Shipment Price Prediction

Keval Parmar
AU1940133

Deep Patel
AU2040250

Roshni Navdiya
AU2040114

Aditya Padhariya
2040151

***Abstract*—Outlining the development and evaluation of a shipment price prediction model using machine learning techniques. The model utilizes historical shipment data to predict the price of future shipments and achieved high levels of accuracy in its predictions.**

Keywords—Prediction, Machine learning, Accuracy, Cost Analysis, cross-validation.

I. Introduction

Predicting shipment prices is a crucial responsibility in the logistics sector since it enables businesses to plan their shipping expenses better and manage their budgets. In this project, we created a machine learning model to forecast shipping costs depending on a number of variables, including the distance between the origin and destination, the weight of the item, and the mode of transportation. It may be challenging for organizations and individuals to predict their shipping prices effectively, which can result in unexpected fees. Shipping rates can vary greatly depending on a number of factors. We set out to create a service that can help people and companies save time and money by precisely forecasting shipping costs. The major goals of this project are to gather and preprocess shipping-related data, choose pertinent features, and create a machine-learning model that can reliably estimate shipment pricing. When the characteristics have been identified, the following step is to select the best machine learning algorithm that can be used to forecast the price. The project has employed numerous machine learning models, including Decision Tree Regressor, Random Forest Regressor, Linear Regression, and Gradient Boosting Regressor. A model known as a "Decision Tree Regressor" employs a tree-like structure to describe the decision-making process, with each node denoting a choice made in response to a feature value. Several decision trees are used in an ensemble model called the Random Forest Regressor to reduce overfitting and boost accuracy. An assumption made by the linear regression model is that the relationship between the input characteristics and the output variable is linear. Another ensemble model that combines many weak learners to produce a stronger model by repeatedly reducing the errors of the preceding models is the gradient-boosting regression.

A. Choice of Algorithms for the Problem:

1) Linear Regression: We used the linear regression model to get the general idea of our dataset, specifically to understand that at what extent our data is linear. The creation of a linear regression model is just the beginning.The linear regression technique makes the assumption that the parameters of the independent variables and the dependent variable Y have a linear relationship. As our dataset is big which contains about 33 features and each features has about 8,600 values so our particuar dataset will be needing complex function to fit the curve and reduce the loss function and we also checked the adjusted R^2 value, it is very low about less than 7%. Hence, We cannot apply the model if the real connection is not linear since the accuracy will be drastically decreased.

2) Random Forest Regressor: We selected this model because of its capacity to handle the big datasets effectively and because of its a high-accuracy ensemble of decision trees that employs randomization on two levels. These two levels be: 1) A subset of characteristics that can be employed as candidates at each split are chosen at random by the algorithm. As a result, different decision trees are not able to depend only on the same collection of attributes. 2) A random sample of training data is drawn by each tree. . However, they pose a major challenge that is that they can't extrapolate outside unseen data and for data with categorical variables having a different number of levels, random forests are found to be biased in favor of those attributes with more levels.We used this as our data has a non-linear trend and extrapolation outside the training data is not important. As per the calculated adjusted R^2 it is about 85.55% accurate.

3) Decision tree Regressor and Gradient Boost Algorithm: A decision tree algorithm will be used to split dataset features through a cost function. The decision tree is grown before being optimised to remove branches that may use irrelevant features. Parameters such as the depth of the decision tree can also be set, to lower the risk of overfitting or an overly complex tree. We used Gradient boost as it improves the accuracy of the model by sequentially combining weak trees to form a strong tree. In this way it achieves low bias and low variance. For these two mentioned algorithms adjusted R^2 of Decision tree regression is 90.6% accurate & Gradient boost is 95% accurate.

II. Literature Survey

Shipment price prediction is a crucial aspect of logistics management, and various machine learning techniques have been applied to this problem. A literature survey revealed that the majority of studies used regression-based models, including linear regression, support vector regression, and random forest regression. Feature engineering was also an essential aspect, with distance, weight, shipment mode, and location being the most frequently used features. However, there is still room for improvement in terms of model performance, feature selection, and data quality. According to references [1] & [2], in order to address the issue of fee setting on a freight transportation brokerage platform, where there is no predefined shipping cost, the primary variables influencing shipping cost setting were identified in this study, and a machine learning model for price prediction was created. Among a total of 73 parameters, including factors gathered from the freight brokerage process, factors that impact the cost of shipping were chosen using correlational analysis and the stepwise technique. The chosen criteria were environmental conditions, vehicle owner characteristics, and cargo characteristics. These criteria were used to build a shipping cost prediction model, and the effectiveness of each model was assessed. Freight weight, loading/unloading location, and duration were among the aspects that made up the characteristics of the cargo. "Vehicle tonnage" and "vehicle type" were included as owner-specific characteristics. An environmental factor was precipitation. The investigation's findings demonstrated that machine learning models such as DNN, XGBoost, and LightGBM outpaced more established analytical techniques such as linear regression.

III. Implementation & data preprocessing & Handling of missing data

In the dataset first step is to create and delete variables such as dropping unneeded columns such as vendor detail variables and Item description. Then cleaning the data for example in our dataset we are removing the rows where the first line designation is equal to no, and also removing the rows where 'Weight (Kilograms)' column is equal to 'Weight Captured Separately'. The second important step is to convert data object to numeric. This is done to ensure that the column contains numerical values that can be used in calculations and analysis. Variables that were anticipated to have an impact on shipping costs were short-listed in order to determine the elements that influence shipping costs and improve the predictive capacity of the shipping cost prediction model. In order to exclude outliers from the data, the interquartile range (IQR) was utilised. Only continuous variables were used for the outlier elimination process. As mentioned above to manage the missing values we deleted the rows which has some NA's in the row and that particular attribute did not play important role in prediction and imputing missing values for categorical variable. When missing values come from categorical columns (string or numerical), the most prevalent category might be used to fill in the blanks. Here in our dataset there is no such need to predict the missing values as they are either deleted or filtered. In short there is no need arise which demanded extrapolation of data. Then the third and final step is to select the parameters on the basis of which algorithms will be implemented in our dataset we selected 9 features 'Unit Price', 'Pack Price', 'Line Item Quantity' ,'Unit of Measure (Per Pack)','Freight Cost (USD)','Weight (Kilograms)','ID', 'Line Item Value', 'Line Item Insurance (USD)'. And then each parameters were visualised and studied individually with each other with the help of scatter plots and histograms.

IV. RESULTS & CONCLUSION

The performance of the Linear Regression and Random Forest Regression models can be compared using metrics such as R^2, mean squared error and root mean squared error. The performance of linear regression model on Y1 train and Y2 train set are as follows:

On training set Y1:

- R^2 value = 0.072260
- Root mean squared error = 51.063126
- Mean squared error = 2607.442830

On training set Y2:

- R^2 value = 0.024017
- Root mean squared error = 2.375077
- Mean squared error = 5.640988

The performance of Random Forest regression are as follows:
On training set Y1:

- R^2 value = 0.855034
- Root mean squared error = 20.184923
- Mean squared error = 407.431112

On training set Y2

- R^2 value = 0.907654
- Root mean squared error = 0.730576
- Mean squared error = 0.533742

Testing on the edge cases where the model gives the best accuracy will be our further implementation and predicting the unique values which can provide best price is aim in the project. We concluded that Linear Regression will be a very bad choice for algorithm implementation in our project and to improve the accuracy we would need an algorithm which is capable of handling multiple variables and predict a value by doing analysis on different small chunks of data. Hence, we choose Decision Tree Regressor and Gradient Boost Regressor for achieving more accuracy in further implementation.

V. References

[1] K. L. Keung, C. K. M. Lee and Y. H. Yiu, "A Machine Learning Predictive Model for Shipment Delay and Demand Forecasting for Warehouses and Sales Data," 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, Singapore, 2021, pp. 1010-1014, doi: 10.1109/IEEM50564.2021.9672946.B. Rieder, *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

[2] Jang, H. S., Chang, T. W., & Kim, S. H. (2023). Prediction of Shipping Cost on Freight Brokerage Platform Using Machine Learning. Sustainability, 15(2), 1122.

[3] Ubaid, A., Hussain, F. K., & Charles, J. (2020). Machine learning-based regression models for price prediction in the Australian container shipping industry: case study of Asia-Oceania trade lane. In *Advanced Information Networking and Applications: Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA-2020)* (pp. 52-59). Springer International Publishing.

[4] Rushikeshkalkar. (2023, January 18). Supply_chain_shipment. Kaggle. Retrieved*A machine learning predictive model for shipment delay and demand ...* (n.d.). Retrieved March 11, 2023, from https://www.researchgate.net/publication/357967446_

[5] Machine_Learning_Predictive_Model_for_Shipment_Delay_and_Demand_Forecasting_for_Warehouses_and_Sales_Dat March 11, 2023, from https://www.kaggle.com/code/rushikeshkalkar/supply-chain-shipment

[6] Our Dataset Link to Kaggle:
https://www.kaggle.com/code/divyeshardeshana/supply-chain-shipment-price-data-analysis/input#:~:text=calendar_view_week-,SCMS_Delivery_History_Dataset,-.csv