Shipment Price Prediction

Deep Patel, Keval Parmar, Roshani Navdiya, Aditya Padhariya

***Abstract*— The construction of a machine learning model for forecasting shipment costs using input characteristics including package weight, dimensions, method of transportation, and distance is discussed in this study. On a dataset comprising data on more than 100,000 shipments, the model was trained and assessed using a variety of regression approaches. With an R-squared score of 0.86 on the validation set and 0.83 on the testing set, the gradient boost was the method that performed the best overall. Logistics businesses may utilize the established model to optimize their pricing plans.**

Keywords—Supply chain, Transportation, Cost analysis, Cross-validation, Preprocessing, Logistics, Pricing strategies.

I. Introduction

The logistics industry is a significant contributor to the global economy, facilitating the movement of goods and services across the world. The pricing of shipments is a crucial factor for businesses and individuals, as it can significantly impact their bottom line. Accurate prediction of shipment prices is essential for companies to manage their budgets and optimize their pricing strategies. In this project, we aimed to develop a machine-learning model that can predict shipment prices based on various factors such as distance, transportation mode, package weight, and dimensions. We collected a large dataset of shipment information to train and evaluate our model. The first step of our project was data preprocessing, where we cleaned and preprocessed the dataset to handle missing values and outliers. We also performed feature engineering to extract relevant features that could impact the shipment price. We then used the processed data to train and evaluate several regression algorithms, including Decision Tree Regressor, Random Forest Regressor, Linear Regression, and Gradient Boosting Regressor. Our analysis revealed that the Gradient Boost Regressor algorithm outperformed the other algorithms, achieving an R-squared value of 0.86 on the validation set and 0.83 on the testing set. The model was then fine-tuned using hyperparameter tuning techniques to improve its performance. The developed model can help logistics companies optimize their pricing strategies and accurately predict their shipment costs. The model can also assist individuals in planning and budgeting their shipping expenses, ultimately leading to cost savings and improved efficiency. The findings of this project have several implications for the logistics industry. Accurate shipment price prediction can help logistics companies optimize their pricing strategies and increase their profitability. It can also assist in managing their budgets, leading to better financial planning and decision-making. The model can be integrated into logistics software systems to provide accurate and timely pricing information to customers. Overall, this project demonstrates the potential of using machine learning algorithms for shipment price prediction. We hope that our work will contribute to the ongoing efforts to improve the efficiency and effectiveness of the logistics industry. The limitations of the model and future research directions are also discussed in this report.

II. Literature Survey

As per the literature survey, shipment price prediction is a crucial aspect of logistics management, and several studies have been conducted to explore the use of machine learning techniques in this area. Regression-based models, such as linear regression, support vector regression, and random forest regression, have been extensively used to predict shipment prices. These models have shown promising results, with factors such as 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. Deep learning models, such as recurrent neural networks and convolutional neural networks, have also been applied to the problem of shipment price prediction, and have shown promising results in improving prediction accuracy. Feature engineering is another crucial aspect of successful shipment price prediction. While distance, weight, shipment mode, and location have been widely used features, there may be other factors that can impact shipment prices, such as weather conditions, traffic, and fuel prices. Therefore, further research is needed to identify and select relevant features that have a significant impact on shipment prices. The quality of the data used to train the models can significantly affect their performance. It is essential to ensure that the data used is accurate and representative of real-world scenarios. The use of biased or incomplete data can lead to inaccurate predictions, and researchers should strive to ensure that the data used is of high quality. In conclusion, while several studies have focused on the application of machine learning models to predict shipment prices, there is still room for improvement in terms of model performance, feature selection, and data quality. Future research in this area could explore alternative features that may impact shipment prices and investigate novel techniques to improve the accuracy of the predictions. Furthermore, researchers should strive to ensure that the data used is of high quality to avoid bias and inaccuracies in the predictions.

III. Implementation

The implementation of shipment price prediction involves several essential steps, starting with data collection and cleaning, followed by data analysis and model building. The first step is to collect data on shipment delivery history, which includes various parameters such as product details, shipment mode, delivery date, and cost details. The data is then cleaned to remove unnecessary columns, filter out irrelevant rows, and convert object data types to numeric

data types to ensure that the data is in a suitable format for analysis.

Once the data is cleaned, the next step is to perform data analysis to explore the dataset's descriptive statistics, identify correlations between variables, and visualize the data using scatterplots, histograms, and boxplots. The data analysis provides valuable insights into the dataset, which can be used to identify trends, patterns, and anomalies.

The next step is to build predictive models to predict shipment cost. Several machine learning models can be used for this purpose, including linear regression, decision tree regression, random forest regression, and gradient boosting regression. These models are built using popular machine-learning libraries such as Scikit-learn and TensorFlow. The models are trained on a training dataset and validated using a testing dataset to ensure that they accurately predict the shipment cost.

The performance of the model is evaluated using various metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared. These metrics provide insights into the accuracy, precision, and reliability of the model's predictions.

The model can be deployed using various deployment methods, such as API integration, web applications, or mobile applications. The models can be continually updated and improved using new data, which ensures that the predictions are accurate and up-to-date.

There are some limitations and challenges faced during the implementation of the shipment price prediction model. The quality of the dataset used for model training and validation can significantly impact the model's performance. Therefore, it is necessary to ensure that the dataset is comprehensive and unbiased. Moreover, the scalability and generalizability of the model are crucial for real-world applications, and they must be evaluated to ensure that the model can handle large datasets and new, unseen data.

The implementation of shipment price prediction involves several critical steps, including data cleaning, data analysis, and model building. The performance of the model can be evaluated using various metrics, and the model can be deployed using various deployment methods. However, it is crucial to address the limitations and challenges faced during the implementation process to ensure that the model is accurate, reliable, and scalable.

IV. Results

Our aim is to develop a model to predict the unit price and pack price of a product based on various attributes. We used a dataset containing information on product attributes such as brand, pack size, and weight, along with the corresponding unit price and pack price. After cleaning and preprocessing the data, we split it into training and testing sets and applied various regression models.

For pack price prediction, we used linear regression, decision tree regression, random forest regression, and gradient-boosting regression models. The results showed that the decision tree regression and gradient boosting regression models had the best performance, with an R2 score of approximately 0.95. For unit price prediction, we applied the same regression models and found that the decision tree regression model had the best performance, with an R2 score of approximately 0.95.

The results indicate that our models can effectively predict the unit price and pack price of a product based on its attributes. These models can be useful for businesses to estimate prices and optimize pricing strategies. By using the predicted prices, businesses can adjust prices to increase revenue and profit margins.

It is important to note that the performance of our models could be further improved by incorporating additional features such as competitor prices, customer demographics, and economic indicators. However, the current models provide a strong foundation for predicting prices based on product attributes alone.

Our project demonstrates the effectiveness of regression models in predicting the unit price and pack price of a product. The decision tree regression model was found to be the best-performing model for unit price prediction, while the gradient boosting regression model and decision tree regression model performed the best for pack price prediction. The models have the potential to be useful tools for businesses to optimize pricing strategies and increase profits.

The performance decision tree regressor and gradient boost as shown as below:
Decision Tree of Unit Price:
MSE: 0.219296
RMSE: 0.468
R2_value : 0.962

Decision Tree of Pack Price:
MSE: 183.45
RMSE: 13.544
R2_value : 0.9347

Gradient Boost of Unit Price:
MSE: 0.4635
RMSE: 0.6808
R2_value : 0.9197

Gradient Boost of Pack Price:
MSE: 140.67
RMSE: 11.86
R2_value : 0.9499

V. Conclusion

Based on the performance metrics and analysis conducted in the end-semester report, it can be concluded that the decision tree model is the better choice compared to the

Linear Regression and gradient boost model for predicting the target variables Y1 and Y2. This conclusion is reaffirmed by the results of the end-semester report, which provides further insight into the models' performance on the validation set.

The Decision tree model outperforms the Linear Regression model on both Y1 and Y2 validation sets, with higher R2 values, lower root mean squared errors, and lower mean squared errors. The R2 values of 0.866 and 0.911 for the Random Forest model on the Y1 and Y2 validation sets, respectively, indicate that the model can explain a significant proportion of the variance in the data. The lower root mean squared errors of 19.722 and 0.700 and mean squared errors of 386.611 and 0.490 indicate that the Random Forest model has a better fit and produces smaller errors in its predictions.

However, it is important to note that model selection should not be based solely on these metrics. Factors such as model interpretability, computational complexity, and data size should also be taken into consideration. In terms of interpretability, the Linear Regression model may be easier to understand and explain due to its simplicity. On the other hand, the Random Forest model may be more computationally complex, especially with larger datasets, but it's high accuracy and ability to handle nonlinear relationships make it a better choice for predicting the target variables.

References

[1] 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_

[2] 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

[3] Github Link: https://github.com/DeepPatel4052/Shipment-Price-Prediction_Machine-Learning.git