

### **Bigmart Sales Data**

**Abstract:** This project attempts to model and predict sales data for the supermarket chain Bigmart. Machine learning methods including Random Forest and Lasso CV were used in an attempt to create applicable regression models. Data from a massive Kaggle data set containing feature data for a variety of metrics was used for model construction. Features included information such as Item MRP, and year of outlet establishment.  $R^2$  was used as the primary evaluation metric, with mean squared error of predictions being considered as well. The random forest regression model was ultimately best, and Item MRP, Outlet Type (size), and years since original opening were the most important features in the final model.




**Overview:** The data set I will be using is the “BigMart Sales Data” set under the “Economic” domain within the course page, downloaded from Kaggle. The data is from 10 different stores and has 1559 different products. The objective is to construct a model to predict sales for each item. The data’s features depict characteristics of the items sold, including item weight, fat content, a visibility value, and the type of item sold. Originally, the goal had been to predict sales within 1% of the actual, but upon reconsidering, the goal is now to create a model with an  $R^2$  value above 0.55, which is more attainable and reasonable given the data set. 0.55 is the highest  $R^2$  value that was found from other models on Kaggle, so the goal is to outperform them by producing a model with an  $R^2$  value greater than 0.55.

This sales model will be of use to BigMart, as a model for sales could help them to decide what should be focused on, and it could also be used to project future revenue, which is helpful for future business planning. Answering “What information is most important in predicting item sales?” is just as interesting and important as “What should the expected sales of an item be?” Studying feature importance is therefore going to be as much of a focus as developing a capable model. Through this project, I also hope to develop something that may be presented to a future employer to showcase my abilities and talent.

**Related Work:** Since this data set was taken off of Kaggle, numerous attempts to predict sales from the data already exist. There are over 100 submitted notebooks of code attempting to answer the question at hand. One submission “Big-Mart-Prediction and Deployment” by Shashank52ez (shashank52ez, 2022) includes a thorough exploratory analysis, and three models; Linear Regression, Random Forest Regressor, and Lasso, using  $r^2$  as the primary scoring metric. Their Random Forest Regressor model proved to be the best, and HyperParameter tuning did not prove to make a significant difference in the predictive capabilities of the model. Their best  $R^2$  value was approximately 0.55, making that a good benchmark to shoot for in my own model. I will be sure to try different models to set my work apart. For one, I believe KNN for regression could be powerful here. Also, within their notebook there is very little in terms of explaining the results, which I believe is just as important as the model itself. I will be sure to include in-depth, but approachable explanations of my findings.

Kaggle user Aishwarya Wuntkal’s Evaluation Metrics for ML Regression Models (Wuntkal, 2020) attempted the same problem, and her work is the “Hottest” rating on Kaggle. Her best model was also constructed using a Random Forest algorithm, and featured a similar  $r^2$  value. Her explanations are more thorough throughout, which makes her report much more digestible. She also makes excellent use of graphics throughout her presentation to keep the audience engaged. Both her model and the previous feature an RMSE of around 1250 for their best model, which is good to keep in mind as another benchmark to shoot for. To keep my project unique, I will attempt to be

more thorough in my conclusion. While her quick 3-bullet-point summary in her conclusion makes for an easy read, I aim to have a more thorough analysis here. I will have a brief summary of the findings towards the end of my report, but I also want to more thoroughly explain the significance of my findings and how they should be interpreted and applied.

For a third reference, I chose another submission to the Kaggle page “Prediction  of Sales  using XGBoost ” by Padmavathi D. (D, 2022) Their introduction is the best of the three here, where they actually explain the problem statement, as well as how their results will be useful. In doing so, they have already broadened their audience, and more solidly affirmed their value. Throughout their notebook, they do an excellent job explaining what the purpose of each step is, and what they hope to achieve, making this notebook also an excellent educational tool for those with less knowledge about Machine Learning. They set their model apart by utilizing the efficient XGBoost to construct their model. Ultimately, they did not achieve a higher  $R^2$  than the other two users had with their random forest models. Also, none of the 3 projects I have documented here did much in the way of visually presenting their models predictions vs the actual. I will be sure to do this in my presentation in order to set my work apart.

### **Data Acquisition:**

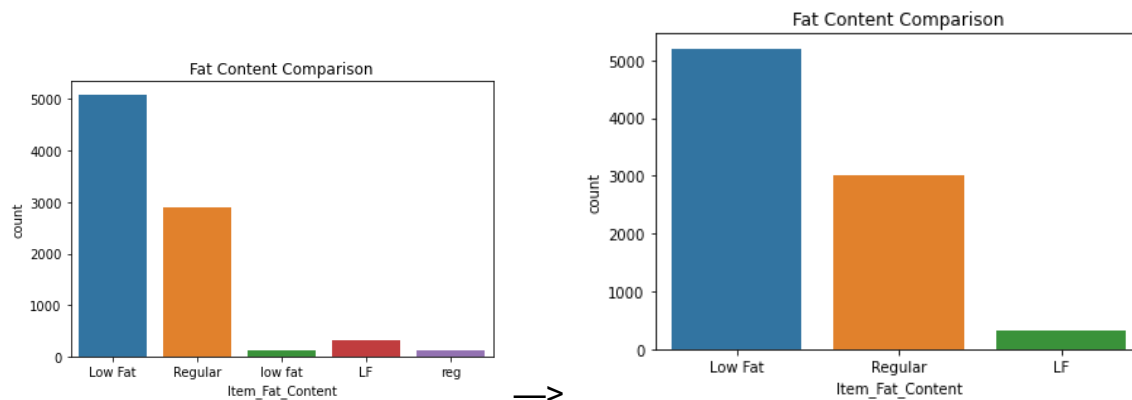
The data includes the features “Item\_Identifier, Item\_Weight, Item\_Fat\_Content, Item\_Visibility, Item\_Type, Item\_MRP, Outlet\_Identifier, Outlet\_Establishment\_Year, Outlet\_Size, Outlet\_Location\_Type, Outlet\_Type, Item\_Outlet\_Sales.” “Item Outlet Sales” is the target variable, and each of the others could be considered for the predictive model. Certain variables such as “Item\_Identifier” may be disregarded during the final model building, since it would be impractical to create thousands of dummy variables to represent thousands of different ID numbers.

The data is from 10 different stores and has 1559 different products. The data exists in the form of two .CSV files (a training file and a testing file) which were uploaded to Kaggle, and obtained from public domain. Since the file is public domain, there should not be any issues or limitations in sharing of the data or my work. As far as use, while the set is free to use and explore for anyone, the data set was uploaded 5 years ago, so its usefulness for modern predictions is likely outdated, and the model should be viewed more as a practice exercise, an example predictive model and evidence of ability, rather than a model that should be applied in its current state.

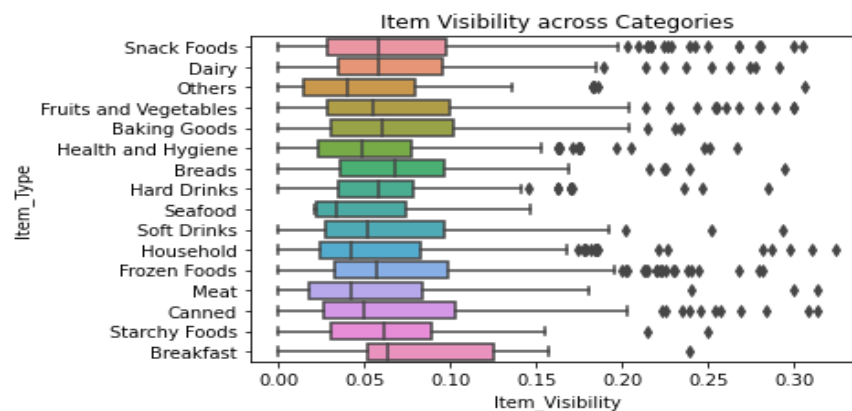
**Preprocessing:** Fortunately, this data set featured very little missing data. The item weight and outlet size featured 17% and 28% missing data, respectively, for both the training and test sets. All other features had no missing values. For the item weight feature, missing values were imputed by using the mean item weight from across the

whole data set. For the outlet size feature, missing values were imputed via the mode of the data set, figuring that using the most common outlet size for missing values was a fair substitution with minimal risk.

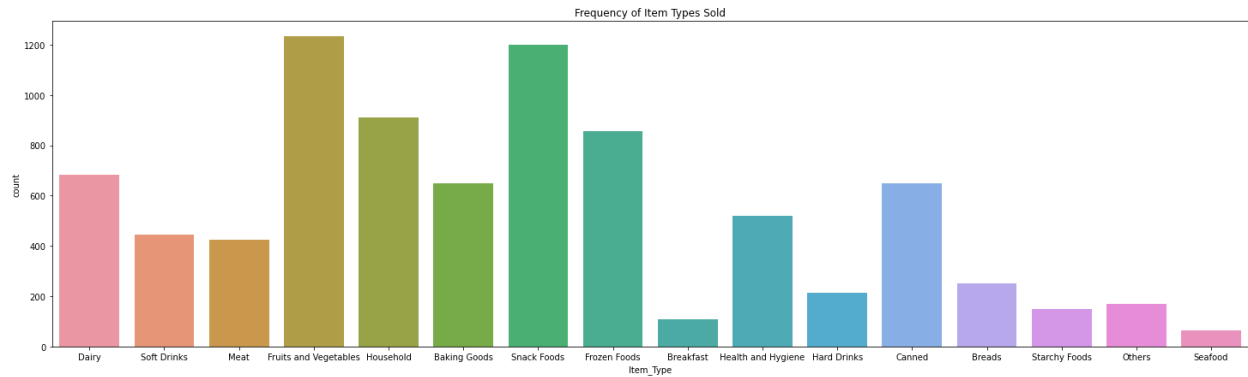
In the original data frame, several of the features were not in the proper format. For example, item fat content was not being considered as a categorical variable. These issues were corrected. The feature “Outlet\_Establishment\_Year” was converted into a more useful “Years since established” feature by subtracting the establishment year from 2015. Also, certain features, such as fat content, required minor cleaning, as it had featured separate values for both “LF” and “Low Fat”, which were clearly meant to indicate the same thing.



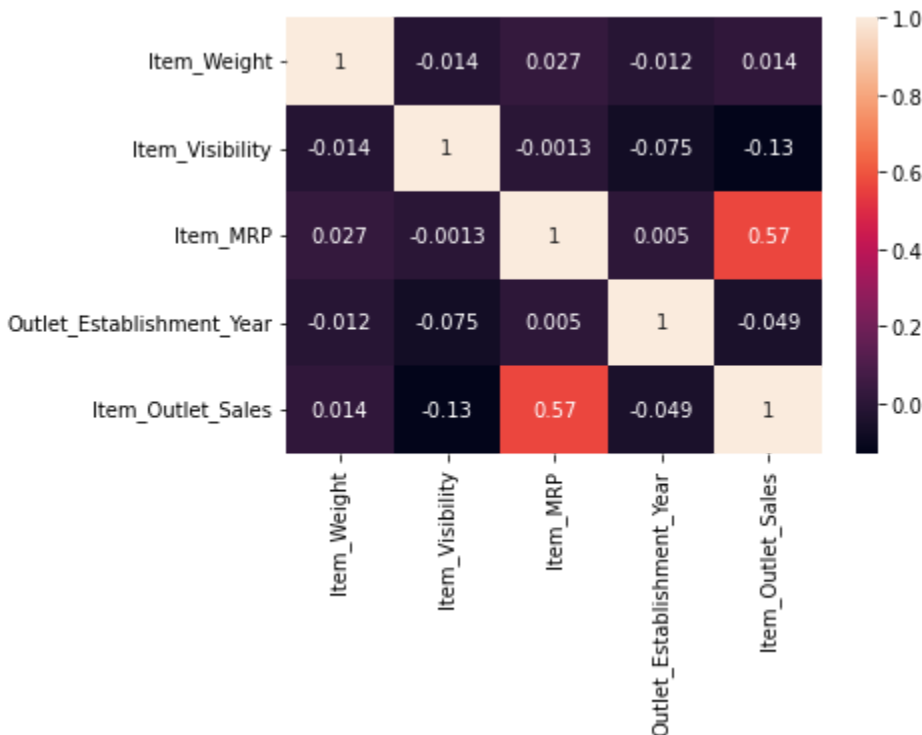
From exploration, I learned of certain interesting patterns. For example, Seafood has generally lower visibility than other categories, but the minimum seafood visibility was above 0, unlike all of the other categories.



Visualizing frequency of item sales was also interesting. As expected, snack foods were one of the highest categories here, but it was interesting to note that fruits and vegetables were the highest category. This is likely due to their being counted by item, and it is common to buy several or dozens of pieces of fruit during a grocery trip, while items like milk are generally only purchased one or two at a time.



A Heat Map for correlation was created so that the correlation statistics could all be visualized together. From this, we see that Item Market Retail Price stands far above the rest of our features in terms of its correlation with our target variable Item Outlet Sales.



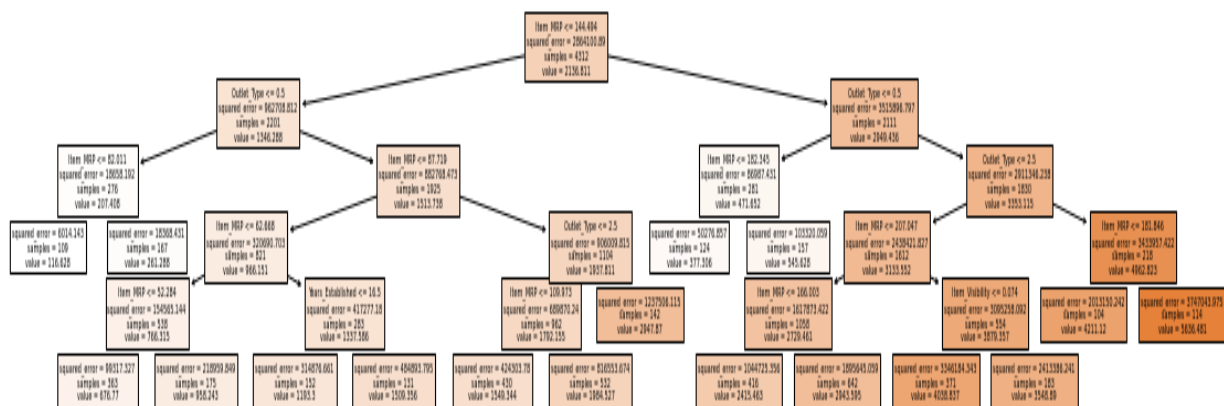
**Model Selection:** Before diving into the primary three models that were constructed, I want to summarize the other efforts that I made that were ultimately left out. After the creation of my models, I went back and performed a Principal Component Analysis on the data, and retrained the models. This proved to be unnecessary since the data set was small enough that the computation time was minimal, and the feature reduction made a completely minimal change in terms of model performance. Also, during the time since the last draft, I created a few new features, most notably, a categorical feature for item visibility. After creating dummy variables from this new feature,

dropping the original, and re-running the models, there was no significant improvement in performance, so these features were excluded.

Three primary models were constructed for evaluation. Others, such as a linear regression model based solely on mutual information score, were considered, but due to poor performance, they were left out during final consideration. The first was a KNN regression model, which was tested for k values 1 through 20, and the best model here was found using a K of 6, where the  $R^2$  value was 0.428.

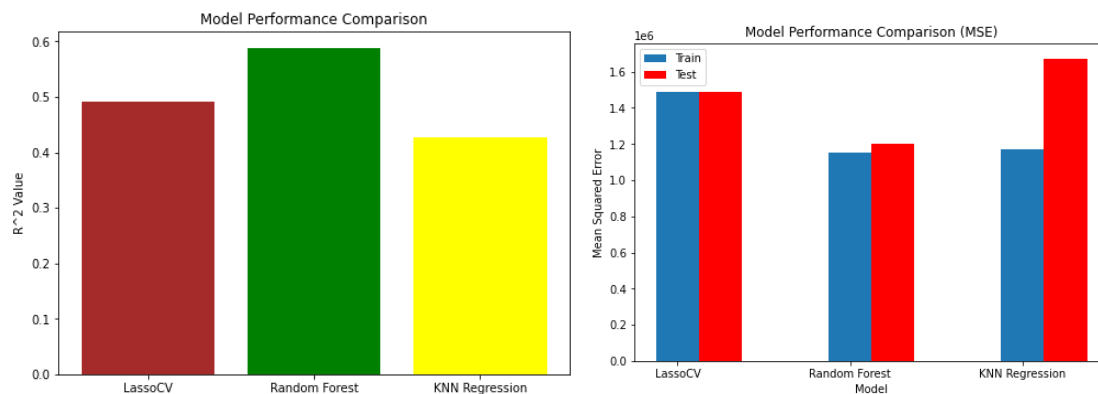
The second model constructed was a significant improvement over the first, the LassoCV model featured a training  $R^2$  of 49.24, and a testing  $R^2$  of 0.4944. After running the same model with PCA applied, the test  $R^2$  dropped to 0.491, which was a major factor in the decision to exclude PCA from the final model.

Finally, a Random Forest regression model was constructed. Due to the robust nature of Random Forest, and my success with it in other projects, I was confident that this model would perform well. Below is the tree printout.

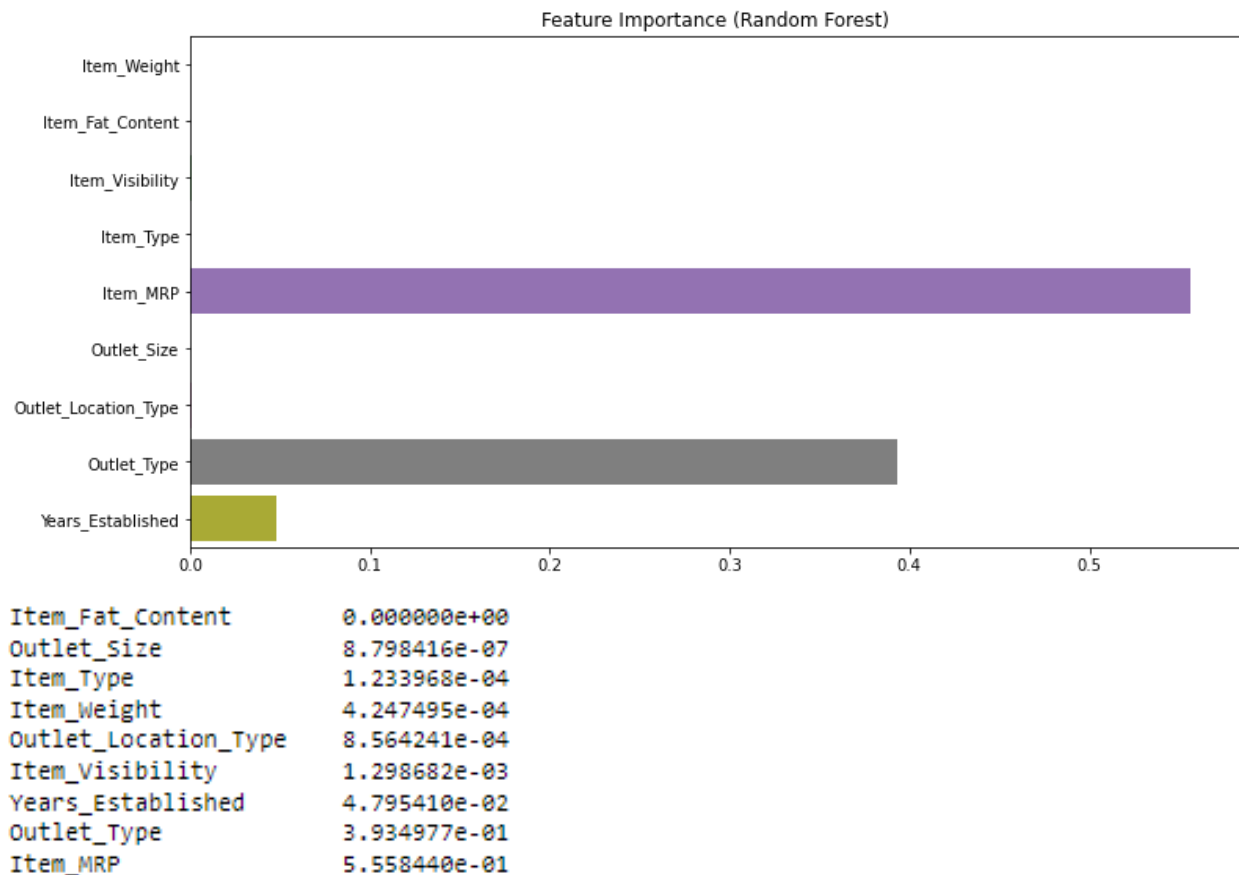


This model significantly outperformed the previous two, with an  $R^2$  value of 0.589 on the training set, and 0.589 on the test set. The model featured the following hyperparameters;  $n\_estimators=200$ ,  $max\_depth=5$ ,  $min\_samples\_leaf=100$ ,  $n\_jobs=4$ .

## Results and Evaluation:

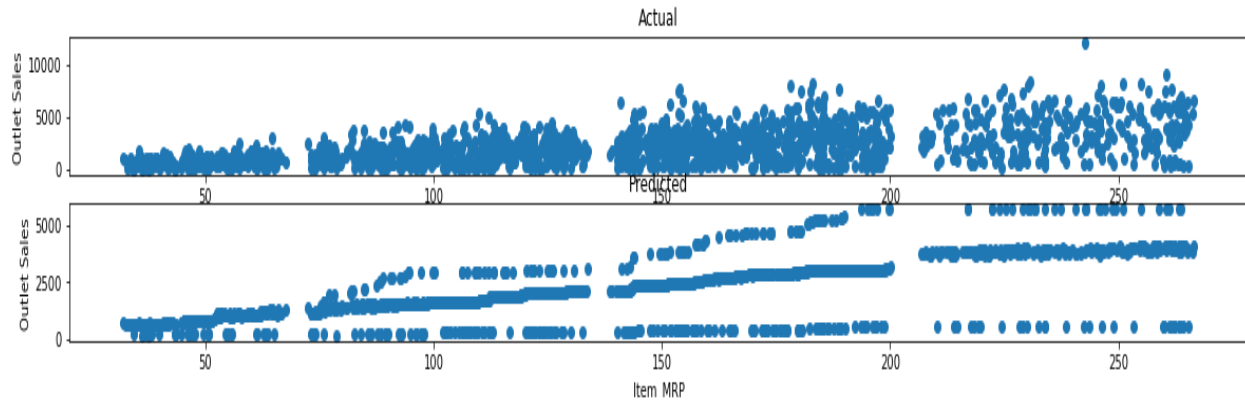


As you can see, the Random Forest is clearly the best in terms of both  $R^2$ , which shows us that it has a stronger fit than the others, and in terms of Mean Squared Error, showing us that it has generally smaller residuals than the others tested. Since the test  $R^2$  was equivalent to the training  $R^2$  for the Random Forest Model, there seem to be no issues in terms of overfitting.



Here we can see that only three major features were critical in the random forest model. By a wide margin, Item Market Retail Price was the most important feature for the model, followed by outlet type, then years established. All of the other features were virtually irrelevant.

Below is a plot of the actual Item Outlet Sales plotted against the most important feature, Item Market Retail Price, with how my model would have predicted these points below.



As you can see, while my model is not perfect (it tends to try to push the points into more distinctly linear patterns) the general trend for prediction is apparent; higher Item Market Retail Price tends to lead to higher outlet sales.

In all, the model performed phenomenally. While in the context of other data sets an  $R^2$  of 0.589 may seem low, it is very strong for this data set. Other, more complicated models such as one that included XGB on Kaggle reached  $R^2$  values of around 0.53, and the best I had seen prior to my own work was 0.55. I am extremely satisfied with the performance of my model, and if this had been for a Kaggle competition, I would have placed very highly.

**Ethics:** In studying the data for a supermarket chain, we are uncovering trends within society, eg; where do people spend their money on groceries? Certain variables, like the visibility of items, indicate whether the market is intentionally trying to promote certain categories. For example, we could have uncovered that snack foods tend to have much higher visibility when compared to healthier alternatives. In uncovering these potential trends, BigMart would have information that they can use when making deals with brands. If we had found that, for example, a high visibility score is a critical factor when determining total sales, BigMart could auction off the best spots to brands, and certain products would be purchased more frequently simply because they were more visible.

Another consideration when it comes to ethics here is the question of whether trying to forecast sales, and determine important factors in prediction from past data, is ultimately a productive goal. While yes, this information will undoubtedly be valuable to BigMart, should we be making these models based off of previous data? If, for example, supermarkets have classically found that people will purchase more chips when they've been placed in an additional, special location outside of the aisle, but have not attempted this practice with more healthy alternatives, and we prove that higher visibility is critically important to the sales of snack foods, the market will continue this practice. If the market has never tested whether the same behavioral patterns would also apply to healthier alternatives such as food and vegetables, then the model



wouldn't be able to prove whether the same practice works across categories. These are just the first examples that come to mind when thinking about how applying data science to a supermarket could be ethically cumbersome, and the implications of all results should be considered, as well as what dangers may arise when taking action.

**Future Work:** Since the last draft, I had wanted to experiment with feature engineering. As stated previously, this was done, but it made a completely negligible difference in terms of the model's performance, so it was left out for this document. If I were to spend more time on this project, I would probably continue down the route of feature engineering; with only 3 features having any major predictive value, the model is quite simple, and it feels like the other features may have had more mutual information if they were dug into a bit deeper. One example could have been to convert the "Item Type" into a dummy variable, so that we might know that a particular type of item tends to lead to higher sales. Constructing a neural network for prediction was also something I had toyed with, and it could have potentially led to stronger results, but I was pleased with the presentability and performance of my Random Forest Model. In terms of expanding this project, I would be curious to see whether the model holds up for other supermarkets, but that would require obtaining a vast amount of data that other companies may not be as eager to expose.

## Works Cited

D, P. (2022, November 26). *Prediction of sales using XGBoost*. Kaggle. Retrieved December 5, 2022, from <https://www.kaggle.com/code/aishwarya2210/prediction-of-sales-using-xgboost>

shashank52ez. (2022, June 12). *Big-Mart-prediction-and-deployment*. Kaggle. Retrieved December 5, 2022, from <https://www.kaggle.com/code/shashank52ez/big-mart-prediction-and-deployment/notebook>

Wuntkal, A. (2020, October 2). *Evaluation metrics for ML Regression Models*. Kaggle. Retrieved December 5, 2022, from <https://www.kaggle.com/code/aishu2218/evaluation-metrics-for-ml-regression-models>