Sales Prediction

Ryan Bolduan

Coding Dojo - Data Science 8.30 Cohort

Oct 10, 2021

Overview of Sales Prediction Project

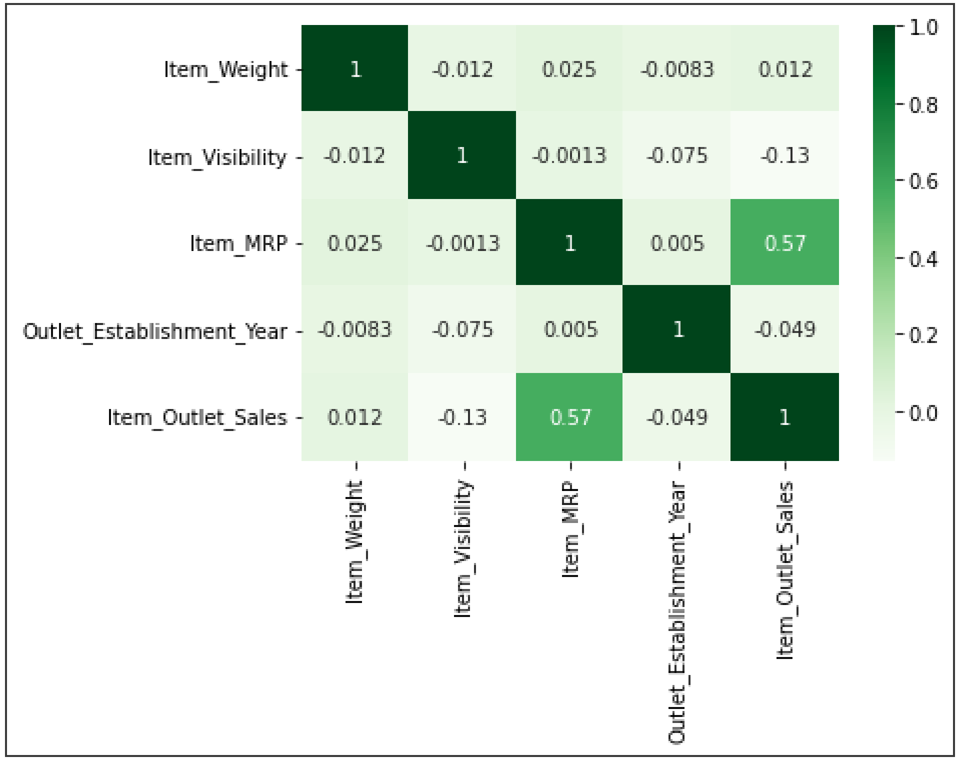
- Project goal was to predicting food item outlet sales
- Data was uploaded (see below)
- Data was compared with Data Dictionary to better understand the data

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.1380
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.2700
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3	Grocery Store	732.3800
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052

Data Cleaning and Exploratory Correlation

- Once the data was loaded, it was cleaned of:
  - Duplicates
  - Missing Data
  - Inconsistent Column Headers

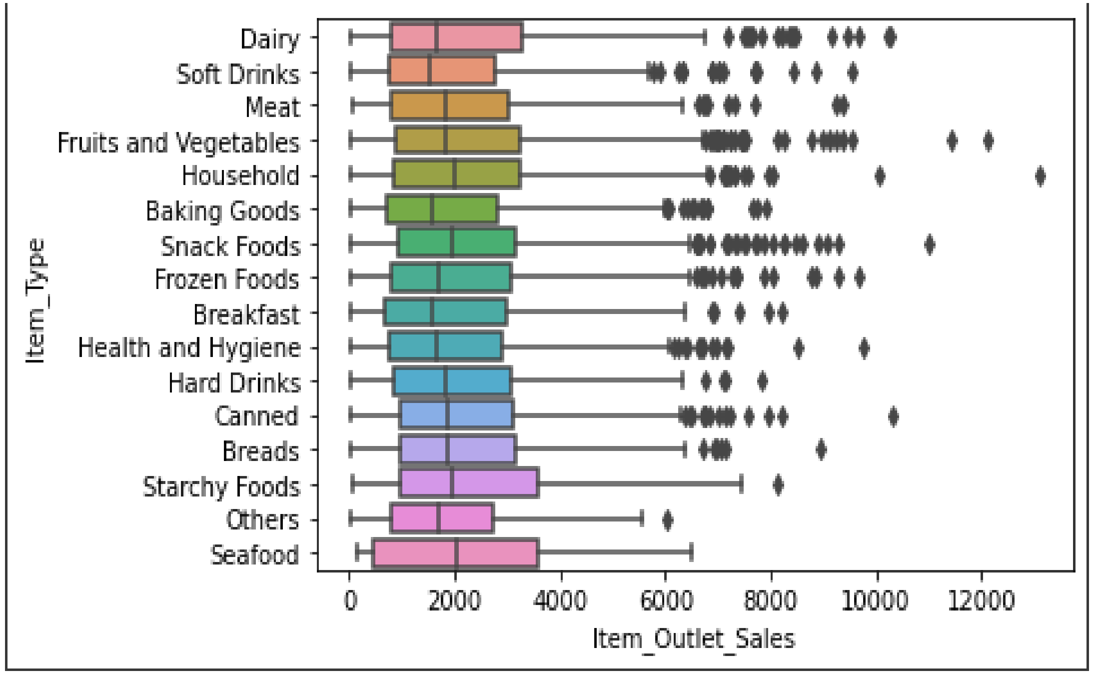
Once the data was cleaned, the data was analyzed using a heat map to look for trends that would help us identify correlations within the data that have an impact on sales.



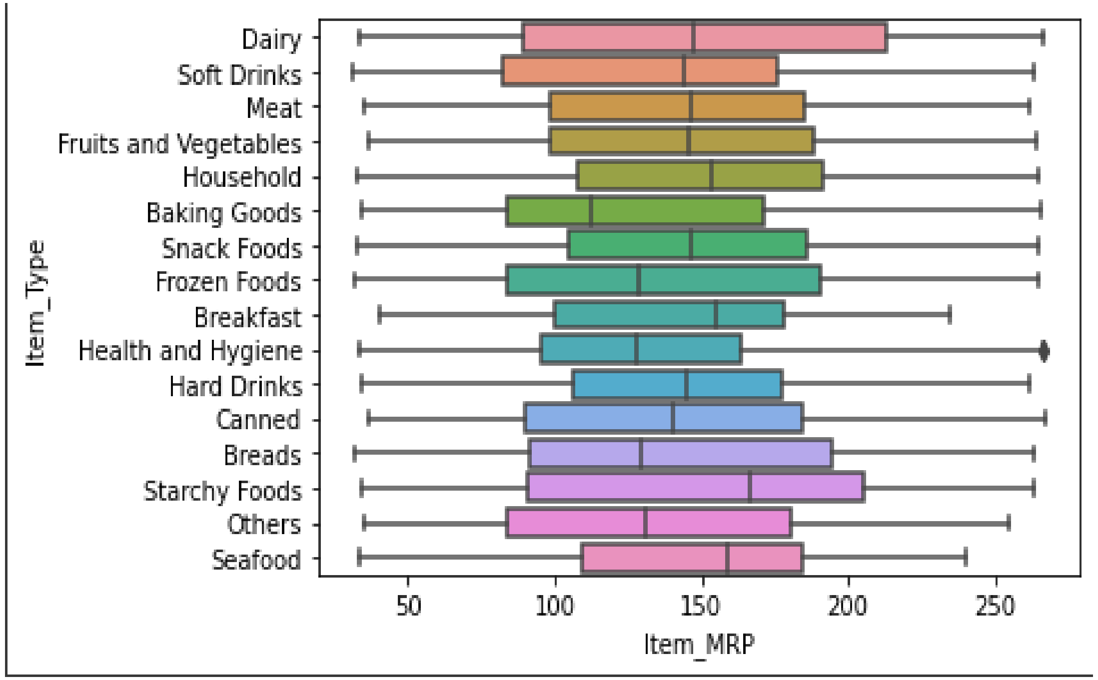
Heat map shows strongest correlation between Item_Outlet_Sales and Item_MRP (0.57)

Further explorations of relationship between Item_Outlet_Sales and Item_MRP, given the correlation on previous slide

Median Sales by Item Type did not provide obvious insights into sales predictions



Median max retail price did provide obvious visual insights into sales predictions



Given that the data visualizations did not offer any obvious insights into sales predictions, The next step was to turn to ML regression models to look for insights

The next step was to turn to ML regression models to look for insights

1) Linear Regression Model
2) Simple Decision Tree Model
3) Bagged Tree Model
4) Random Forest Model

Data was prepared so that the computer could perform ML models,

- one hot encoding

Data was then run through the above supervised learning models.

Random Forest Model, despite being overfit, yielded the best fit of the four models

Linear Regression r2 values (train/test)

0.6716976476073483
2.9682683729842308e+16

Bagged Trees r2 values (train/test)

0.9184773073967285

0.5265376546818851

Regression Tree r2 values (train/test)

0.6122318361448813
0.588890582401477

Random Forests r2 values (train/test)

0.9370377920925759

0.551683795847306

Findings

- Random forest model had the lowest variation of the other models (linear regression, decision tree and bagged trees)
  - Predictions were highly correlated on training data, but not strongly correlated on testing data

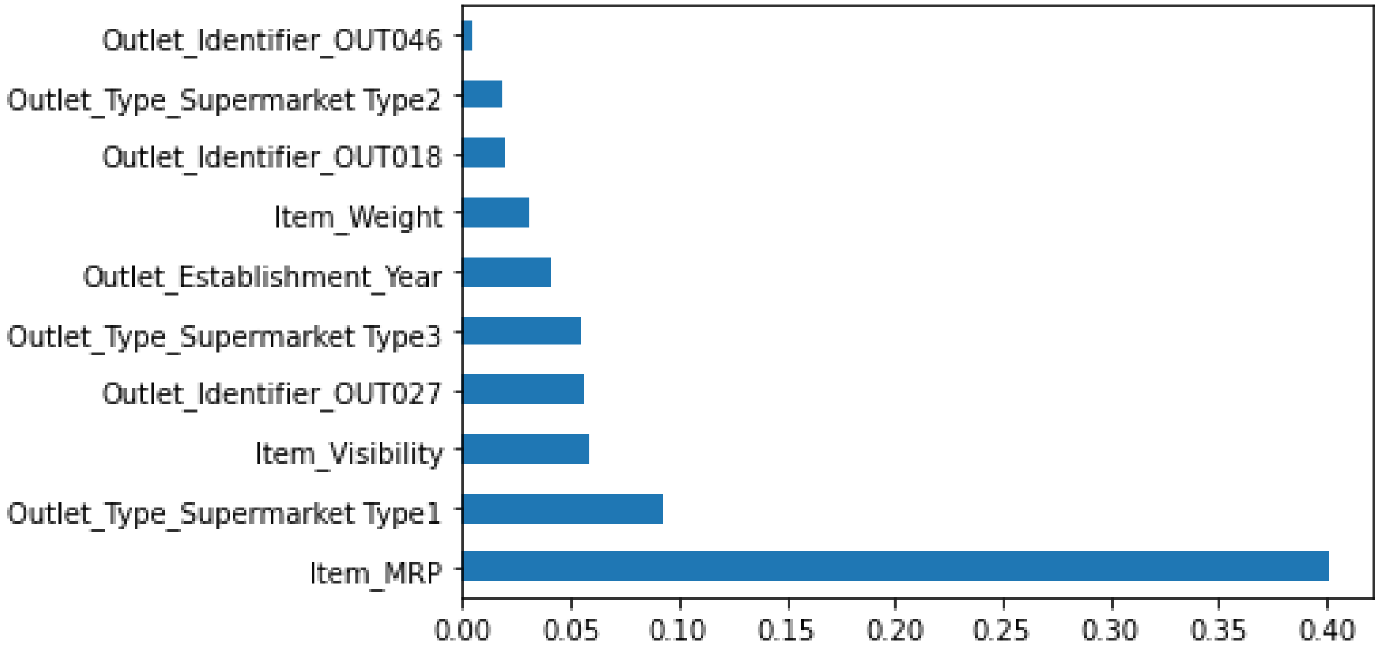
- Of all the models, Random Forests had the best r2, so most effective model to use for predicting sales.

- Translated to lowest variance (measured by RMSE) was $670, which was not very precise given the range of Item Outlet Sales
  - Min = $33
  - Max = $13,087
  - Mean = $2,181

```
Random Forest

Best fit of the four model
s for predicting sales

r2 (train/test)

0.9370377920925759

0.551683795847306

RMSE:

670.0781825468398
```

Advantages of Random Forest model

Random forest models allow for randomization of <u>features,</u> which can help identify the features in a data set that have the most impact on correlation, reducing the variance of the model.



Item_MRP had highest correlation (0.40) explaining an estimated 40% of the variation.