Predicting Rich's transportation rates of unknown markets

Boya Zhang, Data Science and Analytics,

Mathematics Department, SUNY Buffalo State

Buffalo State
The State University of New York

Student Research and Creativity Conference

Abstract

Rich Products Corporation (also known as Rich's) is a privately held, multinational food-products corporation headquartered in Buffalo, New York. The problem of predicting transportation rates is receiving considerable attention with the establishment of new markets in the United States and Canada. Expanding new markets requires the creation of a new distribution center (DC), and its associated cost of transportation to the point of delivery.

This project is to (1) analyze historical transportation rates in the south-east and south-central regions, (2) determine appropriate weight brackets and transportation modes (I'm mainly responsible for TruckLoad transportation), (3) predict base rates for new Origin-Destination combinations incorporating additional data like population may improve the accuracy of formula and (4) use Market rates from the TransPlace database to test formula.

Introduction

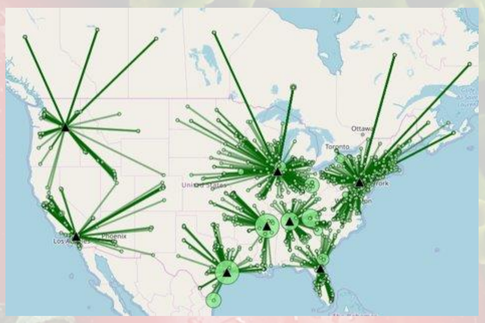

Figure1 Existing DC distribution and transportation lines

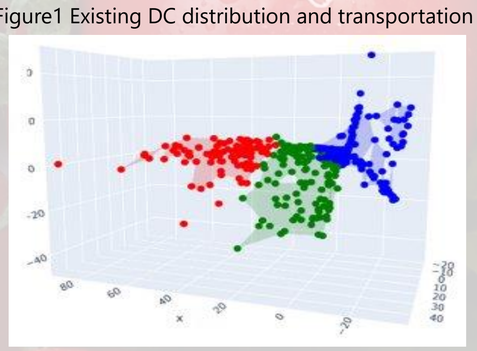

Figure2 three clusters of product groupings

As you can see, Rich's has transportation rates for existing markets. For example, one of their major Distribution Centers (DC). Since this DC distributes to most of the southeast and southcentral regions, they have historical transportation rates into those markets for the weight brackets and transportation modes (Truckload TL, Multi-Stop TL, Less than Truckload LTL, Intermodal IM) that they use.

However, once they want to add a new DC, they don't know the cost of shipping, and currently they use TransPlace database to solve this problem, so it's better to calculate the cost of each lane from the known data.

Method or Approach

1.Collect Data

Getting a database from Rich's & using data from Cencus considered as external influence.

2. Data Analysis Tools

Python 3.8.2 & from sklearn.linear_model import LinearRegression

3. Clean Data

According to the steps[1]: Get Rid of Extra Spaces. Select and Treat All Blank Cells. Convert Numbers Stored as Text into Number. Delete all Formatting.

4. Determine The Mode

Multiple Linear Regression (MLR)[2]:Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

Findings to Date

Explore internal data interconnections

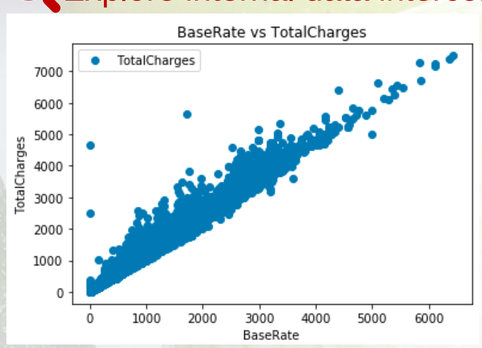

Figure3 The relationship between BaseRate & Totalcharges

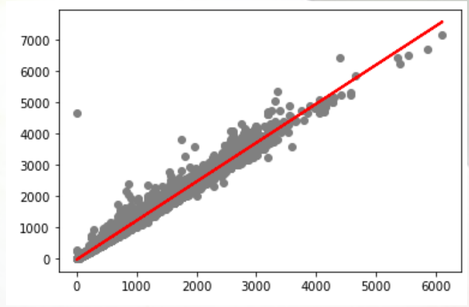

Figure4 Fitting with existing data

Figure5 the correlation matrix for the internal factors with float64 or int64 data type

Explore external data interconnections included in internal

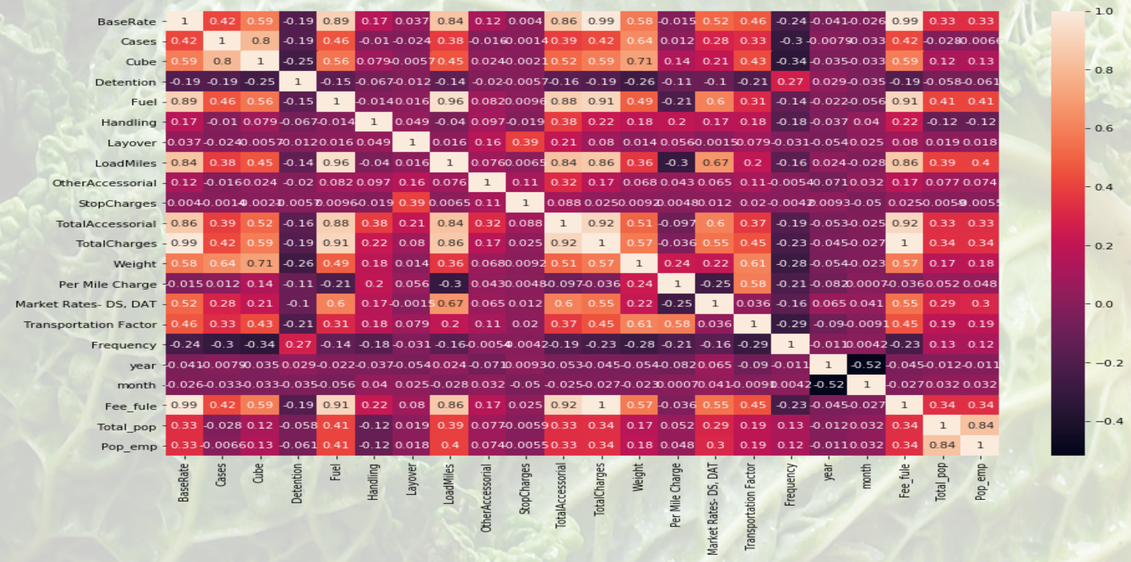

Figure6 the correlation matrix for the internal & external factors with float64 or int64 data type

Therefore, if factors affecting BaseRate are predicted, Totalcharges can be obtained through linear relationship. The necessary factors were screened out by the correlation data and the conclusions were drawn by the training of multiple linear models.

Interpretations of Findings

1. The amount of data.

Data conditions	Amount of data
Raw Data	118,486
Cleaned Data	95,636
Mode Truckload of Data	55,483
Added employment & population data	36,818
Data divided into four seasons(spring to winter)	14414;13825;14434;12810

2. Add the data set into the model and select 20% for the model prediction.

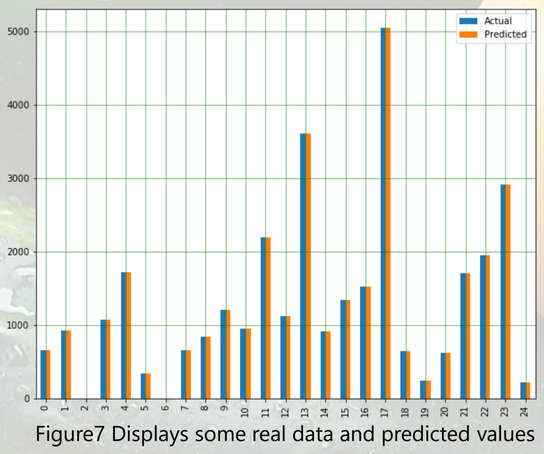

Figure7 Displays some real data and predicted values

We can get r squared, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error after every testing. The larger the value of r squared is, the smaller the value of Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error is, indicating that the prediction method is more accurate.

Conclusions

1.Calculate the one of season cost formula

Totalcharge=-0.04cases+0.1cube+0.56LoadMiles0.08frequency+0.46LoadMiles*PerMileCharge+2.23PerMileCharge+Handling+Layover+StopCharges+OtherAccessorial-8.67580191e-10population-1.89400189e-08employment+152

2. Display of data results

	Coefficients
Cases	-0.038607
Cube	0.107314
Fuel	3.184615
LoadMiles	-0.032528
Weight	0.005512
Frequency	-0.155778
Per Mile Charge	2.235452

```
r squared : 0.9999996537674344
Mean Absolute Error: 0.24386942259804698
Mean Squared Error: 0.2894343208130865
Root Mean Squared Error: 0.5379910043979235
```

References

[1] Image and concept credit, Aaron Tay from Blog – Musings about librarianship, viewed March 2020, <https://www.digitalvidya.com/blog/data-cleaning-techniques/>

[2] Multiple Linear Regression – MLR Definition, By WILL KENTON, Updated Apr 14, 2019, <https://www.investopedia.com/terms/m/mlr.asp/>

Acknowledgements

1.This research was partially supported by The Mathematical Association of America (MAA) and the National Science Foundation (NSF grant DMS-1722275) and the National Security Agency (NSA).

2.Thanks to Dr. Joaquin Carbonara, Dr. Xu Hongliang (Mathematics Department, SUNY Buffalo State) and Catherine March and her team at Rich's (in particular Esha Thorat) for their guidance.