20.8%	82.8%	78.6%	84.8%
Kuwait	29	3,000	19.6%	23.0%	8.1%	67.5%	73.1%	80.1%
Iran	17	1,555	30.2%	41.0%	44.9%	73.7%	76.9%	78.0%
Qatar	26	2,242	40.0%	35.5%	35.6%	66.0%	65.0%	74.1%
S. Arabia	41	9,980	25.9%	33.9%	23.9%	74.8%	73.8%	73.8%
UAE	37	4,485	33.5%	29.7%	14.6%	58.6%	64.4%	72.0%
Angola	35	3,408	48.1%	51.1%	14.4%	56.3%	64.8%	69.7%
Iraq	38	3,213	20.2%	36.0%	15.9%	58.8%	63.5%	68.9%
Nigeria	50	3,469	13.1%	20.9%	5.7%	37.8%	43.9%	51.5%
UK	29	1,227	29.8%	28.3%	18.9%	45.9%	48.5%	47.3%
Turkey	35	1,765	41.5%	38.9%	8.6%	42.9%	46.8%	46.5%
Norway	23	1,707	30.2%	16.6%	10.7%	28.4%	32.3%	45.2%
Russia	49	7,046	14.3%	11.3%	4.9%	39.6%	37.9%	38.3%
Average	33	3,475	36.8%	39.1%	24.9%	60.9%	63.7%	67.9%

Figure 7 shows the variation across countries in terms of the importance of attributes in the GBT model. Overall, information about the cargo plays the most important role in all cases. Information on the vessels employed appears to be of greater importance than economic and geography attributes. It appears that the vessel attribute cluster is more important than the macroeconomic variables for oil trades originating in the Persian Gulf (Saudi Arabia, UAE, Iraq, Kuwait, Iran and Qatar), while the two attribute clusters (vessel and economic) are on par in ex-Africa trades (Nigeria and Angola in Figure 7).

Figure 7: Attribute importance for key countries in GBT-Boosted tree



This suggests that the fleet serving the Persian Gulf oil exports is more heterogeneous, such that knowledge of vessel identities provides some additional valuable information on the shipments' ultimate destination. Overall, the dominance of cargo information attributes in the prediction reflects the presence of long-term offtake agreements and technical processing limitations of the refineries at the destination.

Finally, we note that there are other ways to cross-validate the stability of the prediction models than the above holdout 60/20/20 split of the dataset, especially when the sample is small or imbalanced. K-fold is particularly useful in this case by repeating the holdout partition k times, such that each time one of the k subsets is used as the validation set and the other k-1 subsets are put together as a training set. Leave-p-Out leaves p data points out of training data, with Leave-One-Out (LOO) being the extreme case of this approach (p=1). To illustrate the potential differences in reported accuracy depending on the chosen cross-validation method, Table 4 reports the prediction accuracy for a small sample (Brunei), medium-size sample (Indonesia) and large sample (Mexico), respectively. The experiment suggests that LOO leads to a higher reported accuracy for small size samples, such as the case of Brunei with less than 200 data points. However, it is computationally expensive for large sample sizes.

Table 4: Prediction accuracy in Decision Tree based on different cross validation methods

	Number of shipments	Holdout 60/20/20	k-Fold	Leave-One-Out	
Brunei	197	23%	22%	33%	(0.034)
Indonesia	778	60%	59%	58%	(0.018)
Mexico	3630	90%	68%	88%	(0.005)

Note: Numbers in parenthesis is the standard deviation of the estimated accuracy

6.3 The Case of Saudi Arabia

Saudi Arabia, being the largest crude oil exporter in the world, ships its oil to 40 countries around the world, with the US taking over a quarter of all shipments (26.2%) followed by Japan (16.8%), India (10.8%) and China (9.3%). We note here that as the volume of individual shipments differs this does not match perfectly with the share of the overall traded volume.

Figure 8 reports the test set accuracy for the destination country of Saudi Arabian oil exports as a function of tree depth. We note that the decision tree and random forest algorithms converge at a similar rate and to a similar test accuracy of around 75%, albeit with the later performing better at greater tree depths. The Boosted tree algorithm performs the best for this particular subset with 81% accuracy in the test set. The relative performance is confirmed by measures of Accuracy, Precision, Recall and F1 score, which is shown in Table 5.

The relative importance of the attributes in the classification models is visualized for Saudi Arabian seaborne oil exports by the radar chart in Figure 9. The attributes are first clustered into the four categories of cargo, vessel, geography, and economic indicators, following the criteria presented in Section 4. In the case of DF_1 (Saudi Arabia), attributes related to cargo information play the most important role in the prediction of destinations, followed by attributes for vessels and economic indicators. The location of loading terminals and loading ports is the least important attribute in predicting export destinations in the case of Saudi Arabia. We note that there are only four loading ports (Al Ju Aymah, Ras Tanura, Yanbu, and Khafji) while there are 40 destination countries. Moreover, a ship is often loading in more than one port prior to its international journey (i.e. there are more than one shipment on a single ship, creating dependence) such that the informational content in the load port attribute is limited.

Figure 8: Offtake country prediction accuracy by tree depth (load country=Saudi Arabia)

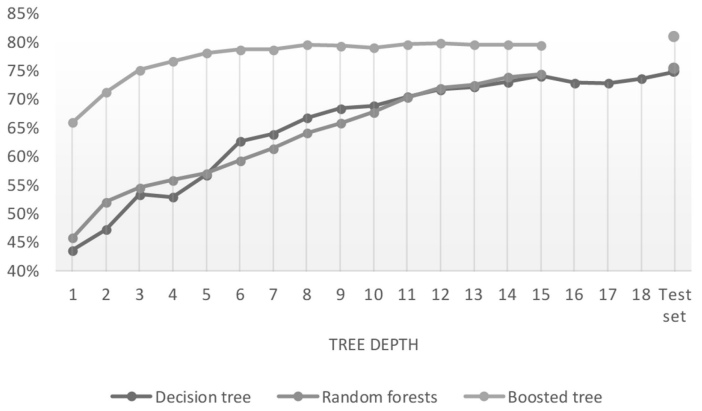
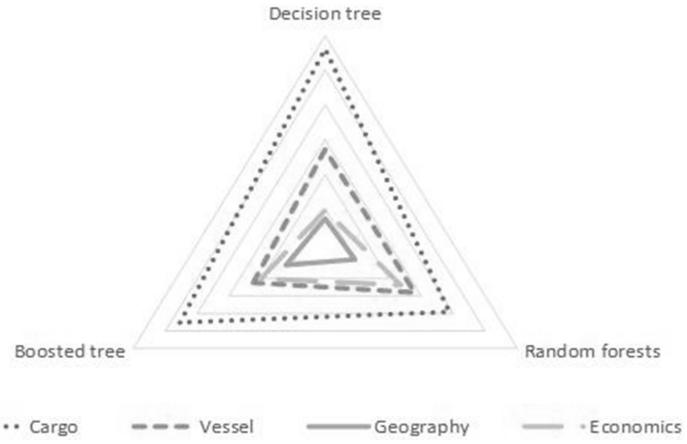


Table 5: Prediction performance measures from the five models (load country = Saudi Arabia)

Performance scores	MNL	NB	DT	RF	GBT
Accuracy	42%	28%	75%	76%	76%
Precision	30%	28%	76%	76%	76%
Recall	42%	28%	75%	76%	76%
F1	34%	24%	75%	74%	74%

Figure 9: Attribute importance for Saudi Arabia subsample



6.4 The Case of Mexico

The destination distribution for Mexico is dominated by two countries—nearly 70% of seaborne oil shipments from Mexico ends up in the US, and another 15% of shipments in Spain. This high degree of concentration results in a high prediction accuracy even at low tree depths, converging to a high 91% average across the three models in the test set, as reported in Figure 10. The overall performance measures are shown in Table 6.

The relative importance of the clusters of attributes also changes for DF_2 (Mexico). The overall importance of the four attribute clusters is presented in Figure 11. Attributes regarding cargo

Table 6: Prediction performance measures from the five models (load country = Mexico)

Performance scores	MNL	NB	DT	RF	GBT
Accuracy	87%	52%	90%	91%	92%
Precision	78%	83%	84%	88%	91%
Recall	87%	52%	90%	91%	92%
F1	82%	63%	86%	89%	91%

Figure 10: Offtake country prediction accuracy (load country=Mexico)

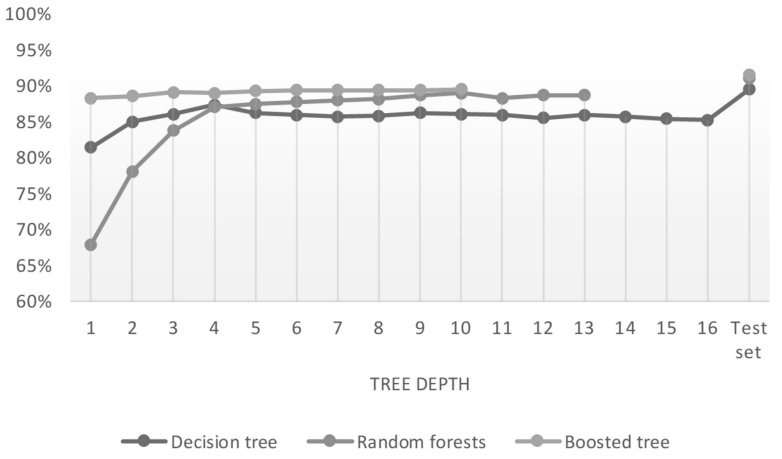
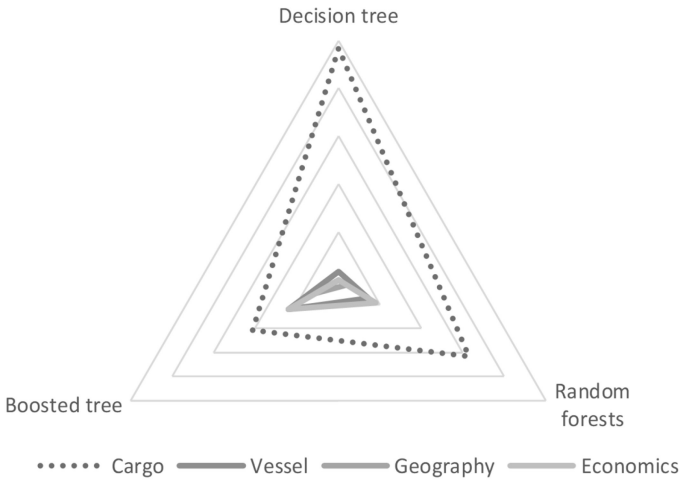


Figure 11: Attribute importance for Mexico subsample



information is still the dominant factor in the three models. For the Boosted tree and Random forests models, vessel attributes as well as economic indicators are more or less equally important. Referring to Section 3.3, GBT models improve on other models particularly for unbalanced dataset by reducing bias and converting weaker learners to more rigorous learners, as is the case here.

Comparing the relative importance of attributes for Saudi Arabia (Figure 9) and Mexico (Figure 11), the prediction of export destinations for Saudi Arabia is based on more dimensions of the data. This naturally reflects the difference in the heterogeneity and complexity of trading patterns

originating from the two countries. Saudi Arabia, being the world's largest oil exporter, has not only more trading partners, but also a more diversified ownership of the oil cargoes and a wider specification of vessels that are employed to carry out the shipment. Conversely, nearly 70% of oil shipments from Mexico are destined for the US, and knowledge of seller information and cargo size/type can yield a highly accurate prediction of destination countries.

7. CONCLUDING REMARKS

We have shown that machine learning can be a powerful and effective tool to predict seaborne oil trade destinations. Gaining visibility of oil trade flows is of great importance and interests to multiple parties, such as oil traders, oil companies or refineries. For many of these players, visibility is currently limited to information gleaned from communication with brokers or in-house data sources. With the increasing availability of data, innovative and advanced data science techniques are increasingly required for information discovery. The prediction of destinations for individual oil cargoes allows better forecasting of regional and local market balances, and also allows for improved prediction of inventory levels and monitoring of the supply chain.

For policy makers, trading patterns affect everything from national energy security and trade balances to the environmental footprint of economic activity. For instance, supranational institutions such as the World Bank may desire a real-time index of global trade in crude oil to monitor changes in economic conditions or the tightness of supply in regional energy markets. Similarly, national governments may want to monitor changes in the sourcing of their energy needs, or the impact of sanctions and trade embargoes. We acknowledge that machine learning models do not adapt quickly to such structural shifts as they learn from historical data, but they nevertheless can be useful to detect departures from what is considered "normal" behaviour according to the recent past.

As destination prediction refers by definition to short-term outcomes, our work is perhaps most applicable to operational decisions. For instance, for maritime port authorities, better predictions of ships' destinations can improve operational planning such as the scheduling of port calls and minimizing port congestion. It is well documented in the literature (see, for instance, Jia, 2018; Jia et al, 2017) that the lack of co-ordination between shipowners, cargo owners and terminals on the availability of cargoes and berths create substantial inefficiencies in the supply chain, with increased waiting time, sailing speeds and emissions as a result.

For shipowners and operators, the destinations of crude oil cargoes directly affect the demand for seaborne transportation. For instance, a short-term increase in the number of predicted long-haul shipments out of a main loading area such as the Persian Gulf will remove ships from the market for longer (reducing future supply), therefore potentially creating a near-term increase in the cost of transportation. The prediction of cargo destinations is therefore an integral part of freight rate forecasting, monitoring of regional the freight market balances and, ultimately, feeds into decisions on the optimal geographical allocation of fleets.

This paper is the first academic research to apply machine learning models in predicting the destinations of seaborne oil trade. We base the training of the models on a rich micro-level dataset of shipments with detailed information on crude quality, oil buyer and seller identity, cargo size and other attributes. However, we have also shown that the "model-form free" approach comes at an analytical cost, as it is necessarily harder to generalize the results from machine learning models than those from traditional statistical analysis, where a universal relationship can be drawn based on regression techniques. Machine learning is data driven, so that feature engineering and results are very much case-dependent. Therefore, it requires domain expertise that is applicable to individual cases (dataset).

Overall, our application to global crude oil exports at the country level results in test sample F1-score with a median of 68% - a strong performance considering the number of possible outcomes. It is possible that the results for certain countries or regions could be further improved by the inclusion of different data. However, even in cases where there are relatively few observations, classical discrete choice regression models do not perform better.

We acknowledge that the relative rigidity of global crude oil trade, with predominance of long-term offtake agreements and national oil buyers and refinery operations should increase predictability relative to other applications of choice models in transportation. However, our work still points to an important application of micro-level data and machine learning models to improve oil and tanker freight market analysis.

We also acknowledge that our machine leaning models suffer from the same inability to predict exogeneous shocks as any other empirical models that are, by definition, based on known, past data only. Examples from the oil market would be the introduction or removal of sanctions on exporting countries, the outbreak of wars or terrorist attacks on refineries and oil processing facilities etc. Once such shocks do occur, the model will gradually incorporate their effects in its estimates. It is beyond the scope of the current paper to investigate the effects of such shocks and we leave this for future research. Future research should also consider the inclusion of other types of data that reflect economic uncertainty or political risk, such as market sentiment, leading economic indicators or textual analysis of news and social media.

ACKNOWLEDGMENTS

This research is partially funded by the Norwegian Research Council, under the project "Smart Digital Contracts and Commercial Management" (project number: 280684).

REFERENCES

Adland, R., H. Jia and S.P. Strandenes (2017). "Are AIS-based trade volume estimates reliable? The case of crude oil exports." *Maritime Policy and Management* 44(5): 657–665. https://doi.org/10.1080/03088839.2017.1309470.

Allenby, G.M. and P.E. Rossi (1998). "Marketing models of consumer heterogeneity." *Journal of Econometrics* 89 (1): 57–78. https://doi.org/10.1016/S0304-4076(98)00055-4.

Alizadeh, A.H., S.P. Strandenes, and H. Thanopoulou (2016). "Capacity retirement in the dry bulk market: a vessel based logit model." *Transportation Research Part E* 92: 28–42. https://doi.org/10.1016/j.tre.2016.03.005.

Andrews, R.L., A. Ainslie, and I.S. Currim (2002). "An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity." *Journal of Marketing Research* 39(4): 479–487. https://doi.org/10.1509/jmkr.39.4.479.19124.

Bellman, R.E. (1961). *Adaptive control processes: a guided tour*. Princeton University Press. https://doi.org/10.1515/9781400874668.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer, ISBN 978-0-387-31073-2.

Breiman, L. (1984). *Classification and regression trees*, New York: Routledge.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone (1998). *Classification and regression trees.* Chapman & Hall/CRC.

Breiman, L. (2001). "Random Forests." *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, T., and C. Guestrin. (2016). "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785.

CIA (2018). Central Intelligence Agency, the World Factbook. https://www.cia.gov/library/publications/resources/the-world-factbook/fields/261.html.

Cirillo, C. and R. Xu (2011). "Dynamic discrete choice models for transportation." *Transport Reviews*: 31(4): 473–494. https://doi.org/10.1080/01441647.2010.533393.

Clarkson Research (2016). Shipping Intelligence Network, www.clarksons.net.

Doyle, J., (1994). *Crude Awakening: The Oil Mess in America: Wasting Energy Jobs and the Environment*. Friends of the Earth, Washington, DC.

Cruz, J.A. and D.W. Wishart (2006). "Applications of machine learning in cancer prediction and prognosis." *Cancer Informatics* 2: 59–78. https://doi.org/10.1177/117693510600200030.

Freitag, D. (2000). "Machine Learning for Information Extraction in Informal Domains." *Machine Learning* 39: 169–202. https://doi.org/10.1023/A:1007601113994.

Friedman, J.H. (1999a). Greedy Function Approximation: A Gradient Boosting Machine. IMS 1999 Reitz lecture https://statweb.stanford.edu/~jhf/ftp/trebst.pdf (accessed 1/8/2019).

Friedman, J.H. (1999b). Stochastic Gradient Boosting. https://statweb.stanford.edu/~jhf/ftp/stobst.pdf.

Friedman, J.H. (2001). "Greedy function approximation: a gradient boosting machine." *Annals of statistics* 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J.H. (2002). "Stochastic gradient boosting." *Computational Statistics & Data Analysis* 38(4): 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Greene, W.H. and D.A. Hensher (2003). "A latent class model for discrete choice analysis: Contrasts with mixed logit." *Transportation Research Part B* 37(8): 681–698. https://doi.org/10.1016/S0191-2615(02)00046-2.

Grover, P. (2017). Gradient boosting from scratch. https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.

Hensher, D. and W. Greene (2003). "The mixe3d logit model: the state of practice." *Transportation* 30(2): 133–176. https://doi.org/10.1023/A:1022558715350.

IEA (2019). International Energy Agency, Key world energy statistics. https://www.iea.org/statistics/kwes/.

Jia, H. (2018). "Crude oil trade and green shipping choices." *Transportation Research Part D* 65: 618–634. https://doi.org/10.1016/j.trd.2018.10.003.

Jia, H., R. Adland, V. Prakash, and T. Smith (2017). "Energy efficiency with the application of virtual arrival policy." *Transportation Research Part D* 54: 50–60. https://doi.org/10.1016/j.trd.2017.04.037.

Kampichler, C., R. Wieland, S. Calme, H. Weissenberger and S. Arriaga-Weiss (2010). "Classification in conservation biology: A comparison of five machine-learning methods." *Ecological Informatics* 5(6): 441–450. https://doi.org/10.1016/j.ecoinf.2010.06.003.

Kearns, M. (1988). Thoughts on hypothesis boosting. https://www.cis.upenn.edu/~mkearns/papers/boostnote.pdf.

Kotsiantis, S.B. (2007). *Supervised machine learning: a review of classification techniques*.

Kubat, M., R.C. Holte, and S. Matwin (1998). "Machine learning for the detection of oil spills in satellite radar images." *Machine Learning* 30: 195–215. https://doi.org/10.1023/A:1007452223027.

Manski, C.F. and D.L. McFadden (1981). *A structural analysis of discrete data with econometric applications.* Cambridge: the MIT Press.

Malchow, M.B. and Kanafani, A. (2004). "A disaggregate analysis of port selection." *Transportation Research Part E* 40: 317–337. https://doi.org/10.1016/j.tre.2003.05.001.

McFadden, D. (1974). "Conditional logit analysis of qualitative choice behavior." In P. Zarembhk (ed.). *Frontiers in Econometrics.* Academic Press: Berkeley, California.

McFadden, D. (1987). "Regression-based specification tests for the multinomial logit model." *Journal of Econometrics* 34: 63–82. https://doi.org/10.1016/0304-4076(87)90067-4.

McFadden, D. (1989). "A method of simulated moments for estimation of discrete response models without numerical integration." *Econometrica* 57(5): 995–1026. https://doi.org/10.2307/1913621.

McFadden, D. and K. Train (2000). "Mixed MNL models for discrete response." *Journal of Applied Econometrics* 15: 447–470. https://doi.org/10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1.

McLachlan, G. and D. Peel (2005). *Finite mixture models*. John Wiley & Sons.

OPEC (2016). OPEC Share of World Crude Oil Reserve, http://www.opec.org/opec_web/en/data_graphs/330.htm.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and J. Vanderplas (2011). "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12(Oct), 2825–2830.

Piendl, R., G. Liedtke, and T. Matteis (2017). "A logit model for shipment size choice with latent classes—empirical findings for Germany." *Transportation Research Part A* 102: 188–201. https://doi.org/10.1016/j.tra.2016.08.023.

Rich, J., P.M. Holmblad, C.O. Hansen (2009). "A weighted logit freight mode-choice model." *Transportation research Part E* 45: 1006–1019. https://doi.org/10.1016/j.tre.2009.02.001 .

Samuel, A.L. (1959). "Some studies in machine learning using the game of checkers." *IBM Journal* 3(3): 535–554. https://doi.org/10.1147/rd.33.0210.

Steve, A.B. and T.M. Corsi (2012). "Choosing a port: an analysis of containerized imports into the U.S." *Transportation Research Part E* 48: 881–895. https://doi.org/10.1016/j.tre.2012.02.003.

Timmermans, H.J. (1996). "A stated choice model of sequential mode and destination choice behaviour for shopping trips." *Environment and Planning A* 28(1): 173–184. https://doi.org/10.1068/a280173.

Train, K. (1986). *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*. MIT Press. Chapter 8.

Train, K. (2008). "EM Algorithms for nonparametric estimation of mixing distributions." *Journal of Choice Model* 1(1): 40–69. https://doi.org/10.1016/S1755-5345(13)70022-8.

Train, K. and C. Winston (2007). "Vehicle Choice Behavior and the Declining Market Share of U.S. Automakers." *International Economic Review* 48(4): 1469–1496. https://doi.org/10.1111/j.1468-2354.2007.00471.x.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory (information science and statistics)*. Springer Verlag. https://doi.org/10.1007/978-1-4757-2440-0.

Vij, A. and R. Krueger (2017). "Random taste heterogeneity in discrete choice models: flexible nonparametric finite mixture distributions." *Transportation Research Part B.* 106: 76–101. https://doi.org/10.1016/j.trb.2017.10.013.

Wang, L., and Lo, L. (2007). "Immigrant grocery-shopping behavior: Ethnic identity versus accessibility." *Environment and Planning A* 39(3): 684–699. https://doi.org/10.1068/a3833.

Witten, I.H. and E. Frank (2005) *Data Mining Practical Machine Learning Tools and Techniques*. San Francisco: Elsevier Inc.

Wu, X., V. Kuman, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D.J. Hand, D. Steinberg (2008). "Top 10 algorithms in data mining." *Knowledge and Information System*. 14(1): 1–37. https://doi.org/10.1007/s10115-007-0114-2.