Predicting Shipping Prices Using Machine Learning

Zachary J. Fuller

Northwest Missouri State University, Maryville MO 64468, USA
s553830@nwmissouri.edu and zachfuller1@gmail.com

Abstract. Keywords: data analytics · data science · machine learning · logistics · shipping

1 Introduction

This project will cover the domain of logistics. I currently work for a freight auditing company, so my knowledge of the industry will be helpful in this analysis. Additionally, I am hopeful that what I learn from this study can be applied to the work I do in my professional life. My data will likely from Kaggle, as I have found a couple of decent data sources on there already. Google's Dataset search engine will also be helpful. If any data I find has tracking numbers associated with it, then UPS, FedEx, or the associated carrier's website should also provide some valuable data about the package. I would like to do some predictive analytics using previously shipped packages as a training set. Specifically, predicting the cost to ship a package based on multiple factors like weight, destination, origin, and distance traveled could likely be accomplished using a regression model. I feel that this could be useful as carriers pricing policies are not very transparent unless you are well versed in logistics, so having a pricing estimator could empower small business owners and online sellers to choose the best shipping option for their needs. The steps I will follow to complete this project are:
1. I will first locate my dataset.
2. Then, I will clean the dataset using the python packages Pandas and NumPy, removing any drastic outliers if needed and deciding how to handle missing values.
3. Then, I will use SciKitLearn to separate the data into a train and test set. 4. I will evaluate the results. If needed, I will retrain and test the data, repeating steps 2 and 3.
5. After training and testing is complete, I will use a python plotting package, likely Plotly (my preferred choice) or Matplotlib to display the results.
6. Finally, I will write the report, making sure to examine any possible errors.

The key components of my approach will be finding the right data sources, making sure the data is cleaned in a way so as not skew results, and most importantly, the training and testing process. This will likely be done using multiple linear regression, so making sure that the data is trained correctly is key finding the best fit to account for each independent variable.

2 Methodology

3 Data Collection

My main data source was found on Kaggle [1]. I have also found some supplementary data on Data World [3] [2]. that can help to provide some general context for my data domain of shipping and logistics in general. The data is found in a few CSV files. I did not have to use any data scraping techniques, as my main data source and the supplementary data sources are made available for easy download through each of their host sites. From the "Shipping Optimization Challenge" dataset on Kaggle, I will use the sendtimestamp, sourcecountry, destinationcountry, freightcost, grossweight, and shipmentcharges fields. The dataset has thousands of columns, so I will likely need to fill in some null values. Additionally, I may slice off the timestamps and only use dates, as it appears that a lot of this data involves international shipping, and the timestamp may be less important and take up space/time. Finally, I will need to convert the country fields, as they are simply abbreviations at the moment. I will likely read the CSVs into a pandas dataframe in a Jupyter notebook, merge the necessary tables and fields together, and then perform this cleaning.

4 Data Processing

5 Results and Analysis

6 Limitations

7 Conclusions

[]

References

1. GAUTAM, S.: [1] shipping optimization challenge, https://www.kaggle.com/datasets/salil007/1-shipping-optimization-challenge?select=train$_{2p}r.csv$
2. Griffith, B.: Ups$_{d}aily_{r}ates.xlsx, https$: $//data.world/siyeh/crm - project/workspace/file?filename = UPS_{d}aily_{r}ates.xlsx$
3. Hoov, G.: How freight moves, https://data.world/garyhoov/how-freight-moves
4. Pollack, N.: 3 ways to improve logistics management using shipping data, https://transimpact.com/nextsights/3-ways-to-improve-logistics-management-using-shipping-data/