Sales Prediction

for a large supermarket

Using Machine Learning

ML models developed by Lorenzo Martinelli
at www.RealWorldAI.co summer, 2020

Problem Statement

The key problem concerned determining an outlet's number of sales for a certain item.

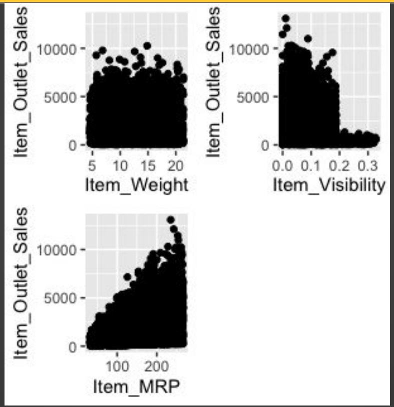
Data Import and Wrangling

There were multiple missing and NA values in the data which had to be determined through visualization (for categorical variables) or mean of similar rows (continuous). There were also many false unique values that had to be changed.

Methodology

I used a regression model to predict the sales of an item per outlet. Also, I created many new features and expanded other features using label and one hot encoding.

Visualizations



Item Weight matters little to item sales but lower visibility and higher maximum retail price have a clear correlation

Algorithms Used

I only used a linear regression algorithm for my predictions, but many other types of algorithms can be tested in the future.

Challenges

NAs/missing values
Repeat unique values for a column
Encoding (label + one hot)
Using common sense to change certain values

Significance

A good predictive model can improve the entire retail industry by providing information on what contributes to high sales and how much inventory of an item a store should have. This info can help maximize the profits of that business.

conclusions

I was able to create a final model and predictions with an RMSE of 1121. The creation of this model helped me understand the importance and value of multiple encoding techniques and the wide variety of data wrangling that must be done to achieve a strong and correct model.