PREDICTING RICH´S TRANSPORTATION RATES OF UNKNOWN MARKETS

by

Boya Zhang

May 2020

A dissertation submitted to the

Faculty of Department of Mathematics

SUNY Buffalo State

&

Department of Computer Science and Technology

Beijing Union University

in partial fulfilment of the requirements for the

degree of

Bachelor of Science

Acknowledgments

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 (Mathematics Department, SUNY Buffalo State), Dr. Xu Hongliang (Mathematics Department, SUNY Buffalo State) for their guidance.

3. This work is part of the PIC Math program, I am a member of MAA PIC Math team from BuffaloState College and worked with Rich company, which works to Predict transportation rates of unknown markets. Our industry liaisons at RICH were Catherine March and her team at Rich (in particular Esha Thorat)

Table of Contents

List of Tables

List of Figures

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 am 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.

Keywords: prediction of transportation rates, multiple linear regression, data analysis, freight's calculation

Chapter 1

Introduction

1.1 Background of Rich Company

Rich Products Corporation (also known as Rich's) is a privately held, multinational food products corporation headquartered in Buffalo, New York.

Figure 1.1: 1959: the nationâs first non-dairy creamer hits supermarkets

They were founded in 1945 by Robert E. Rich, Sr., after his development of a non-dairy whipped topping based on soybean oil. In the 1940s their annual revenue was 344K. Since then, they had expanded their non-dairy frozen food offerings, and also supplies products to retailers, in-store bakeries, and food service providers. Today Richâs Annual revenue is 4B, with 11k associates worldwide, operating in 100+ companies. They operate under private label, marketplace, and consumer brands.Recent acquisitions to grow market in

cookies (Christieâs 2019), GF pizza (Rizuttos 2019), bread capacities (TreeHouse 2020), and seafood (Mauryâs 2020). They even expanded business into non-food subsidiaries.

1.2 Existing Supply Chain

Many supply chains exist as an evolution of individual decisions–expansions, closures, acquisitions, and organic growth. In most cases, the current state is not the optimal state in terms of cost, efficiency, or service. So, they had a Network modeling which enables them to diagnose the current state and then examine what alternative scenarios might look like. Using the Distribution Model, they segmented the west coast customers and identified

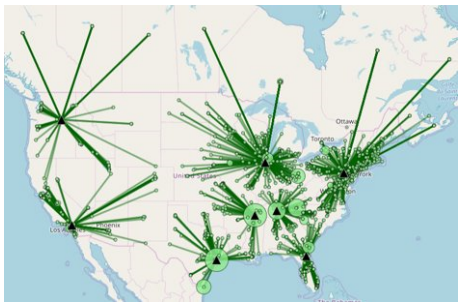


Figure 1.2: Existing DC distribution and transportation lines

two potential sites as alternatives to the current structure. Used on customer demand to determine products that were frequently sold together regardless of production location.

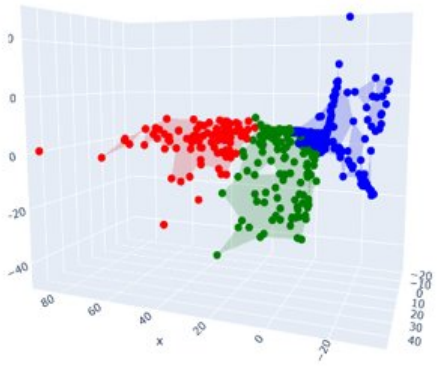


Figure 1.3: Three clusters of product groupings

1.3 Transportation Modes

Transportation modes include Truckload TL, Multi-Stop TL, Less than Truckload LTL, Intermodal IM. (This is Team Work Project, so I am mainly responsible for TL mode)

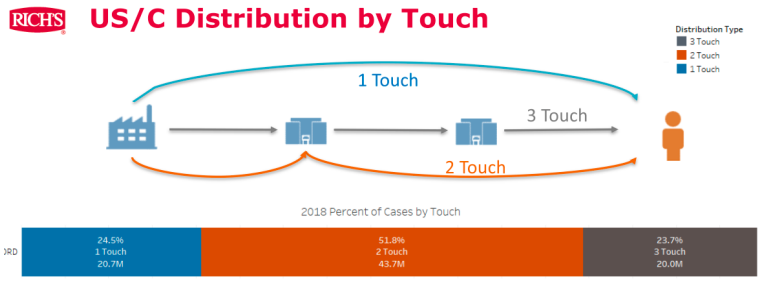


Figure 1.4: The different distribution type determines different transportation modes

Truckload (TL): A term that refers to a shipment that requires the use of an entire truck for shipments. TL is priced per mile within specific designated lanes, regardless of the size of the shipment as long as it fits in the truck. These shipments are also often less expensive per unit than LTL freight. Less Than Truckload (LTL): Involves shipments that donât need the use of an entire truck for shipping goods. LTL pricing is based on freight class, weight and total lane mileage. LTL companies normally utilize a complete network of consolidation and de-consolidation points to make sure shipments can fit within available truck space, carrying several shipments from multiple customers in a single truck.

1.4 New Business Needs

As you can see, Richâs 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 that they use. When they do network design problems and try to assess if there should be a new DC added, they donât have historical transportation rates to use for the potential lanes. For example, adding a DC in Texas to

Figure 1.5: New DC added in TX instead of old DC in TN

service Texas and Oklahoma customers from instead of the Murfreesboro TN DC. Because they donât have historical rates for those lanes, they use market rates from a TransPlace database. To keep âapples to applesâ comparison they also substitute those market rates in for the known historical rates (such as Murfreesboro TN to customers in Texas). However, once they want to add a new DC, they don't know the cost of shipping, and currently, they use the TransPlace database to solve this problem, so it's better to calculate the cost of each lane from the known data.

Chapter 2

Method and Approach

2.1 General Steps for Analyzing Data

There are several steps for analyzing data.

Usually, the first step is defining questions. In this project, I need to find and calculate the transportation cost per route on an existing basis, then gradually add external factors to the calculation.

The second step is setting clear measurement priorities. By analyzing the influence of each factor on the cost, the necessary factors are selected for analysis

The third step is collecting data. In this step, Iâll determine what information could be collected from existing databases or sources on hand. (Rich provided) Then, explore differences in high-density lanes (frequency of shipments), variations in load size, and market regions (urban to rural and vice versa). Incorporate external data (population density, employment rate, etc.) to the prediction. And keep collected data organized in a log.

The fourth step is analyzing data. Begin by manipulating data in several different ways, such as plotting it out and finding correlations or by creating a pivot table. The relationship between the data was found by visualizing the data, and then external factors were added to improve the accuracy of the model. During this step, data analysis tools and software are

extremely helpful. Visio, Minitab, and Stata are all good software packages for advanced statistical data analysis. However, in most cases, nothing quite compares to Microsoft Excel in terms of decision-making tools.

The fifth step is interpreting results. Find the solution by defining the problem clearly and estimate the accuracy of the solution.

2.2 Theoretical Method

2.2.1 Statistical Theory

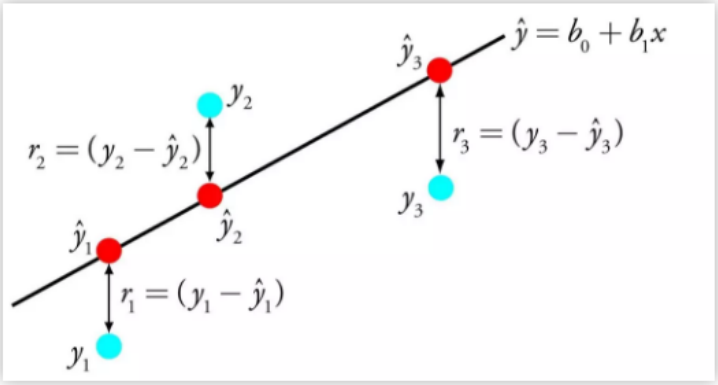
a) Correlation analysis Study whether there is a certain dependency relationship between phenomena, and explore the relevant direction and degree of the specific dependency phenomenon. 1. Single correlation: The correlation between the two factors is called a single correlation, that is, only one independent variable and one dependent variable are involved in the study; 2. Complex correlation: The correlation between three or more factors is called complex correlation, ie The research involves the correlation between two or more independent variables and the dependent variable; 3. Partial correlation: When a certain phenomenon is related to multiple phenomena when other variables are assumed to be unchanged, the correlation between the two variables The relationship is called partial correlation.

b) Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for p explanatory variables $x_1, x_2, \ldots, x_p$ is defined to be $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$. This line describes how the mean response μ_y changes with the explanatory variables. The observed values for y vary about their means $\mu_y$and are assumed to have the same standard deviation σ . The fitted values $b_0, b_1, \ldots, b_p$ estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$of the population regression line. Since the observed

values for y vary about their means μ_{y}, the multiple regression model includes a term for this variation. In words, the model is expressed as DATA = FIT + RESIDUAL, where the "FIT" term represents the expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$. The "RESIDUAL" term represents the deviations of the observed values y from their means μ_y, which are normally distributed with mean 0 and variance σ . The notation for the model deviations is ε . Formally, the model for multiple linear regression, given n observations, is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i$ for $i = 1, 2, \ldots n$ In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. The least-squares estimates $b_0 b_1, \ldots b_p$ are usually computed by statistical software. The values fit by the equation$b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}$ are denoted $\hat{y}_i$, and the residuals ei are equal to yi - $\hat{y}_i$, the difference between the observed and fitted values. The sum of the residuals is equal to zero. The variance σ^2may be estimated by $s^2 = \frac{\sum e_i^2}{n-p-1}$, also known as the mean-squared error (or MSE). The estimate of the standard error s is the square root of the MSE.[**1**]

2.2.2 Machine Learning Theory

a) Linear regression is probably one of the most well-known and understandable algorithms in statistics and machine learning. Predictive modeling focuses on minimizing model errors or making the most accurate predictions at the expense of interpretability. We borrow, reuse, and misappropriate algorithms from many different fields, which involve some statistical knowledge. Linear regression is represented by an equation that describes the linear relationship between the input variable (x) and the output variable (y) by finding the specific weight of the input variable (B). Example: y = B0 + B1 * x Given the input x, we will predict y. The goal of the linear regression learning algorithm is to find the values of the coefficients B0 and B1. Different techniques can be used to learn linear regression



Linear Regression

Figure 2.1: Linear Regression

models from the data, such as linear algebra solutions for ordinary least squares and gradient descent optimization. Linear regression has been in existence for more than 200 years, and extensive research has been conducted. If possible, some rules of thumb when using this technique are to remove very similar (correlated) variables and remove noise from the data. This is a fast and simple technique and a good first algorithm.

b) Bagging and the random forest is one of the most popular and powerful machine learning algorithms. It is an integrated machine learning algorithm called Bootstrap Aggregation or Bagging. Bootstrap is a powerful statistical method used to estimate a certain amount from a data sample, such as the average. It will draw a large number of sample data, calculate the average, and then average all the averages to estimate the true average more accurately. The same method is used in bagging, but the most commonly used is the decision tree, rather than estimating the entire statistical model. It multiplies the training data and then builds a model for each data sample. When you need to predict new data, each model will make predictions and average the prediction results to better estimate the true output value. Random forest is an adjustment to the decision tree. Compared with selecting the best split point, random forest achieves sub-optimal splitting by introducing randomness.

Therefore, the differences between the models created for each data sample will be

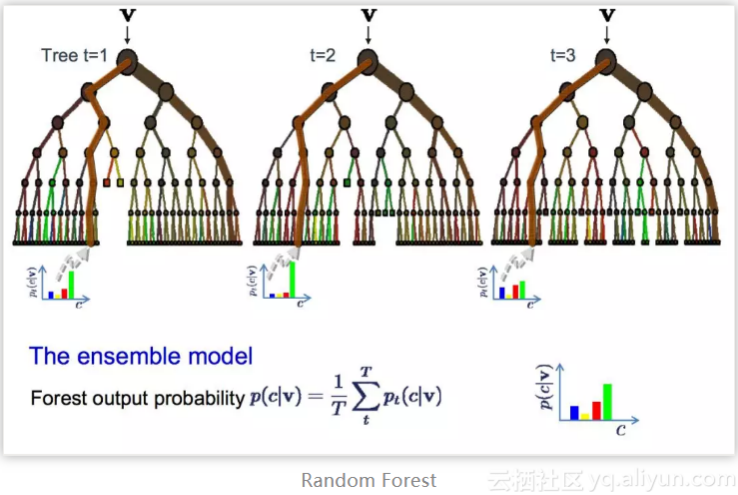


Figure 2.2: Random Forest

greater, but they are still accurate in their sense. Combining the prediction results can better estimate the correct potential output value.

If you use a high variance algorithm (such as a decision tree) to obtain good results, then the effect will be better after adding this algorithm.

2.3 Technology Method

a) Python Language Introduction Python is a widely-used general-purpose, high-level programming language. It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code.

Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions: Python 2 and Python 3. Both are quite different.[**2**]

b) I mainly imported the sklearn library to analyze the data. Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering

algorithms including support vector machines, random forests, gradient boosting, k-means, and DBSCAN, and is designed to interoperate with the Python numerical and scientific

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as seabornInstance
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
```

Figure 2.3: The libraries I used

libraries NumPy and SciPy.[**3**]

c) The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter.

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.[**4**]

Chapter 3

Process

3.1 Understanding Existing Data

Here are the internal data items provided by RICH.The final dependent variable to be predicted is TotalCharge, in which BaseRate is the most important dependent variable that constitutes TotalCharge, and then BaseRate is used as the dependent variable for analysis, in which Fuel is the most important dependent variable that constitutes BaseRate.

The selected independent variable is highly correlated with its dependent variable, such as: Detention, Handling, Layover, StopCharge, OtherAccessorial, Cases, Cube, Load-Miles, and Weight.

Factors	DataTypes	Explanation
BaseRate	float64	BaseRate was supplied to Rich by a shipping company, so Rich wanted to know expenses incurred.
Carrier_Key	object	Unique code identifying an authorized carrier
Cases	int64	Total casesâ volume of the inventory on the carrier move
Cube	int64	The total cubic volume of the inventory on the carrier move
Currency	object	Currency in use
DestCity	object	City of the last stop location
DestCountry	object	Country of the last stop location
DestPostal	object	Postal of the last stop location
DestState	object	State of the last stop location
Detention	float64	Detention fee
Fuel	float64	Fuel fee
Handling	float64	Handling fee
Layover	float64	Layover fee
LeanID	int64	The ID
LoadMiles	int64	Total miles traveled
Mode	object	TL, MTL, LTL, IM
OrderType	object	MSC, ORT, STO
OriginCity	object	City from which the carrier move originates.
OriginCountry	object	Country from which the carrier move originates.
OriginID	object	ID from which the carrier move originates.
OriginState	object	State from which the carrier move originates.
OtherAccessorial	float64	The charge of other Accessorial fee
PaidDate	datetime64	Date of payment
RichOrder	object	The number of order id
ShipDate	datetime64	Date of shipment
StopCharges	float64	Charges for stops
Stops	int64	Number of stops
TotalAccessorial	float64	The total charge of another Accessorial fee
TotalCharges	float64	Total charges
Weight	int64	Total weight of the inventory on the carrier move.
source-test	object	The initial city and state names (TransPlace database)
destination-test	object	The destination city and state names (TransPlace database)
OrigPostal	object	Postal from which the carrier move originates.
Per Mile Charge	float64	Per Mile Charge
Market Rates- DS DAT	float64	(TransPlace database)
Transportation Factor	float64	Market Rates/Base Rates

Table 3.1: Interpretation of data items

3.2 Processing internal data

a) Clean the data step by step:

1. Select useful columns

2. Delete the wrong data which TotalCharches $<$ Additive factor

3. Delete the rows which have null values

4. Select my mode is TL

b) Adding the required new columns and splitting the existing column entries

1. Count the number of rows with the same origin and destination, and name them the "Frequency" column item. And the update is still the $df4$-TL data set.

2. Divide the ShipDate into years and months(to analyze by season)

3. Add new column named$Fuel_Fee = LoadMiles * PerMileCharge$

3.3 Processing external data

a) The data source

The data is collected by the Census Bureau, whichâs mission is to serve as the nation's leading provider of quality data about its people and economy. b) Add data

1. Add demographic data to the dataset (Source: U.S. Census Bureau, 2018 American Community Survey 1-Year Estimates; tableâs name: ACS DEMOGRAPHIC AND HOUSING ESTIMATES) named Total-pop and whichâs unit is thousands people/city.

2. Add employment data to the dataset (Source: U.S. Census Bureau, 2018 American Community Survey 1-Year Estimates; tableâs name: EMPLOYMENT STATU) named Pop-emp and whichâs unit is thousands people/city.

3.4 Preliminary visualization of data

a) Statistics of the mathematical attributes of each column item

	BaseRate	Cases	Cube	Detention	Fuel	Handling	Layover	LoadMiles	OtherAccessorial	StopCharges
count	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000
mean	985.638838	1481.767803	1591.680244	0.080513	180.210472	27.451641	2.704091	552.000180	6.058961	1.449543
std	777.841643	902.744175	846.815536	1.444385	191.113032	78.821067	35.695591	530.941739	48.822764	23.021107
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	4.000000	-55.000000	0.000000
25%	320.000000	1084.000000	1072.000000	0.000000	17.760000	0.000000	0.000000	199.000000	0.000000	0.000000
50%	812.700000	1491.000000	1814.000000	0.000000	142.020000	0.000000	0.000000	422.000000	0.000000	0.000000
75%	1340.000000	1848.000000	2271.000000	0.000000	257.750000	0.000000	0.000000	747.000000	0.000000	0.000000
max	6438.470000	6675.000000	5993.000000	200.000000	1156.740000	750.000000	1800.000000	2976.000000	4655.000000	475.000000

Figure 3.1: Statistics of the mathematical attributes of each column item (1)

TotalAccessorial	TotalCharges	Weight	Per Mile Charge	Rates- DS, DAT	Transportation Factor	Frequency	year	month	Fee_fule
55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000	55483.000000
217.955221	1203.579713	27919.285277	4.303226	923.523643	1.117289	566.850260	2018.911162	6.517420	1203.581059
225.442222	974.197106	12979.286327	5.388871	814.058349	0.801241	685.944064	0.284576	3.419785	974.102541
-37.240000	0.230000	0.000000	0.000000	0.848700	0.000000	1.000000	2017.000000	1.000000	0.000000
24.420000	332.600000	21973.000000	1.870000	245.000000	0.820000	77.000000	2019.000000	4.000000	332.640000
168.330000	1008.000000	30830.000000	2.490000	831.660000	1.060000	275.000000	2019.000000	7.000000	1008.000000
302.480000	1644.720000	39633.000000	3.620000	1344.420000	1.270000	733.000000	2019.000000	10.000000	1645.000000
4655.000000	7500.000000	74623.000000	81.630000	5581.620000	17.140000	2617.000000	2019.000000	12.000000	7488.090000

Figure 3.2: Statistics of the mathematical attributes of each column item (2)

b) Find a linear relationship between BaseRate and TotalCharges

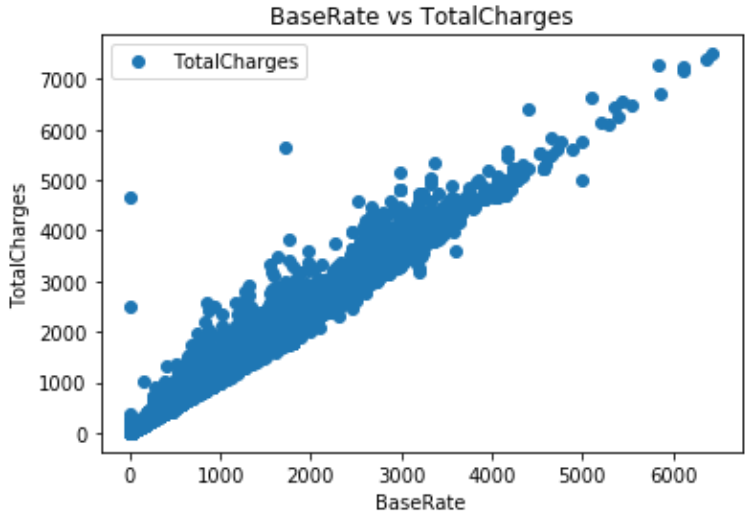


Figure 3.3: Relationship between BaseRate and TotalCharges (1)

c) Observe the overall distribution of TotalCharges

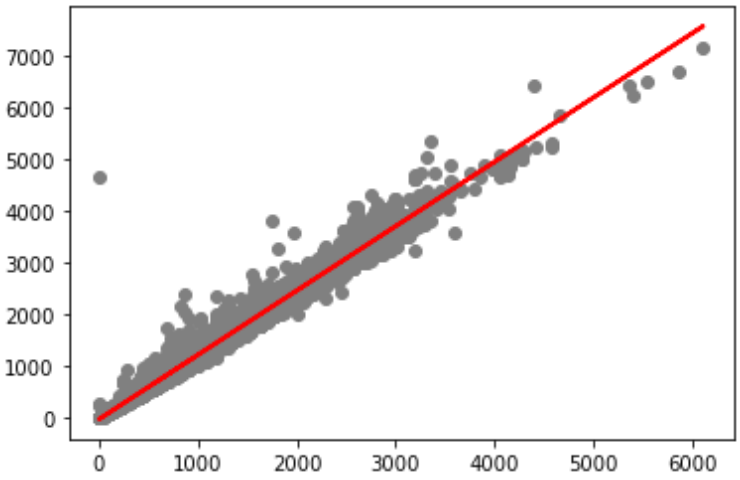


Figure 3.4: Relationship between BaseRate and TotalCharges (2)

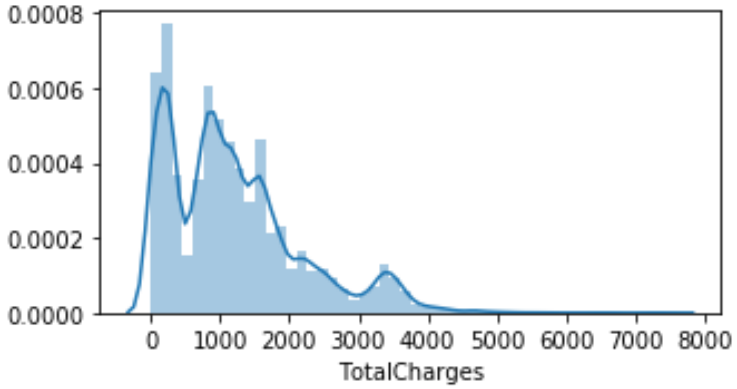


Figure 3.5: Distribution of TotalCharges

3.5 Preliminary conclusions from observational data

It can be seen from the data that there is a linear relationship between TotalCharge and BaseRate. To predict TotalCharge, we must know the influence of other factors on TotalCharge and BaseRate.

Since there is a linear relationship between them, the method I choose is a multiple linear regression model.

Chapter 4

Result

4.1 Explore the relationship between TotalCharge and BaseRate, Detention, Fuel, Handling, Layover, StopCharges, OtherAccessorial, Fuel-Fee

I split 80 % of the data to the training set while 20% of the data to test. The test-size variable is where we specify the proportion of the test set. After splitting the data into training and testing sets, finally, the time is to train the algorithm. For that, I need to import Linear Regression class, instantiate it, and call the fit() method along with our training data. In the case of multi-variable linear regression, the regression model has to find the most optimal coefficients for all the attributes.

a) intercept slope the intercept: 0.1466980567831797 the slope: [0.91759522 0.08795729 0.91816986 0.91770732 0.91778511 0.91762806 0.91777436 0.08223755]

	Coefficients
BaseRate	0.917595
Detention	0.087957
Fuel	0.918170
Handling	0.917707
Layover	0.917785
StopCharges	0.917628
OtherAccessorial	0.917774
Fee_fule	0.082238

Figure 4.1: The coefficients of the independent variables

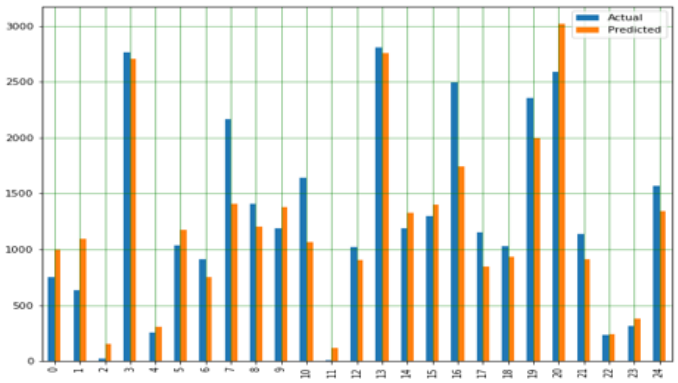


Figure 4.2: Comparison of Actual and Predicted value (25 records)

b) Criteria for evaluating models r squared: 0.9999998254420235 Mean Absolute Error: 0.15760308010153284 Mean Squared Error: 0.16808910095386995 Root Mean Squared Error: 0.4099867082648777

c) Conclusion TotalCharges=0.99BaseRate+Fuel+0.99Handling+Layover+0.99StopCharges +0.99OtherAccessorial+0.09fee-fule+0.088Detention+0.187

4.2 Consider the appropriate parameters

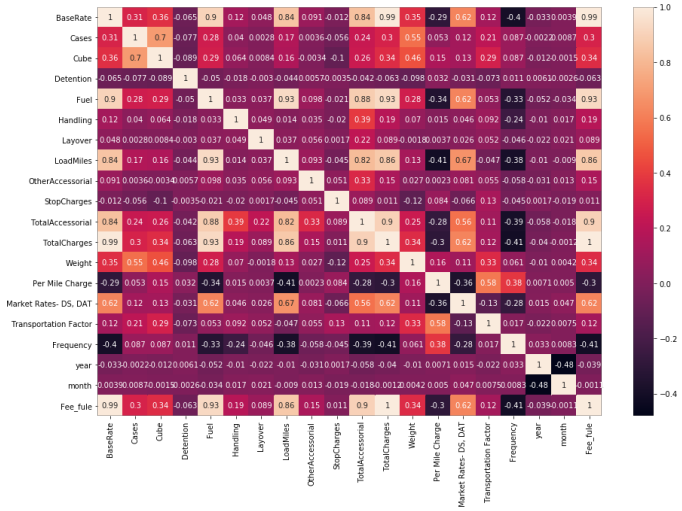


Figure 4.3: The correlation matrix for the columns with float64 or int64 data type of Internal data

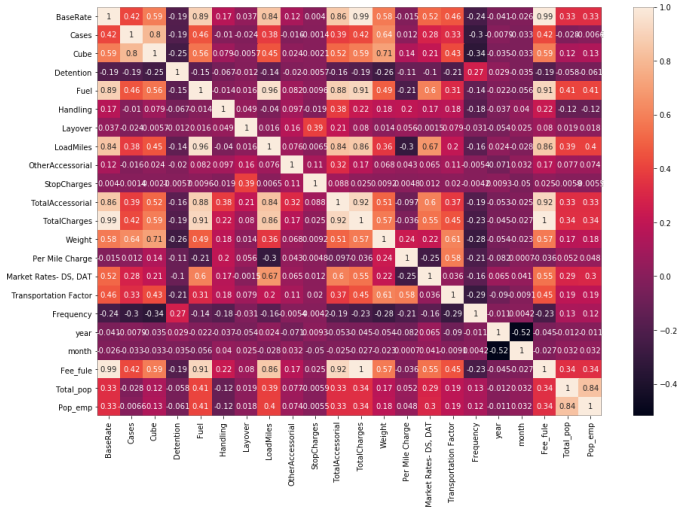


Figure 4.4: The correlation matrix for the columns with float64 or int64 data type of External data

From the above two graphs, the factors with high correlation can be considered into the formula. The way I choose the independent variable is based on the correlation matrix, I

chose factors with a correlation of more than 0.2 as independent variables.

4.3 Data processing flow results

Figure 4.5: Internal & External Data flow chart (the circle represents the data, the rectangle represents the formula)

According to the above data flow chart, in the overall data processing process, the data set is divided into four data streams of different seasons, namely spring, summer, autumn, and winter, and 24 formulas will be generated.

In conclusion, there will be 4 formulas for calculating the total charge generated by internal data and 4 formulas for calculating the total charge generated by external data. By comparing the two, we can know whether the added external data improves the accuracy of the formula.

4.3.1 The influence of each internal factor on BaseRate (Take spring as an example)

a) Explore the relationship between BaseRate and Cases, Cube, Fuel, LoadMiles, Weight, Frequency, Per Mile Charge the intercept: 184.4920513036459 the slope: [-0.03860693 0.10731449 3.18461492 -0.03252795 0.00551237 -0.15577762 2.23545167]

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

Table 4.1: The coefficients of the independent variables

b) Criteria for evaluating models

r squared : 0.8618312258690974

Mean Absolute Error: 198.14514326691926

Mean Squared Error: 80048.36288596975

Root Mean Squared Error: 282.9281938689917

c) Conclusion

$BaseRate = -0.04cases + 0.1cube + 3.18fuel - 0.03LoadMiles - 0.15frequency + 2.23PerMileCharge + 184.5$

4.3.2 The influence of each internal factor on Fuel (Take spring as an example)

$Fuel_Fee = LoadMiles * PerMileCharge$

a) Explore the relationship between Fuel and Cases, Cube, LoadMiles, Weight, Frequency, Fee-fule

the intercept: -80.56416836123549

the slope: [0.00299093 0.00383216 0.20816467 0.00075002 0.01747389 0.09783572]

	Coefficients
Cases	0.002991
Cube	0.003832
LoadMiles	0.208165
Weight	0.000750
Frequency	0.017474
Fuel-Fee	0.097336

Table 4.2: The coefficients of the independent variables

b)Criteria for evaluating models

r squared : 0.9474702631854132

Mean Absolute Error: 29.87232435801521

Mean Squared Error: 2104.644144488171

Root Mean Squared Error: 45.8764007359794

c) Conclusion

$Fule = 0.21LoadMiles + 0.017Frequency + 0.09Fuel - Fee - 80.56$

Chapter 5

Conclusion

5.1 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

Table 5.1: The amount of data in each data set

5.2 Formula composition

a) Total formula explored from internal data

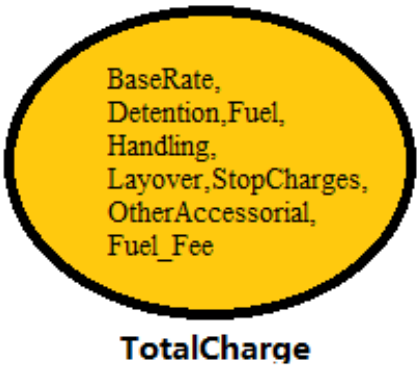


Figure 5.1: The variables of the TotalCharge formula

b) Select BaseRate and Fuel as the dependent variables

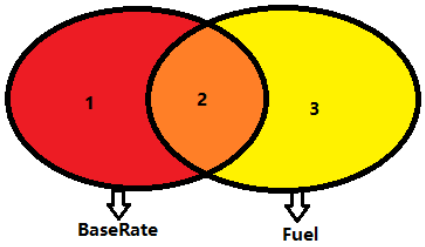


Figure 5.2: The variables of the BaseRate and Fuel formulas

The selected independent variable:

1: Per Mile Charge

2: Cases, Cube, LoadMiles, Weight, Frequency

3: Fuel-Fee

5.3 Formula about TotalCharge (every season)

All the results were based on 80% of the data to the training set while 20% of the data to test.

1. Spring

a)The intercept & the slope

the intercept: -0.00858844464255526

the slope: [0.99491946 0.99245067 0.99490278 0.9949145 0.99490723 0.99492336 0.99491959 0.00508839]

b)Coefficients

	Coefficients
BaseRate	0.994919
Detention	0.992451
Fuel	0.994903
Handling	0.994914
Layover	0.994907
StopCharges	0.994923
OtherAccessorial	0.994920
Fuel-Fee	0.005088

c) R squared&Errors

r squared: 0.9999999968852412

Mean Absolute Error: 0.008116283693174827

Mean Squared Error: 0.002900086904719412

Root Mean Squared Error: 0.053852454955363

2. Summer

a)The intercept & the slope

the intercept: 0.1368782037513938

the slope: [0.91762702 0.03303517 0.91815733 0.91774141 0.91777018 0.91763261 0.91765564 0.08221802]

b)Coefficients

	Coefficients
BaseRate	0.917627
Detention	0.033035
Fuel	0.918157
Handling	0.917741
Layover	0.917770
StopCharges	0.917633
OtherAccessorial	0.917656
Fuel-Fee	0.082218

c) R squared&Errors

r squared: 0.9999998394755417

Mean Absolute Error: 0.14363521814561236

Mean Squared Error: 0.15845196734088754

Root Mean Squared Error: 0.3980602559172261

3.Fall

a)The intercept & the slope

the intercept: -0.000555746352802089

the slope: [9.99801814e-01 9.99938113e-01 9.99799873e-01 9.99801549e-01 9.99801307e-01 9.99802084e-01 9.99801068e-01 1.98756229e-04]

b)Coefficients

	Coefficients
BaseRate	0.999802
Detention	0.999938
Fuel	0.999800
Handing	0.999802
Layover	0.999801
stopCharges	0.999802
OtherAccessorial	0.999801
Fuel-Fee	0.000199

c)R squared&Errors

r squared : 0.9999999999998089

Mean Absolute Error: 0.0003215841057208946

Mean Squared Error: 1.6549852748580476e-07

Root Mean Squared Error: 0.00040681510233250286

4.Winter

a)The intercept & the slope

the intercept: -0.00858844464255526

the slope: [0.99491946 0.99245067 0.99490278 0.9949145 0.99490723 0.99492336 0.99491959 0.00508839]

b)Coefficients

	Coefficients
BaseRate	0.994919
Detention	0.992451
Fuel	0.994903
Handling	0.994914
Layover	0.994907
StopCharges	0.994923
OtherAccessorial	0.994920
Fuel-Fee	0.005088

c)R squared&Errors

r squared: 0.9999999968852412

Mean Absolute Error: 0.008116283693174827

Mean Squared Error: 0.002900086904719412

Root Mean Squared Error: 0.0538524549553631

5.4 Formula of BasteRate adds a comparison of internal and external factors

All the results were based on 80% of the data to the training set while 20% of the data to test.

1) Spring

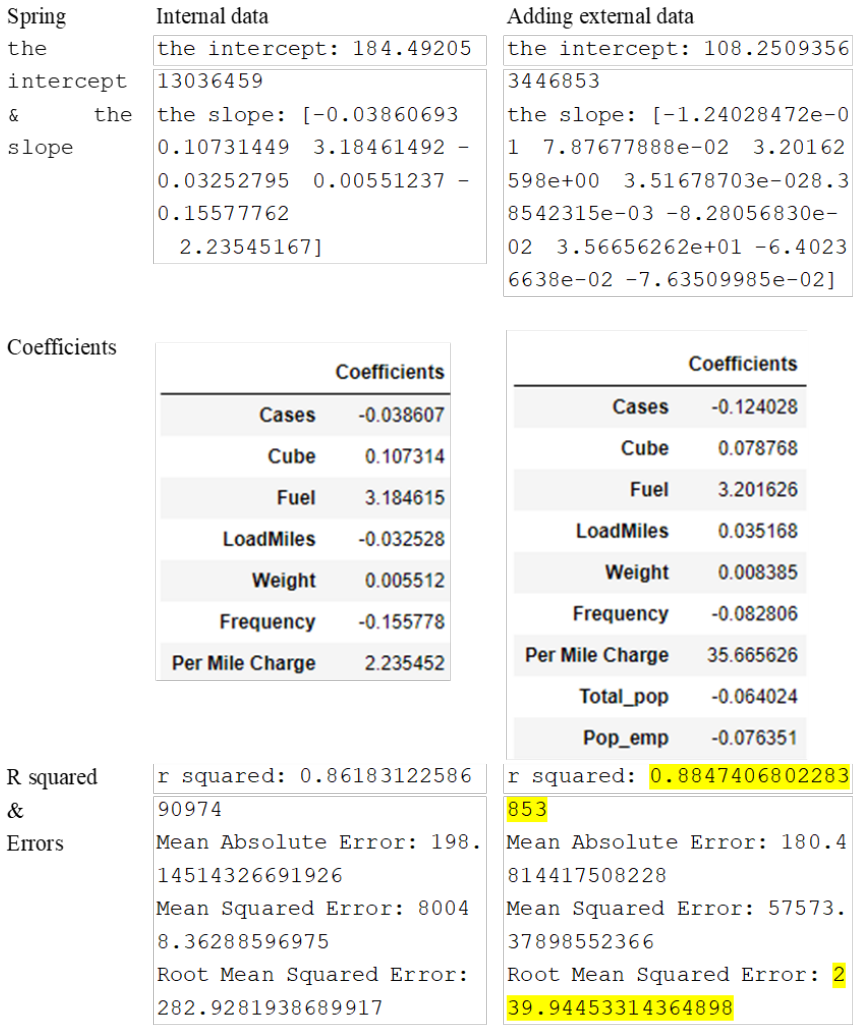


Figure 5.3: The spring formula adds a comparison of internal and external variables

2)Summer

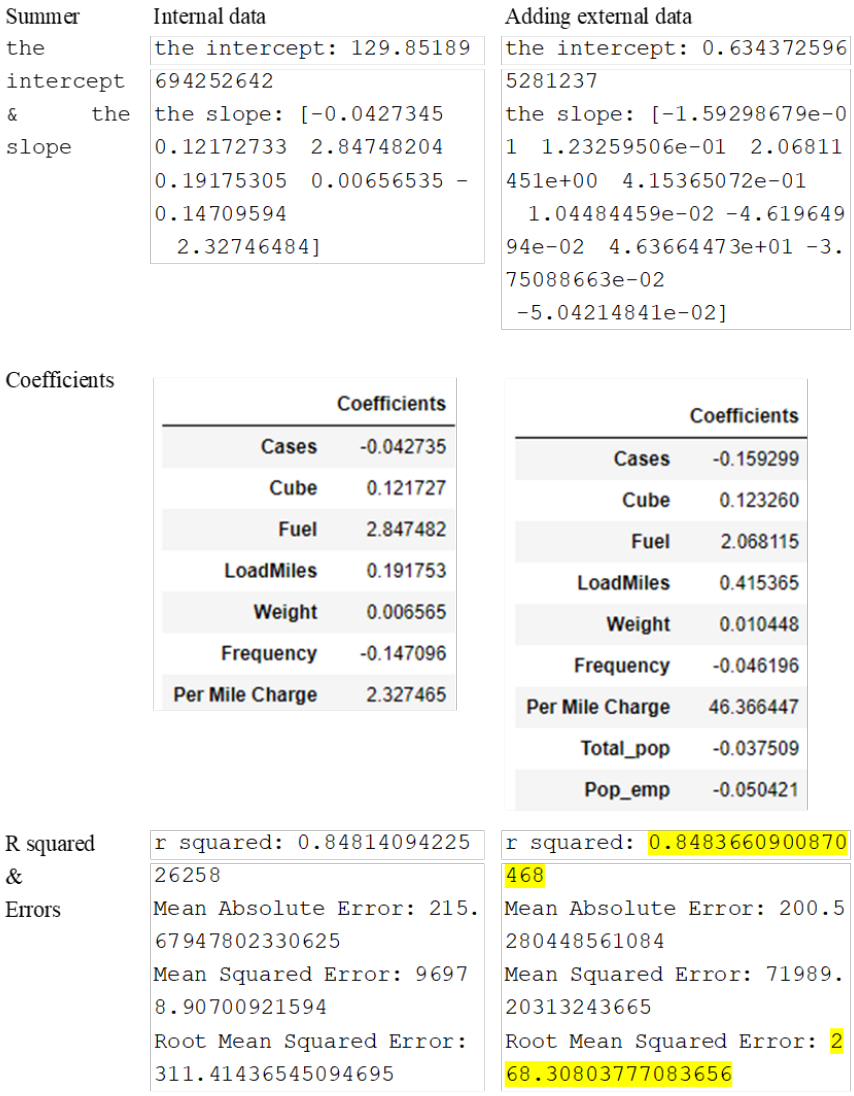


Figure 5.4: The summer formula adds a comparison of internal and external variables

3) Fall

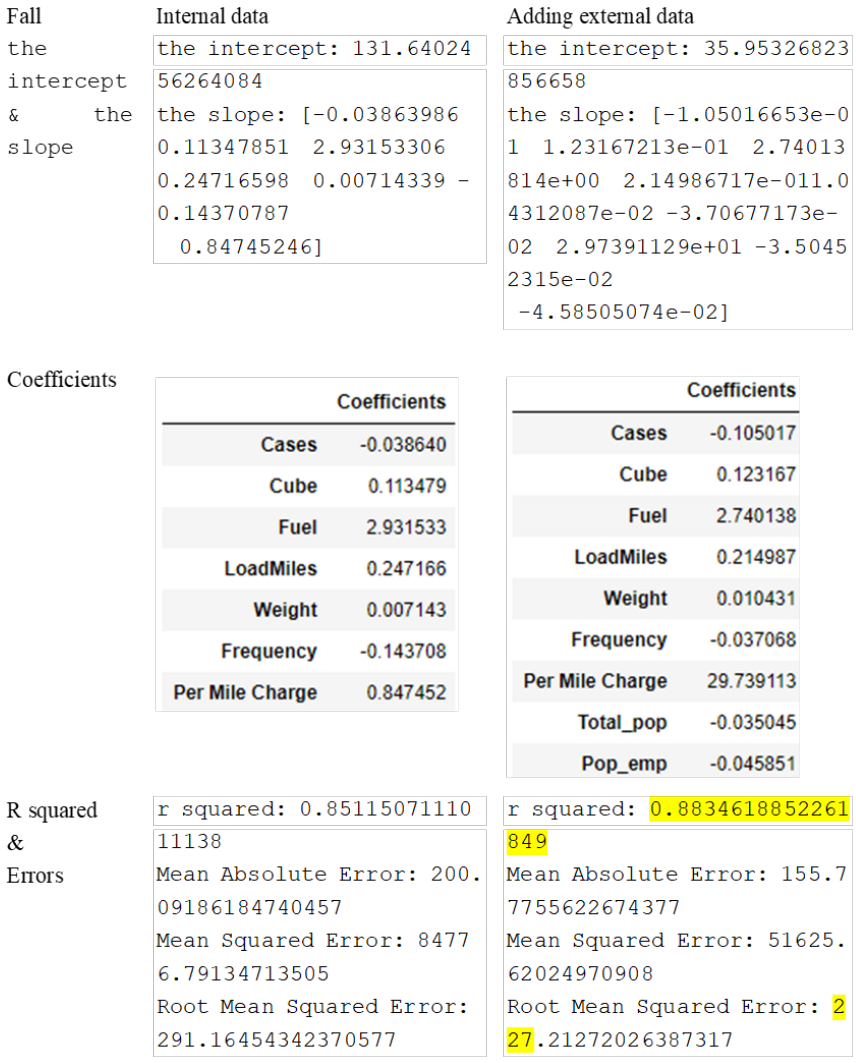


Figure 5.5: The Fall formula adds a comparison of internal and external variables

4) Winter

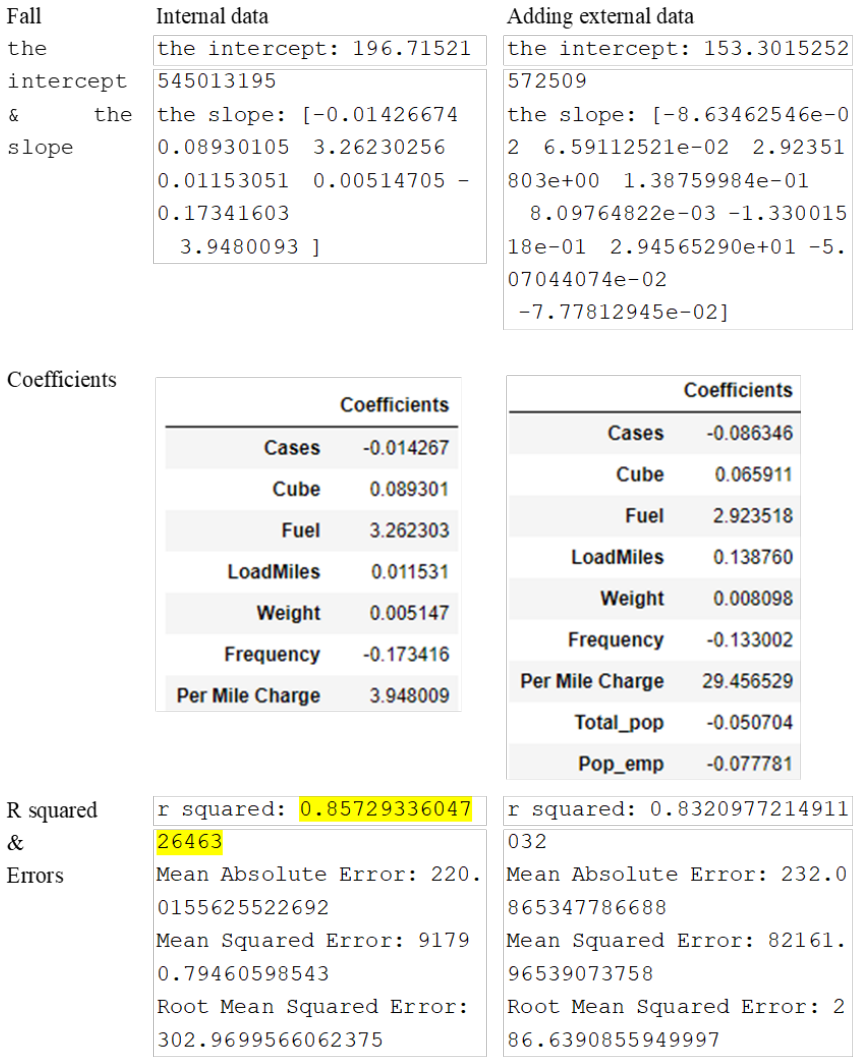


Figure 5.6: The Winter formula adds a comparison of internal and external variables

5.5 Formula about Fuel (every season)

All the results were based on 80% of the data to the training set while 20% of the data to test.

1)Spring

Spring Internal data

the intercept & the slope

```
the intercept: -80.56416836123549
the slope: [0.00299093 0.00383216 0.20816467 0.00075002 0.01747389 0.09783572]
```

Coefficients

	Coefficients
Cases	0.002991
Cube	0.003832
LoadMiles	0.208165
Weight	0.000750
Frequency	0.017474
Fuel_Fee	0.097836

R squared & Errors

```
r squared: 0.9474702631854132
Mean Absolute Error: 29.87232435801521
Mean Squared Error: 2104.644144488171
Root Mean Squared Error: 45.8764007359794
```

Figure 5.7: The spring formula of fuel

2)Summer

Summer Internal data

the intercept & the slope

```
the intercept: -69.18872752618549
the slope: [0.00348998 0.00193927 0.19452377 0.00062104 0.0151632  0.08930827]
```

Coefficients

	Coefficients
Cases	0.003490
Cube	0.001939
LoadMiles	0.194524
Weight	0.000621
Frequency	0.015163
Fuel_Fee	0.089308

R squared & Errors

```
r squared: 0.9324426718813983
Mean Absolute Error: 32.011791638082826
Mean Squared Error: 2430.535489707883
Root Mean Squared Error: 49.30046135390502
```

Figure 5.8: The summer formula of fuel

3) Fall

Fall the intercept & the slope

Internal data

```
the intercept: -58.682763225402
the slope: [0.00132269 0.00245676 0.20268993 0.00
056214 0.01234827 0.06736189]
```

Coefficients

	Coefficients
Cases	0.001323
Cube	0.002457
LoadMiles	0.202690
Weight	0.000562
Frequency	0.012348
Fuel_Fee	0.067362

R squared & Errors

```
r squared: 0.9406554004848217
Mean Absolute Error: 23.08946242580253
Mean Squared Error: 1652.1739321000161
Root Mean Squared Error: 40.64694246926841
```

Figure 5.9: The fall formula of fuel

4)Winter

Winter the intercept & the slope

Internal data

```
the intercept: -76.596036982372
the slope: [-0.00036385  0.00706889  0.22291372
0.00084974  0.01611731  0.07923226]
```

Coefficients

	Coefficients
Cases	-0.000364
Cube	0.007069
LoadMiles	0.222914
Weight	0.000850
Frequency	0.016117
Fuel_Fee	0.079232

R squared & Errors

```
r squared: 0.9477911245200481
Mean Absolute Error: 29.20430573312556
Mean Squared Error: 2148.130413807043
Root Mean Squared Error: 46.34792782646321
```

Figure 5.10: The winter formula of fuel

5.6 Final Conclusion

1) Spring

$$TotalCharge = Detention + Handling + Layover + StopCharge + OtherAccessorial + 0.015Fuel_Fee - 0.12Cases + 0.08Cube + 0.635LoadMiles + 0.010Weight - 0.023Frequency + 35.67PerMileCharge - 0.064Total_pop - 0.076Pop_emp - 149.542$$

2) Summer

$$TotalCharge = Detention + Handling + Layover + StopCharge + OtherAccessorial + 0.16Fuel_Fee - 0.16Cases + 0.12Cube + 0.82LoadMiles + 0.011Weight - 0.016Frequency + 46.37PerMileCharge - 0.038Total_pop - 0.035Pop_emp - 142.59$$

3)Fall

$$TotalCharge = Detention + Handling + Layover + StopCharge + OtherAccessorial + 0.067Fuel_Fee - 0.11Cases + 0.12Cube + 0.61LoadMiles + 0.011Weight - 0.017Frequency + 29.74PerMileCharge - 0.035Total_pop - 0.046Pop_emp - 85.52$$

4)Winter

$$TotalCharge = Detention + Handling + Layover + StopCharge + OtherAccessorial + 0.084Fuel_Fee - 0.01Cases + 0.11Cube + 0.61LoadMiles + 0.0076Weight - 0.143Frequency + 3.95PerMileCharge - 53.00$$

Appendix A

This is Appendix

A.1 Part of the core code

(Take for example the TotalCharge)

$X = df4_TL[['BaseRate','Detention','Fuel','Handling','Layover','StopCharges',$

$'OtherAccessorial','Fuel_Fee']]$

$y = df4_TL['TotalCharges'].values$

Next, we split 80% of the data to the training set while 20% of the data to test set using below code.

$X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)$

Now lets train our model.

$regressor = LinearRegression()$

$regressor.fit(X_train, y_train)$

To retrieve the intercept:

$print("theintercept : ", regressor.intercept_)$

For retrieving the slope:

$print("theslope : ", regressor.coef_)$

in the case of multivariable linear regression, the regression model has to find the most

optimal coefficients for all the attributes.

To see what coefficients our regression model has chosen, execute the following script:

$coeff_df = pd.DataFrame(regressor.coef_, X.columns, columns = ['Coefficients'])$

$coeff_df$

$y_pred = regressor.predict(X_test)$

from sklearn.metrics import accuracy$_score, r2_score$

print('r squared :',r2$_score(y_test, y_pred))$

$print('Mean Absolute Error :', metrics.mean_absolute_error(y_test, y_pred))$

$print('Mean Squared Error :', metrics.mean_squared_error(y_test, y_pred))$

$print('Root Mean Squared Error :', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))$

Bibliography

[1] Course List for 1997-98, Department of Statistics and Data Science, Yale University, Multiple Linear Regression http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm.

[2] Shivamsaraswat, Lalitshankarch, Python Language Introduction, https://www.geeksforgeeks.org/python-language-introduction/.

[3] Fabian Pedregosa; GaÃl Varoquaux; Alexandre Gramfort; Vincent Michel; Bertrand Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vincent Dubourg; Jake Vanderplas; Alexandre Passos; David Cournapeau; Matthieu Perrot; Ãdouard Duchesnay (2011). ”Scikit-learn: Machine Learning in Python”. Journal of Machine Learning Research. 12: 2825â2830

[4] Mike Driscoll, Jupyter Notebook: An Introduction, https://realpython.com/jupyter-notebook-introduction/