

University of Louisville

## ThinkIR: The University of Louisville's Institutional Repository

---

Electronic Theses and Dissertations

---

5-2020

### A machine learning approach for allocating route cost to customers for transportation and logistics services.

Alison Davis  
*University of Louisville*

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

 Part of the [Industrial Engineering Commons](#)

---

#### Recommended Citation

Davis, Alison, "A machine learning approach for allocating route cost to customers for transportation and logistics services." (2020). *Electronic Theses and Dissertations*. Paper 3355.  
Retrieved from <https://ir.library.louisville.edu/etd/3355>

This Master's Thesis is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact [thinkir@louisville.edu](mailto:thinkir@louisville.edu).

A Machine Learning Approach for Allocating Route Cost to Customers for  
Transportation and Logistics Services

By

Alison Davis  
B.S., University of Louisville, 2019

A Thesis  
Submitted to the Faculty of the  
University of Louisville  
J. B. Speed School of Engineering  
as Partial Fulfilment of the Requirements  
for the Professional Degree

MASTER OF ENGINEERING

Department of Industrial Engineering

May 2020



A Machine Learning Approach for Allocating Route Cost to Customers for  
Transportation and Logistics Services

Submitted by:    
Alison Davis

A Thesis Approved on

4/1/2020

---

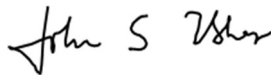
(Date)

by the Following Reading and Examination Committee:



---

Lihui Bai, Thesis Director



---

John Usher



---

Dar-Jen Chang

## ACKNOWLEDGEMENTS

I'd like to express my gratitude to my advisor and professor, Dr. Lihui Bai, for her continuous support and encouragement throughout this thesis. I'm indebted to my friends and family, especially my mom, for their vital moral support throughout this process. I'm forever grateful for the collaboration with the industry partners that made this research possible.

## ABSTRACT

Advancements in big data enabled management practices inspire logistics companies to study deeper into their transportation operations with a data driven approach. One such question asks: How can a logistics firm identify high-cost customers in their service network? In the presence of rich data on routes involving many customers, this thesis develops a framework to allocate a route cost among customers that the route serves, where each route is associated with multiple route features related to the transportation cost. Cost is allocated using the proportional allocation approach in combination with the random forest method in machine learning. First, this framework ensembles random forest regression models to determine the importance values of all route features. Next, the importance values of route features are used to allocate cost among customers. Finally, posterior analysis identifies customers in a route or in general that are most costly to serve. Several additional analyses are performed to show potential uses of this cost allocation output. Results of the framework and analyses on three simulated case and two industry cases show the validity of the model and the potential for actionable operational analysis and changes.

## TABLE OF CONTENTS

|  | <u>Page</u> |
|--|-------------|
| APPROVAL PAGE .....  | ii          |
| ACKNOWLEDGEMENTS .....   | iii         |
| ABSTRACT .....   | iv          |
| LIST OF TABLES .....   | vii         |
| LIST OF FIGURES .....  | viii        |
| I. INTRODUCTION .....  | 1           |
| II. LITERATURE REVIEW.....                                       | 4           |
| A. Shapley Allocation .....                                      | 4           |
| B. Proportional Allocation .....                                 | 6           |
| C. Machine Learning studies in Logistics and Transportation..... | 7           |
| III. PROBLEM STATEMENT .....                                     | 9           |
| A. Overview of Customer Characteristics.....                     | 9           |
| B. Overview of Route Characteristics.....                        | 11          |
| C. Machine Learning.....   | 12          |
| D. Cost Allocation Algorithm .....                               | 13          |
| IV. AN INTEGRATED PREDICTION AND COST ALLOCATION FRAMEWORK. 14   |             |
| A. Preliminaries on Machine Learning .....                       | 14          |
| 1. Regression Tree .....   | 14          |
| 2. Random Forest Regression.....                                 | 17          |
| B. Random Forest Prediction Model.....                           | 18          |
| 1. Feature Development.....                                      | 18          |
| 2. Feature Selection .....                                       | 20          |
| 3. Modeling and Feature Importance.....                          | 21          |
| C. Cost Allocation Algorithm .....                               | 21          |
| D. An Integrated Decision Support System .....                   | 22          |
| V. COMPUTATIONAL RESULTS.....                                    | 24          |
| A. Simulated Cases .....   | 24          |
| 1. General and Case Specific Route Features.....                 | 26          |
| 2. Feature Analysis .....  | 26          |
| 3. Hyperparameter Tuning.....                                    | 33          |

|   |    |
|---|----|
| 4. Model Performance .....  | 35 |
| 5. Feature Importance .....   | 37 |
| B. Overview of Industry Cases .....                                   | 40 |
| C. A Decision Support System Using the Cost Allocation Framework..... | 42 |
| 1. Visualization of Cost Allocation .....                             | 42 |
| 2. Customer Grouping .....  | 43 |
| 3. New Customer Pricing.....  | 45 |
| VI. CONCLUSION AND FUTURE RESEARCH .....                              | 48 |
| A. Conclusion.....  | 48 |
| B. Future research .....  | 49 |
| REFERENCES .....  | 50 |



## LIST OF TABLES

|   |    |
|---|----|
| TABLE I TABLE OF CHARACTERISTICS AND FUNCTION OF DERIVATION ... | 19 |
| TABLE II TIME WINDOW FUNCTION EXAMPLE.....                      | 20 |
| TABLE III DESCRIPTION OF SIMULATED CASES.....                   | 25 |
| TABLE IV FEATURE-TO-FEATURE CORELATIONS FOR BASELINE.....       | 27 |
| TABLE V FEATURE IMPORTANCE RANKINGS .....                       | 40 |
| TABLE VI COMPANY A AND B SPECIFIC FEATURES.....                 | 41 |
| TABLE VII CUSTOMER GROUPING EXAMPLE FOR CUSTOMER 1 .....        | 45 |

## LIST OF FIGURES

|  |    |
|--|----|
| FIGURE 1 - Customer Distance Characteristic .....  | 10 |
| FIGURE 2 - Customer Proximity Characteristic .....   | 11 |
| FIGURE 3 - Methodology Flow Chart .....  | 13 |
| FIGURE 4 - Example Regression Tree .....   | 16 |
| FIGURE 5 - Violin plot of the Average of Customer Time Windows route feature for Baseline for (a) train data and (b) test data ..... | 29 |
| FIGURE 6 - Violin plot of the Total Deliveries route feature for Baseline for (a) train data and (b) test data .....                 | 30 |
| FIGURE 7 - Violin plot of the Total Customer Stop Time route feature for Baseline for (a) train data and (b) test data .....         | 30 |
| FIGURE 8 - Violin plot of the Total Product Amount (Gallons) route feature for Baseline for (a) train data and (b) test data.....    | 31 |
| FIGURE 9 - Violin plot of the Sum of Customer Distances route feature for Baseline for (a) train data and (b) test data .....        | 31 |
| FIGURE 10 - Violin plot of the Sum of Customer Proximities route feature for Baseline for (a) train data and (b) test data.....      | 32 |
| FIGURE 11 - Feature correlation to route cost for Baseline .....   | 33 |
| FIGURE 12 - Baseline Tree Depth Tuning .....   | 34 |
| FIGURE 13 - Alternative 1 tree depth tuning.....   | 34 |
| FIGURE 14 - Alternative 2 tree depth tuning.....   | 35 |
| FIGURE 15 - Model performance for the Baseline case.....   | 36 |
| FIGURE 16 - Model performance for Alternative 1 .....  | 36 |
| FIGURE 17 - Model performance for Alternative 2 .....  | 37 |
| FIGURE 18 - Feature importance values for Baseline.....  | 39 |
| FIGURE 19 - Feature importance values for Alternative 1 .....  | 39 |
| FIGURE 20 - Feature importance values for Alternative 2.....   | 39 |
| FIGURE 21 - Company A and B Model Performance .....  | 41 |
| FIGURE 22 – Baseline case customer cost/gallon .....   | 43 |
| FIGURE 23 – Baseline case network focused on the Atlanta department colored by a customer’s average cost/gallon.....                 | 43 |
| FIGURE 24 - Prediction settings for a new customer .....   | 46 |
| FIGURE 25 - Prediction results based on the prediction settings .....  | 46 |

## I. INTRODUCTION

In the area of transportation and logistics planning there exists the long-standing problem of how to best allocate costs among customers on a route, stemming from both academic and practical interests. For example, in fundamental research from 1984, Samet, et al. (1984) studied the application of cost allocation to the transportation problem using Auman-Shapley Prices, an allocation method rooted in game theory. Driving the interest in this cost allocation problem is route optimization. Minimizing route costs has been used as a common objective in most industry-focused and well-studied vehicle routing problem (VRP) (e.g., Psaraftis, et al., 2016). In the latter, although minimizing the overall costs on routes has been widely adopted and used as standard practice in logistics planning (e.g., (Desrochers, et al., 1992) and (Fabri and Recht, 2006 )), the understanding of individual customers' contribution to the overall route cost is disproportionately understudied. Nevertheless, knowledge of individual customer cost and thus profitability is not only critical to management but has become accessible in this data era. The availability of, perhaps even real-time granular route data such as order sizes, time windows, and other real-time route and customer specific characteristics has allowed and motivated companies to mine deeper into their cost at the customer level instead of the overall cost. Hence, the current thesis attempts to address this gap in the literature, with the aim to provide methodological guidance towards customer-centered cost analysis practices.

In particular, this research will focus on developing a model to fairly allocate costs to customers on the same route. It is envisioned that such a model will be of great value to not only companies in transportation logistics, but others in related fields such as energy markets, airlines, and telecommunications. Of particular importance to the proposed

methodology and models is the ease of usage by potential industry users, who would rely on the actionable model results for developing management solutions and making business decisions.

In the literature, there are two main streams of research on cost allocation in transportation. One stream studies allocation using the Shapley value. Relevant literature such as Dror (1990) and Frisk et al. (2010) uses this value along with game theory principles to allocate costs among cooperating players in a “game”, or a route in the case of this thesis. This method, although theoretically sound, lacks the ability to deal with practical scale and complex features in today’s logistics industry. The second stream studies proportional allocation models including works such as Fishburn and Pollak (1983), Sun, et al. (2016); and Dror (1990). Compared to the first method, the proportional allocation method is easier to implement in practice and produces results that can be interpreted by management; however, the fairness of the output depends greatly on the parameters used to make the allocation.

Therefore, the current thesis focuses on the proportional allocation method due to the simplicity in interpreting the results but expands on current research by integrating machine learning techniques to develop a data-driven model for fair cost allocation. Machine learning models such as decision tree regressors and random forest regressors are studied as methods to produce inputs to the proportional cost allocation method.

In addition to the cost allocation model, the development of a decision support system (DSS) to utilize the results of the model is also presented in this thesis. Management overseeing business decisions requires actionable results. This thesis will expand on the cost allocation model to discuss potential data transformations that can be

integrated with a management centered DSS. The goal of this DSS is to analyze historic customer-based decisions to make improvements to managerial decision-making processes in the future. Specifically, this thesis intends to expand on the allocation model to develop an analysis to isolate high-performing and low-performing customers.

The thesis will continue as follows. It will first review pertinent literature in Section II. Section III will define the problem statement of this thesis including a descriptive overview of the proposed methodology and models. Section IV will present the proposed integrated prediction and cost allocation framework. This will include the development of a machine learning model used by the subsequent cost allocation algorithm. It will also elaborate on the integration of a decision support system using outputs of the allocation algorithm. Section V will follow with the computational results of several simulated and industry cases. It will detail the data preparation methods, the performance of the machine learning model, and the results of several DSS analyses. The thesis will conclude with Section VI summarizing findings and pointing to future research.

## II. LITERATURE REVIEW

The literature review will introduce works related to cost allocation (CA) in the transportation and logistics field. It will focus on two main streams: Shapley value allocation and proportional allocation. They are the most widely used, traditional CA methods according to a survey paper by Guajardo and Rönnqvist (2016) on CA in collaborative transportation.

### A. Shapley Allocation

The first stream of literature is Shapley allocation. Using game theory, Shapley developed a formula to assign a value to each player in a game based on their expected marginal contribution to their coalition. (Shapley, 1953) Shapley allocation is rooted in cooperative game theory where players enter a coalition so that all players benefit from participating. In the case of a transportation problem, the customers (players) enter a route (coalition), where the total coalition cost must be “efficiently” distributed among the players. Additionally, the allocation ensures that the cost allocated to a player is less than the cost they would incur outside of the coalition or in a different coalition. This methodology assumes that the players can make the decision to enter or leave the coalition.

In this stream of research, Engevall et. Al study applied Shapley allocation, also in a cooperative game setting, to a traveling salesman problem at an oil and gas company to allocate costs to customers on a tour (1998). Vanovermeire et al. use the Shapley value to allocate costs among a horizontal alliance composed of three partners to show increased flexibility (2014a). Agarwal and Ergun, instead of analyzing a given customer network, consider an optimal design of a coalition use Shapley value allocation (2010). The

allocation scheme is a parameter in the optimal design of a collaborative coalition of carrier alliances in liner shipping. In fact, using the Shapley allocation in designing coalition is of interest to many other researchers as well. Cruijssen et al. use a Shapley allocation procedure to design a methodology to create more synergistic shipping coalitions and applies the procedure to the Dutch grocery transportation sector (2010). Vanovermeire et al. develop a combined operational plan and cost allocation method for a generalized collaborative bundling problem to satisfy all agents by planning on-time deliveries and ensuring balanced profits (2014b). Zakharov and Shchegryaev develop a cost minimization model for a VRP considering the customer cost distribution described by the Shapley value (2015). Krajewska et al. analyze horizontal cooperation among cost centers in a freight forwarding company to show that cooperation can reduce overall transportation costs using Shapley cost allocation (2008). Computationally, it is shown that Shapley methods are complex and expensive. Further, Shapley allocation-based methods often require assumptions that relax practical business considerations.

Finally, Shapley allocation is also applied to other topics. For example, Fiestras-Janeiro et al. develop a methodology based on Shapley allocation to distribute order cost among agents who place joint orders based on an EOQ model in a joint inventory and transportation problem (2012). In an environmental application, Petrosjan and Zaccour study the allocation of pollution reduction costs among cooperating countries using Shapley allocation (2003).

## B. Proportional Allocation

The second stream of CA literature focuses on proportional allocation (PA). PA methods are computationally simplistic compared to Shapley methods. This method allocates a fraction of the route cost to each customer on the route. Dror offers a proportional allocation method where all customers on a route are allocated equal costs (1990). Fairer proportional methods are those that allocate based on external factors. For example, in a collaborative transportation problem in the forestry industry, Frisk et al. allocate cost to customers based on their proportional demands (2010). This same study finds that the proportional method is more likely to be accepted in industry. A fundamental study by Fishburn and Pollak examines the allocation of cost on a multi-stop trip by airplane where each destination is allocated a cost based on their willingness to pay (1983). Nguyen et al. develop a model to consolidate transportation for suppliers of low-demand products in the agricultural industry in combination with a methodology to allocate costs proportional to the demand of the supplier (2014). Ozener and Ergun develop various cost allocation schemes for routes previously designed for minimal cost (2008). One scheme is a proportional method where costs are allocated to shippers proportional to the cost of the standalone routes. Sun et al. performs a comparative computational study of various allocation methods, including PA, on routes of 5 to 20 customers (2015). The study claims that PA has poor performance when considering fairness but provides a good tradeoff solution when considering practicality and computational efficiency. Studies also consider the allocation of emissions due to transportation among route participants. Özener develops a framework for allocating cost and emissions responsibilities to customers (2014). The research studies proportional models where distance and product amount are the factors for



allocation. Kirschstein and Bierwirth solve the TSP problem and allocate the route emissions proportional to the distance to each stop (2018).

Fields outside of transportation and logistics consider cost allocation as well. Henriot and Moulin consider the allocation of cost in a communication network and develop a model where cost is allocated to users proportional to their traffic, or usage in hours of the network (1996). Baroche et al. develop a model to allocate cost among users in a peer to peer electricity network and develop a proportional cost allocation policy based on so-called “electrical distance” between peers (2019). Moreover, in a manufacturing/remanufacturing setting, Toktay and Wei develop a model where the remanufacturing department assumes a fraction of manufacturing costs (2011).

Given that the proportional application method is simple and has been well accepted by a variety of industries, the current thesis will employ this method considering a wide range of factors (e.g., distance, shipment amount, proximity measure) in such proportional allocation of cost to customers. Fairness of the proportional allocation model is often a concern in literature according to Frisk et al (2010). To consider the fairness of the model, the methodology is integrated with machine learning techniques to methodically select proportionality parameters in this thesis.

### C. Machine Learning studies in Logistics and Transportation

This literature review shows that a machine learning approach to this specific cost allocation problem is novel, however; the use of data science techniques including machine learning is studied in logistics applications. Ma et al. utilizes a data-driven approach to gain

understanding of ripple effects in traffic due to congestion using deep learning techniques (2015). Lin et al. also use deep learning techniques to predict delivery demand to build more efficient logistics models (2018). Similarly, Knoll et al. develop a methodology for predicting future inbound logistics using a generalized machine learning approach (2016). Another common problem addressed in machine learning literature is real-time identification of transportation mode using smart phone captured acceleration data. Shafique and Hato study this machine learning application by comparing the performance of models such as decision trees and random forests, which they find to perform best in predicting the transportation mode (2015). This thesis intends to supplement this field of research by providing an example of machine learning applied to a transportation and logistics problem.

### III. PROBLEM STATEMENT

This thesis develops a model to determine the cost impact of individual customers as well as to identify the significant factors of route-based transportation costs. The goal of the model is to fairly allocate route costs among customers on the route. The analysis requires historic route data consisting of several components. The first component is the cost per route, which includes mileage and labor costs. It is important to note at this stage that each route can be made up of many customers with different attributes on the route. This necessitates the second set of components for a route, the attributes of individual customers on a route. Examples of these customer characteristics include individual customer distance from the depot, projected customer duration from the depot, product amount, stop time at a customer, etc. A comprehensive explanation of these characteristics follows. Finally, a third component, route characteristics, is required and is derived from customer characteristics.

#### A. Overview of Customer Characteristics

**1. Customer Distance** For each customer on a route the distance from the starting location of the route, referred to as the depot, to the customer is used as a key characteristic of the customer. This distance is calculated as over-the-road distance. FIGURE 1 illustrates an example route that visits three customers. Their customer distance characteristics are denoted as  $D_1$ ,  $D_2$ , and  $D_3$ , respectively.

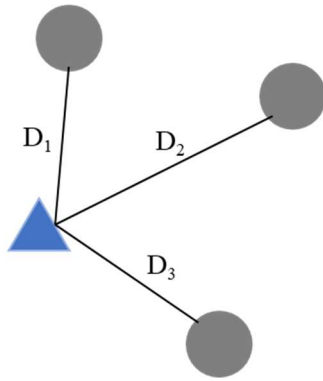


FIGURE 1 - Customer Distance Characteristic

**2. Customer Duration.** This feature is aligned with the Customer Distance characteristic. It is the projected duration from the depot location to the customer location. It provides additional information not provided by the Customer Distance characteristic, namely information about traffic and congestion. FIGURE 1 logic applies to this feature as well.

**3. Customer Product Amount.** For each customer on a route, this characteristic is equal to the quantity of product delivered to the customer. This can be measured in pounds, gallons, units, etc. depending on the business.

**4. Customer Stop Time.** This characteristic is equal to duration of time spent stopped at the customer and is comprised of unloading time and waiting time.

**5. Customer Proximity.** Proximity measures the closeness of each customer compared to the other customers on the route. The proximity measurement is found by first locating the centroid of all customers on the route and second calculating the distance from each customer to the centroid. In FIGURE 2, for each customer,  $C_i$ , on this example route the proximity metric is the distance from the customer to the centroid, i.e.,  $P_1$ ,  $P_2$ ,  $P_3$ , and  $P_4$  as noted in the figure.

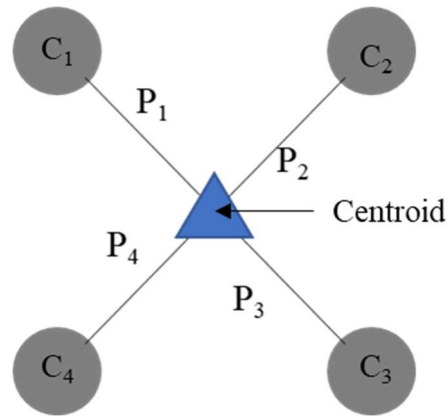


FIGURE 2 - Customer Proximity Characteristic

**6. Customer Time Window.** The customer time window characteristic is a measure of the flexibility of the delivery. For each customer, there is a time period during which the delivery has permission to be made. The customer time window metric is the total duration of this time window. A greater customer time window indicates that the customer is more flexible for the purpose of route planning and vice-versa.

**7. Customer Deliveries.** This characteristic is a count of the number of deliveries per customer on the route. Each customer will be visited once per route, but the number of orders/SKUs/etc. will vary. The customer deliveries measurement describes the variety of the products being delivered to the customer as it distinguishes between different deliveries on the same route.

The customer characteristics described above are commonly used characteristics in most cases, or “general route features”. For some cases, though, additional “case specific” characteristics may be introduced, and they will be discussed in Section V as applied to two industry cases.

## B. Overview of Route Characteristics

Route level characteristics, or features as they are called in machine learning, make up the third and final input required for the model. The route level features correspond to the customer level characteristic from all customers on the route. In other words, one route observation corresponds to one set of route characteristics. For customer distance, customer duration, customer product amount, customer stop time, customer proximity, and customer deliveries the corresponding route-level characteristics are the sum of customer-level characteristics over all customers on the route. For customer time window the corresponding route-level characteristic is the average of the customer-level characteristic. A detailed description of the feature design will be discussed in Section IV. At a high-level the direct connection between the customer characteristic and the route feature is vital for the machine learning prediction model and the subsequent allocation of route cost among customers.

### C. Machine Learning

After the customer and route characteristics are prepared, the route features are used as inputs in developing a machine learning model to determine the level of importance of each route feature in predicting the route cost. In this thesis, the methodology is to train and test a random forest machine learning model to predict the route cost from the route features. Unlike the conventional use of a predictive algorithm where the major output is the prediction, in this research, the intended output of the prediction algorithm is the feature importance of each route feature, a number between 0 and 1. This importance index will be used as the weight assigned to each route feature, called “feature weight,” all of which sum to 1. Subsequently, this feature weight will be used to proportionally allocate route cost to individual customers based on their customer-level characteristics. The feature

importance measurement calculates the decrease in node impurity and will be discussed in further detail in Section IV.

#### D. Cost Allocation Algorithm

The final step of the model is to allocate the route costs among customers on the route. Components required for this step are route cost, customer characteristics, and feature importance weights. The methodology applies a proportional allocation algorithm to these inputs to produce the customer cost per route where customer costs on a route sum to the total route cost. FIGURE 3 shows an overview of the proposed methodology.

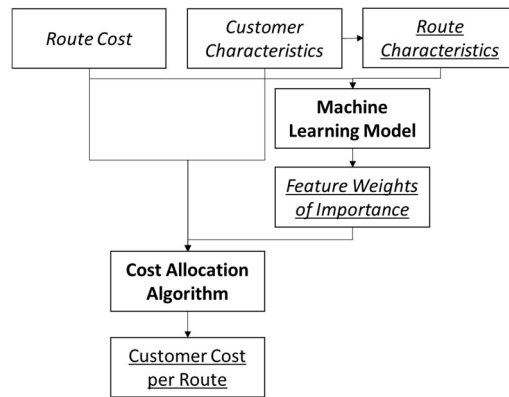


FIGURE 3 - Methodology Flow Chart

## IV. AN INTEGRATED PREDICTION AND COST ALLOCATION FRAMEWORK

### A. Preliminaries on Machine Learning

The field of machine learning (ML) that this thesis considers is supervised machine learning. Supervised machine learning techniques develop a model to map a set of features (X data set) to a corresponding Y variable. Within supervised learning the thesis is focused on regression algorithms to predict a continuous Y variable. The three supervised machine learning algorithms considered are linear regression, regression trees, and random forest regression. For the algorithms, it is important to split data into a test and train set of data. A standard split, used in this thesis, designates that a randomly selected 80% of data be contained in the train set and the remaining 20% be contained in the test set. The train data set is used to develop the prediction model and the test data set is used to test the model on a separate, non-biased data set. Additionally, this thesis takes advantage of the output of feature importance in order to determine among multiple factors, related to the total route cost, (e.g., distance, shipment amount, stop time), which should be given higher weights than others.

#### 1. Regression Tree

The regression tree is similar to the more common decision tree classifier but predicts a continuous variable instead of a discrete variable. The regression tree is made up of nodes (leaves) and splits (branches) where the top node of the tree contains all train data. The algorithm progresses by making true/false splits on the feature variables. The split



decisions are made by minimizing the mean square error (MSE) among potential splits at node  $n$  defined as

$$MSE(n) = \frac{1}{N(n)} \sum_{i \in T_n} (y(i) - \hat{y}_n)^2,$$

where  $N(n)$  is the number of samples at node  $n$ ,  $T_n$  is the train data subset at node  $n$ ,  $y(i)$  is the actual value for observation  $i$  in  $T_n$ , and  $\hat{y}_n$  is the predicted value calculated from the mean of all observations at node  $n$ .

See FIGURE 4 for a simplified example of a regression tree predicting route cost using features such as distance, shipment amount, and stop time duration, among others. Based on this sample regression tree, if the distance is less than or equal to 84 miles and the shipment amount is less than 500 gallons the predicted route cost is \$230. In contrast, if travel distance is greater than 84 miles and stop time is greater than or equal to 150 minutes, the predicted route cost is \$430. On the other hand, if the distance is less than or equal to 84 and if the shipment amount is less than 500, then the predicted route cost is \$230. This is intuitive as routes with longer travel distances and longer stop time durations will incur a greater cost. A regression tree in practice is much more complex, but the data-driven results may still be intuitive.

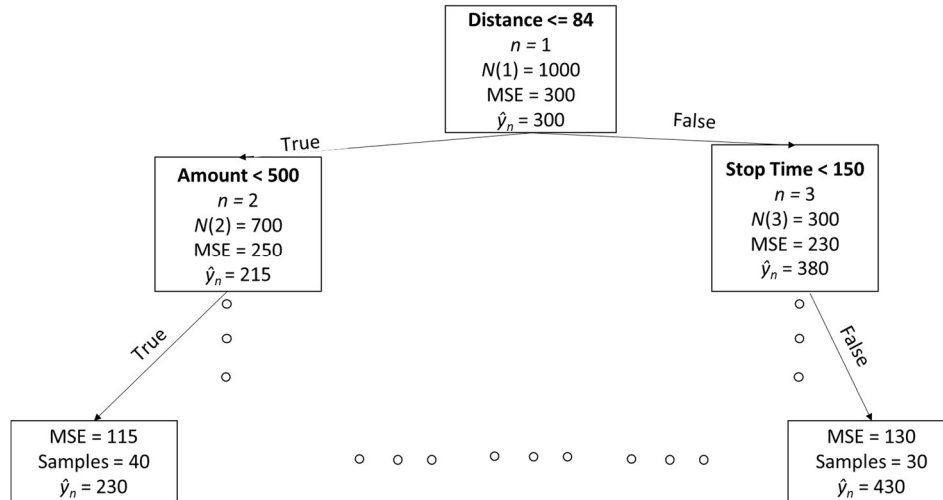


FIGURE 4 - Example Regression Tree

The regression tree makes splits based on the above-mentioned MSE, but additional hyperparameters play a role in the splits as well. The first hyperparameter available for tuning is the maximum depth parameter. This parameter indicates the maximum number of levels that the tree can traverse from the starting level. Tuning of this hyperparameter is important to prevent overfitting because as the allowable depth increases past a point, the performance of the model will generally decrease. This is because the model can specifically describe the train input data; but when tested with unseen data, the model will perform poorly. The minimum-samples-for-splitting parameter is another hyperparameter often considered. This parameter constrains the number of samples that must be present to split at an internal, or intermediate node. This parameter can lead to underfitting if the minimum samples of splitting is set too high because the model may be provided with limited information. Minimum samples at leaf is a third hyperparameter. This hyperparameter sets the minimum number of samples allowed at a leaf node, or a node at the bottom of the tree. This hyperparameter behaves like the previous parameter.

## 2. Random Forest Regression

The random forest regression algorithm is an ensemble machine learning algorithm extended from the regression tree algorithm. The random forest model is a bagging technique that ensembles multiple regression trees in parallel. The prediction result of the random forest model is the mean prediction of the trees contained within it. A few benefits of this aggregate method are increased prediction accuracy, ability to process many input variables, and capability to handle large data sets. According to a fundamental study on random forest models, this algorithm is “highly accurate”, “robust to outliers”, and “gives useful internal estimates of error, strength, correlation, and variable importance” (Breiman, 2001). In a comparative study of the random forest and decision tree algorithms applied to multiple data sets (Ali, et al., 2012), researchers determine that the random forest model has a higher level of performance than the decision tree model for large data sets. They find that the random forest model is a significantly more precise prediction tool.

There are many methods for developing the individual regression trees within the random forest. The tree development in random forest models depends on the concepts of bootstrapping/bagging and the random subspace method (Xu, 2013). The random forest algorithm creates bootstrapped samples where subsets of data are sampled from the train data with replacement. Multiple regression trees can then be developed from the bootstrapped samples. In addition to bootstrapping, the random subspace method is used by selecting a random subset of features to use to develop each tree. Next the many regression trees must be ensembled to generate the prediction for each observation. The most common method to reconcile the results is to take the average prediction across all

trees. Additionally, more complex ensemble techniques are utilized in practice. Some literature address methods for bias correction (Xu, 2013). Other techniques ensemble trees based on seasonality and/or tree performance (Booth et al, 2014).

## B. Random Forest Prediction Model

Based on the preliminary assessment of various machine learning methods like the regression tree and the random forest, this thesis continues with the random forest model as the machine learning component of the framework. The goal of the machine learning problem in this research is to determine the level of importance of various route features in predicting route cost. Therefore, route-level features must be developed, calculated, and selected as inputs to the model.

### 1. Feature Development

The next step in the prediction framework is the development of features calculated at the route level. The route features are derived from the customer characteristics discussed in Section III. There are two methods for calculating the route level feature: by summation

$$R_{i,r} = \sum_c C_{i,c,r},$$

or by average

$$R_{i,r} = \frac{\sum_c C_{i,c,r}}{N_r},$$

where  $R_{i,r}$  is the route feature  $i$  for route  $r$ ,  $C_{i,c,r}$  is the customer characteristic  $i$  for customer  $c$  on route  $r$ , and  $N_r$  is the number of customers visited on route  $r$ .

The route level characteristics are designed as functions of their corresponding customer characteristics. TABLE I specifies the customer characteristics, corresponding route features, and the route features' function formulation. All features are derived by summation except for Average of Customer Time Windows. The function is chosen based on the purpose of the route feature. Average of Customer Time Windows is an exception to the summation rule because the route feature is a measure of flexibility and greater values indicate a more flexible route. A summation of customer time windows would skew the feature result and incorrectly indicate that the route with more customers is more flexible than a route with few customers. For example, consider the sample routes in TABLE II. In this example, the average case is more representative of the measurement since both routes should result in equal flexibility.

TABLE I

TABLE OF CHARACTERISTICS AND FUNCTION OF DERIVATION

| <b>Customer Characteristic (i)</b> | <b>Route Feature (i)</b>         | <b>Function</b> |
|------------------------------------|----------------------------------|-----------------|
| Customer Distance                  | Sum of Customer Distances        | Summation       |
| Customer Duration                  | Sum of Customer Durations        | Summation       |
| Customer Product Amount            | Total Product Amount             | Summation       |
| Customer Stop Time                 | Total Customer Stop Time         | Summation       |
| Customer Proximity                 | Sum of Customer Proximities      | Summation       |
| Customer Time Window               | Average of Customer Time Windows | Average         |
| Customer Deliveries                | Total Deliveries                 | Summation       |

TABLE II  
TIME WINDOW FUNCTION EXAMPLE

|                                   | <b>Route A</b> | <b>Route B</b> |
|-----------------------------------|----------------|----------------|
| Customer 1 Time Window            | 10 mins        | 10 mins        |
| Customer 2 Time Window            | 10 mins        | 10 mins        |
| Customer 3 Time Window            | 10 mins        | 10 mins        |
| Customer 4 Time Window            | -              | 10 mins        |
| Route Feature with Summation      | 30             | 40             |
| <b>Route Feature with Average</b> | <b>10</b>      | <b>10</b>      |

The two input components for the machine learning model are the route cost and the route characteristics. These two input components are used to fit models to the three machine learning models considered in this research: linear regression, regression trees, and random forest regression. From the models, the feature importance of each feature can be found and is used for the next step in this framework, cost allocation. The detailed results of the machine learning modeling are discussed in the Section V.

## 2. Feature Selection

The last step before modeling is to select the features to be used as inputs for the random forest regression model. In this thesis, a correlation analysis is performed on the features to identify overlapping features in order to eliminate redundant variables. Feature selection is especially important when modeling with simple machine learning models. Redundant features in more complex models, like random forest, are less likely to interfere with the results of the model and do not necessarily need to be eliminated. Nonetheless,

this framework will identify and eliminate redundant variables using a correlation test in Section V.

### 3. Modeling and Feature Importance

The model is ensembled using the out-of-the box random forest regression model from Python's scikit-learn ensemble package. The regression model takes the feature and predictor data from 80% of a dataset to develop the model. The random forest model in the current thesis only tunes the hyper-parameter of tree depth to provide high quality test results. In preliminary studies, it was found that tuning other two hyper-parameters did not yield significantly different results thus was dropped in the main study. The output required from this model is the level of importance of each input feature between 1 and 0. The feature importance values are calculated using the pre-existing feature importance functionality in the scikit-learn package. The complex calculations behind this functionality combine the impurity measures at the nodes and the probability of reaching nodes, where train observations have a higher probability of reaching nodes earlier in a tree before many splits have been made. This generally means that more important features will be used to split earlier in the tree (the top of the tree).

#### C. Cost Allocation Algorithm

Next, this framework distributes route cost among customers using the ML results. The formulation utilizes a proportional allocation technique where customers are assigned a weighted cost based on their individual contribution to each feature, each feature's contribution to total route cost, and each feature's correlation to total route cost. Customer cost on a route is calculated using the following:

$$C_{c,r} = X_r * [(\sum_i \frac{F_{i,c,r}}{\sum_c F_{i,c,r}} * y_i * \frac{W_i}{\sum_i W_i}) + (\sum_i \frac{N_{i,c,r}}{\sum_c N_{i,c,r}} * (1 - y_i) * \frac{W_i}{\sum_i W_i})] \quad \forall r, c \in r \quad (1)$$

and

$$N_{i,c,r} = 1 - \frac{F_{i,c,r}}{\sum_c F_{i,c,r}} \quad \forall r, c \in r, \quad (2)$$

where  $C_{c,r}$  is the cost allocated to customer  $c$  on route  $r$ . There is a set of significant features,  $i \in I$ , where each feature has a weight,  $W_i$ , between 0 and 1. It is important to note that some features used to fit the model could be excluded from the allocation formulation due to insignificant importance. Therefore, the sum of  $W_i$  may not equal 1, so a weighted average of the importance weights is incorporated into the equation so that the entirety of the route cost is allocated. For each feature and customer on a route there is a customer characteristic value,  $F_{i,c,r}$  where customer  $c$  is visited on route  $r$ .  $N_{i,c,r}$ , calculated with Equation 2, is the helper variable for feature  $i$  when the correlation of feature  $i$  to route cost is negative. This variable essentially reverses the impact of the feature importance so that a feature with negative correlation to cost will impact the allocation fairly. The proportion of route cost is multiplied by the route cost,  $X_r$ , to determine the customer cost contribution per route. Additionally, the indicator variable  $y_i$  denotes the correlation of feature  $i$  with route cost where 1 indicates a positive correlation and 0 indicates a negative correlation. This ensures that cost is allocated to customers dependent on both the magnitude of feature importance ( $W_i$ ) but also direction of feature importance ( $y_i$ ).

#### D. An Integrated Decision Support System

The framework results in the calculation of a cost for each customer on a route. With this calculation, the opportunities for a decision support system are countless. A few



beneficial decision support concepts that can utilize this cost-to-serve result are showing data visualizations, predicting new customer costs to develop service agreements, and isolating opportunities for operational improvements.

The visualization of the cost-to-serve results can provide an overview of the cost of the transportation network, especially individual customers to upper management. Data aggregation techniques can highlight costly geographical regions and can provide insight about decisions such as network expansion or relocation. In addition, historical customer cost data developed by this framework can be used to predict the cost of new customers based on their characteristics. This can be performed by simply comparing the cost of customers with similar features to the potential customer or can be taken a step further with more advanced prediction methods. In this case, the results of this allocation model would be the predictor for tuning customer agreements.

To continue, further data manipulations can provide insights into opportunities for operational improvements. For example, one analysis to perform is a customer grouping exercise to determine customers that have significantly higher costs than customers with whom they are grouped. In this thesis, the customer grouping is performed using the K-nearest neighbors implemented with the Python Scikit-Learn package. The results of the decision support system analyses applied to two cases is discussed in more detail in Section V.

## V. COMPUTATIONAL RESULTS

So far, this thesis has described in detail the context of the cost allocation problem and the framework for the solution. This computational results section will describe the model as applied to three simulated cases, show results of the framework on these cases, and provide insights into a few potential managerial uses of the cost allocation framework. Additionally, it will describe two industry cases, Company A and B, and will overview the application of the framework to the cases. Due to confidentiality restrictions, the results of these industry cases will not be discussed.

### A. Simulated Cases

To analyze the performance of this model, data is simulated for three cases: Baseline Case, Alternative 1, and Alternative 2. The Baseline Case is the case from which the other cases are systematically modified and eventually compared. Alternative 1 and Alternative 2 vary from Baseline based on the route cost calculation as described in TABLE III. Modifications are made to the route cost to serve as an experimental variable. The remainder of the data simulation acts as a control; all other parameters remain the same. The route cost variability allows the experimentation to check if the results of the machine learning model respond to changes as expected. The Baseline route cost is composed of hourly and mileage-based costs. Alternative 1 adds to the Baseline route cost by including the measure of customer proximity multiplied by a scaling factor of 10%. This serves to penalize routes that contain more remote customers. Similarly, Alternative 2 adds a cost based on the amount of product. First, a tiered-pricing structure per pound is created where larger loads are assigned a lower rate per gallon. Next, the

product amount cost component is multiplied by a scaling factor of 1/6000, a smaller factor to allow for comparable values among the measurements since gallon values are larger than distance and duration values. Therefore, routes that deliver more product will be more costly but at a lower rate per gallon for Alternative 2. Therefore, it is hypothesized that the Sum of Customer Proximities route feature will be an important feature in Alternative 1 and the Total Product Amount route feature will be an important feature in Alternative 2. TABLE III also gives an overview of the size of the data sets for all cases.

TABLE III  
DESCRIPTION OF SIMULATED CASES

| <b>Simulated Case</b> | <b>Route Cost Calculation</b>  | <b>Features Related to Route Cost</b>   | <b># of Unique Customers</b> | <b># of Routes</b> |
|-----------------------|--|---|------------------------------|--------------------|
| Baseline              | Route Distance • Cost per Distance + Route Duration • Cost per Duration  | Sum of Customer Distances, Sum of Customer Durations                              | 500                          | 14,021             |
| Alternative 1         | Route Distance • Cost per Distance + Route Duration • Cost per Duration + Sum of Customer Proximities • .1             | Sum of Customer Distances, Sum of Customer Durations, Sum of Customer Proximities | 500                          | 13,390             |
| Alternative 2         | Route Distance • Cost per Distance + Route Duration • Cost per Duration + Total Product Amount • Route Rate per Gallon | Sum of Customer Distances, Sum of Customer Durations, Total Product Amount        | 500                          | 14,413             |

## 1. General and Case Specific Route Features

Before analyzing the route features, the first step is to specify any additional features included in the analysis. The features introduced in Section III are the “general route features” which are standard for the prediction and the allocation. There are, however; additional “case specific route features” that can contribute to prediction power and to gaining an understanding of the network. The additional case-specific features can either be used to allocate costs (they differ per customer on the route) or cannot be used to allocate costs (the feature is the same across all customers on the route). In the simulated cases, all “case-specific features” are the same across customers on a route and do not have the potential to be used for cost allocation.

There are two additional features included in the simulated cases. The department feature specifies what department/depot is responsible for the route. The department feature in these cases are distinguished by the geographical regions that they serve. Trailer type indicates the type of trailer/truck used to deliver product on the route.

## 2. Feature Analysis

For the implementation of the proposed framework, an in-depth analysis of the route features is crucial. This includes eliminating redundant features and comparing the distribution of the selected features across both the test and the train data sets. The first step is to eliminate redundancies.

This is performed using a correlation matrix analysis where the feature-to-feature correlation is calculated. The resulting correlation matrices (limited to a selection of significant interactions) is displayed in TABLE IV for the Baseline Case. From the

correlation matrix, it is evident that the customer distance and customer travel duration features are redundant, and one should be removed. Therefore, this analysis continues without the travel duration feature. The stop time feature is closely correlated with both product amount and the number of deliveries, but it is decided that all three features will still be included to maintain a variety of features. A similar conclusion is drawn from the correlation analysis for the two alternative cases.

TABLE IV  
FEATURE-TO-FEATURE CORELATIONS FOR BASELINE

|                                | Route Deliveries | Route Trailer Capacity | Route Stop Time | Route Product Amount | Route Customer Distance | Route Proximity | Route Travel Duration | Route Time Window |
|--------------------------------|------------------|------------------------|-----------------|----------------------|-------------------------|-----------------|-----------------------|-------------------|
| Route Deliveries               | 1.00             | 0.79                   | 0.97            | 0.92                 | 0.81                    | 0.80            | 0.81                  | 0.11              |
| Route Trailer Capacity         |                  | 1.00                   | 0.84            | 0.85                 | 0.66                    | 0.64            | 0.66                  | -0.01             |
| Route Stop Time                |                  |                        | 1.00            | 0.98                 | 0.79                    | 0.79            | 0.79                  | 0.12              |
| Route Product Amount           |                  |                        |                 | 1.00                 | 0.75                    | 0.74            | 0.75                  | 0.12              |
| <b>Route Customer Distance</b> |                  |                        |                 |                      | 1.00                    | 0.94            | 1.00                  | 0.02              |
| Route Proximity                |                  |                        |                 |                      |                         | 1.00            | 0.94                  | 0.06              |
| Route Travel Duration          |                  |                        |                 |                      |                         |                 | 1.00                  | 0.02              |
| Route Time Window              |                  |                        |                 |                      |                         |                 |                       | 1.00              |

Next, feature analysis must compare the distributions of features across the test and train data. This is meant to ensure that the train and test data have similar distribution so that the model testing procedure is valid. The distribution analysis is performed using violin plots which graph the probability density of the feature values on the x axis vs. the feature values on the y-axis. The violin plot for a given feature can be compared visually across the test and train data sets. A visual analysis of the feature distributions show that the test/train split is sufficiently uniform across all pertinent features and the model can continue with the current data split. Major differences between all corresponding test/train plots highlight only differences in outliers. This type of difference can be disregarded because the robust nature of the random forest algorithm eliminates bias due to outliers. See FIGURE 6 to **Error! Reference source not found.** FIGURE 10 for violin plots for the Baseline. The alternative cases show comparable results, and it can be concluded that the distribution of test and train data sets are similarly distributed.

As an illustration, FIGURE 5 shows the violin plot for the Average of Customer Time Windows route feature. The violin plot shows a normally distributed route feature centered around 300 minutes, the most likely average stop time on a route for both the test and train data set.

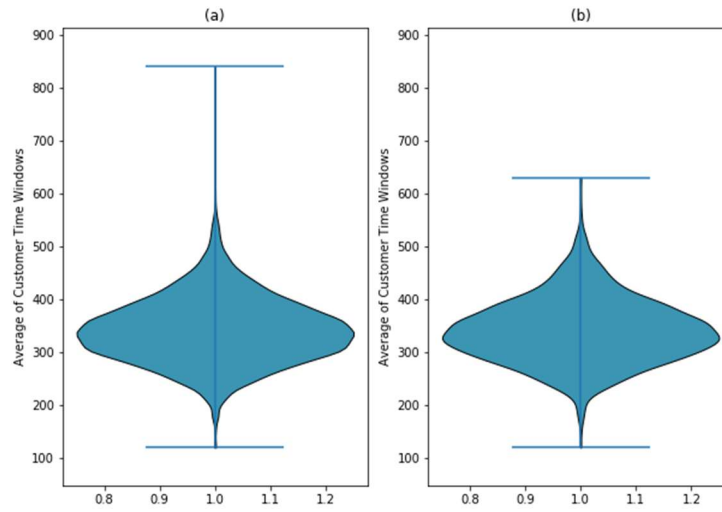


FIGURE 5 - Violin plot of the Average of Customer Time Windows route feature for Baseline for (a) train data and (b) test data

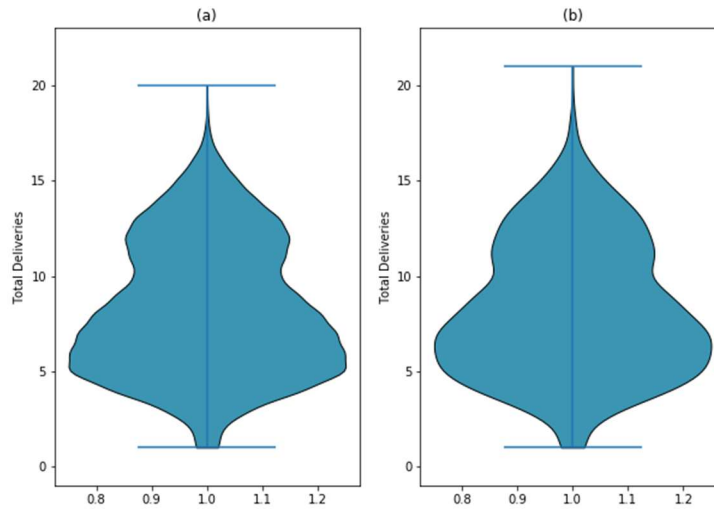


FIGURE 6 - Violin plot of the Total Deliveries route feature for Baseline for (a) train data and (b) test data

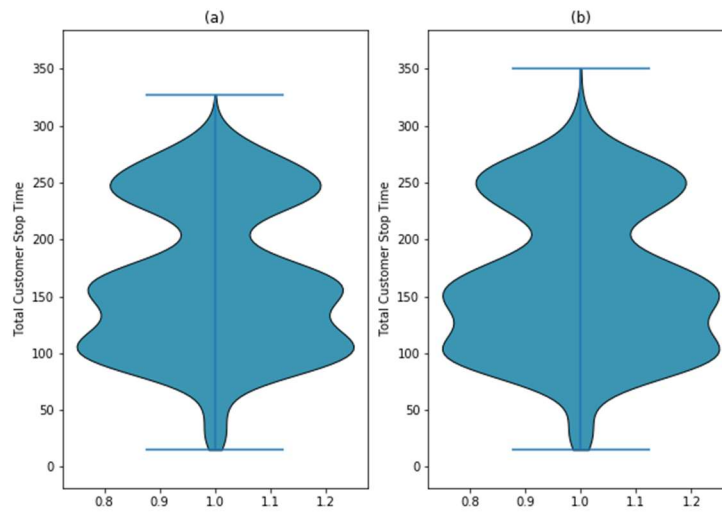


FIGURE 7 - Violin plot of the Total Customer Stop Time route feature for Baseline for (a) train data and (b) test data



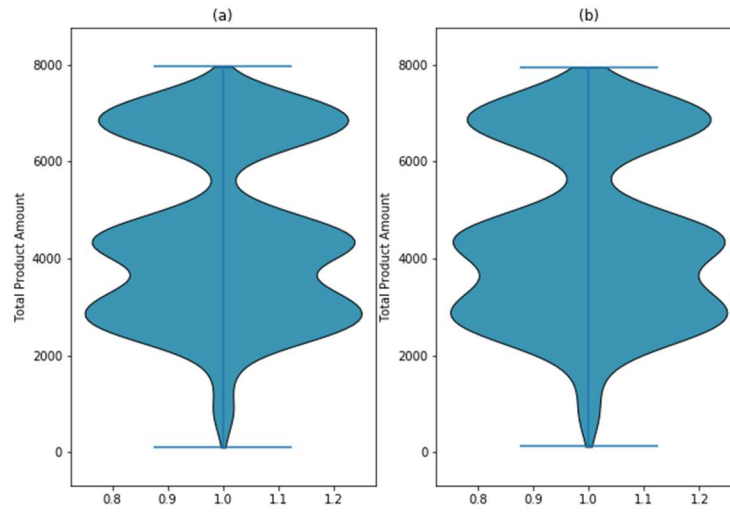


FIGURE 8 - Violin plot of the Total Product Amount (Gallons) route feature for Baseline for (a) train data and (b) test data

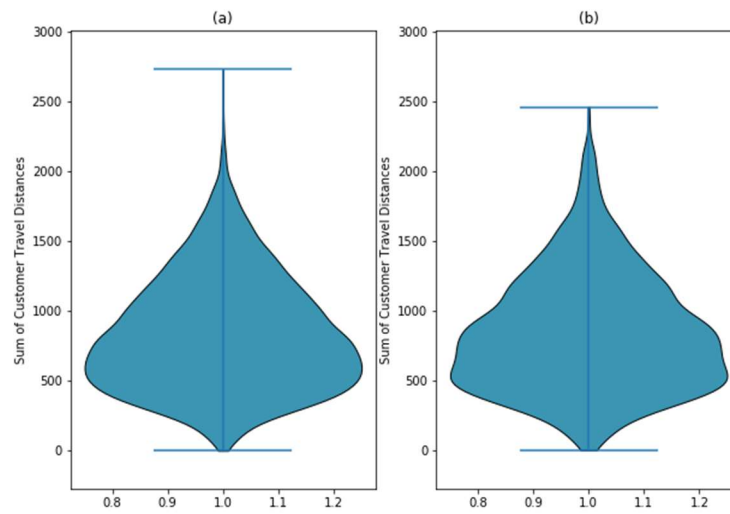


FIGURE 9 - Violin plot of the Sum of Customer Distances route feature for Baseline for (a) train data and (b) test data

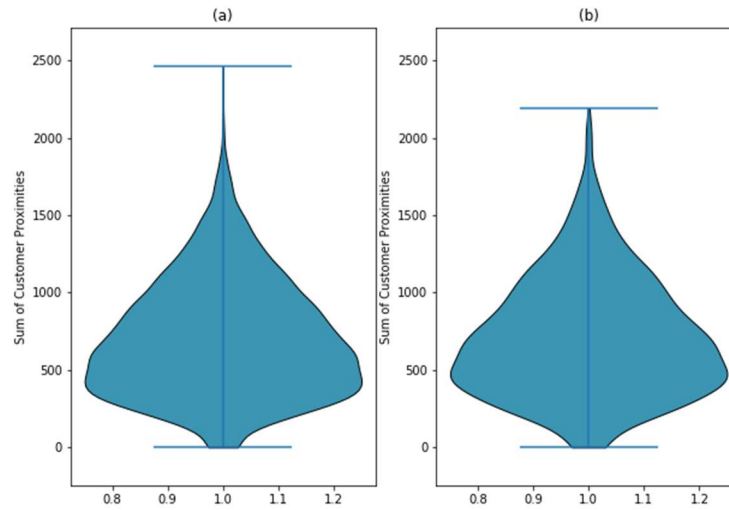


FIGURE 10 - Violin plot of the Sum of Customer Proximities route feature for Baseline for (a) train data and (b) test data

Finally, the correlation of each feature to the route cost must be considered. This is crucial to the cost allocation in order to ensure that customers are allocated cost fairly and dependent on how the route feature influences the route cost. FIGURE 11 shows the feature-to-route cost correlation for the baseline. The results of this correlation analysis show that most features are correlated positively with route cost. The only feature that is correlated in the negative direction is the Department - Nashville route feature. This indicates that routes originating from Nashville are generally cheaper than those from other departments. This conclusion does not affect the results of the allocation, however, because all customers on a single route originate from the same department. The correlation between route costs and various features (impacting the allocation) for Alternative 1 and Alternative 2 cases is similar to the baseline case. FIGURE 11 represents the  $y_i$  input values for Equation 1. Since all features used for the allocation are positive,  $y_i = 1 \forall i$ , for all three simulated cases.

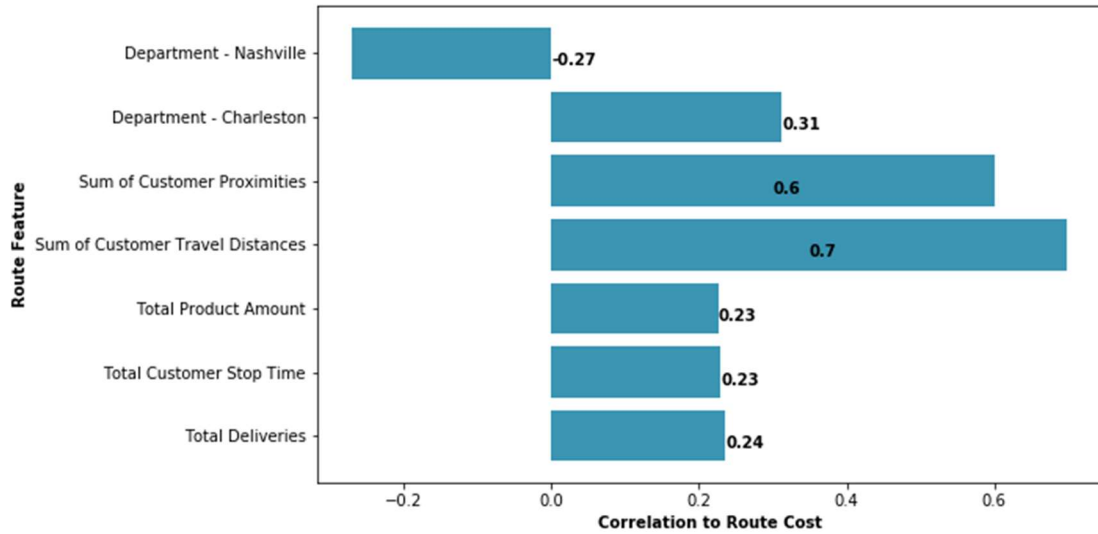


FIGURE 11 - Feature correlation to route cost for Baseline

### 3. Hyperparameter Tuning

Hyperparameter tuning was initially discussed in Section IV. This section will continue with the results of the hyperparameter tuning applied to the simulated cases. This thesis considers only the tree depth parameter applied to the random forest model. FIGURE 12, FIGURE 13, and FIGURE 14 show the effect of the tree depth setting on the outcome of the predictions on both the train data and the test data for the Baseline, Alternative 1, and Alternative 2 respectively. Generally, as the tree depth increases the model performance of the train data increases, but a point may be reached where the performance using the test data decreases seen in FIGURE 12 or reaches a limit as seen in FIGURE 13. Therefore, it can be concluded that the maximum tree depth for all simulated cases should be approximately 10 to avoid overfitting.

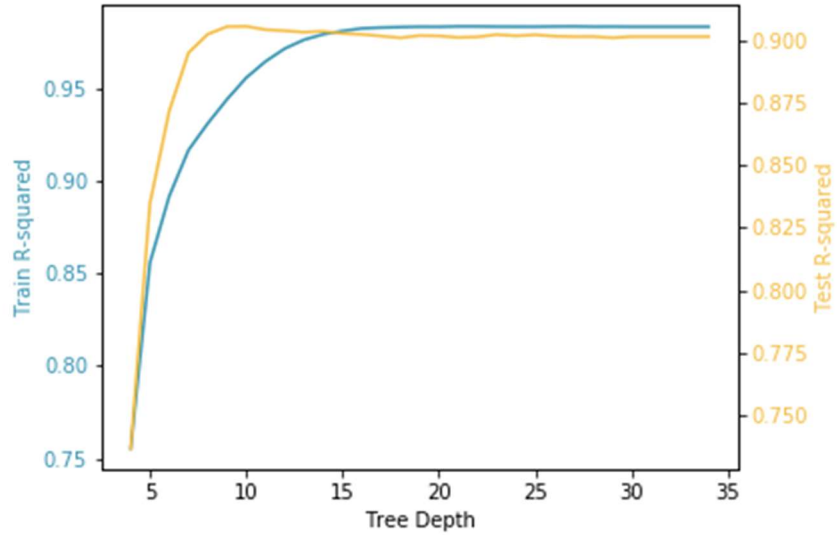


FIGURE 12 - Baseline Tree Depth Tuning

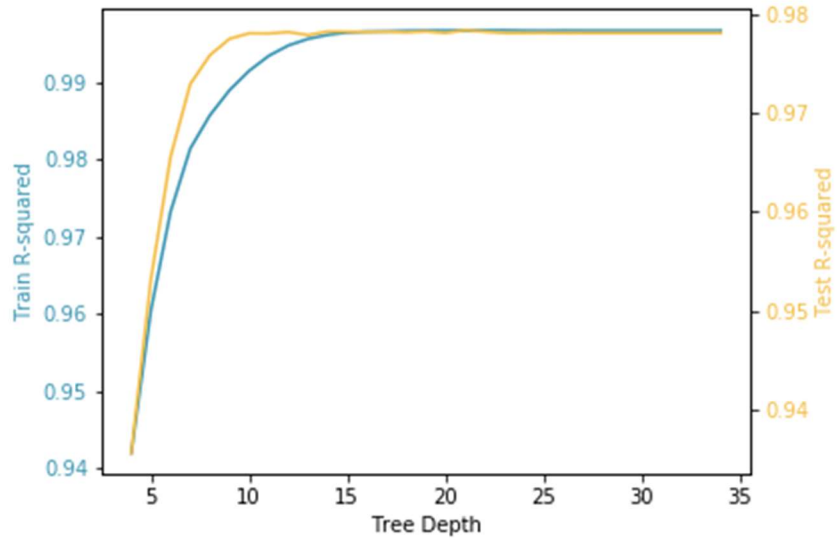


FIGURE 13 - Alternative 1 tree depth tuning

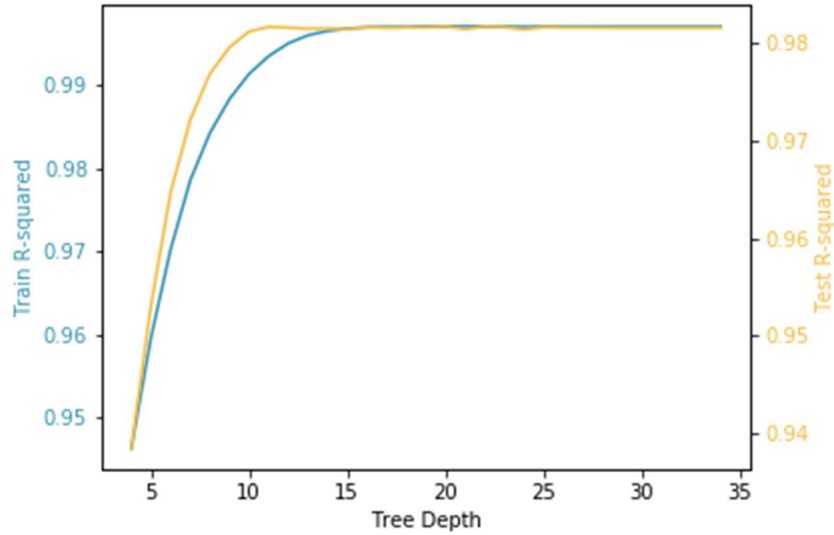


FIGURE 14 - Alternative 2 tree depth tuning

#### 4. Model Performance

The framework relies on the use of a prediction model, more specifically a machine learning model to determine the level of importance of each feature that contributes to the total route cost. The preliminary experimentation is done in choosing a suitable prediction technique between linear regression, decision tree regression and random forest regression, using data from three cases. The performance of the prediction is measured by the coefficient of determination of the prediction ( $R^2$ ). The calculation for the coefficient of determination is

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where  $i$  is an observation,  $y_i$  is the actual value of observation  $i$ ,  $\hat{y}_i$  is the predicted value of  $i$ , and  $\bar{y}$  is the mean of all  $y_i$ . FIGURE 15, FIGURE 16, and FIGURE 17 show the  $R^2$  results evaluated on both the test and train data sets for all three simulated cases. This evaluation places a higher importance on the results of the test score because of the removal of bias

between the model and data. As expected, the results show that the random forest regression model outperforms the linear regression and decision tree regression for all cases. It is interesting to note that the difference between the three models for the Alternative cases is small compared to the baseline. This can be explained because linearity is introduced by directly manipulating the route cost response variable.

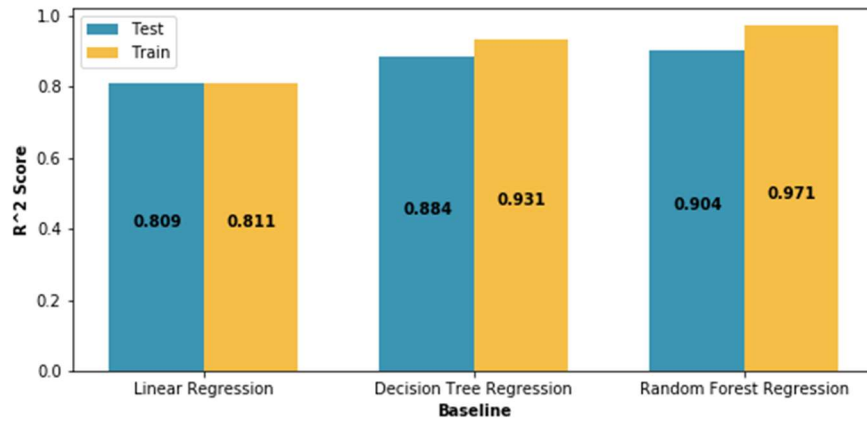


FIGURE 15 - Model performance for the Baseline case

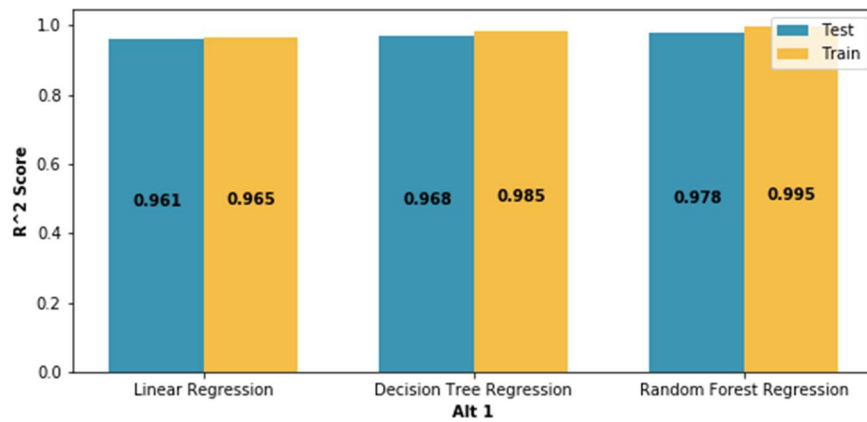


FIGURE 16 - Model performance for Alternative 1

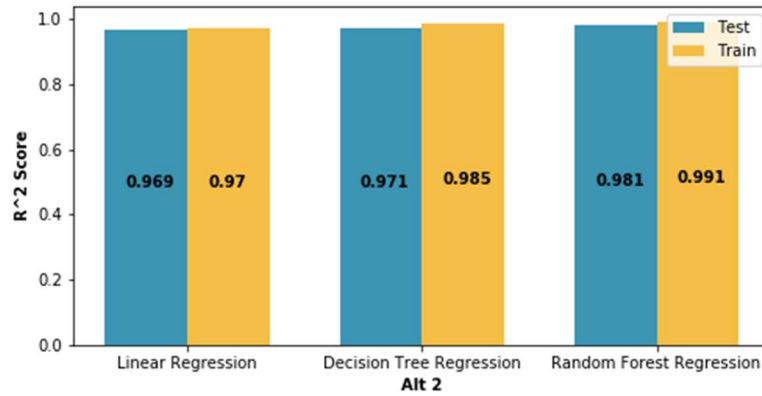


FIGURE 17 - Model performance for Alternative 2

## 5. Feature Importance

The evaluation of the levels of feature importance across the three simulated cases is the next step. The previous section confirmed the use of a random forest regression algorithm to determine these importance levels. The feature importance calculation, overviewed in Section IV, is provided in the SciKit learn package. The feature importance values for the Baseline, Alternative 1, and Alternative 2 cases are detailed in FIGURE 18, FIGURE 19, FIGURE 20, respectively. These figures show the level of importance for features in the model, which ranges between 0 and 1. If the level of importance is determined to be less than 1%, it is truncated for the purpose of this thesis because the impact on the cost allocation would be negligible. From these graphs, the feature importance values as part of the outputs of the random forest prediction model correctly portray the relationship between total route cost and respective features. In addition, as the definition of the total route cost is carefully varied/controlled in Alternative 1 and Alternative 2, compared to the baseline case, the resulting feature importance values have changed as expected. TABLE V shows the feature importance rankings of all features

across the three simulated cases. The italicized rankings are the features that are considered “important” and are used in the cost allocation.

The Baseline shows that the Sum of Customer Distances route feature is the most important followed by Total Deliveries, Sum of Customer Stop Time, and Sum of Customer Proximities. When comparing this to Alternative 1, the proximity feature becomes overwhelmingly important and displaces the distance and subsequent features. This is to be expected since proximity was included as a factor influencing route cost. The proximity feature becomes much more important because the scaling factor is exaggerated for the purposes of this demonstration. Recall that the proximity component of route cost makes up 10% of the cost and is directly related to route feature of the same name resulting in a more important feature.

When comparing the Baseline to Alternative 2, the outcome is also as hypothesized. The product amount feature becomes important for the first time and displaces the other features. The introduction of this new important feature is expected because the product amount is used to calculate route cost in this scenario. The feature importance for product amount overwhelms the customer distance feature because of the exaggeration of the scaling factor. Recall that for the product amount component of route cost the scaling factor is  $1/6000$ . The Sum of Customer Travel Distances and Total Deliveries features remain important, but the rest are now of negligible importance.



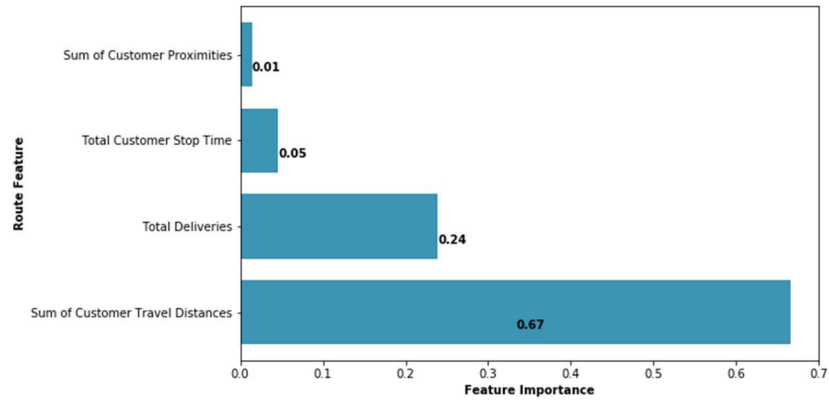


FIGURE 18 - Feature importance values for Baseline

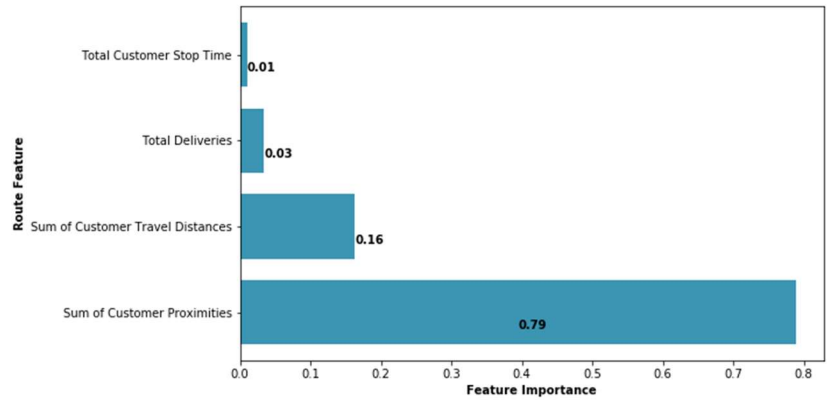


FIGURE 19 - Feature importance values for Alternative 1

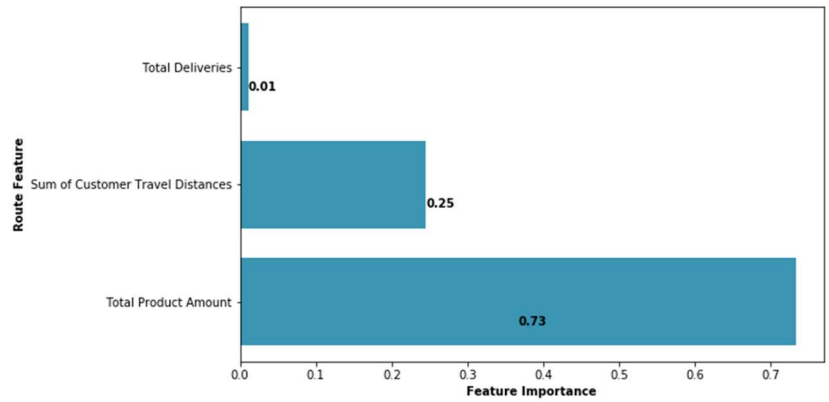


FIGURE 20 - Feature importance values for Alternative 2

TABLE V  
FEATURE IMPORTANCE RANKINGS

| <b>Baseline</b>                           | <b>Alternative 1</b>                       | <b>Alternative 2</b>                       |
|---|--|--|
| <i>1.Sum of Customer Travel Distances</i> | <i>1.Sum of Customer Proximities</i>       | <i>1. Total Product Amount</i>             |
| <i>2.Total Deliveries</i>                 | <i>2. Sum of Customer Distances</i>        | <i>2. Sum of Customer Travel Distances</i> |
| <i>3.Total Customer Stop Time</i>         | <i>3. Total Deliveries</i>                 | <i>3. Total Deliveries</i>                 |
| <i>4.Sum of Customer Proximities</i>      | <i>4. Total Customer Stop Time</i>         | <i>4. Total Customer Stop Time</i>         |
| <i>5.Total Product Amount</i>             | <i>5. Total Product Amount</i>             | <i>5. Sum of Customer Proximities</i>      |
| <i>6.Average of Customer Time Windows</i> | <i>6. Average of Customer Time Windows</i> | <i>6. Average of Customer Time Windows</i> |

B. Overview of Industry Cases

The methodology introduced in this thesis was applied at two industry cases. The first cost-to-serve industry case is applied to an oil and gas distribution company, Company A. Company A routes their vehicles from various starting depot locations to various customers and makes routing decisions using optimization techniques. This case considers three months of historic routing data to perform the cost allocation study using the framework presented in this thesis. During these three months, the case considers over 3000 routes delivering over 150 different products to 3000 customers. The second case considers a much larger distribution network for a national distributor, Company B. The second case considers only one month of historic routing data. During this time the distributor manages over 30,000 routes visiting over 30,000 customers. The two cases introduce additional “case-specific features” to be considered in the models (TABLE VI).

The application of the framework shows performances comparable to the performance of simulated cases (FIGURE 21). With the real data, the random forest regression model performs significantly better than the other models compared to the results of the simulated cases (Baseline, Alternative 1, and Alternative 2). This further confirms the use of the random forest model to determine the feature importance levels.

TABLE VI  
COMPANY A AND B SPECIFIC FEATURES

| Company A - Specific Features | Company B - Specific Features |
|-------------------------------|-------------------------------|
| Department                    | Department                    |
| Department type               | Route trailer capacity        |
| Route trailer capacity        |                               |
| Inventory management type     |                               |
| Trailer type                  |                               |
| Product type                  |                               |

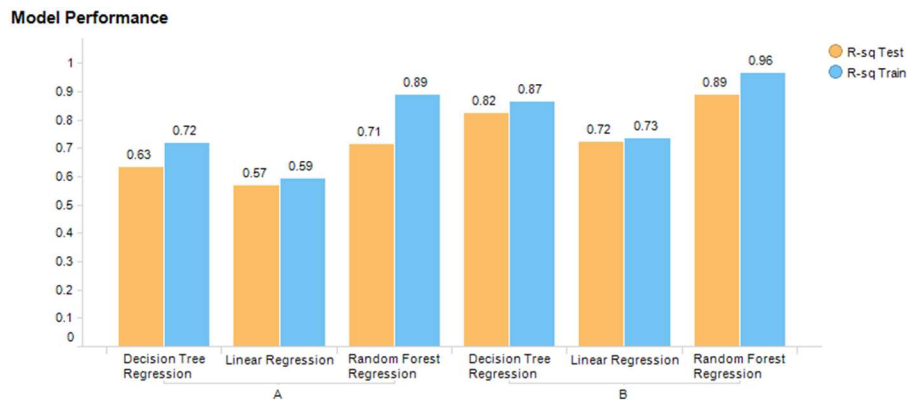


FIGURE 21 - Company A and B Model Performance

### C. A Decision Support System Using the Cost Allocation Framework

The Cost Allocation Framework results in a customer route cost using Equation 1. This output can provide value to managerial decision making through the development of decision support systems. DSS analyses can provide information on customer and operational performance. Three analyses using the Cost Allocation Framework are described in the following sections. The first analysis visualizes average customer costs geographically, the second groups customers to identify costly customers, and the third provides a method for predicting the cost of new customers. From these three examples, the DSS can be extended based on firm needs using the previously presented framework.

#### 1. Visualization of Cost Allocation

One potential analysis involves the visualization of average customer cost on a map for spatial comparison. FIGURE 22 shows the network for the simulated Baseline case where each circle represents a customer and each grey diamond represents a depot location. The customer circles are colored by the average cost/gallon of product delivered where green represents low cost and red represents high cost. FIGURE 23 shows the network focusing only on customers serviced from the Atlanta department. The customer cost visualizations provide management with indications of customer costs. It is easy to see that customers located far from the depot result in a more costly allocation. This can provide insight into potential depot additions or relocations and can help to identify costly customers.

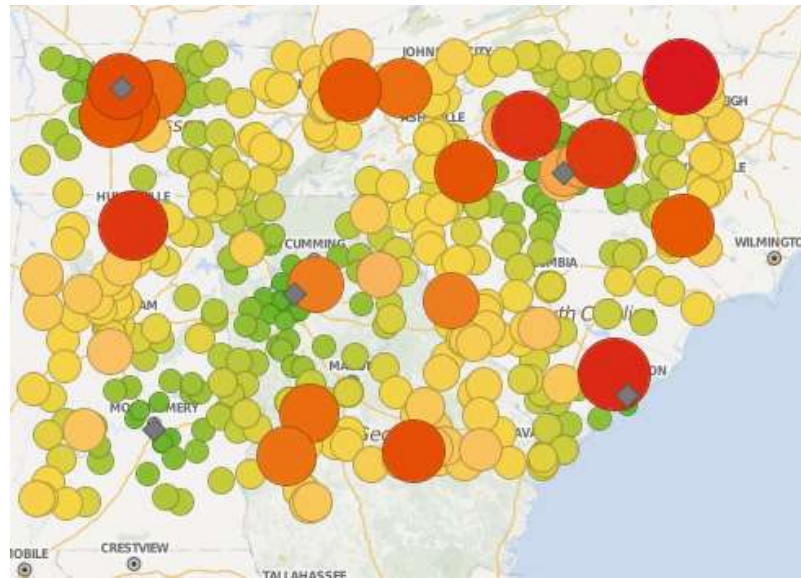


FIGURE 22 – Baseline case customer cost/gallon

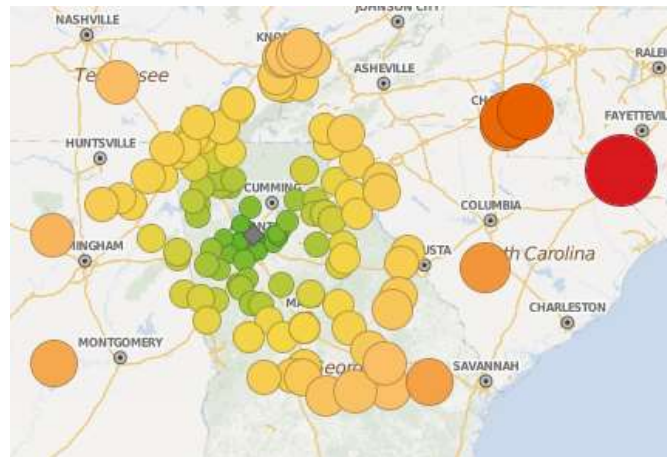


FIGURE 23 – Baseline case network focused on the Atlanta department colored by a customer’s average cost/gallon

## 2. Customer Grouping

Another potential DSS feature derived from the cost allocation results is the identification of costly customers that show improvement potential. This can be performed using a grouping methodology. The grouping methodology first separates all customers

into groups based on the departments they are served by. Next, for each customer in a department the method finds the  $n$  ( $n=14$  in this case) most similar customers using the  $K$ -nearest neighbors algorithm based on their demand and distance from depot, called the customer group. The  $K$ -nearest neighbors algorithm uses the Euclidean Distance in this methodology. Once the customer groups are determined, the methodology calculates the average cost/gallon for each group. For each customer, the difference between the average cost/gallon of the customer and the average cost/gallon of the customer group is calculated. This allows management to identify individual customers that are not performing as expected (they have a large difference between customer and group average cost/gallon). For this DSS feature, an example will be shown for Company A allocation results.

TABLE VII shows an example for Customer 1 and displays the group for Customer 1. Customer 1 is the most costly customer of its group. One observation for this costly customer is that the customer on average received less product per delivery. In fact, this customer received small delivery sizes which required more trips than would be expected. Additionally, a measure of customer utilization is provided as well. On average, deliveries to Customer 1 only utilize 26% of the available tank storage capacity at the customers. This is an indication to management that an increased delivery size could increase the delivery efficiency to this customer by decreasing the number of trips to the customer.

TABLE VII

CUSTOMER GROUPING EXAMPLE FOR CUSTOMER 1

| Customer Number | Standardized Average Cost/Gallon per Trip | Standardized Average Demand per Trip | Standardized Average Customer Distance per Trip | Standardized Average Delivery Amount per Trip | Average Delivery Amount/Tank Capacity | Number of Trips to Customer |
|-----------------|---|--------------------------------------|---|---|---------------------------------------|-----------------------------|
| 1               | 2.22                                      | 1.01                                 | 1.07  | -0.71   | 0.26                                  | 64                          |
| 2               | 1.51                                      | -0.42                                | 0.84  | -0.69   | 0.52                                  | 20                          |
| 3               | 1.37                                      | -0.86                                | 0.14  | -0.67   | 0.67                                  | 8                           |
| 4               | 0.60                                      | -0.46                                | -1.02   | -0.71   | 0.27                                  | 21                          |
| 5               | 0.32                                      | -0.81                                | -0.79   | -0.57   | 0.50                                  | 6                           |
| 6               | -0.08                                     | -0.81                                | 1.07  | -0.27   | 0.30                                  | 3                           |
| 7               | -0.33                                     | 1.12                                 | -0.09   | -0.47   | 1.34                                  | 26                          |
| 8               | -0.41                                     | -0.89                                | 1.07  | 0.58  | 0.61                                  | 1                           |
| 9               | -0.48                                     | -0.10                                | 0.84  | -0.49   | 0.64                                  | 13                          |
| 10              | -0.51                                     | -0.67                                | -1.25   | -0.52   | 0.29                                  | 7                           |
| 11              | -0.62                                     | 0.65                                 | -1.49   | 0.06  | 0.39                                  | 9                           |
| 12              | -0.79                                     | 0.11                                 | -0.79   | -0.33   | 0.44                                  | 11                          |
| 13              | -0.79                                     | 0.06                                 | -1.25   | 0.27  | 0.15                                  | 5                           |
| 14              | -1.01                                     | -0.59                                | 1.07  | 1.92  | 0.39                                  | 1                           |
| 15              | -1.01                                     | 2.69                                 | 0.60  | 2.60  | 0.58                                  | 5                           |

### 3. New Customer Pricing

The cost allocation data can also be used to predict costs for new customers. If customer features such as delivery quantities, distance from depot, and department are known, historic data can be queried to give a cost prediction for new customers entering an existing network. See FIGURE 24 for the prediction settings and FIGURE 25 for the results of the settings using data from the Baseline case..

**Cost Prediction Settings**

**Delivery Amount - Gallons**

900 1,000

Distance

150 200

**Department**

Type to search in list Q

(All) 5 values

Atlanta

Charleston

Charlotte

Montgomery

Nashville

FIGURE 24 - Prediction settings for a new customer



FIGURE 25 - Prediction results based on the prediction settings

This example shows what Baseline case can predict as the average cost/gallon, average cost/mile and average total cost for a new customer based on the historic data of existing customers and the prediction settings of the new customer. This information can be used to determine pricing structures and service agreements for new customers within the same network. For the example in FIGURE 24, if the average delivery size is increased to a range of 1,500 to 1,750 for a new customer, the Average Cost/Gallon decreases to \$.0105. Managers can use this functionality to make decisions about pricing for a new



customer, or to provide incentives to existing customers for adjusting their service agreements, by increasing delivery sizes for example.

## VI. CONCLUSION AND FUTURE RESEARCH

### A. Conclusion

This thesis attempts to answer the question: what do individual customers cost the business when costs are tracked at the route level? This understudied cost allocation problem is driven by the lack of customer visibility when a transportation network is designed optimally, and costs are shared among customers in a network. Improved data collection techniques and big data trends allow for granular visibility into customer cost utilizing this framework.

In literature, cost allocation methods have been widely studied. Focused in two streams, researchers mostly investigate methods related to Shapley allocation and proportional allocation methods. While Shapley methods are considered fairer than proportional methods, the computational and theoretical complexity discourage this thesis from utilizing Shapley methods in order to provide a framework for managerial oversight

The problem is addressed by (1) developing a high-level machine learning and cost allocation framework detailing the steps to derive a customer cost, (2) applying the framework to simulated and industry cases, and (3) providing a few example analyses, that can be applied to a DSS, utilizing the results of the allocation. Specifically, the framework presented in this thesis utilizes three inputs to generate the magnitude and direction of importance of various features in predicting route cost: (1) route-level cost, (2) route-level features, and (3) customer characteristics. The feature importance output is generated using the random forest algorithm which is shown to have the best performance when compared to linear regression and regression trees To continue, a cost allocation formula takes into

account (1) route cost, (2) feature importance output, and (3) customer characteristics to produce customer cost on a route.

Application of the framework to simulated cases and industry cases show similar results and support the validity of the model. Random forest regression shows the greatest prediction power and is used to generate feature importance levels. The feature importance rankings generated for the simulated cases confirm the original hypothesis that certain variables would become importance after the route cost is manipulated.

Lastly, the thesis provides examples of further analyses that can utilize the results of the cost allocation to aide management in decision making. The visualization of customer cost metrics, customer grouping and costly customer isolation, and new customer pricing are the three described analyses in this thesis. A case example provides an instance where the average delivery size to a customer should be increased to improve cost effectiveness. The opportunities for analyses using customer cost are numerous and should be further studied.

#### B. Future research

One potential improvement to this framework involves an advancement of the new customer pricing methodology. The cost allocation calculated values and customer characteristics could be utilized as machine learning model inputs to then develop a model to predict customer prices. This methodology would be more advanced than the historic data querying method. Regression models, decision tree regressors, and random forests could be studied.

## REFERENCES

- Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), 272.
- Agarwal, R., & Ergun, Ö. (2010). Network design and allocation mechanisms for carrier alliances in liner shipping. *Operations research*, 58(6), 1726-1742.
- Baroche, T., Pinson, P., Latimier, R. L. G., & Ahmed, H. B. (2019). Exogenous Cost Allocation in Peer-to-Peer Electricity Markets. *IEEE Transactions on Power Systems*.
- Booth, A., Gerding, E., & Mcgroarty, F. (2014). Automated trading with performance weighted random forests and seasonality. *Expert Systems with Applications*, 41(8), 3651-3661.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Crujssens, F., Borm, P., Fleuren, H., & Hamers, H. (2010). Supplier-initiated outsourcing: A methodology to exploit synergy in transportation. *European Journal of Operational Research*, 207(2), 763-774.
- Desrochers, M., Desrosiers, J., & Solomon, M. (1992). A new optimization algorithm for the vehicle routing problem with time windows. *Operations research*, 40(2), 342-354.
- Dror, M. (1990). Cost allocation: the traveling salesman, binpacking, and the knapsack. *Applied Mathematics and Computation*, 35(2), 191-207.
- Engevall, S., Göthe-Lundgren, M., & Värbrand, P. (1998). The traveling salesman game: An application of cost allocation in a gas and oil company. *Annals of Operations Research*, 82, 203-218.
- Fabri, A., & Recht, P. (2006). On dynamic pickup and delivery vehicle routing with several time windows and waiting times. *Transportation Research Part B: Methodological*, 40(4), 335-350.
- Fiestras-Janeiro, M. G., García-Jurado, I., Meca, A., & Mosquera, M. A. (2012). Cost allocation in inventory transportation systems. *Top*, 20(2), 397-410.
- Fishburn, P. C., & Pollak, H. O. (1983). Fixed-route cost allocation. *The American Mathematical Monthly*, 90(6), 366-378.
- Frisk, M., Göthe-Lundgren, M., Jörnsten, K., & Rönnqvist, M. (2010). Cost allocation in collaborative forest transportation. *European Journal of Operational Research*, 205(2), 448-458.

- Guajardo, M., & Rönnqvist, M. (2016). A review on cost allocation methods in collaborative transportation. *International transactions in operational research*, 23(3), 371-392.
- Henriet, D., & Moulin, H. (1996). Traffic-based cost allocation in a network. *The Rand Journal of Economics*, 332-345.
- Kirschstein, T., & Bierwirth, C. (2018). The selective Traveling Salesman Problem with emission allocation rules. *OR Spectrum*, 40(1), 97-124.
- Knoll, D., Prüglermeier, M., & Reinhart, G. (2016). Predicting Future Inbound Logistics Processes Using Machine Learning. *Procedia CIRP*, 52, 145-150.
- Krajewska, M. A., Kopfer, H., Laporte, G., Ropke, S., & Zaccour, G. (2008). Horizontal cooperation among freight carriers: request allocation and profit sharing. *Journal of the Operational Research Society*, 59(11), 1483-1491.
- Lin, Y. S., Zhang, Y., Lin, I. C., & Chang, C. J. (2018, March). Predicting logistics delivery demand with deep neural networks. In *2018 7th International Conference on Industrial Technology and Management (ICITM)* (pp. 294-297). IEEE.
- Ma, X., Yu, H., Wang, Y., & Wang, Y. (2015). Large-scale transportation network congestion evolution prediction using deep learning theory. *PloS one*, 10(3), e0119044.
- Nguyen, C., Dessouky, M., & Toriello, A. (2014). Consolidation strategies for the delivery of perishable products. *Transportation Research Part E: Logistics and Transportation Review*, 69, 108-121.
- Özener, O. Ö. (2014). Developing a collaborative planning framework for sustainable transportation. *Mathematical Problems in Engineering*, 2014.
- Özener, O. Ö., & Ergun, Ö. (2008). Allocating costs in a collaborative transportation procurement network. *Transportation Science*, 42(2), 146-165.
- Petrosjan, L., & Zaccour, G. (2003). Time-consistent Shapley value allocation of pollution cost reduction. *Journal of economic dynamics and control*, 27(3), 381-398.
- Psaraftis, H. N., Wen, M., & Kontovas, C. A. (2016). Dynamic vehicle routing problems: Three decades and counting. *Networks*, 67(1), 3-31.
- Samet, D., Tauman, Y., & Zang, I. (1984). An application of the Aumann-Shapley prices for cost allocation in transportation problems. *Mathematics of Operations Research*, 9(1), 25-42.

- Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation*, 42(1), 163-188.
- Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28), 307-317.
- Sun, L., Rangarajan, A., Karwan, M. H., & Pinto, J. M. (2015). Transportation cost allocation on a fixed route. *Computers & Industrial Engineering*, 83, 61-73.
- Toktay, L. B., & Wei, D. (2011). Cost allocation in manufacturing–remanufacturing operations. *Production and Operations Management*, 20(6), 841-847.
- Vanovermeire, C., Sörensen, K., Van Breedam, A., Vannieuwenhuysse, B., & Verstrepen, S. (2014a). Horizontal logistics collaboration: decreasing costs through flexibility and an adequate cost allocation strategy. *International Journal of Logistics Research and Applications*, 17(4), 339-355.
- Vanovermeire, C., & Sörensen, K. (2014b). Integration of the cost allocation in the optimization of collaborative bundling. *Transportation Research Part E: Logistics and Transportation Review*, 72, 125-143.
- Xu, R. (2013). Improvements to random forest methodology.
- Zakharov, V. V., & Shchegryaev, A. N. (2015). Stable cooperation in dynamic vehicle routing problems. *Automation and Remote Control*, 76(5), 935-943.