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

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. Pollack, N.: 3 ways to improve logistics management using shipping data,
   https://transimpact.com/nextsights/3-ways-to-improve-logistics-management-using-
   shipping-data/