ISE 5103 Intelligent Data Analytics

Final Project

Daniel Carpenter, Sonaxy Mohanty, & Zachary Knepp

December 2022

Contents

1 General Data Prep

For general data preparation, please see conceptual steps below. See `.rmd` file for detailed code.

1.1 Creating the CSV Dataset

- Note that the original training and test data found here contains two zipped files totaling around 55,000 `.txt` files
- In order to convert this data into a usable format, we created a function that:
  - Reads all `txt` contained within a specified folder
  - Cleans whitespace, variable naming conventions, and converts `Time_Ordered` and `Time_Ordered_Picked` from HH:MM string time.
  - All variables are cast to their correct data types
  - Finally, the data is exported to a single CSV.
  - This function is applied to to the training and test data
  - This R file containing the function is located here
- Note that the function is not run within this file due to the time required to run the code. Since there are so many files, it takes a large amount of time.

1.2 Read Training and Test Data

- Read training and test data CSV files from GitHub
- Clean data to ensure each read variable has the correct data type (factor, numeric, Date, etc.)

1.3 Create `numeric` and `factor` *base* `data frames`

Make data set of `numeric` variables called `df.train.base.numeric`

Make data set of `factor` variables called `df.train.base.factor`

2 Data Understanding

Create a data quality report of `numeric` and `factor` data
Created function called `dataQualityReport()` to create factor and numeric QA report

2.1 Numeric Data Quality Report

Num_Numeric_Variables	Total_Observations
9	45593

variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Restaurant_Latitude	0	1.00	17.0	8.19	-30.91	12.9	18.6	22.7	31
Restaurant_Longitude	0	1.00	70.2	22.88	-88.37	73.2	75.9	78.0	88
Delivery_Location_Latitude	0	1.00	17.5	7.34	0.01	13.0	18.6	22.8	31
Delivery_Location_Longitude	0	1.00	70.8	21.12	0.01	73.3	76.0	78.1	89
Time_Order_Picked	0	1.00	18.3	4.54	8.25	16.0	19.5	21.8	24
Target_Variable	0	1.00	26.3	9.38	10.00	19.0	26.0	32.0	54
Time_Ordered	1731	0.96	18.1	4.54	8.17	15.8	19.3	21.7	24
Delivery_Person_Age	1854	0.96	29.6	5.82	15.00	25.0	30.0	35.0	50
Delivery_Person_Ratings	1908	0.96	4.6	0.33	1.00	4.5	4.7	4.9	6

2.2 Factor Data Quality Report

Num_Factor_Variables	Total_Observations
11	45593

variable	n_missing	complete_rate	n_unique	top_counts
Id	0	1.00	45593	100: 1, 100: 1, 100: 1, 100: 1
Delivery_Person_Id	0	1.00	1320	JAP: 67, PUN: 67, HYD: 66, JAP: 66
Vehicle_Condition	0	1.00	4	2: 15034, 1: 15030, 0: 15009, 3: 520
Type_Of_Order	0	1.00	4	Sna: 11533, Mea: 11458, Dri: 11322, Buf: 11280
Type_Of_Vehicle	0	1.00	4	mot: 26435, sco: 15276, ele: 3814, bic: 68
Festival	0	1.00	2	1: 45365, 0: 228
Name	0	1.00	45593	0: 1, 1: 1, 2: 1, 3: 1
Road_Traffic_Density	601	0.99	4	Low: 15477, Jam: 14143, Med: 10947, Hig: 4425
Weather_Conditions	616	0.99	6	Fog: 7654, Sto: 7586, Clo: 7536, San: 7495
Multiple_Deliveries	993	0.98	4	1: 28159, 0: 14095, 2: 1985, 3: 361
City	1200	0.97	3	Met: 34093, Urb: 10136, Sem: 164

2.3 Expected Approach for Cleaning Data

2.3.1 Missingingness

- To handle missingness, we will likely take the following approach for `numeric` and `factor` data:
  - `Numeric`: Impute missing values using predictive mean matching with the `mice` package
  - `Factor`: Leverage k-nearest neighbors to impute missing factor data. This is likely possible because there is not a significant portion of the factor data that is missing, so it should not be computationally extensive.

2.3.2 Outliers

1. We will prioritize limiting outliers of the target variable.
2. We will also analyze each numeric independent variable to discover any outliers. If there are few outliers, then we will likely omit that data. If outliers persist in a large portion of the data, then we will limit the removal of outlying data.

2.3.3 Skews

2.3.3.1 Target Variable

- The below exploratory analysis shows that the $Target\_Variable$ is skewed.
- However, see that the $\log(Target\_Variable)$ is close to being normal, so we will not need to transform this data.

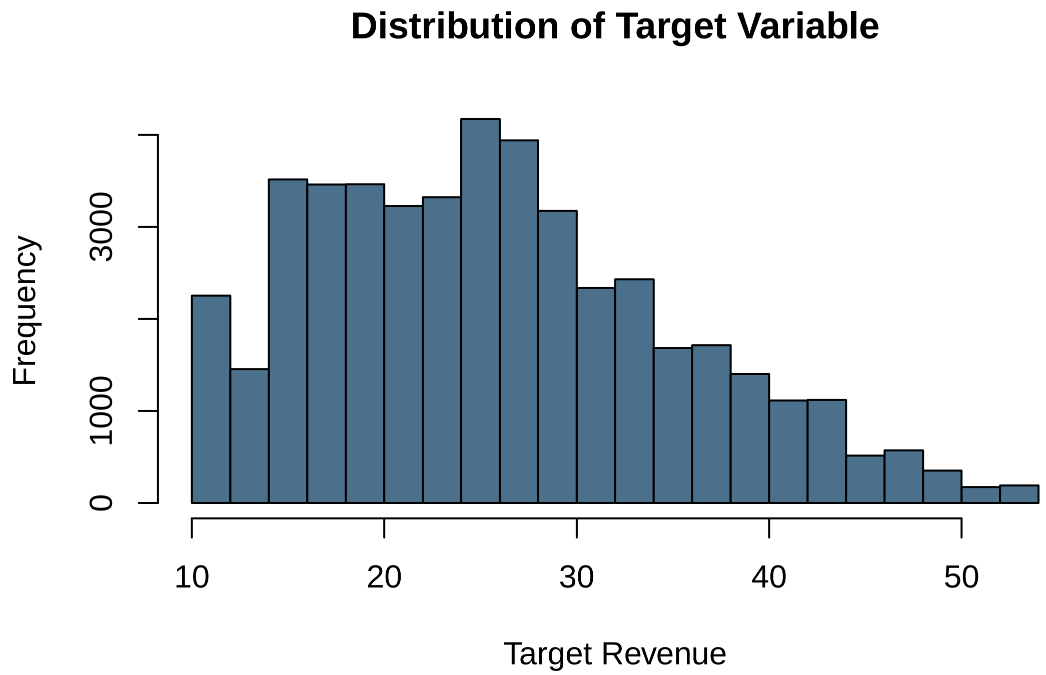
2.3.3.2 Other Numeric Predictors If other numeric variables are highlight skewed (within the test and train data), then we will likely use the `boxcox` function to normalize the test and training variables associated.

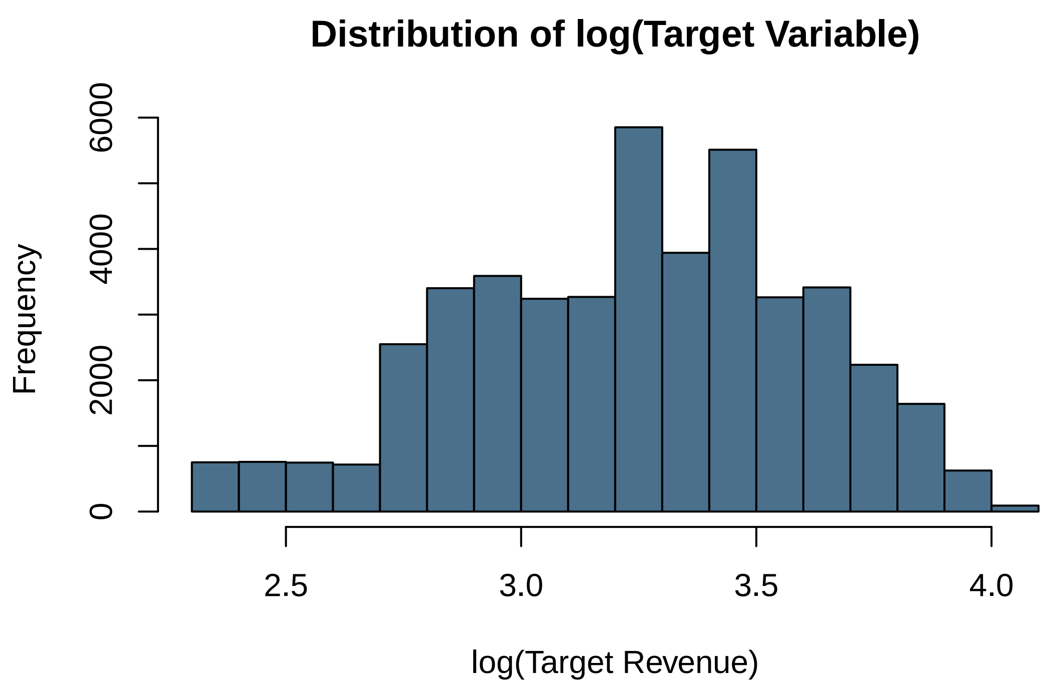
2.3.4 Factors

Since the factor data contains few unique values, we will not need to factor lump the data. If there was many unique values in related variables, then we would factor lump to help fit the models more efficiently.

2.4 Exploratory Analysis and Visualizations

2.4.1 Analysis 1: Exploring the Target Variable





2.4.2 Analysis 2: Correlation of Numeric Data

2.4.2.1 Correlation between numeric features in the dataset