# **Optimizing Ship from Store in Omnichannel Retailers**

Jordan Nushart

Data Science, University of Wisconsin La-Crosse

DS 785 - Capstone

August 8, 2021

# **Abstract**

With the continuous increase in online purchases and shipping costs, omnichannel retailers are trying to find ways to decrease their shipping costs while still providing fast delivery. This study focuses on the recent shift towards ship from store as an option to utilize their brick-and-mortar locations to access more inventory, ship products from closer to the customer to reduce delivery costs, delivery time, and improve customer satisfaction, and loyalty. As Tractor Supply Company’s ship from store program has started and expanded in the last year, data has been collected that was used to train and test Random Forest machine learning models to predict which stores should be activated, as well as the expected savings per order for each category of items to limit the program to items that will decrease shipping costs the most. The outputs are a clear map that allows needs to be seen with ease along with a table with the estimated savings for each category. The findings show specific regions of the country that need stores to be activated into the program and that savings per order can double the current amount by filtering the categories available for ship from store. Insights from this study show omnichannel retailers that their fulfillment options can be optimized through machine learning to decrease costs and deliver products faster.

**Contents**

[**Optimizing Ship from Store in Omnichannel Retailers** 1](#_Toc79325900)

[**Abstract** 2](#_Toc79325901)

[**Introduction** 4](#_Toc79325902)

[**Background of the Problem** 4](#_Toc79325903)

[**Statement of the Problem** 5](#_Toc79325904)

[**Purpose of the Study** 6](#_Toc79325905)

[**Significance of the Study** 6](#_Toc79325906)

[**Methods of the Study** 8](#_Toc79325907)

[**Literary Review** 9](#_Toc79325908)

[**Shipping Fee Strategy** 9](#_Toc79325909)

[**Shipping Policy** 13](#_Toc79325910)

[**Ship from Store Strategy** 15](#_Toc79325911)

[**Data Collection & Transformation** 17](#_Toc79325912)

[**Methodology** 20](#_Toc79325913)

[**Results** 24](#_Toc79325914)

[**Store Model & Recommendations** 24](#_Toc79325915)

[**Item Model & Recommendations** 28](#_Toc79325916)

[**Summary** 31](#_Toc79325917)

[**Insights** 31](#_Toc79325918)

[**Impact** 32](#_Toc79325919)

[**Future Research** 32](#_Toc79325920)

[**Conclusions** 32](#_Toc79325921)

[**References** 34](#_Toc79325922)

[**Appendix A** 39](#_Toc79325923)

[**List of Tables** 39](#_Toc79325924)

[**List of Figures** 40](#_Toc79325925)

[**Appendix B** 41](#_Toc79325926)

[**Store Model Python Code** 41](#_Toc79325927)

[**Item Model Python Code** 44](#_Toc79325928)

# **Introduction**

## **Background of the Problem**

E-commerce sales are continuing to grow year over year and are continuing to take share from brick-and-mortar stores. eMarketer is expecting e-commerce sales to make up 15.5% of total retail sales, increased from 13.2%, which eMarketer had projected before the COVID-19 pandemic ([Droesch](https://www.emarketer.com/analysts/blake-droesch), 2021). eMarketer expects e-commerce sales to make up 19% of total retail sales by 2023 (Droesch, 2021), proving that shipping costs are not a short-term problem.

With e-commerce shares continuing to increase, retailers need to create loyalty with their customers to ensure a consistent base of sales before focusing on capturing new customers. Tractor Supply Company has an extremely strong base of loyal customers who are consistently purchasing animal feed, farm supplies, apparel, tools, etc. repeatedly and prefer to buy from a retailer in their community that they trust because many products are very specialized. Since a portion of the Tractor Supply Company’s products are used for farming, they are often larger, heavier, and more expensive on average compared to a company like Amazon or Bed Bath & Beyond.

Henry Ren (2021) from Bloomberg stated that extremely high shipping costs are being included in contracts for the next year as containers from China are staying near record highs for months at a time. Ren (2021) continues by expecting this increase to not be short-term since the carrier holds more power in negotiations going forward. With import costs increasing, it raises the total cost of shipping, which is often passed on to the customer. More retailers are moving to a shipping membership, such as Amazon Prime and Walmart+. However, this system is not viable for all retailers. To compete with these memberships, Tractor Supply Company needs to keep its shipping costs as low as possible.

## **Statement of the Problem**

With e-commerce continuing to grow, small parcel freight carriers are continuing to try and capitalize by raising costs every year to meet the demand. This creates a difficult balancing act between accepting the extra cost or passing it onto the customer. Bypassing the entire cost onto the customer can often cause the loss of loyal customers due to the retailing behemoths of Amazon and Chewy, among others, offering free shipping.

These circumstances force e-commerce retailers to constantly analyze their strategy and effectiveness to minimize their costs as much as possible. Retailers have tried countless strategies to offset this cost. Having the customer pick the order up at a store with buy online pick up in-store, which essentially translates to the customer’s store putting their items aside for them and is normally available if a product is available in a store’s inventory which has zero shipping charges besides restocking the item, or ship to store, where the customer picks the items up at the store and the items could be sent from a drop-ship vendor or a distribution center for the retailer but removed the residential surcharge that is included for any delivery to home from small parcel carriers. Other examples include optimizing box sizes, improving inventory by being proactive, or getting the product as close to the customer as possible.

## **Purpose of the Study**

By getting the product as close to the customer as possible, the retailer is going to pay less in shipping costs, the customer will get the product faster, the stores help offset the need for more distribution centers, among other benefits. One way that Tractor Supply Company has been trying to get the product closer to the customer has been their ship from store program. Ship from store uses the true omnichannel experience that Tractor Supply Company is trying to create. By leveraging the stores’ location, inventory, and staff, Tractor Supply Company can offset some of the cost and avoid passing the added cost on to the customer. The study aimed to improve the ship from store program by recommending stores to activate and limiting items to those that will allow for higher savings.

## **Significance of the Study**

Tractor Supply Company launched a ship from store program to a minimal number of stores in early July 2020 with a hard cap of 10 orders per store per day. Instantly this program was being maxed out by Tractor Supply Company’s order management system every day and enough data was collected to see how successful the program was. Even with around 10 stores active and no optimizations, the savings was over $1.00 per order. This has created a frenzy to try and improve the program and expand but optimizations are required to increase the savings as much as possible.

By correctly adding more stores into the program, TSC can allow 10 more orders per added store to receive the savings of the ship from store program, under the current restrictions per weekday. By adding 15 more stores, they can increase their weekly capacity by 750 orders. Under current savings, Tractor Supply could expect to see yearly savings of around $60,000 from those 15 added stores if the stores hit their capacity each day. However, the savings per order can be increased by filtering down the items available for the program, which can lead to doubling the savings for the new stores and the existing stores alike. If the estimated savings were to be doubled with the current allocation of stores, that would increase savings by over $250,000 yearly. The potential savings by this program is truly meaningful to Tractor Supply Company in their efforts to offset shipping costs and keep customers engaged by charging fair shipping costs, compared to other retailers.

Miller (2020) looked at how omnichannel retailers have been using ship from store during the COVID-19 pandemic. In that publication, Jeanette Barlow of IBM Sterling said that ship from store is now a necessity to minimize inventory concerns and keep customers trust that their orders can be fulfilled. If orders are not able to be fulfilled due to inventory concerns, the customer will likely look elsewhere for the same product or a replacement product. According to Corsten and Gruen (2004), 40% of customers do not delay their purchase or find a substitute at the same store when their product is out of stock. These 40% instead leave the store, 31% of them to buy the product at another store, while the rest do not purchase an item at all. With the increase in e-commerce shopping, it becomes much faster and easier for customers to find products elsewhere. If the problem is consistent, customer loyalty is likely to disappear and customers will find a different first-choice option, which could result in millions of dollars in lost sales every year for a billion-dollar retailer (Corsten & Gruen, 2004).

The COVID-19 pandemic, along with a U.S.-China trade war has caused suppliers to look elsewhere for their products, such as a “ ’China plus one’ strategy that will split production between China and another Southeast Asian country” (Shih, 2020). Chinese manufacturing and importing have decreased throughout the pandemic. Suppliers are looking towards other countries to import goods but adding new suppliers can take time and have a slow ramp-up time causing products to not be available in mass quantities. By being able to access inventory from stores, will decrease the chance that a product is not available, since it can pull from more locations inventory compared to just distribution centers. While the COVID-19 pandemic will not always be around, it has proven how important ship from store can be in attempts to keep inventory online and purchases flowing.

## **Methods of the Study**

In this paper, we will seek to explore cost avoidance and profit improvement opportunities by creating multiple machine learning models. The first model will use Random Forest classification, historical data, and demographic data based on zip codes to figure out what attributes of the current stores are making them successful. The inactive stores will then be fed through the model to recommend which inactivated stores should be activated.

The second model will analyze the SKUs that are available for the program. By using historical costs from UPS and item size data, we will be able to analyze which categories of items are most successful and choose which items will be available for the program. By pairing down the SKU list to only the most profitable SKUs, Tractor Supply Company can benefit the most from the orders before the capacity is hit and save the company and their customers as much money as possible.

# **Literary Review**

Shipping is a constant issue for online retailers and the pain points continue to grow as more retailers are going to fixed-rate or even free shipping. This adds pressure from the customer to do the same. Retailers such as Tractor Supply Company are consistently shipping heavy and large items, which makes it hard to provide free shipping. We first look at the impact of shipping prices and shipping procedures on customer satisfaction before moving onto a more direct ship from store analysis.

## **Shipping Fee Strategy**

According to Chen and Ngwe (2018), some retailers will try to recoup delivery charges by charging the full cost of shipping, a portion of the cost, giving free shipping if their cart is above a certain value, adding the cost into the price of the product, or just providing free shipping entirely. Tractor Supply currently uses a combination of all these methods to give a better customer experience on individual items or categories. One finding from Chen and Ngwe (2018), showed that “shoppers are less sensitive to shipping fees than they are to product prices, which echoes the results from mental account literature.” However, Chen and Ngwe (2018) continue by saying free shipping contingent on basket size is a successful motivator for increasing average order value and that free shipping memberships, like Amazon Prime, can increase demand but severely hurt profitability. This information will help Tractor Supply Company understand how to price their items and the shipping charges to make the ship from store, along with other shipping programs, as effective and profitable as possible.

Ma (2016) studied the differences in customer perception between fast and free shipping. Ma concluded, “Customers perceived ambiguity when they paid a higher shipping fee and the delivery time was long.” If customers paid a higher shipping fee, they expected their delivery time to be faster and were unsatisfied. These customers would be slightly more tolerant to slow shipping with low or free shipping. Ma also stated that “customers were less likely to make purchases if shipping was long, as they perceived more risk if the delivery was long.” If Tractor Supply Company continues to expand their Ship from store program, it will allow for lower shipping costs, and they can pass these savings on to their customers by lowering shipping charges. Lowering shipping costs will improve customer perception, especially when combined with decreased shipping times that will be inherently available when the product continues to get closer to the customer. The combination of these two factors can lead to improved customer satisfaction and a higher customer return rate.

In *Examining the Impact of Shipping Charge Fairness on Customer Satisfaction and Behavior,* Jones et al. (2019) concludes that customers are aware of the amount to ship products, and the amount they pay directly influences their satisfaction. Retailers increasing their shipping costs to increase profits will struggle with customer satisfaction. However, allowing customer visibility to see how segments of shipping costs add up to their total cost can improve satisfaction for higher costs (Jones et al., 2021). Adding visibility to show customers that their shipping cost is higher because it is going to a state like California, which has higher restrictions and costs for shipping, could provide more clarity to customers. By providing clarity to customers, they can better understand that the retailers are not over-charging them but instead just trying to recoup costs. This is in direct correlation with Ma in *Fast or Free Shipping Options in Online and Omni-Channel Retail? The Mediating Role of Uncertainty on Satisfaction and Purchase Intentions*. Here, Ma spoke about increased ambiguity when costs were higher, or shipping times were longer. By removing ambiguity on shipping price or delivery time, retailers can improve customer satisfaction and gain customers through word of mouth.

Koukova et al. (2012, as cited in Nguyen, 2016) looked at customers’ responses to flat-rate and cart value-based free shipping and found that customers’ understanding of shipping fees is centered around the order value. Customers were more likely to understand and accept higher shipping charges for higher value carts and show satisfaction. With customers more willing to accept shipping charges for higher-value purchases, the goal for retailers should be to push customers into making higher-value purchases. One way that this can be done is through shipping. While some retailers can implement value-based free shipping, such as free shipping over $50 cart value, another solution is creating a flat shipping rate based on the amount in the customers’ cart. A promotion such as $10 shipping for a $40 order but $7.50 shipping for a $60 order could push customers into converting on higher value carts. This would cost the retailer shipping revenue but could potentially be made up by the gross margin of the increased sales.

Customer satisfaction plays an important part in creating loyal customers who repurchase. Curtis (2011) combines the studies of Zeithaml et all. (1996, as cited in Curtis, 2011), concluding that highly satisfied customers were more likely to repurchase, and Yu and Dean (2001, as cited in Curtis, 2011), who concluded that it is cheaper to retain current customers than win over new customers. These findings can be combined to show retailers that taking a small loss in shipping can increase customers’ satisfaction, loyalty, and likelihood to repurchase. This small loss can cost less than it would be to capture new customers. Kamakura (2001, as cited in Curtis, 2011) fights this point by saying that customer satisfaction does not always lead to customer loyalty and repurchases, but instead is impacted by multiple factors despite an overall positive relationship. Winning new customers can often take sizeable sales or promotions to catch their attention. Tractor Supply Company has done this with their April 2021 changes to their loyalty program. They boast over 20 million customers in their loyalty program and are now providing discounts, free trailer rental, or free shipping based on their loyalty level (Warner, 2021). These benefits show that Tractor Supply Company is willing to take a small profit hit to keep their customers loyal and push customers to spend more money with Tractor Supply Company, compared to retailers like Amazon and Chewy.

Shipping fees are an extremely important factor in capturing and keeping e-commerce customers. With the largest retailers switching to free shipping based on membership or cart value, other retailers need to find a method that will lower their shipping costs so that they can continue to be profitable while minimizing orders lost to the e-commerce giants. Omnichannel retailers, such as Tractor Supply Company have been using their brick-and-mortar locations to minimize shipping costs, which allows them to charge their customers less for shipping or open themselves up for promotions for free or reduced shipping more often. Customers prefer free or reduced shipping and have higher satisfaction when they receive this benefit for orders. However, this is not the only factor that impacts customer satisfaction. Customers’ satisfaction drops when orders are not delivered when promised or expected delivery times are too long.

## **Shipping Policy**

Chang et al.’s (2021) research builds nicely from Ma (2106) and Chen and Ngwe’s (2018) by looking at the importance of free shipping compared to pricing and return policy. Chang et al. (2021) concluded that free shipping was more enticing than lower pricing, countering Chen and Ngwe’s (2018) research. Chang et al. (2021) also added that free returns were also less important than free shipping. The journal also concluded that “low priced retailers should adopt free shipping. High priced retailers should use calculated shipping and focus customers attention onto product quality and brand reputation.” Low-priced items are normally smaller in size, which would lead to shipping costs being less and making the products more available for free shipping. Customers also tend to struggle when the shipping price is near the product price, continuing to add to the point of free or fixed rate shipping for small and low-cost items.

Guo et al. (2020) found that 42% of online shoppers in Asia increased their shopping cart value to qualify for free shipping. This pattern has caused e-commerce retailers to use free shipping, or free shipping based on cart value, to improve their site’s performance. Guo et al. (2020) mentions that while shipping fees can be used as a tool to convert customers, it can create a no-win scenario. Retailers can either charge shipping and recoup their shipping costs to maintain a balance or they can offer free shipping to increase sales but decrease profits. Amazon loses billions on shipping every year due to its Amazon Prime membership. In the third quarter of 2019, Amazon profited $2.1 billion, with Amazon Web Services profiting $3.6 billion and their advertising division profiting $3 billion (Statt, 2019). This shows that Amazon Prime makes up a large portion of Amazon losing over $4 billion in that same quarter. This loss will keep increasing as Amazon continues to move towards same-day or next-day shipping. Amazon is fine to lose money on shipping since they have web services and advertisement departments that are extremely profitable. In the meantime, they can put pressure on other retailers and attempt to claim a stranglehold on e-commerce.

According to Fisher (2019), increasing the speed at which customers receive their products benefits not only the online store but also the brick-and-mortar locations as well. Fisher concluded that making improvements to online delivery speed “is always statistically significant and increases over time and is the largest between 25-36 weeks after the treatment begins.” The journal also states that brick-and-mortar locations will see a slight increase in traffic for a few weeks before declining back to their previous levels. The ship from store program at Tractor Supply Company is designed to decrease shipping times but moving the product multiple zones, and likely days, closer to the customer. If Tractor Supply Company can commit to and follow through with three business day shipping for ship from store orders and it was available in enough regions, then customers would provide good word of mouth to capture new customers and online sales would increase with higher retailer loyalty. According to Fisher (2019), the brick-and-mortar locations would also see increases. This is essential for omnichannel retailers to avoid shipping charges anywhere possible while providing same-day fulfillment for their customers.

In *Understanding Customers’ Adoption of Express Delivery Service for Last-Mile Delivery in the U.K.*, Zhong et al. (2021) surveyed e-commerce shoppers to gauge how long they are willing to wait for an order to arrive. Zhong et al. (2021) found that promising products within 3 days were not statistically significant, but more effort should be focused on delivery reliability since customers are more frustrated at missed promises. Overpromising and under-delivering on delivery times can lead to customers not trusting the express delivery in the first place and could push them to seek out other retailers that can accurately predict delivery dates. In the current e-commerce world, products can often be purchased from multiple sources. Customers will often choose the fastest delivery if they need it by a certain date. If the retailer cannot fulfill their delivery promise, the customer will be very upset, especially if they paid for the express shipping. Tractor Supply Company can change their shipping estimates on their site during checkout to consider the product location and the customers' address to provide a more accurate estimate and improve customer satisfaction.

With seemingly endless amounts of retailers to order products from, customers can shop around to find the best price and fastest or cheapest shipping, depending on their needs. If customers need products rapidly, they are likely to pay more to get quick delivery. When delivery date promises are not met, especially when paying a premium, customers are very dissatisfied and are increasingly likely to look elsewhere for their product the next time they are about to make an order. By increasing the speed of delivery, both sides of the omnichannel business should see boosts.

## **Ship from Store Strategy**

Yang (2020) analyzed an omnichannel retailer to analyze how Ship from store affected the retailer’s profits. Yang concluded that while ship from store did not cause a significant decrease to store volume, it can leave the company worse off because they must replace the item at the store and the customer likely would have bought the product anyway. Yang also mentions that customers will no longer be going into the store to make that purchase and therefore, the store loses out on potential revenue from impulse buys. However, Yang is thinking about ship from store in a different way than Tractor Supply Company has implemented the program. Tractor Supply Company customers do not choose if they will receive the product from a store or a distribution center. Instead, a sourcing algorithm will determine which shipping location is closer and will use the closer shipping location if inventory or capacity has not been reached. Tractor Supply Company’s Ship from store program does not take away from store sales since the customer is just selecting ‘Standard Delivery’ and would not be entering the store regardless.

According to Li (2020), Yang’s thoughts that stores could be worse off is not always correct. Li brings up the idea of Ship from Store to Store. Ship from store allows stores to move overstocked products to other stores in attempts to move the struggling products more effectively, restock stores faster using nearby inventory, or shipping from one store to another so that a customer can pick the order up in-store. Moving products between stores for inventory purposes is a certain benefit to adapt quickly and move product to where it is selling best. Ship from store for customer pickup enjoys all the benefits of ship from store that have been mentioned previously. This also has an added benefit of potential cross-selling or impulse buys once the customer enters the store. This benefit is commonly known across the omnichannel space and is why so many retailers push for Buy Online Pickup in Store (BOPIS).

Bayram (2021) investigated the benefits of Ship from store program implementations and how more stores affected the profitability of the program. Bayram brings up previously mentioned benefits of ship from store such as “faster delivery, lower shipping costs, higher in-stock probability, increased sales, and customer satisfaction, etc.” With all these available benefits, it seems obvious that omnichannel retailers should be implementing ship from store. However, Bayram mentions that implementing it and integrating it effectively can be time-consuming and costly. The journal continues to say that once implemented, the program is more profitable with 200 stores than 20 stores. Being more profitable as stores increase makes sense as it will help inventory concerns and continue to move the product closer to the customer.

The best ways to improve customer perception and experience are to lower shipping times and costs. The Ship from store program is proven to be successful in both goals. As the program continues to expand, products will continue to get closer to the customer and make the benefits more noticeable. This will result in lower shipping times and costs, which can then be passed on to the customer to increase customer satisfaction, loyalty, and likelihood to repurchase.

# **Data Collection & Transformation**

Prior analysis on the ship from store has been completed throughout the implementation of the progress with a similar process for data collection and transformation but did not include any regression or classification models. Instead, it was based on a quick look and analysis based on grouping data. The following process will allow a much more in-depth and improved view aimed at making the Ship from store process as streamlined and optimized as possible by discovering what makes stores most successful in this program and which items cause the largest savings.

Interviews were held with the Omnichannel and Store Operations teams to understand the process for adding new stores to the program and how items are determined to be eligible for ship from store and if Tractor Supply can add a specific SKU to be eligible or if it is at a higher level, such as category or department. According to Lee Stewart (2021), Director of Omnichannel Operations, the best process is to ideally set the available items at the category level because the items in a category should be very similar and are often bought together.

Order line level, unique order number, and item, data were collected from Tractor Supply’s Order Management System (OMS) database, along with UPS invoice data, and the U.S. Census data. From Tractor Supply’s OMS, order line data from the start of the ship from store program, 7/7/2020 through 7/7/2021, along with fact tables for the item and store information to help with predictions.

From here, the OMS data was joined with the UPS invoice data based on the tracking number assigned to the order line and the shipment. Order lines without a matching tracking number were filtered out in the Alteryx workflow. Each order line was assigned a percentage of the total shipment weight. The percentage of the shipment weight was used to divide the total shipment cost across multiple order lines, for orders with multiple unique items. UPS provides the number of zones the order was shipped, which can be combined with the weight to be found on Tractor Supply’s UPS rate table to get a base cost for that item. This base cost does not include any of the accessorial fees or surcharges applied to the shipment. The assumption here is that the surcharges would be applied regardless of where it was shipped from. Next, an estimated zone for the order coming from a distribution center was calculated based on the nearest distribution center with available inventory. As an example, if a ship from store order went four zones to California, but the Arizona distribution center is only two zones away, we ignored the Arizona distribution center and instead compared the order to if the order came from Kentucky. This was done based on the sourcing logic for the ship from store program, which will find the closest shipment center and attempt to source the order to that location. If the closest location cannot fulfill the order, it continues down the line. In the prior example, it could be assumed that Arizona did not have stock of that item and could not fulfill the order, so it should not be compared. The estimated zone was then used to calculate an estimated base cost, the same way as the actual zone, before being compared to the actual base cost to create estimated savings, and estimated zones improved. This created a data set with 38,685 orders, as seen in Table 1, that could be the foundation for both models.

**Table 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Orders** | **Order Lines** | **Unique Items** | **Active Stores** |
| 38,685 | 50,931 | 6,502 | 97 |

*Note:* Summary table of information used in the dataset and models.

Once estimated savings were calculated, the data could be grouped by the item id and store in two separate datasets and joined with the appropriate fact tables, and the store dataset was joined with U.S. Census data to add demographics such as median household income, percent of the population in agricultural jobs, population, and population density. Sales percentage by zip code and miles to the nearest store were also added to assist in the classification of what makes a successful store. Stores with data that could not be accessed were removed. Once the data sets were complete, the data could be analyzed.

# **Methodology**

After the datasets were complete, the data could be brought into Python, where the analysis and modeling could begin. Random Forests were used for both models due to the ability to handle both classification and regression. Random Forests are flexible models that explicitly tell users what predictors are most important, provide high accuracy through bagging, and minimizes the ability to overfit the model.

Figure 1 shows which predictors were most important in the decision of the model. This allows for users to easily engage in supervised learning by removing low importance predictors. By retesting the model after removing variables, the final model importance can be seen in Figure 2, with median household income removed. Importance charts easily explain to stakeholders what drives the model and allows them to be able to use this information in other business decisions, such as where to place new brick-and-mortar locations.

**Figure 1**

Chart, bar chart

Description automatically generated

*Note:* First factor importance graphs to determine the store model.

**Figure 2**

Chart, bar chart

Description automatically generated

*Note:* Final factor importance graphs to determine the store model.

Random Forest models use bootstrap aggregation, or bagging, to build the dataset used to train the model. Bagging uses bootstrapping to create unique data samples by randomly selecting points and replacing the points after selection, which are then trained in parallel and independent before the classifier with the most predictions is taken to give a classifier value to a row of data (Breiman, 1996). Bootstrapping with replacement means that values can be selected multiple times, while some values can be ignored. For Random Forest Regressor models, the process is the same but instead of a classifier based on majority voting, an average of all outputs is taken to provide an estimate. Both the models in this research were given 10,000 decision trees. For the store model, if a store received 5,001 or more yes or no votes, that would be the output. The item model would be an average score from the 10,000 decision trees. The number of decision trees for each model can be changed, however adding more will likely not improve the accuracy anymore, while decreasing the amount could cause underfitting the data if enough decision trees are removed.

Logistic and Linear Regression were considered for the store and item models respectively. Logistic Regression would provide a similar product to the Random Forest Classification model. However, Random Forests handle categorical data better and provide more visual assistance for non-technical users to understand, such as an importance chart and a decision tree chart, according to Chitroda (2020). Logistic Regression handles larger numbers of predictor variables better than Random Forests (Chitroda, 2020), but we were interested in using a minimal number of predictors to keep the data from being over-fit or suffering from multi-collinearity. Since there are minimal variables and categorical variables, the Random Forest Classification model was more appropriate.

Random Forest Regressor was chosen over Linear Regression since categorical data was being used and other predictors did not have a strong linear relationship to the dependent variable. Linear Regression models were not able to get above a 0.67 coefficient of determination, or R², with Akaike Information Criterion, AIC, in the thousands. Since 33% of the variance in the dependent variable was not able to be predicted based on the independent variables and the AIC was well above an acceptable range (Stattrek, 2021), other options were tried. Random Forest Regressor was the next model to be tried since all the requirements were met, only one model type would need to be explained to the non-technical stakeholders.

K-Nearest Neighbors was seen as a viable solution for both models since it can be used for classification and regression. K-Nearest Neighbors models are used in recommendation systems and decision-marking models by comparing the test data to the most similar points in the training data (Schott, 2019). According to Schott (2019), k-Nearest Neighbor models struggle when data points are placed on a boundary between groupings and could be placed in either group. While these models would have been applicable and valid to use, it was believed that decision trees and importance graphs would be easier to understand the model process, given the target audience.

Beginning with the store classification model, a classifier field was created to determine if a store’s estimated savings per shipment were in the top 25% of stores. First, a random 80% of the dataset was selected as the training set and the remaining 20% was used to test against. The data was then put into a Random Forest Classifier model which would return an accuracy score, mean squared error (MSE), and the importance factor for each of the predictor variables that are used in the model. After the output, through supervised learning, the lowest impact predictors were then removed, and the Random Forest was used again to improve the accuracy of the model.

After the model was created and tested, then predictions for the inactive stores could be run against the Random Forest model to create recommendations of which stores should be made active soon. The output then will pass through to Tableau where a map will be created that will highlight the recommended stores to activate.

A Random Forest Regressor model was used with the predictors of item type, unit weight, zone improvement per order, largest side (in.), smallest side (in.), and dimensional weight. UPS uses zones instead of miles to calculate distance. UPS uses zones to group distances from the shipment point to help determine shipping cost and shipment time (“Understanding FedEx and UPS Shipping Zones”, 2020). This process will be repeated until low importance predictors are removed, and the model is as accurate as possible. Tractor Supply Company’s item table, filtered based on Ship from store eligibility, will then be fed into the model to create predictions. The predictions will then be grouped by category to recommend which categories should be active, as to not waste the limited number of shipments on low savings items.

# **Results**

## **Store Model & Recommendations**

The store Random Forest Classification model was first run with the predictors based on the zip code of the stores for the percentage of Tractor Supply Company’s shipped sales, miles to the nearest store, median household income, percentage of the population working in agriculture, total population, and population density with 10,000 decision trees in the random forest.

**Figure 3**

Timeline

Description automatically generated

*Note:* Example of 1 of the 10,000 decision trees in the Random Forest.

This Random Forest returned with an accuracy score of 0.8 and a mean squared error of 0.2. Median household income and miles to the nearest store were the features with the lowest importance on the accuracy of the model.

Other combinations of the predictors were used to check if removing the lowest importance predictors would improve the accuracy of the model. Removing only median household income had no change in accuracy for the random forest model. Using the model without median household income, the stores that are not currently active in the Ship from store program were fed into the model to create a prediction for each store that had all the predictors available. 124 out of the 1,879 inactive stores were predicted that they would be recommended to turn active. The output was then loaded into Tableau and graphed on a map, Figure 4, along with the current active stores. The output showed five areas of the United States where stores should be implemented into the program.

**Figure 4**

Map, scatter chart

Description automatically generated

*Note:* Map of active and recommended stores.

First, is the western section of Washington state, from Seattle to Eugene. This area only has 3 active stores, and the nearest distribution center is in Arizona. The next closest distribution center is in Kentucky, so this area makes sense to add more stores to since it is the farthest from any of the distribution centers and Tractor Supply Company would see a massive increase in costs if the Arizona distribution center was out of stock of the products being ordered.

Continuing down the west coast, California is a hotspot for the recommended stores for many of the same reasons as Washington. With a strong cluster of recommendations around the Sacramento area, stores should be activated in this area due to the high population, population density, and percentage of the population working in agriculture all being above the mean. There is a smaller cluster around Fresno, along with a few more sporadic recommendations heading south towards the Mexican border.

Working east, Texas is the next clear state where stores should be activated. From the Dallas-Fort Worth area down to Houston, there are 31 recommended stores. With 25% of the recommended stores being placed within half of a state, this is a great opportunity for improvement and cost avoidance. The zip codes in the eastern half of Texas are some of the highest spending customers, based on the percentage of sales to those zip codes. With so much demand in this area, heavy items being shipped (feed, cattle equipment, etc.), nearest active stores being over 120 miles away, and being equidistant between the Arizona and Kentucky distribution centers there is a large amount of opportunity in this area of the country.

Continuing, central Florida is the next region that has a cluster of recommendations, especially in the area between Orlando and Gainesville. This region has seven recommended stores despite having four active Ship from store stores already. This area is consistently in the top half of sales percentage, population density, and agricultural workers. Like Washington, the nearest distribution center is not close, and the second closest distribution center is going to lead to very expensive shipments.

Lastly, the mid-Atlantic region from North Carolina to Pennsylvania is the final region to have a cluster of recommended stores. Positioned between the Kentucky and New York distribution centers, these stores are in zip codes that have percent of sales, population, population density, and percentage working in agriculture all above the mean of the active store dataset. By having above-average numbers for the 4 most important predictors, this region is a high-volume area. While adding stores to this region may not save the most distance per order, compared to the previous four regions, this region receives enough orders that save enough in shipping to be recommended.

Tractor Supply Company should add multiple stores in each of these 5 designated regions. These regions all have high demand and are located far away from distribution centers. By adding stores in these regions, Tractor Supply Company will increase the likelihood of having inventory nearby customers, have a higher capacity of orders that can be shipped by stores, save money in shipping, increase customer satisfaction, and decrease the likelihood of an order shipping across the country.

## **Item Model & Recommendations**

The item Random Forest Regressor model took the predictors of item type, unit weight, largest side, smallest side, dimensional weight, and zone improvement per order to predict the expected savings when that item is shipped by feeding the data through 10,000 decision trees. When using all the previously mentioned predictors, the mean squared error of the Random Forest model was 0.309 with unit weight being by far the most important predictor, followed by zone improvement per order, largest side, dimensional weight, and smallest side respectively. Item type had around 0.002 importance and was removed from the model, as can be seen in Figures 5 & 6.

**Figure 5**

Chart

Description automatically generated

*Note:* First factor importance graphs to determine the item model.

**Figure 6**

Chart, waterfall chart

Description automatically generated

*Note:* Final factor importance graphs to determine the item model.

The result led to a mean squared error of 0.310 and the predictors' importance remained relatively unchanged. The predictions on the test data can be seen in Figure 7. There is a strong positive linear pattern on the graph which shows how accurate the model is. There are a few outliers, which almost all seem to be where the model is underestimating the expected savings.

**Figure 7**

Chart, scatter chart

Description automatically generated

*Note:* Scatter plot showing the accuracy between actual and predicted savings.

After using a prediction on our entire item table, there will be predicted estimated savings per order for each item and this table can be grouped by department or category. Grouping the items to a higher level is essential to limit the amount of work when changes need to be made and make sure that eligibility can be managed by picking through 20 departments or 130 categories instead of 100,000 individual items. When looking at the data averaged across the category level, the expected 15 most successful categories averaged just under 47 lbs., as seen in Figure 9. The top 15 categories included a lot of feed (equine, livestock, etc.) categories, which often ship at 35 or 50 lbs. Other items include grilling fuels, barb wire, walk-behind mowers, dog houses, or toolboxes. The savings in the top 15-20 categories are far better than the average across all the active items. By removing the items with lower predicted savings, Tractor Supply Company can limit their categories to optimize savings, since their stores have a limit on the number of items they can ship daily.

**Table 2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Category Description** | **Average Unit Weight** | **Average Dim Weight** | **Average Predicted Savings** |
| GRILLING FUELS | 40.00 | 69.00 | $8.59 |
| CAT LITTER | 40.35 | 57.35 | $7.35 |
| NET WRAP | 84.75 | 93.64 | $7.23 |
| BARB WIRE | 41.00 | 58.50 | $7.22 |
| WALK-BEHIND MOWERS | 69.58 | 123.11 | $6.91 |
| 3PT GROUND MAINT | 38.00 | 82.00 | $6.47 |
| EQUINE DUMOR FEED | 50.00 | 72.00 | $5.95 |
| WRAPPED FORAGE | 50.00 | 77.00 | $5.55 |
| EQUINE ECON FEED | 50.20 | 72.85 | $5.31 |
| OUTDOOR STRUCTURES | 83.25 | 93.52 | $4.98 |
| LVSTK DUMOR FEED | 42.02 | 63.00 | $4.80 |
| EQUINE BRANDED FEED | 48.92 | 72.80 | $4.77 |
| PROTEIN & MINERAL | 38.10 | 58.51 | $4.65 |
| LVSTK BRANDED FEED | 36.25 | 63.07 | $4.50 |
| STRAW | 37.50 | 78.00 | $4.26 |
| **Grand Total** | **46.85** | **71.66** | **$5.08** |

*Note:* Summary of the top 15 predicted categories.

# **Summary**

Shipping costs are continuing to become more important to omnichannel retailers. Many options are being weighed to combat increasing costs, one of which, being Ship from Store. The research analyzes if ship from store programs are effective at cutting costs and improving customer relations, along with how Tractor Supply Company and other omnichannel retailers can use machine learning to improve their existing programs.

## **Insights**

Tractor Supply Company should tweak the current ship from store system to make it more profitable. First, more stores should be added to the program. Washington State, Northern California, East Texas, Central Florida, and the Rust Belt are the 5 regions that should be immediately added to the Ship from store program. These regions match the demographics of other successful regions, fulfill specific geographic gaps, and limit the number of long-distance shipments. In each of these regions, multiple stores should be added to meet the demand for these regions and increase the capacity that can be shipped from the regions instead of distant distribution centers.

Second, the number of items allowed in the program should be limited. Since each store can only fulfill ten orders each day, Tractor Supply Company needs to maximize those orders by shipping the heavier categories from a store, leaving the small, light items to be shipped from distribution centers at longer distances. By filtering down the categories to around 25, Tractor Supply Company can increase the average order savings from around $1.50 to well over $3.00. Since customers do not choose Ship from store, this will have minimal impact on them.

## **Impact**

Currently, a process is being established to refresh the data and models monthly to get up-to-date recommendations on stores and categories to activate into the program. The initial insights were presented to the Omnichannel team’s leadership and the Store Operations team. The teams are seeking approval to add the requested stores and filter down the items to only the top 30 categories and their “most frequently bought together” as this will focus most of the sales towards the recommended categories but allow customers to increase their cart value by allowing add on items, such as dog treats to go with a bag of dog food without forcing the order to a distribution center.

## **Future Research**

The models described in this research can be run with supervised learning as more stores have opted into the program and orders are shipped. As more data is collected, the model accuracy will improve, and more insights can be gained. More demographic data can be collected to be used for the store models to check for importance in variables that the current model is missing. After the recommendations of this research are implemented, we will be able to analyze the actual outputs and compare the results to the predictions and tweak the models accordingly.

## **Conclusions**

From the models and literary review, it was apparent that Tractor Supply Company can improve its ship from store program to cut costs and improve customer experience through strategic placement of items to minimize shipping distance and time. Heavier items save more in shipping costs than light items. By focusing the shipments on categories with heavier items and minimizing zones shipped, Tractor Supply Company could improve their current ship from store success by multiple dollars per order. Orders will also be closer to the customers and be delivered faster, leading to higher customer satisfaction and increased online sales.

# **References**

Bayram, Armagan & Cesaret, Bahriye, 2021. "Order fulfillment policies for ship-from-store

implementation in omni-channel retailing," European Journal of Operational Research, Elsevier,

vol. 294(3), pages 987-1002.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140.

https://doi.org/10.1007/bf00058655.

Chang, S., Dong, Y., & Wang, X. (2021). Optimal shipping policy in retail competition and its effect on

customers. Electronic Commerce Research and Applications, 45, 101020.

https://doi.org/10.1016/j.elerap.2020.101020.

Chen, C., & Ngwe, D. (2018, October 16). Shipping Fees and Product Assortment in Online Retail.

Harvard Business School. https://www.hbs.edu/ris/Publication%20Files/19-034\_b2382177-

a462-447e-86f8-690d1ea7af18.pdf.

Chitroda, C. (2020, January 12). Logistic regression vs random forest classifier. LinkedIn.

https://www.linkedin.com/pulse/logistic-regression-vs-random-forest-classifier-chintan-

chitroda/.

Corsten, D., & Gruen, T. (2014, August 1). Stock-outs cause walkouts. Harvard Business Review.

https://hbr.org/2004/05/stock-outs-cause-walkouts.

Curtis, T., Abratt, R., Rhoades, D. L., & Dion, P. (2011). Customer Loyalty, Repurchase and Satisfaction: A

Meta-Analytical Review. Journal of Consumer Satisfaction, Dissatisfaction and Complaining

Behavior, 24(). Retrieved from https://commons.erau.edu/db-management/.

Droesch, B. (2021, May 17). How will the pandemic affect US ecommerce sales in 2021? Insider

Intelligence. https://www.emarketer.com/content/how-will-pandemic-affect-us-ecommerce-

sales-2021.

Guo, Y., Hu, X., Zou, Y., Li, S., Cheng, W., & Li, Z. (2020). Maximizing E-Tailers’ Sales Volume through the

Shipping-Fee Discount and Product Recommendation System. Discrete Dynamics in Nature and

Society, 2020, 1–14. https://doi.org/10.1155/2020/7349162.

Jones, A., Griffis, S., Schwieterman, M., & Daugherty, P. (2019). Examining the Impact of Shipping Charge

Fairness on Consumer Satisfaction and Behavior. Transportation Journal, 58(2), 101-125.

doi:10.5325/transportationj.58.2.0101.

Li, Rong. (2020). Reinvent Retail Supply Chain: Ship‐from‐Store‐to‐Store. Production and Operations

Management. 29. 10.1111/poms.13195.

Ma, Siqi. (2016). Fast or free shipping options in online and Omni-channel retail? The mediating role of

uncertainty on satisfaction and purchase intentions. The International Journal of Logistics

Management. 28. 10.1108/IJLM-05-2016-0130.

Miller, J. A. (2020, November 17). Ship from store makes sense in a pandemic. does it make sense long

term? Supply Chain Dive. https://www.supplychaindive.com/news/ship-from-store-warehouse-

backroom-fulfillment-distribution-retail/589117/.

Nguyen, D. H., de Leeuw, S., & Dullaert, W. E. H. (2016). Consumer behaviour and order fulfilment in

Online Retailing: A systematic review. International Journal of Management Reviews, 20(2),

255–276. https://doi.org/10.1111/ijmr.12129.

Schott, M. (2020, February 27). K-Nearest neighbors (KNN) algorithm for machine learning. Medium.

https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-

learning-e883219c8f26.

Shih, W. (2020, August 18). Global supply chains in A POST-PANDEMIC WORLD. Harvard Business

Review. https://hbr.org/2020/09/global-supply-chains-in-a-post-pandemic-world.

Stat trek (2021). Stat trek. Linear Regression. https://stattrek.com/regression/linear-regression.aspx.

Statt, N. (2019, October 24). Amazon is spending billions on shipping as it makes prime ONE-DAY

delivery a reality. The Verge. https://www.theverge.com/2019/10/24/20931055/amazon-prime-

one-day-delivery-shipping-q3-2019-earnings-spending.

Stewart, L., & Nushart, J. (2021, June 7). Ship From Store Item Limitations. personal.

Understanding FedEx and UPS SHIPPING zones - blog. Blog | AuditShipment.com. (2020, December 24).

https://www.auditshipment.com/blog/fedex-and-ups-shipping-zones/.

Warner, P. (2021, April 6). Tractor supply upgrades its neighbor's club loyalty program to provide

customers even more rewards. Tractor Supply Company - Tractor Supply Upgrades Its

Neighbor's Club Loyalty Program to Provide Customers Even More Rewards.

https://corporate.tractorsupply.com/newsroom/news-releases/news-releases-

details/2021/Tractor-Supply-Upgrades-Its-Neighbors-Club-Loyalty-Program-to-Provide-

Customers-Even-More-Rewards/default.aspx.

U.S. Census Bureau (2019). Industry by Class of Worker for the Civilian Employed Population 16 Years

and Over American Community Surveys. Retrieved from

https://data.census.gov/cedsci/table?q=agriculture%20workers&tid=ACSDT1Y2019.B24070&hid

ePreview=true.

U.S. Census Bureau (2010). Population Density Decennial Census. Retrieved from

https://blog.splitwise.com/2014/01/06/free-us-population-density-and-unemployment-rate-by-

zip-code/.

U.S. Census Bureau (2010). Total Population Decennial Census. Retrieved from

https://data.census.gov/cedsci/table?q=population%20and%20density&tid=DECENNIALCD1162

010.P1&hidePreview=true.

Yang, Daojian & Zhang, Xiujie. (2020). Omnichannel operations with ship-from-store. Operations

Research Letters. 48. 10.1016/j.orl.2020.03.009.

Zhong, Shuya, Lomas, Carl, Worth, Tracy. (2021) Understanding customers' adoption of express delivery

service for last-mile delivery in the UK. International Journal of Logistics Research and

Applications 0:0, pages 1-18.

# **Appendix A**

## **List of Tables**

[**Table 1 Dataset Summary Table** 19](#_Toc79318829)

[**Table 2 Optimal Categories for Ship from Store** 30](#_Toc79318830)

## **List of Figures**

[**Figure 1 First Importance Chart for Store Model** 20](#_Toc79318845)

[**Figure 2 Final Importance Chart for Store Model** 21](#_Toc79318846)

[**Figure 3 Store Decision Tree Example** 24](#_Toc79318847)

[**Figure 4 Map of Recommended and Active Stores** 25](#_Toc79318848)

[**Figure 5 First Importance Chart for Item Model** 28](#_Toc79318849)

[**Figure 6 Final Importance Chart for Item Model** 28](#_Toc79318850)

[**Figure 7 Accuracy Scatter Plot for Item Model** 29](#_Toc79318851)

# **Appendix B**

## **Store Model Python Code**

#Import packages

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

from sklearn import linear\_model

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn import metrics

import matplotlib.pyplot as plt

import seaborn as sns

#Load Data

stores = pd.read\_excel(r'C:\Users\jnushart\OneDrive - Tractor Supply Co\Downloads\Capstone.xlsx', sheet\_name = 'Stores')

inactivestores = pd.read\_excel(r'C:\Users\jnushart\OneDrive - Tractor Supply Co\Downloads\Capstone.xlsx', sheet\_name = 'Stores Not in Program')

#Data Prep & run the model the first time to find the lowest important predictors

stores['classifier'] = np.where(stores['Savings per Shipment'] >= np.percentile(stores['Savings per Shipment'],75),1,0)

x=stores[['Sales%', 'Median HH Income','% in ag','Nearest Store','population', 'density']]

y=stores[['classifier']]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=190)

clf=RandomForestClassifier(n\_estimators=10000)

clf.fit(X\_train,y\_train)

y\_pred=clf.predict(X\_test)

print('Mean Absolute Error:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('Mean Squared Error:', metrics.mean\_squared\_error(y\_test, y\_pred))

print('Root Mean Squared Error:', np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred)))

print("Accuracy:",metrics.accuracy\_score(y\_test, y\_pred))

clf=RandomForestClassifier(n\_estimators=10000)

clf.fit(X\_train,y\_train)

feature\_imp = pd.Series(clf.feature\_importances\_,index=x.columns.values).sort\_values(ascending=False)

feature\_imp

%matplotlib inline

sns.barplot(x=feature\_imp, y=feature\_imp.index)

plt.xlabel('Feature Importance Score')

plt.ylabel('Features')

plt.legend()

plt.show()

#Run model second time with less predictors, by removing low importance predictors to improve accuracy

x=stores[['Sales%','% in ag','Nearest Store','population', 'density']]

y=stores[['classifier']]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=190)

clf=RandomForestClassifier(n\_estimators=10000)

clf.fit(X\_train,y\_train)

y\_pred=clf.predict(X\_test)

accuracy = metrics.accuracy\_score(y\_test, y\_pred)

print('Mean Absolute Error:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('Mean Squared Error:', metrics.mean\_squared\_error(y\_test, y\_pred))

print('Root Mean Squared Error:', np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred)))

print("Accuracy:",metrics.accuracy\_score(y\_test, y\_pred))

clf=RandomForestClassifier(n\_estimators=10000)

clf.fit(X\_train,y\_train)

feature\_imp = pd.Series(clf.feature\_importances\_,index=x.columns.values).sort\_values(ascending=False)

feature\_imp

%matplotlib inline

sns.barplot(x=feature\_imp, y=feature\_imp.index)

plt.xlabel('Feature Importance Score')

plt.ylabel('Features')

plt.legend()

plt.show()

#Run inactive stores through the model to create prediction

inactivestores2 = inactivestores

inactivestores2 = inactivestores2.dropna()

inactivestores2.reset\_index(drop=False, inplace=True)

inactivestores3 = inactivestores2[['Sales%', '% in ag','population','density','Nearest Store']]

predict = clf.predict(inactivestores3)

#Prepare data for output

inactivestores4 = inactivestores2[['Store No', 'Zip Code']]

inactivestores4 = pd.merge(inactivestores4,inactivestores3,right\_index = True, left\_index=True)

inactivestores4['prediction'] = predict

yes = inactivestores4['prediction'] ==1

recommendedStores = inactivestores4[yes]

activestores = stores[['Shipment Ship Node', 'Zip Code']]

activestores['Status'] = 'Active'

recommendedStores['Shipment Ship Node'] = recommendedStores['Store No']

recommendedStores = recommendedStores[['Shipment Ship Node', 'Zip Code']]

recommendedStores['Status'] = 'Recommended'

#Output Data

stores\_final = activestores.append(recommendedStores, ignore\_index=True)

stores\_final.to\_csv(r'M:\Omnichannel\JNushart\Store Recommendation.csv', index=False)

## **Item Model Python Code**

#Import packages

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

from sklearn.ensemble import RandomForestRegressor

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_squared\_error, r2\_score

import seaborn as sns

from sklearn import metrics

from sklearn import tree

#Load data and run first Item Model

items = pd.read\_excel(r'C:\Users\Jordan\Desktop\Capstone\Capstone.xlsx', sheet\_name = 'Items')

items['Dim Weight'] = items['Largest Side'] + 2\*items['Middle Side'] + 2\*items['Smallest Side']

x1= items[['Item Type']]

x2= items[['Unit Weight','Largest Side', 'Smallest Side','Dim Weight','Zone Improvement per Order']]

#x= items[['Unit Weight','Dim Weight']]

y= items [['Savings Per Order']]

x = pd.get\_dummies(data=x1, drop\_first=True)

x = x.merge(x2,left\_index=True, right\_index=True)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=.25, random\_state=768)

model = RandomForestRegressor(n\_estimators=10000)

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

mse = mean\_squared\_error(y\_test,y\_pred)

rmse = np.sqrt(mse)

rmse

feature\_imp = pd.Series(model.feature\_importances\_,index=x.columns.values).sort\_values(ascending=False)

print(feature\_imp)

print('Mean Absolute Error:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('Mean Squared Error:', metrics.mean\_squared\_error(y\_test, y\_pred))

print('Root Mean Squared Error:', np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred)))

%matplotlib inline

sns.barplot(x=feature\_imp, y=feature\_imp.index)

plt.xlabel('Feature Importance Score')

plt.ylabel('Features')

plt.legend()

plt.show()

#Start second model after removing lowest importance variables and increasing accuracy

x= items[['Unit Weight','Largest Side', 'Smallest Side','Dim Weight','Zone Improvement per Order']]

y= items [['Savings Per Order']]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=.25, random\_state=768)

model = RandomForestRegressor(n\_estimators=10000)

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

mse = mean\_squared\_error(y\_test,y\_pred)

rmse = np.sqrt(mse)

rmse

feature\_imp = pd.Series(model.feature\_importances\_,index=x.columns.values).sort\_values(ascending=False)

print(feature\_imp)

print('Mean Absolute Error:', metrics.mean\_absolute\_error(y\_test, y\_pred))

print('Mean Squared Error:', metrics.mean\_squared\_error(y\_test, y\_pred))

print('Root Mean Squared Error:', np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred)))

%matplotlib inline

sns.barplot(x=feature\_imp, y=feature\_imp.index)

plt.xlabel('Feature Importance Score')

plt.ylabel('Features')

plt.legend()

plt.show()

#Accuracy Plot & output

plt.scatter(y\_test, y\_pred)

plt.xlabel('Actual')

plt.ylabel('Predicted')

predictions = model.predict(X\_test)

items['predictions'] = predictions.tolist()

items.to\_csv(r'C:\Users\Jordan\Desktop\Capstone\Items Recommendation.csv', index=False)