Jumpei Fujimoto

DSCI 301

Prof. Teka Kubrom

Dec 12 2019

Final project

In this assignment, I will do the following; Introduction, Exploratory Data Analysis, fit a model, model diagnostics, Give prediction of the model, and Summary.

1. Introduction

I wanted to predict the “Sales of the product in the particular store” on Black Friday. I got the data from <https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/>. They have train and test data set, and train data set has target variable( Item\_Outlet\_Sales), but test data has not the target variable.

Data detail is following:

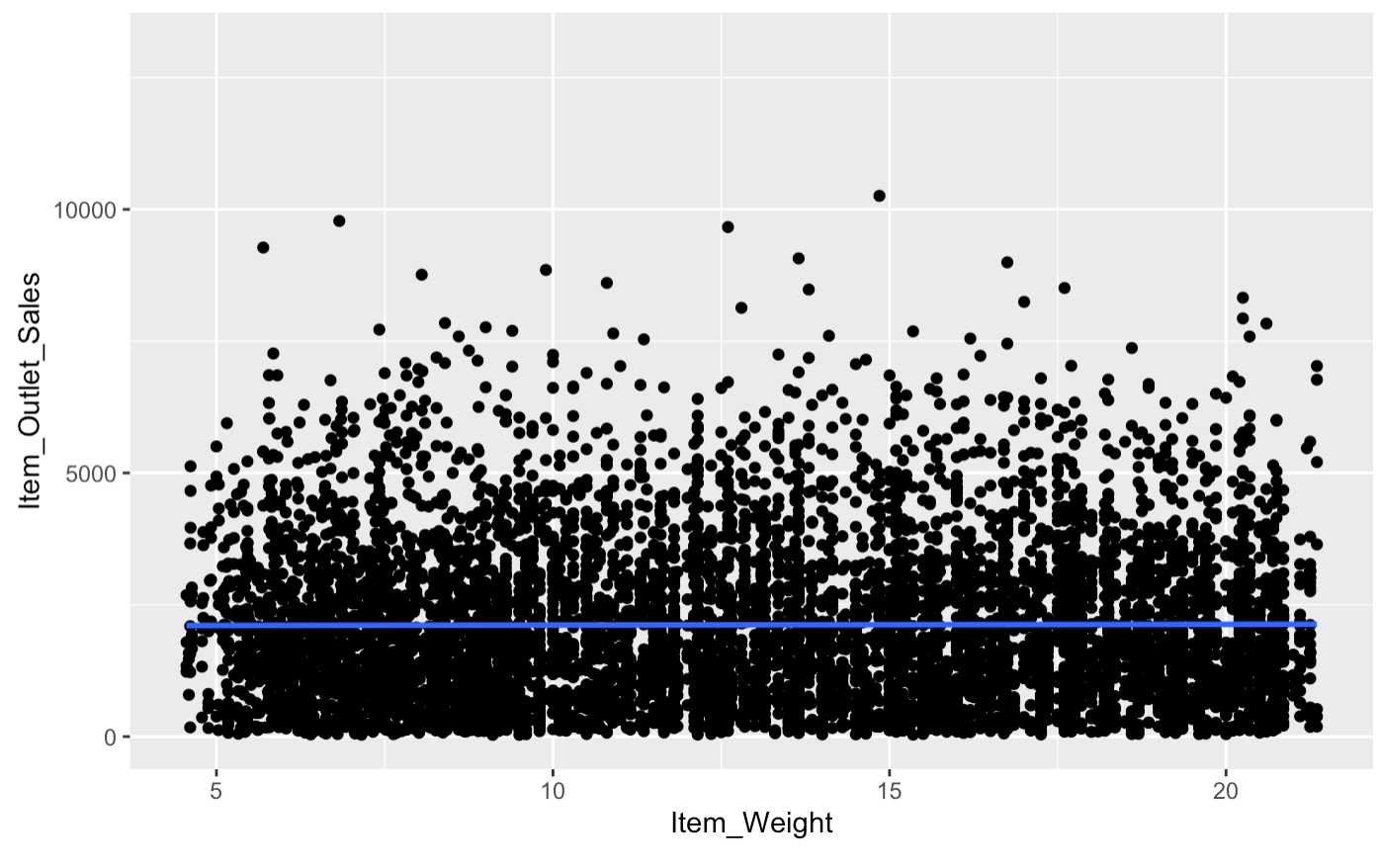
A screenshot of a cell phone

Description automatically generated

1. Exploratory Data Analysis and Fit a model

First, I saw first 6 rows and I realized that Item\_Identifier does not help to predict for target variable. Then I dropped that column and it called df. Next, I saw the summary of the data and recognize what is the categorical variable and numerical variable. Then I checked all categorical variable how they make label.

Next, I plotted as x = Item\_Weight and y = Item\_Outlet\_Sales.



We cannot see any pattern from this graph. And then check the p-value for making sure that this model is not significant. So, I have to find another variable which has a good relationship for target variable.

Then, I make the linear model as y = target variable and x = every column.

A close up of text on a white background

Description automatically generated

From this graph, I could see the significant p-value for the target variable. Then, I focused these variables, Item\_MPR and Outlet\_IdentifierOUT018, on more. So, let’s make the plot as x = Item\_MPR and y = target variable.A picture containing photo

Description automatically generated

This plot seems a good relationship. As Item MPR goes heavier, Item\_Outlet\_Sales increases. Thus, let’s keep this model for base model. Next, check how much each Outlet Identifier has for bar plot. Then put Outlet Identifier for the model. A close up of text on a white background

Description automatically generated

This model has a lot of improvement from the previous model. After I found this model, I used the other variable to put in model.

A screenshot of text

Description automatically generated

But this model is the best one, and tried to the other one as much as I can, but I could not find so I keep the previous model to diagnostics and predict

1. Model diagnostics

A screenshot of a social media post

Description automatically generated

As prediction gets higher, residual divorces, which means that this model did not predict properly, but I will keep this model because this is the best model that I could find in my level.

1. Give prediction of the model

I predict the Item\_Outlet\_Sales of a model was 1098.838 with a Item\_MRP of 60 and Outlet\_Identifier is OUT049.

This equation is y = -1851.2591+ 60\* 15.561+ 2016.4706.

1. Summary

To be honest, this prediction was really challenging, and I am not satisfied about the result. But for the process to get the insight of the data, I could enjoy and tried to do my best by using the skills what I learned in this class and actually improved the model. I could find the relationship item weight and store is the important to predict the sales of the product in black Friday. However, I strongly realized that I need to study more and improve my skills in the future. Anyway, I am satisfied what I learned in this class. Thank you for teaching this semester. I am really lucky to learn from you!