Latin American Oil Export Destination Choice: A Machine Learning Approach

Haiying Jia[1], Roar Adland[1], Yuchen Wang[2]
[1]Norwegian School of Economics, Bergen, Norway
[2]Operations Research Center, Massachusetts Institute of Technology, Cambridge, USA
Haiying.jia@nhh.no

Abstract - **We implement machine learning techniques to predict the destination for Latin American crude oil exports. Utilizing a unique dataset of micro-level crude oil shipment data, derived from the Automatic Identification System (AIS) for ship tracking, we investigate the micro- and macro-level determinants of the destination choice. We use decision tree, Random Forests and boosted trees techniques in training a model to predict the export destinations which can help to identify seller/buyer groups with similar oil trade requirements. The results show that while macro data, such as regional oil price differences and crack spreads, impacts the crude oil flow, micro level information about the oil shipment are key attributes in the destination prediction. Our research has practical implications, particularly with regards to prediction of oil transportation demand, spatial price arbitrage and short-term forecasting of regional crack spreads.**

Keywords - **big data, crude oil, machine learning**

I. INTRODUCTION

Latin America as a region may not always catch the world's attention on the global energy scene, yet some of the countries in the region possess some of the largest oil and gas reserves in the world. Venezuela and Mexico have been large oil producers for decades, and oil discoveries ten years ago in the Tupi fields off the coast of Brazil has made the country one of the major oil and gas producers in the world. According to the World Factbook [1], seven countries in Latin America are net oil and gas exporters as of 2017, with total exporting volume of 8.3 million barrels per day (mbbl/day). Though the picture of the region's economic outlook is mixed – recessions in Brazil and Argentina are said to come to an end; while the situation in Venezuela continue to be disturbing – the region's dependence on oil will continue to be an essential feature of its integration into the world economy.

The eventual destination of oil exports is the result of a complex and dynamic system including, for instance, trade agreements (long-term bilateral agreement and short-term commercial contracts), political relationships (sanctions or restrictions), supply and demand, and price fluctuations. Discrete choice models, see for instance, Malchow and Kanafani [2], Rich *et al.* [3], Steven and Corsi [4], and Piendl et al. [5], provide a theoretical foundation for this research in terms of the choices are statistically related to some attributes. However, we opt for utilizing machine learning techniques which have the advantage of dealing with high dimensionality, mixture date type and nonstandard data structure [6]. As there has been no academic research on the topic, the contribution of this paper is to provide an in-depth investigation of the attributes that determine the destination of seaborne oil exports using machine learning algorithms. Based on actual micro-level crude oil shipment data for the period January 2013 through mid-March 2016, we investigate how the destinations are determined based on cargo information (such as seller's identity, loading port, and cargo grade etc.) and economic data (e.g. oil prices and crack spread). We train the machine learning algorithm based on historical data and test the out-of-sample prediction performance of the model. Our results are potentially important as a building block in commercial applications that deal with oil and freight market analysis.

II. METHODOLOGY

We are given data (w_i, y_i), $i=1,2,\ldots,n$ with $w_i \in \mathbb{R}^p, y_i \in \mathbb{Z}$. Here output y_i is a categorical variable with values 1, 2, …, k. The goal is to build a classification model for predicting the values of y from new x values. In this paper three types of machine learning models are utilized to find solutions to the destination classification problem: decision trees, Random Forests and Boosted trees.

2.1 Decision trees

A "divide-and-conquer" approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree [2]. 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.

2.2 Random Forests

The Random Forest [6] 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.

2.3 Gradient Boosted trees

Similar to Random Forest, Gradient Boosted trees (GBT), see Friedman [7] [8], 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

978-1-7281-3804-6/19/$31.00 ©2019 IEEE345

a previously trained tree. Boosting is an ensemble technique in which the predictors are not made independently, but sequentially [9]. 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) [10].

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 [11] [12].

2.4 Gaussian Naïve Bayes Classification

For the purpose of benchmarking the performance of the machine learning algorithms, we include a Gaussian Naïve Bayes (GNB) Classification model, in which a simple and strong assumption is levied on the data: independence among the attributes. 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.

III. DATA

We use a dataset comprised of 14,745 oil shipments loading from 43 ports in 17 Latin American 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.); grade country/region; grade api (e.g. light, medium or heavy), sulfur content (e.g. sweet or sour), size (bbls).

2. Seller and buyer identity: load_owner, offtake_owner

3. Trade information: load_port/country/region, load_time, offtake_port/country/region, offtake_time.

4. Vessel information: Details about the vessel undertaking each shipment (e.g. vessel name, IMO, flag, class, Year-of-Built, Deadweight)

The cargo data is provided by Clipper Data and enriched with technical vessel data from the World Fleet Register of Clarkson Research. 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).

In an attempt to proxy some of the macro-level attributes that affect the destination of crude oil exports we six categories of economic 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 we include 38 time series.

IV. RESULTS

The prediction is done at the exporting country level, i.e. 17 sets. Each of the dataset is divided into (1) 60% as the training set, (2) 20% as the unbiased cross validation set, and (3) 20% as the out-of-sample test set. The division is purely by random allocation of certain percentage of observation to each set. In each export country case, we train three models as illustrated in section II.

For the purpose of illustration, we present a partial decision tree classification result for the predicted destination ports for Venezuela oil exports in Fig. 1. In this case, the oil grade is the first classifier (level 0) – if the oil cargo grades belong to certain 13 grades (out of 16 in total) shipments data is classified to the Left branch; otherwise to the Right branch. For the level-1 left branch, the next classifier is load terminal – if shipments are loaded in either Puerto De La Cruz Refinery or Puerto Miranda Terminal, the data is further classified to the level 2-Right branch, which is then further classified by oil grade, LPG inventory level or Cargo size, and so on. Eventually, those shipments, for instance, which are “diluted crude oil, loaded in Puerto Miranda Terminal, when LPG inventory level is over 142 mln barrels”, with about 65% confidence, our models suggest that the destination port is Port Aalborg in Denmark.

978-1-7281-3804-6/19/$31.00 ©2019 IEEE

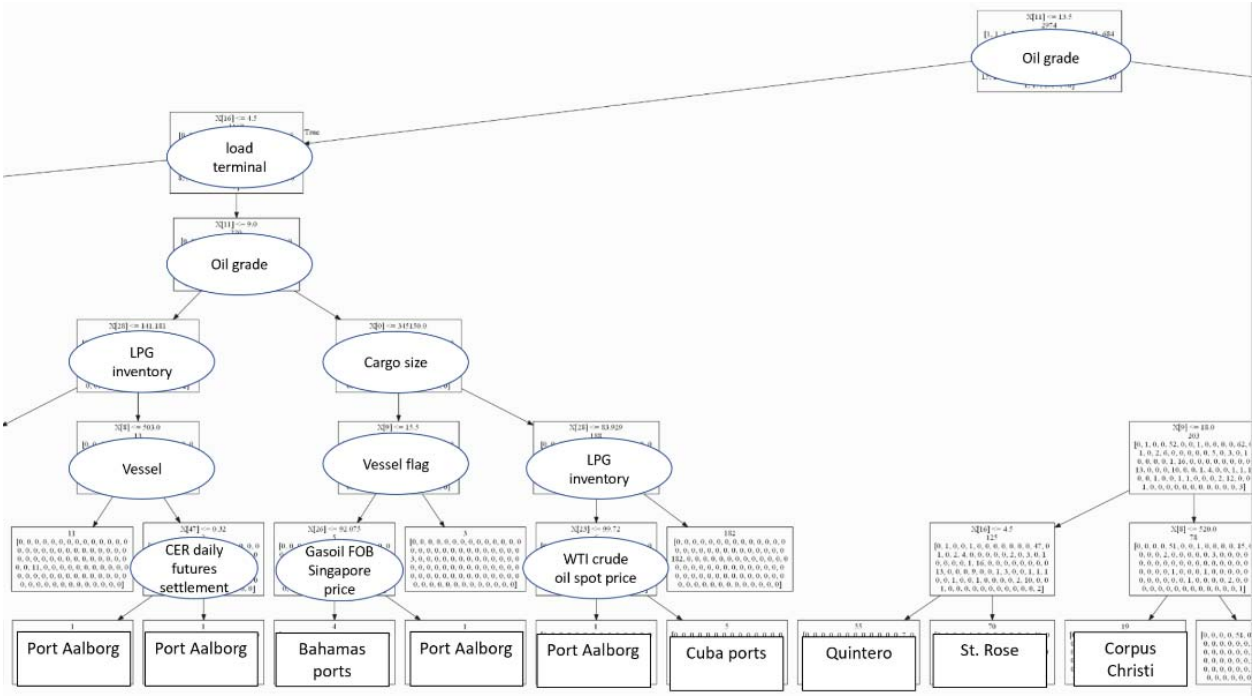

Fig. 1. Case example: *Partial* Decision tree predicted destination ports Venezuela oil export

Table 1. Test set accuracy for predicting offtake countries by load country

Load country	Training accuracy			Test accuracy			Naïve Bayes	% shipments	
	Dtree	Random forests	GBR	Dtree	Random forests	GBR	Classification	to the US	Shipment obs
Venezuela	83.20%	84.01%	86.42%	84.68%	83.47%	86.42%	16.40%	44.20%	3719
Mexico	86.50%	88.98%	89.26%	89.67%	91.05%	92.29%	52.20%	68.29%	3630
Colombia	68.36%	73.16%	75.71%	70.14%	72.96%	74.93%	12.11%	50.85%	1772
Ecuador	75.58%	81.78%	86.43%	76.74%	82.95%	88.37%	24.81%	27.24%	1290
Brazil	65.09%	70.28%	70.28%	66.98%	71.70%	73.11%	19.81%	63.02%	1061
Panama	90.64%	90.64%	90.64%	90.64%	91.23%	90.06%	14.62%	73.19%	854
Netherland Antilles	72.35%	77.06%	75.29%	69.41%	76.47%	75.29%	34.71%	26.53%	849
Bahamas	73.95%	73.95%	78.15%	75.63%	81.51%	79.83%	20.17%	53.69%	596
Aruba	72.88%	74.58%	76.27%	71.19%	71.19%	69.49%	38.98%	45.05%	293
Uruguay	88.37%	90.70%	90.70%	93.02%	95.35%	90.70%	58.14%	56.86%	216
St. Lucia	68.18%	68.18%	68.18%	68.18%	68.18%	90.91%	86.36%	38.89%	109
Argentina	80.00%	80.00%	80.00%	80.95%	80.95%	80.95%	52.38%	75.23%	102

In many applications, the prediction of importing country would be sufficient. Therefore, we report the results for prediction offtake countries for various exporting countries in Table 1.

There are a few takeaways from Table 1. Firstly, comparing to the benchmark Gaussian Naïve Bayes Classification, the machine learning models perform exceptionally well. Secondly, the test accuracy does not decline materially out of sample. As the training and testing 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. It also presumably reflects the fact that trading patterns are somewhat stable. Thirdly, we also see that countries that are not in fact oil exporters, but locations for large tank storage facilities in the Caribbean (St. Lucia, Netherlands Antilles, Bahamas and Aruba falls into this category), have lower prediction performance than the remaining countries. This reflects the fact that transshipments from these countries will typically be opportunistic and driven by the oil trading companies that own the storage and transshipment facilities. As such, destinations are likely to be more random in nature and therefore harder to predict. Panama is an outlier here as it is not an oil producer, only a transshipment hub, yet destinations appear to be highly predictable. Venezuelan and Mexican exports have a high prediction accuracy, driven by the historical dominance of the United States as a recipient.

V. CONCLUSIONS

This paper is the first academic research to apply cutting-edge machine learning models in predicting 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. Our application to Latin American crude oil exports at the country level results in test accuracies ranging between 70 and 90% - a strong performance. Predicting oil export destinations allows for better forecasting of regional and local market balance, improved knowledge of inventory levels and monitoring of the supply chain.

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 in an effort to improve oil and tanker freight market analysis.

Further research in this area should look at improving prediction accuracy by diving further into the details, such as vessel type, vessel ownership, and trade characteristics.

ACKNOWLEDGMENT

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

REFERENCES

[1] CIA, "The World Factbook," pp. https://www.cia.gov/library/publications/resources/the-world-factbook/fields/261.html, 2018.

[2] M. a. K. A. Malchow, "A disaggregate analysis of port selection," *Transportation Research Part E ,* vol. 40, pp. 317-337 , 2004.

[3] J. H. P. a. H. C. Rich, "A weighted logit freight mode-choice model," *Transportation Research Part E,* vol. 45, pp. 1006-1019, 2009.

[4] A. a. T. C. Steven, "Choosing a port: an analysis of containerized imports into the US," *Transportation Research Part E,* vol. 48, pp. 881-895, 2012.

[5] R. G. L. a. T. M. Piendl, "A logit model for shipment size choice choice with latent classes - empirical findings for Germany," *Transportation Research Part A,* vol. 102, pp. 188-201, 2017.

[6] L. J. F. R. O. a. C. S. Breiman, Classification and regression trees, Chapman & Hall/CRC, 1998.

[7] J. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics,* pp. 1189-1232, 2001.

[8] J. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis,* vol. 38, no. 4, pp. 367-378, 2002.

[9] P. Grover, "Gradient boosting from scratch," 2017. [Online]. Available: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.

[10] M. Kearns, "Thoughts on hypothesis boosting," 1988. [Online]. Available: Thoughts on hypothesis boosting. https://www.cis.upenn.edu/~mkearns/papers/boostnote.pdf.

[11] F. V. G. G. A. M. V. T. B. G. O. .. &. V. J. Pedregosa, " Scikit-learn: Machine learning in Python.," *Journal of machine learning research,* vol. 12, pp. 2825-2830, 2011.

[12] T. a. C. G. Chen, " Xgboost: A scalable tree boosting system," *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* pp. 785-794, 2016.