Sales Prediction Model Presentation

By Tibebu Sime

July 18, 2021

Seattle Metropolitan Area

Coding Dojo: Data Science & Machine Learning

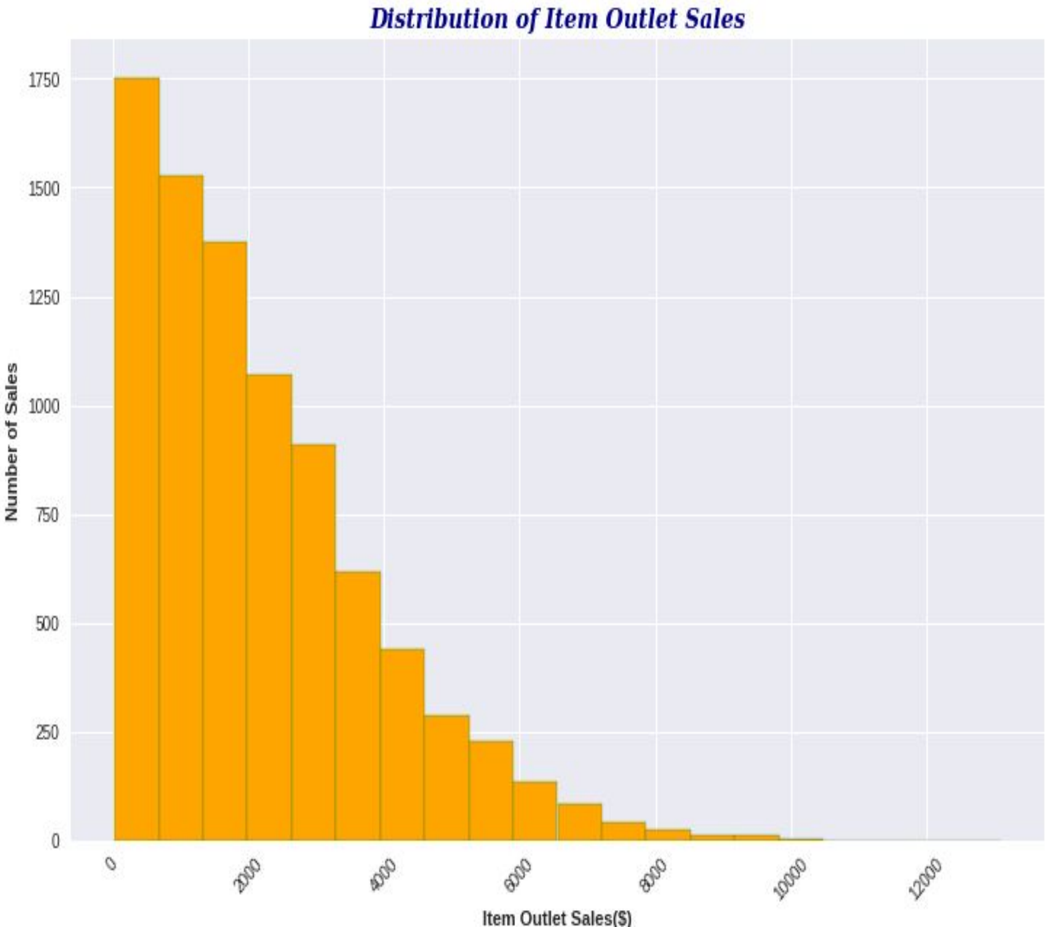
Introduction

The goal of this project was to build a predictive model and forecast the sales of each product at a particular outlet. You know that sales prediction is an essential task for the management of a store. Without proper sales prediction, business planning and decisions will be based on unreliable estimates which can lead to many inefficiencies and missed opportunities.

The predictive model developed can help retailers to answer these kinds of questions:

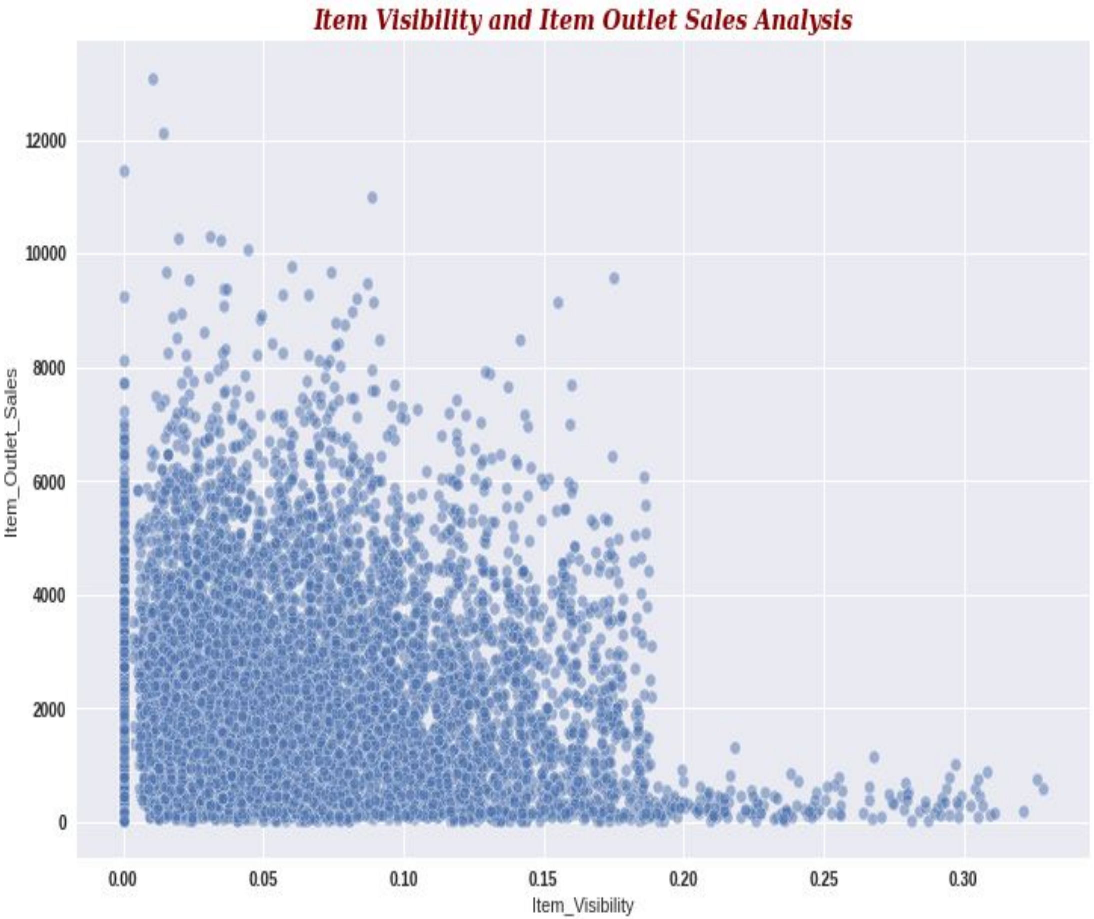
1. How many more staff should be hired?
2. Do we have enough inventory to meet demands each week? If not, how much stock should be ordered to meet the growing demands?
3. How much revenue do we expect in this month or the following month?

Data Visualization: Distribution of the Sales



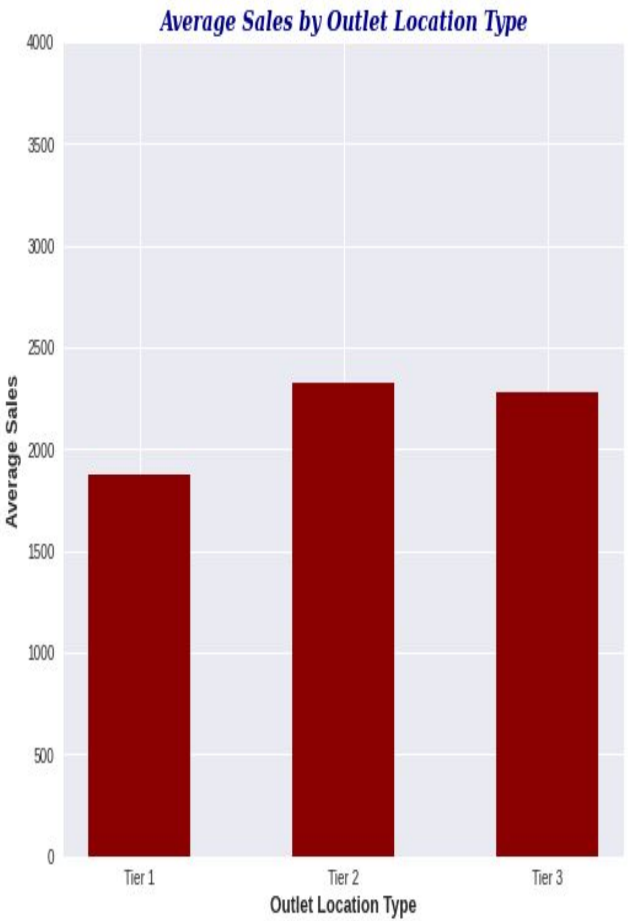
From this graph we can see that the sales distribution deviates from the normal distribution. It is tailed to the right -- meaning lower sales are concentrated on the left side.

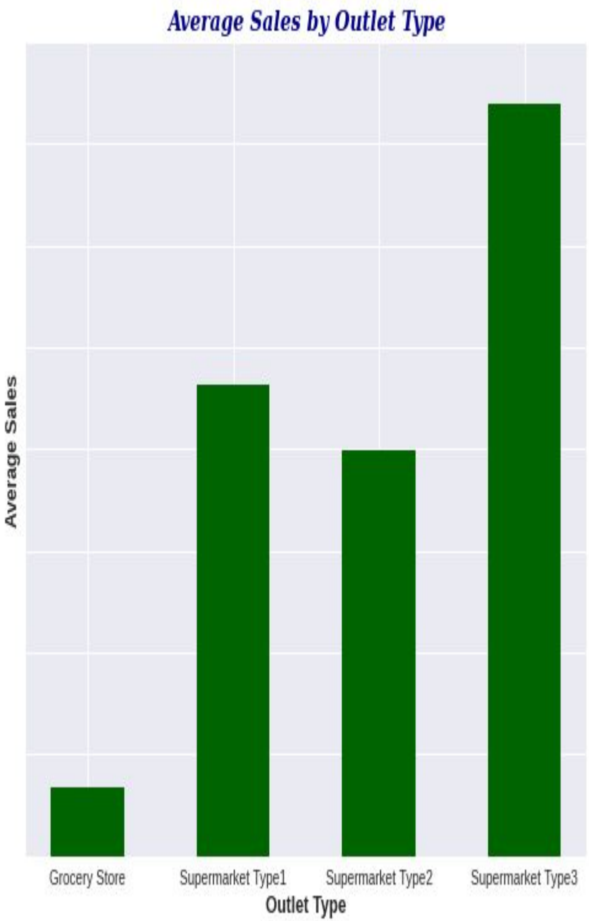
Data Visualization: Impact of Item Visibility of the Sales



The location of product in a store will impact the sales. The obvious assumption is that products with higher visibility are likely to sell more because products which are kept at the entrance of the store will catch the shoppers attention rather than the ones in the back. But from the correlation matrix generated earlier and now this plot, we can see that the more visible a product is the less higher its sales will be. This might be due to the fact that a great number of daily used products, which in fact do not need high visibility, control the top of the sales chart.

Data Visualization: Impact of store locations and outlet types on the sales





1. More items are being sold in stores which are found in less densely populated cities probably because the presence of these stores are higher in those cities.

2. Supermarket Type3 has the highest average sales though the availability of Supermarket Type1 is higher. Grocery Store has the least sales.

Models Performance Evaluation: Performance score

- LG training R2: 0.5544536260526491
- LG testing R2: 0.5900393833521596
- KNN training R2: 0.5870112626717922
- KNN testing R2: 0.5890014065072519
- RFG training R2: 0.6087463007379363
- RFG testing R2: 0.6223139819671555

Interpretation of R2 score(Coefficient of Determination R2): is how well the regression model fits the observed data. Generally a higher R-squared indicates a better fit for the model.

What is the acceptable R2 value? Ideal value is 1 (100%). This is impossible because we may

- ★ 0.3< R2 value<0.5: weak model
- ★ 0.5< R2 value<0.7: moderate model
- ★ R2 value>0.7: strong model

Based on these performance scores:

1. Our models are not overfit since they all perform better on the testing data subset. In all of the cases, the R2 values are higher in the testing data subset than in the training data subset.
2. Random Forest Regressor is our best model because we have the highest R2 value. There is a lot of improvement in our performance score.

Models Performance Evaluation: Error Metrics

- RMSE_training: 1139.8109171305882
- RMSE_testing: 1090.253268820084
- RMSE_knn_training: 1097.376004638942
- RMSE_knn_testing: 1091.6325991731178
- RMSE_rfg_training: 1068.109023283575
- RMSE_rfg_testing: 1046.457951594882

What is the mean error? It is the average of all errors in the dataset, that is the average of the difference between true values and measured values. Its ideal value is 0! This is practically impossible because we may have a model that perfectly predicts our training data, but which is very unlikely to perfectly predict any other unseen data.

1. The RMSE values also confirm that our models are not overfit.

   Note: Overfit model has very low RMSE for training and higher RMSE for testing/validation/unseen data

Summary of the Results & Recommendations

1. All of three models I have developed for the sales prediction project are considered to be moderate but there is a lot of improvement in the coefficient of determination R2 with the Random Forest Regressor model.

2. Contrary to the general hypothesis, products with less visibility in the stores are likely to sell more. I encourage the store management to improve the visibility of the products which are in high demand so that they can easily be noticed by the shoppers. This is how they can more boost the sales.

3. Stores located in less densely populated cities have the highest sales. Of course the presence of stores in those areas is also high. It is also possible to boost the sales in highly populated cities where there could be high income residents by making their visuals appealing to customers, promoting your products on social media or TVs and providing consistently high-standard customer service