Opportunity 1: Forecast Transportation Cost

Author: Gorden Li, Lingyi Chen, Wei Xu, Qinyi Liu

Organization: Wustl Olin Business School

Abstract: In the paper, we include 5979 samples to fit the forecast model after data cleaning. Random Forest is used to do prediction and traditional regression is used to do indication. According to our study, Random Forest Model can reach 0.91 in R^2 and MAPE is as low as 12%. We use Bootstrap and LASSO to do indication of variables. Miles, Route Highway Density, Total Gross Weight, FRUITS Size have positive relationship with Actual Freight Cost, indicating that orders with longer ship distance, better highway system, higher products weight cost more. For Route Highway Density, the well-developed highway system may reduce the time order transportation cost, but the toll may increase, leading to higher total cost. Also, from the coefficients, we can see that in Spring & Summer, the transportation cost is comparatively lower due to supply and demand relation.

Key Words: Random Forest, Bootstrap, LASSO, Miles, Route Highway Density, Total Gross Weight, Season

1 Data Preparation

1.1 Data Cleaning

The original data set combined both Report 2 & Report 3 contains 10078 samples. After deleting in null-value sample, abnormal-value sample and removing samples with possible noise interrupting models, 5979 samples were remained including 80+ predictors. The cleaned data set still cover almost all the information in important dimension like Actual Freight Cost, Miles, Gross Weight etc. Credibility of the cleaned data set is ensured.

1.2 Derived & Additional Variables

To fit the model more accurately, we include some derived and additional variables below.

1.2.1 Season

According to "Weather and Climate Change Implications for Surface Transportation in the USA" [1], climate change will affect the efficiency, safety and reliability of existing transportation infrastructure , and thus demand and supply will change accordingly, leading to transportation cost change. So, we should include season as a factor.

1.2.2 Route Highway Density

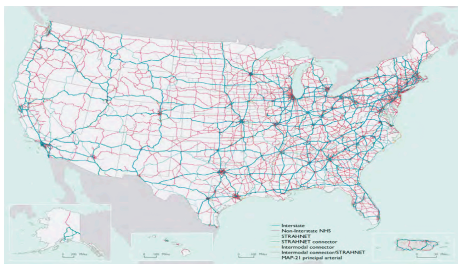

Figure 1 - National Highway System [2]

Figure 1 shows the National Highway System (NHS) and other principal arterials and intermodal connectors, comprising an extensive system of highways supporting densely populated urban centers in the northeast and parts of the Midwest, South, and West. We can see from the Figure 1 that development condition of Highway System in the United States differs by region. It gradually decreases from Northeast area to the Southwest area.

So, we put different weight on each state based on area to get their route highway density.

1.2.3 Weekly Diesel Price

Diesel price is a critical factor when deciding the price of transporting products. After searching for related information, trucks in America will use Diesel No.2.

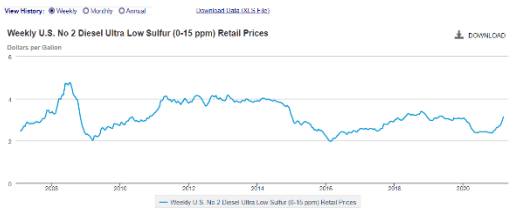

Figure 2- Weekly Diesel Price [3]

As Figure 2, we get diesel price every week based on data from U.S. Energy information Administration.

Also, we found that average distance per week of a fully-loaded truck is about 2000km. [4] , we then assume that price of 7 days before the date of order is the price of the last day, and then we get the price of every order.

2 Feature Engineering

2.1 Why Feature Engineering?

Feature Engineering is a process with which important features, also called predictors, having great relation with the

dependent variable will be chosen into model. Meanwhile, it helps to relieve issues like multicollinearity to some extent to make the result more accurate. Proper choice of predictors decides the ceiling of any model. If the model contains a lot number of unnecessary predictors, not only the accuracy will be impacted, but the fitting speed of model will be lowered.

2.2 Random-Forest Feature Filtering

2.2.1 Why Random-Forest?

Random-Forest is chosen in feature filtering since it is a powerful tool in dealing with filtering involving great number of predictors. Besides, it is also adaptable in solving both linear and non-linear data and data is not necessary to be scaled. **Random-Forest is powerful, stable and no requirement in data preprocess.**

2.2.2 Result of Random-Forest Filtering

From the Random-Forest model with 100 decision trees in it, 15 possible predictors were chosen, leading by **Miles**, **Diesel Price**, **Route Highway Density**, **Gross Weight**, **Season**, **Year**. These feature covers more than 95% of contribution of importance.

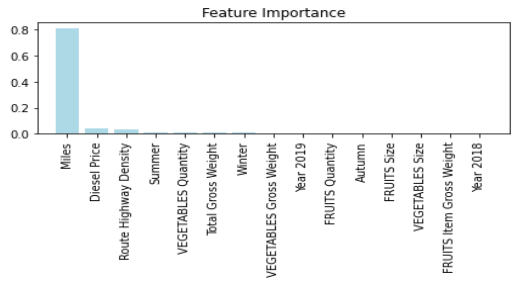

Figure 3-Feature Importance of Random-Forest

2.3 Correlation analysis

To increase credibility, correlation analysis is included as sensitivity check.

	Abs Correlation
Actual Freight Cost	1.00
Miles	0.87
FRUITS Gross Weight	0.52
VEGETABLES Gross Weight	0.30
Route Highway Density	0.29
FRUITS Size	0.29
FRUITS Quantity	0.29
VEGETABLES Quantity	0.24
VEGETABLES Size	0.09

Table 1-Abs correlation with Actual Freight Cost

We can also see from correlation analysis **Miles**, **Gross Weight**, **Route Highway Density** are among main features.

2.4 Significant Variables Description

After Filtering the feature, dimension of predictors was lower significantly from 80+ to 12 most significant feature in Table-2.

Variable Name	Description
Miles	Miles
Route Highway Density	Highway Length(km)/State Area(km^2); Highway including interstate highway, US highway, State highway
Doner Highway Density	Density of Doner State
Receiver Highway Density	Density of Receiver State
Diesel Price	No.2 Diesel Price ($/GAL)
Total Gross Weight	Weight of Goods (lbs)
VEGETABLES Gross Weight	Weight of Vege (lbs)
VEGETABLES Quantity	Unit number of Vege
VEGETABLES Size	Weight of each Vege unit (lbs)
FRUITS Gross Weight	Weight of Fruit (lbs)
FRUITS Quantity	Unit number of Fruit
FRUITS Size	Weight of each Fruit unit (lbs)
Year	Dummy, {2018~2021}
Season	Dummy, {Spring, Summer, Autumn, Winter}

Table 2-Significant Variables Description

3 Forecast Transportation Cost

3.1 Description Analysis

3.1.1 Main Feature Distribution

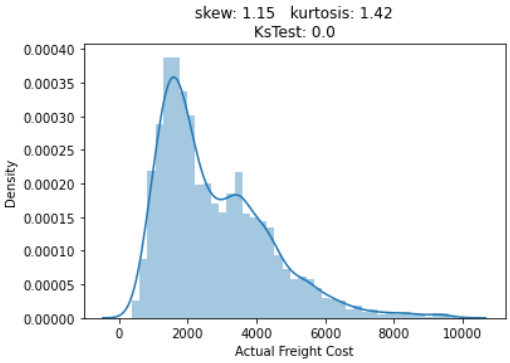

Figure 4-Distribution of Actual Freight Cost

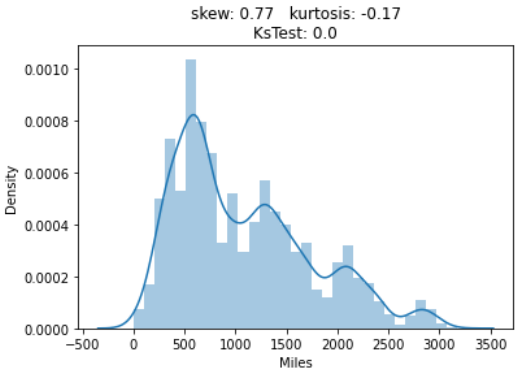

Figure 5-Distribution of Miles

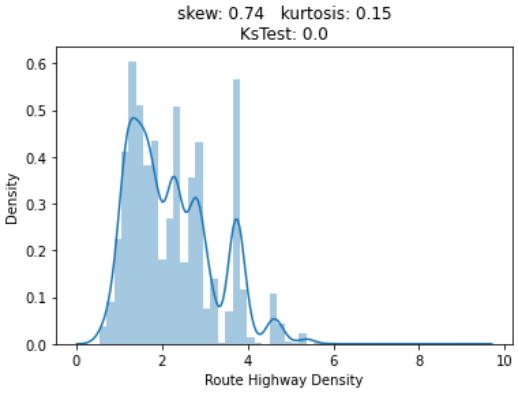

Figure 6-Distribution of Route Highway Density

The main feature description (including skewness, kurtosis, distribution) of Actual Freight Cost, Miles and Route Highway Density shows that these factors' data is not normally distributed. So, we choose **bootstrap** to improve the accuracy of coefficient estimation.

3.1.2 Main Feature Relation Overview

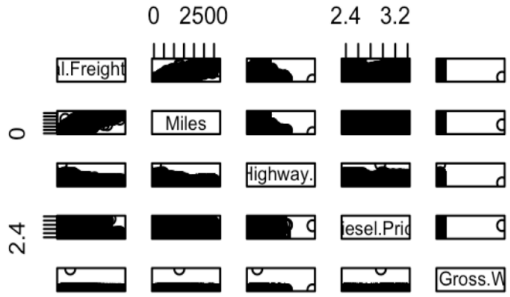

Figure 7- Relationship Between main Variables

According to Figure 7, Actual Freight Cost has a linear relationship with most of the main factors (miles, Route Highway Density, Diesel Price, Total Gross Weight, etc.). Thus, there is no need to do data transformation.

3.2 Random Forest Forecasting

Random Forest is good in fitting non-linear with bootstrap method, high dimension data with high accuracy. Moreover, it does not require data preprocess like transformation, standardization, normalization etc. It' s user friendly.

3.2.1 Hyperparameter adjustment

It' s important to adjust Hyperparameter in Random Forest not only for the accuracy on test data but also for reduction in time consumed. Here we will use **grid search** to adjust **2 hyperparameter**: num of trees in forest (**n_estimators**) & max depth of each tree (**max_depth**)

In Random Forest Model, we still use significant variables in Table-2.

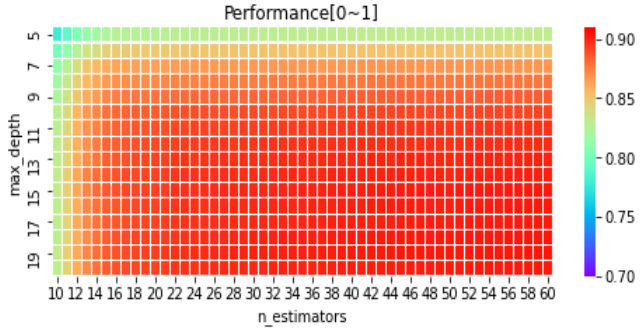

Figure 8-Model Performance Heatmap

Figure 8 describe how performance of model changes as n_estimators & max_depth change. When n_estimators and max_depth increase, performance of explaining the data increase. However, time consumed increased as well. The best balance between performance and computational workload is meted when **n_estimator = 60** and **max_depth = 15,** the model can explain **91%+** of true value with limited time.

3.2.2 Performance of Models

Using the Random Forest Model with n_estimator = 60, max_depth = 15, we can calculate the avg %Error between estimation & true value by using MAPE:

$$MAPE = \frac{\text{Abs(Estimate Cost} - \text{True Cost)}}{\text{True Cost}}$$

Formular 1-MAPE(Error Rate)

$$True\ Cost = \frac{\text{Estimated Cost}}{1 \pm \text{MAPE}}$$

Formular 2-Relation between True Cost & Estimation

	R^2	MAPE
Random Forest	0.91	12.3%

Table 3-Performance of Random Forest

From the comparison of Table 3, we see that Random Forest is significantly much better in fitting the data, with MAPE = 12.3%

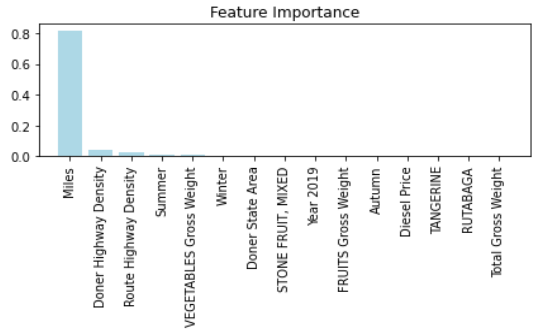

Figure 9-Feature Importance

In Figure 9, Miles, Highway Density, Season and Gross Weight also play important roles.

3.2.3 Whether to take the order

According to Formular 2, we can get the **fair value (True Cost)** of each order. However, the market will fluctuate due to a lot of reason out of the model like price competition within carriers, price strategy etc. Formular 2 help recipient recognize the True Value of each transportation. Based on such fact, we can create a tool for recipient to decide whether to accept the order or not based on the relation between listed cost and fair value (true cost). We set a **Threshold ≤ MAPE**. Threshold is set based on your urgency and financial capability. More urgent, capable on financial status you are, threshold you set can be bigger:

Reject order: $Listed\ Cost = \frac{\text{Estimated Cost}}{1-\text{MAPE}}$

Wait for more choice: $Listed\ Cost \in [\frac{\text{Estimated Cost}}{1-Threshold}, \frac{\text{Estimated Cost}}{1-\text{MAPE}}]$

Accept order: $Listed\ Cost = \frac{\text{Estimated Cost}}{1-\text{Threshold}}$

3.2.4 Limitation of Model

Though Random Forest has a good performance on both training and test set with a R^2 = 0.91, the model still has some limitation. For instance, the Forest count the influence of price competition between carriers into error. If we can have the data of number of carriers in each city, estimated number of trucks owned by them, our model can be more accurate.

3.3 Traditional Model Forecasting

While Random Forest Forecasting can fit the model quite well, it' s a black box. We still need traditional model to identify the **indication** of variables.

3.3.1 Model Choosing

Since there are a large number of

predictors in the original dataset, we use StepWise and LASSO to run the regression. StepWise can filter data, and similarly, LASSO can effectively reduce multicollinearity and unnecessary predictors by using F1-norm function.

3.3.2 Performance of Models

Based on the feature importance showed in Figure 3, we select Miles, Route Highway Density, Diesel Price, Total Gross Weight, VEGETABLES Gross Weight, VEGETABLES Quantity, VEGETABLES Size, FRUITS Gross Weight, FRUITS Quantity, FRUITS Size, Year (as Dummy), and Season (as Dummy) as predictors in the model.

When Building models, we use StepWise to filter the variables, which will not cause multicollinearity. The coefficients of these filtered variables are shown in Table 4.

Variables	Coefficient	Significance
(Intercept)	1838.85***	***
Miles	2.07***	***
Route Highway Density	35.22***	***
Diesel Price	-129.77***	***
Total Gross Weight	0.0027***	***
VEGETABLES Gross Weight	-0.0032***	***
FRUITS Size	0.19***	***
Year 2018	-619.12***	***
Year 2019	-1019.90***	***
Year 2020	-791.26***	***
Spring	-383.85***	***
Summer	-434.00***	***
Autumn	-72.57***	***

Table 4-Coefficient of Variables (StepWise)

Then use Kappa function to test the multicollinearity. In this model, Kappa = 22.73, which shows there exists no multicollinearity. In this model, Year 2021 and Winter are excluded since their coefficient is not significant.

We continue to use variables selected from the StepWise model to bootstrap regression and LASSO regression as sensitive analysis. The performance of models is not as good as Random Forest but they give us implication of predictors shown in Table 5.

	R^2	MAPE
Linear Regression	0.80	22.9%
LASSO Regression	0.80	22.8%
Random Forest	0.91	12.3%

Table 5-Performance between models

The coefficients of these filtered variables are showed in Table 6.

Variables	Coef of Bootstrap	Coef of LASSO
(Intercept)	1987.54***	743.01***
Miles	2.06***	2.06***
Route Highway Density	35.21***	35.79***
Diesel Price	-129.77***	-123.84***
Total Gross Weight	0.0027***	0.0023***
VEGETABLES Gross Weight	-0.0032***	-0.0029***
FRUITS Size	0.19***	0.20***
Year 2018	-619.14***	142.75***
Year 2019	-1020.00***	-235.82***
Year 2020	-791.26***	NA
Year 2021	NA	783.83***
Spring	-384.85***	-75.72***
Summer	-434.00***	-134.84***
Autumn	-72.57***	254.97***
Winter	NA	310.14***

Table 6- Coefficient of Variables (Bootstrap & LASSO)

The outcome shown in Table 4 & Table 6 aligned with each other. Miles, Route Highway Density, Total Gross Weight, FRUITS Size have positive relationship with Actual Freight Cost, indicating that orders with longer ship distance, better highway system, higher products weight or larger product size cost more. For Route Highway Density, the well-developed highway system may reduce the time order transportation cost, but the toll may increase, leading to higher total cost. Also, from the coefficients, we can see that season factors have negative relationship with Actual Freight Cost, and the cost of order transportation will be relatively low in summer.

References

[1] World Meteorological Organization (WMO). 2009. Weather and Climate Change Implications for Surface Transportation in the USA. Available at https://public.wmo.int/ as of 2009.

[2] United States Department of Transportation. 2020. Transportation Statistics Annual Report 2020 Available at https://fdd.bts.gov/freight-data-dictionary/ as of 2020

[3] U.S. Energy information Administration. 2020. Available at https://www.eia.gov/ as of 2020

[4] Quora Answer Board. 2020 Available at https://www.quora.com/How-many-miles-can-the-average-semi-truck-pull-a-full-load-without-refueling as of 2016