Big Mart Sales Prediction using Machine Learning

ML models developed by Ronit Randhawa at www.RealWorldAI.co summer, 2020

Problem Statement

The central problem for the Big Mart sales prediction problem was predicting the number of sales for a specific product depending on the outlet.

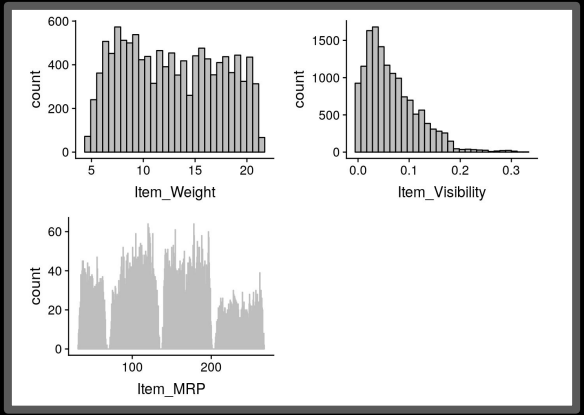
Data Import and Wrangling

Within the dataset, there contained a lot of missing, NA, and false unique values. I determined the missing and NA values through visualization or through mean of similar rows. For the false unique values, I simply changed them.

Methodology

The Methodology I used for this project was a linear regression model to predict the sales of a given item. I additionally added more features using feature selection to increase the accuracy of the model, and other features such as label encoding and one hot encoding.

Visualizations:



Item weight has little relevance and correlation here, but item MRP (maximum retail price) and item visibility have a clear and strong connection.

Algorithms Used

I simply used a linear regression algorithm to predict, although, many other algorithms such as Linear Discriminant Analysis (LDA) and Classification and Regression Trees (CART) can be used/tested in the future.

Challenges

A couple challenges I faced on this project include:

- One hot encoding
- Label encoding
- NA, missing, and false unique values
- Using visualization to impute missing values and checking outliers.

Significance

A linear regression model with high accuracy can help the retail sector and specific stores in increasing profits by returning highly probable data on which products will bring in high sales, which products will fail, and how much supply/inventory each store should keep on-hand of each product.

Conclusions:

I was successful in creating a final model with a good RMSE (1118), and a good R2 (0.85924). This project really helped me solidify my knowledge of data preparation, exploration, visualization, feature selection, and testing/training. I was able to understand and learn other techniques such as one hot encoding, label encoding, and other data wrangling methods.