Sales Prediction Model

Improving Sales Through Data Science

Michael McCann | March 2022

Business Challenge

Fictional Customer is the head of a food producer/manufacturer. They are looking to bring a new item to market and want to optimize item characteristics and outlet partnerships to increase sales.

Our Approach

Data and Scope

- Sales data originally from a datahack challenge from May 2016 on analyticsvidhya.com
- 8523 observations collected from 10 different outlets of varying size, type, and location
- 1559 unique items across 16 item categories

Key Metrics:

- Item Outlet Sales: Sales of a given product in a particular outlet

Analysis

- We conducted Exploratory data analysis (EDA) to find patterns within the data and generate our observations and findings for this report
- Four separate machine learning models were implemented to validate our EDA findings and test feature importance (see appendix)
- A random forest model performed best achieving an RMSE of 1097.597 and R2 of 0.586

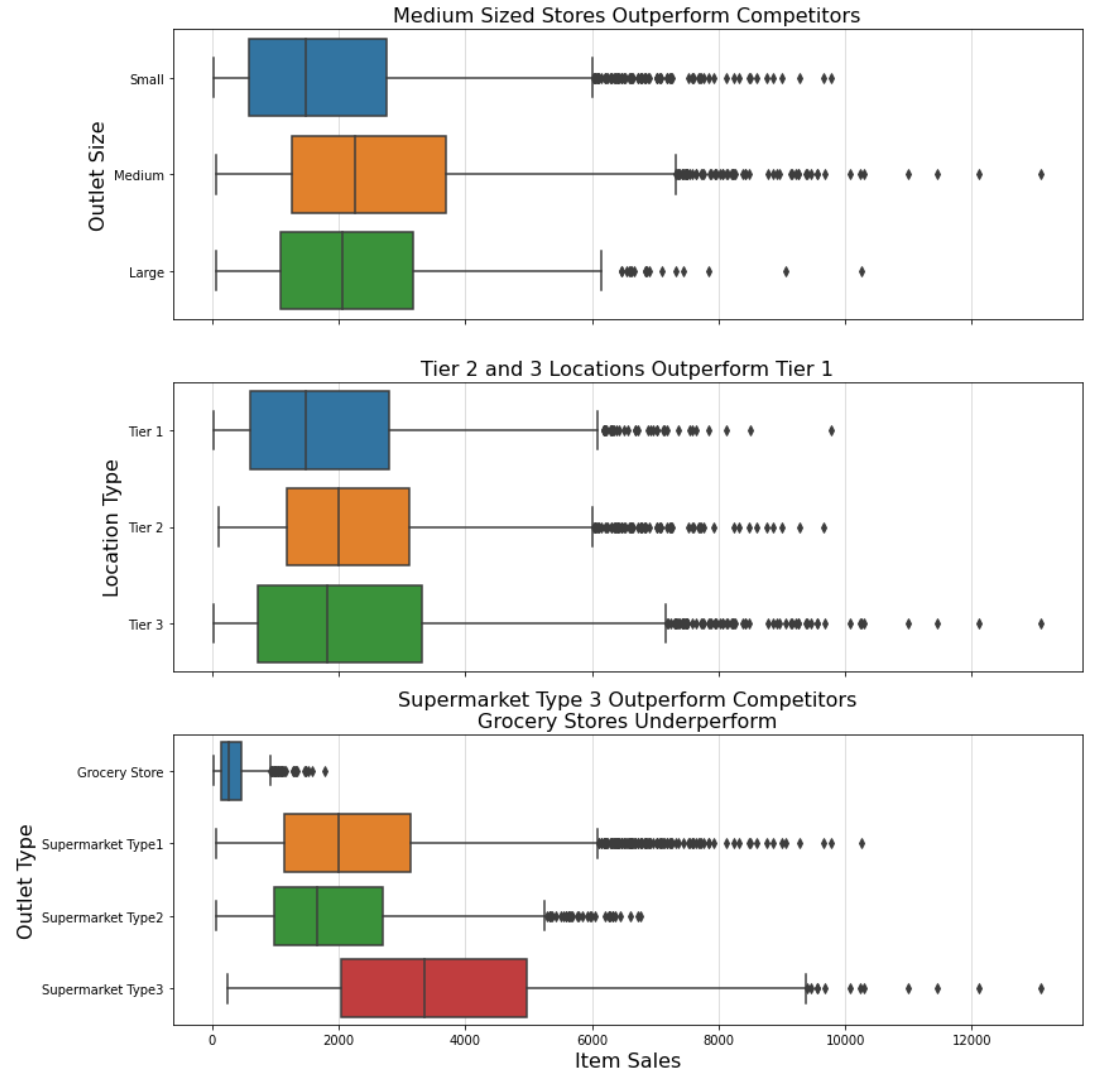
Caveats and Considerations

The dataset documentation does not provide clear guidelines/definitions for the variables Outlet Size, Outlet Locations, and Outlet Type. I find these variables to be important but am limited in my ability to provide insight based on incomplete information.

Size and Location Matter

Our data shows that the size, location, and type of outlet have an impact on item outlet sales. These results were further validated by our machine learning models which suggests outlet type is an important feature.

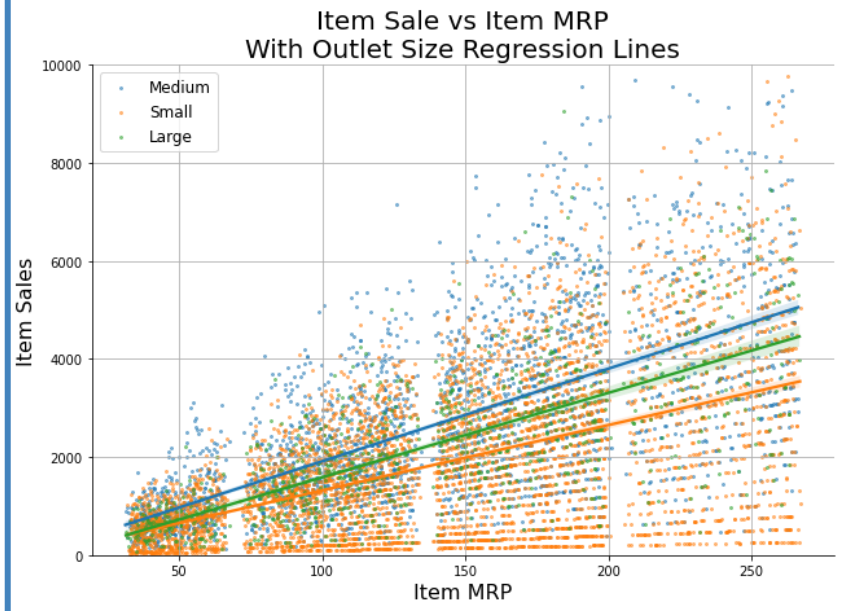
- Medium outlets have a higher median item outlet sales value as well as a higher third quartile and more/larger outliers. Large outlets have a higher median and third quartile than small stores but lack high value outliers.
- Outlets located in tier 3 locations have more high value outliers as well as a higher third quartile. Outlets in tier 1 locations underperform in all categories.
- Supermarket Type 3 outperforms all its competitors, having a median item outlet sale value above the third quartile of its nearest competitor.
- Grocery Stores underperform relative to all supermarket types.

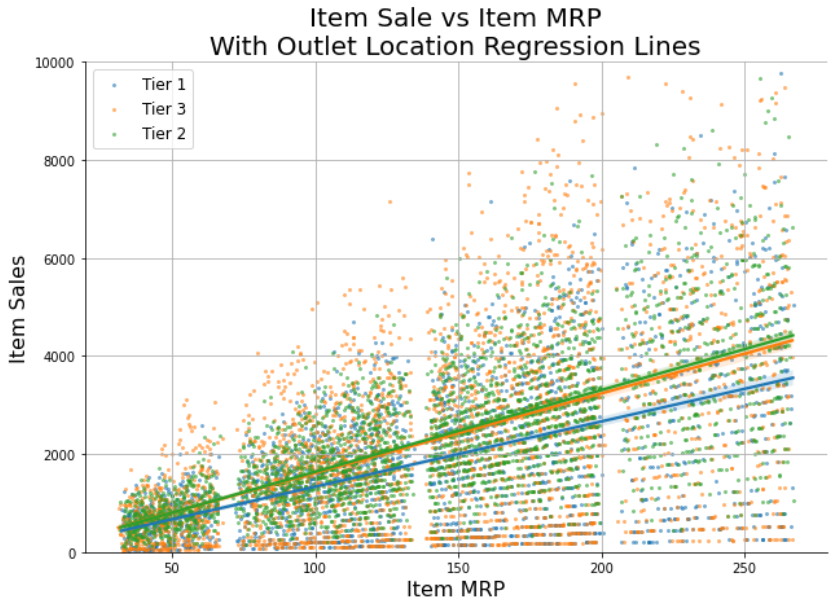


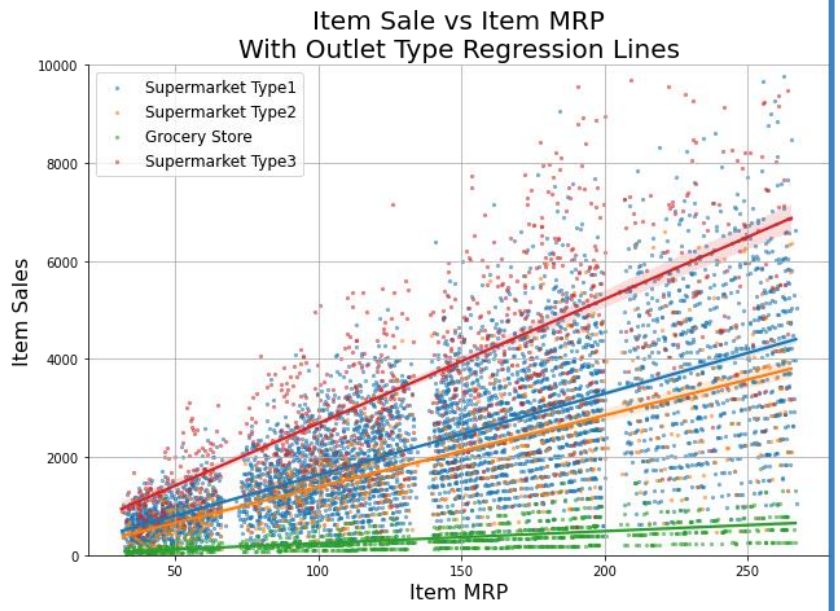
Higher MRP leads to Higher Sales

Maximum Retail Price has an unsurprising correlation with increased Item Outlet Sales. Of note certain outlet sizes, locations, and types appear to boost this relationship.

- Without any tweaking Item MRP has a moderate positive correlation (0.57). This intuitively makes sense as sales should probably have some relationship to the max price of the item.
- When graphed on a scatterplot (see below) the correlation between Item MRP and Sales is clearly evident. However, the same graphs also show that the correlation is more strongly positive with certain outlets.



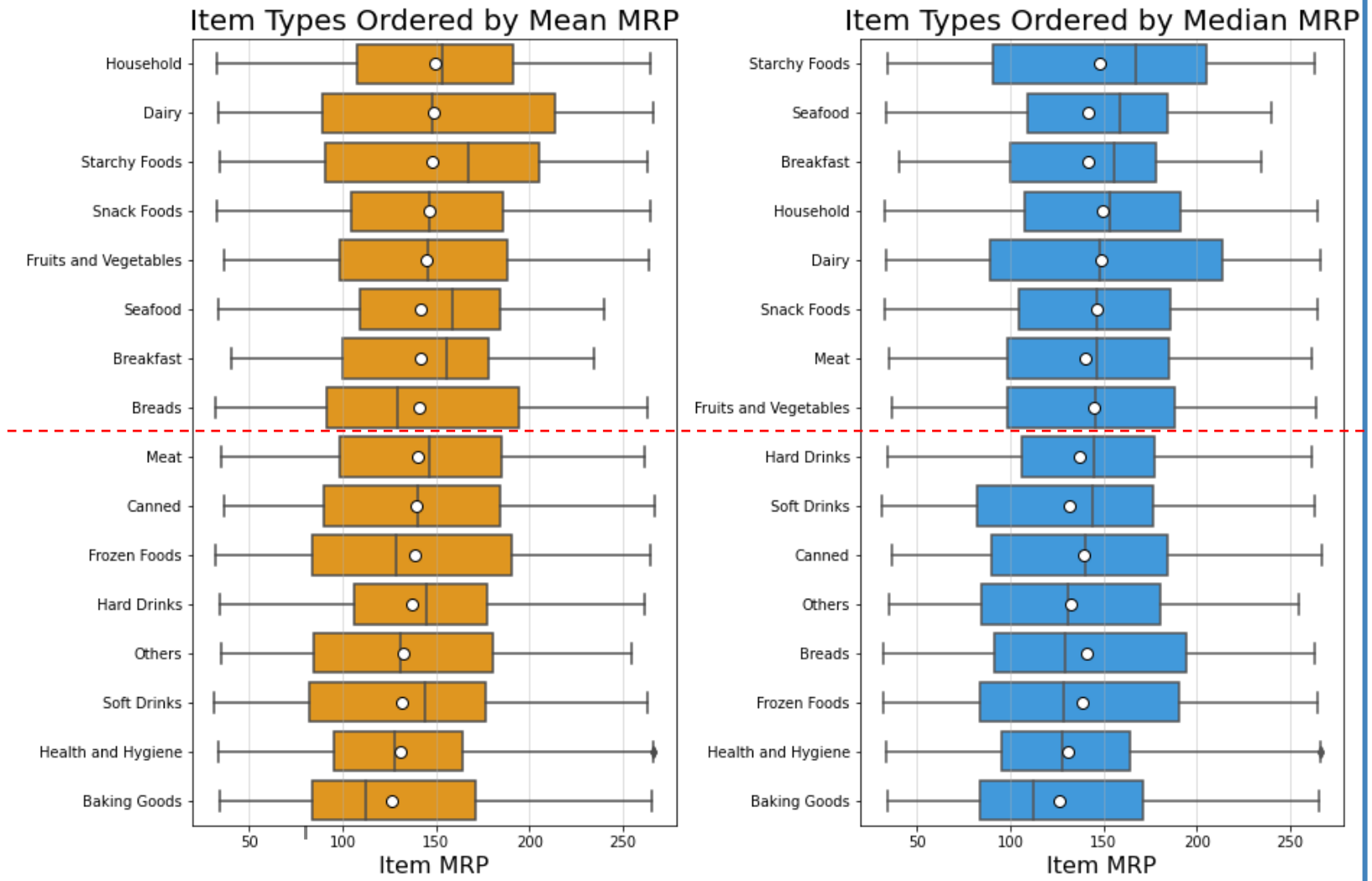




High MRP Item Types

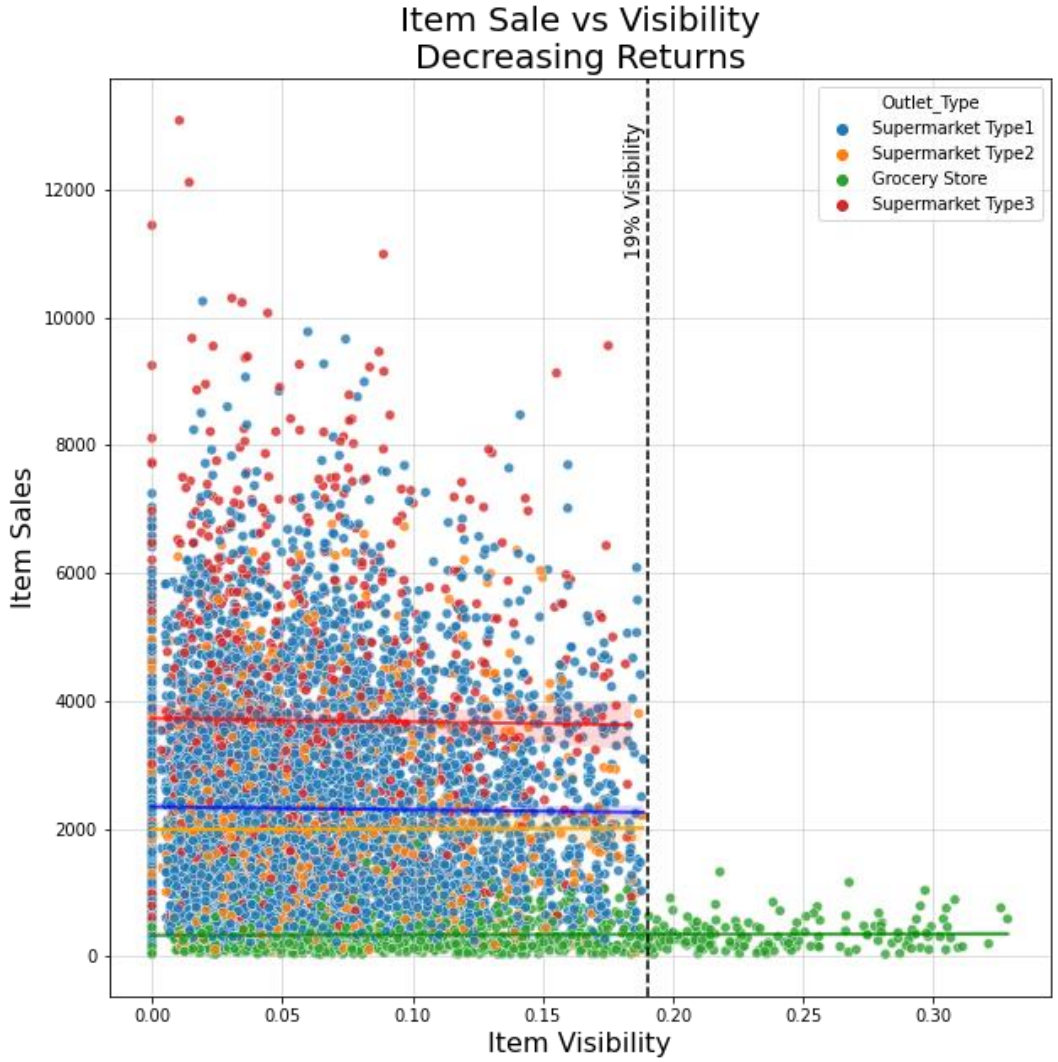
As previously noted high Max Retail Price is positively correlated with higher Item Outlet Sales. Identifying Item Types with high MRP will help maximize our search for items with high sales value.

- Item Types with a high median and mean MRP include: household goods, dairy products, starchy foods, snack foods, fruits and vegetables, seafood, and breakfast foods.
- Baking Goods and Health and Hygiene Items appear at the bottom of both lists.
- These categories should be used as guidelines, however, individual items may have a higher or lower MRP and additional market research should be conducted before production



Visibility Paradox

Conventional wisdom suggests that items given more shelf space would generate more traffic resulting in higher sales. However, our data suggests this is not the case.



- When graphed in a scatterplot (see left) there appears to be no correlation between visibility and Item Sales. The graph depicted shows the relationship with Outlet Type, however, this was also true for Outlet Location and Outlet Size.
- The maximum visibility given to any item in a supermarket was 19%. Grocery stores allowed for a higher percentage with no effect on sales.
- It is noteworthy that some of the items with highest sales are those with the lowest visibility.
- While there appears to be no relationship between the amount/percentage of visibility in an outlet we do not have the data to comment on premium placement within an outlet

Recommendations

Sales can be increased by carefully selecting which outlets to partner with, focusing on high MRP Items, and ignoring metrics with no return on investment.

Partner with Specific Locations, Avoid Grocery Stores:

- Partnering with outlets that maximize sales could optimize returns. Medium and large sized outlets, outlets in tier 2 and 3 locations, and supermarkets.
- Avoid: Grocery Stores

Focus on High MRP Items or Item Types:

- Our model suggests that MRP is the most important feature in item sales and we found a moderate correlation between MRP and item sales.
- Peruse: household goods, dairy products, starchy foods, snack foods, fruits and vegetables, seafood, or breakfast foods
- Avoid: Baking Goods and Health and Hygiene Items

Visibility Matters… But Don’t Pay For Additional Space:

- Our model suggests that visibility is an important feature in item sales. However, there appears to be diminishing returns to purchasing additional shelf space/visibility.

Questions?

Prepared by: Michael McCann

Contact: msmccann10@gmail.com

Appendix

Machine Learning Model Performance

And Model Recommendation

Machine Learning Model Performance

Dummy Model: RMSE: 1678.869 | R2 Train Score: 0.000 | R2 Test Score: -0.000

Linear Regression Model

- RMSE: 1135.226
- R2 Train Score: 0.568
- R2 Test Score: 0.543

Decision Tree Model

- RMSE: 1145.560
- R2 Train Score: 0.938
- R2 Test Score: 0.534

Bagged Tree Model

- RMSE: 1145.236
- R2 Train Score: 0.938
- R2 Test Score: 0.535

Random Forest Model

- RMSE: 1097.597
- R2 Train Score: 0.623
- R2 Test Score: 0.586

Model Recommendation

All of our models outperformed the baseline test, however, we recommend using a Random Forest Model which provided the best root mean squared error (RMSE) and correlation coefficient (R2).