**PR-Milestone 1 Report**

**Table of Contents**

[Preprocessing 2](#_Toc74320651)

[Data Analysis 3](#_Toc74320652)

[Regression Techniques 7](#_Toc74320653)

[Model Acquired Results 8](#_Toc74320654)

[Features Used/Discarded 9](#_Toc74320655)

[Data Size 10](#_Toc74320656)

[Acquired Results 11](#_Toc74320657)

[Conclusion 13](#_Toc74320658)

\

# **Preprocessing**

1. Renaming columns names from X1,X2,… to a better naming convention ( Item ID, Store ID, etc..) by using function rename in pandas.

Renaming in detail:

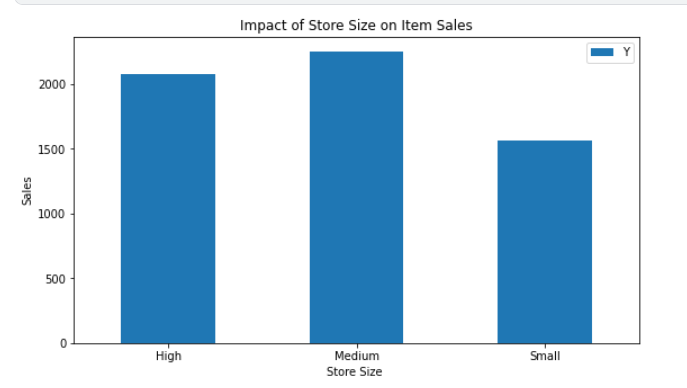
* X1: Item\_ID
* X2: Item\_Weight
* X3: Item\_Fat\_Amount
* X4: Item\_Store\_Allocation
* X5: Item\_Category
* X6: Item\_Price
* X7: Store\_ID
* X8: Store\_Establishment\_Year
* X9: Store\_Size
* X10: Store\_Location\_Type
* X11: Store\_Type

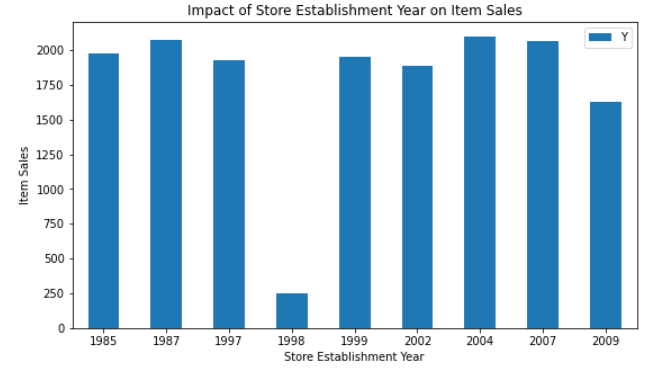
1. Replacing the following in item fat amount column using pandas rename:
   1. LF to Low Fat
   2. Low fat to Low Fat
   3. Reg Regular
2. Filling the NaN’s in item weight using with backward fill– using fillna function found in pandas with a method parameter=’bfill’
3. Filling the 0’s in item store allocation with backward fill -using fillna function found in pandas with a method parameter=’bfill’
4. Filling the NaN’s in store size with backward fill - using fillna function found in pandas with a method parameter=’bfill’
5. We performed label encoding on Store\_Size , Store\_Location\_type, store\_type

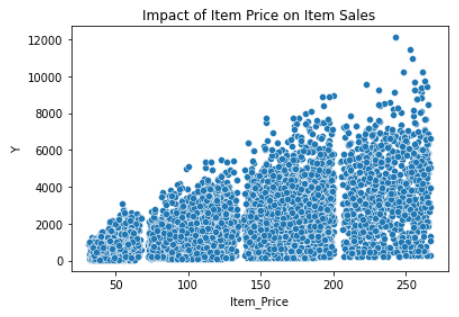
# Data Analysis

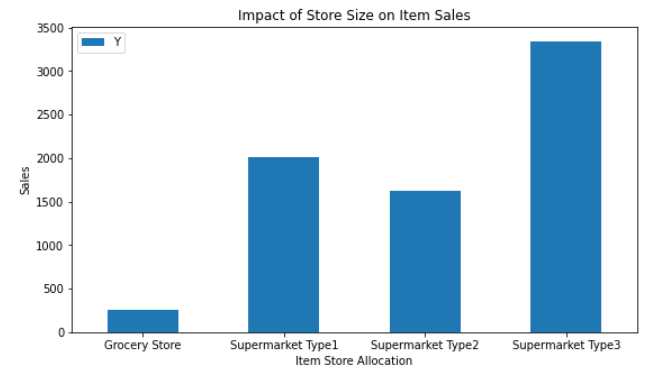


Visually, as we can see above. This is the correlation between each feature with the other. In our case, we will mainly focus on the relation between all features with our target which is “Y”.









Here is a summary table that shows the relation of all features with our target.

|  |  |
| --- | --- |
| Feature | Correlation with target “Y” |
| Item\_Weight | 0.01 |
| Item\_Fat\_Amount | 0.00 |
| Item\_Store\_Allocation | -0.14 |
| Item\_Price | 0.58 |
| Store\_type | 0.39 |
| Store\_Establishment\_Year | -0.05 |
| Store\_Size | -0.09 |
| Store\_Location\_Type | 0.08 |

Based on the table above, our top features are:

* + Item\_Store\_Allocation
  + Item\_Price
  + Store\_Size
  + Store\_Location\_Type
  + Store\_Type

# Regression Techniques

1. Random Forest
2. Gradient Boosting Regressor
3. Polynomial

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Description** | **Advantages** | **Mean absolute error** |
| **Random Forest** | **Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction** | **Can be used for regression & classification** | **764** |
| **GradientBoostingRegressor** | **Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. ... If a small change in the prediction for a case causes no change in error, then next target outcome of the case is zero.** | **provides predictive accuracy that cannot be beaten.** | **755** |
| **Polynomial** | **used in many experimental procedures to produce the outcome using this equation. It provides a great defined relationship between the independent and dependent variables** | **Polynomial provides the best approximation of the relationship between the dependent and independent variable.** | **769** |

# Model Acquired Results

# Features Used/Discarded

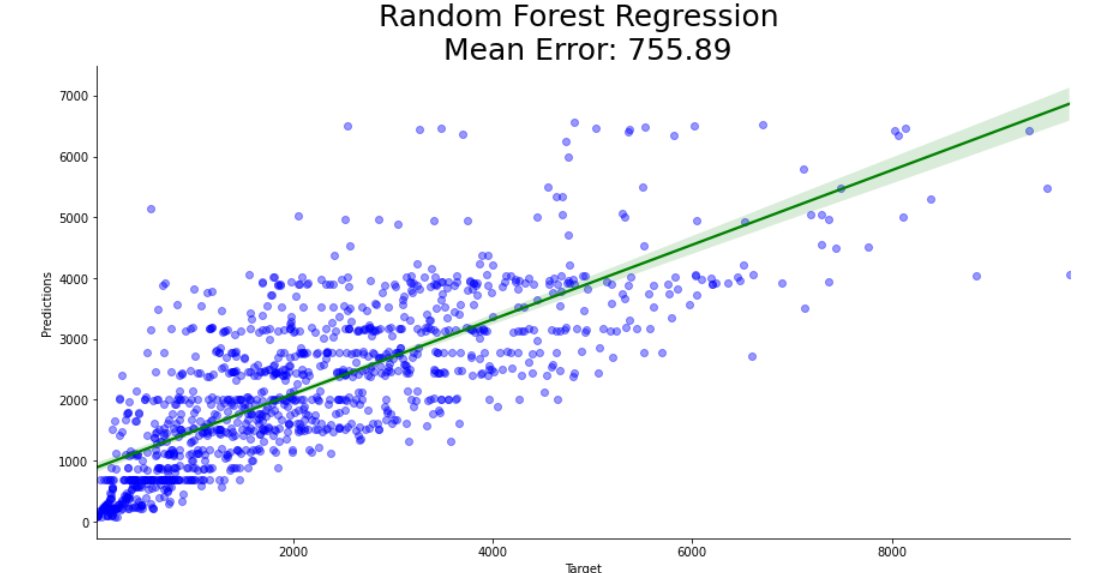
* **Features Used:**
  + Item\_Store\_Allocation
  + Item\_Price
  + Store\_Size
  + Store\_Location\_Type
  + Store\_Type
* **Features Discarded:**
* Item ID
* Item Weight
* Item fat amount
* Item category
* Store ID
* Establishment year

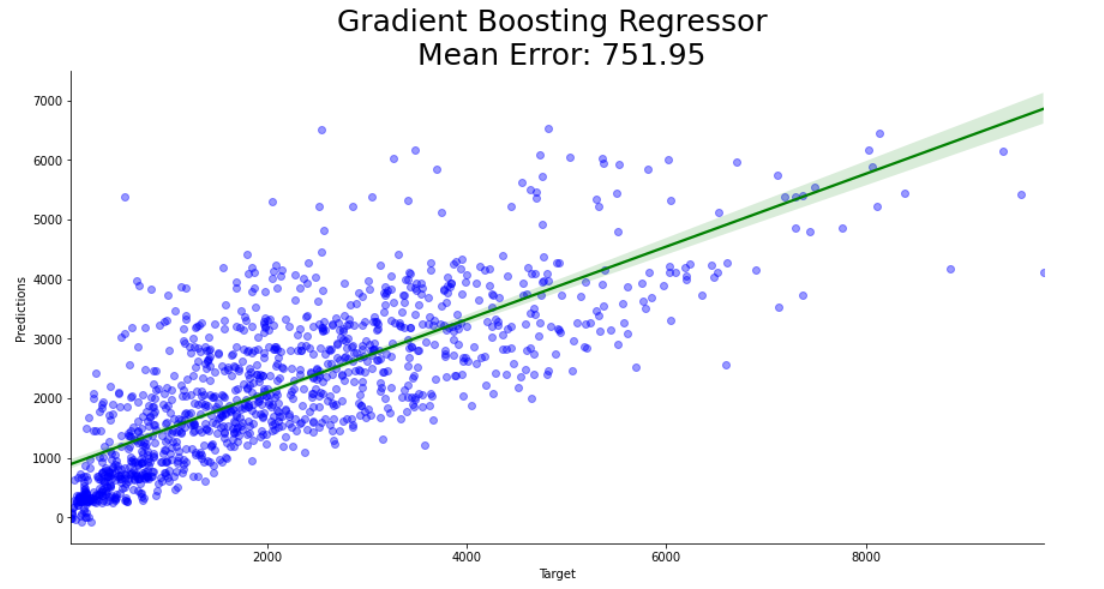
# **Data Size**

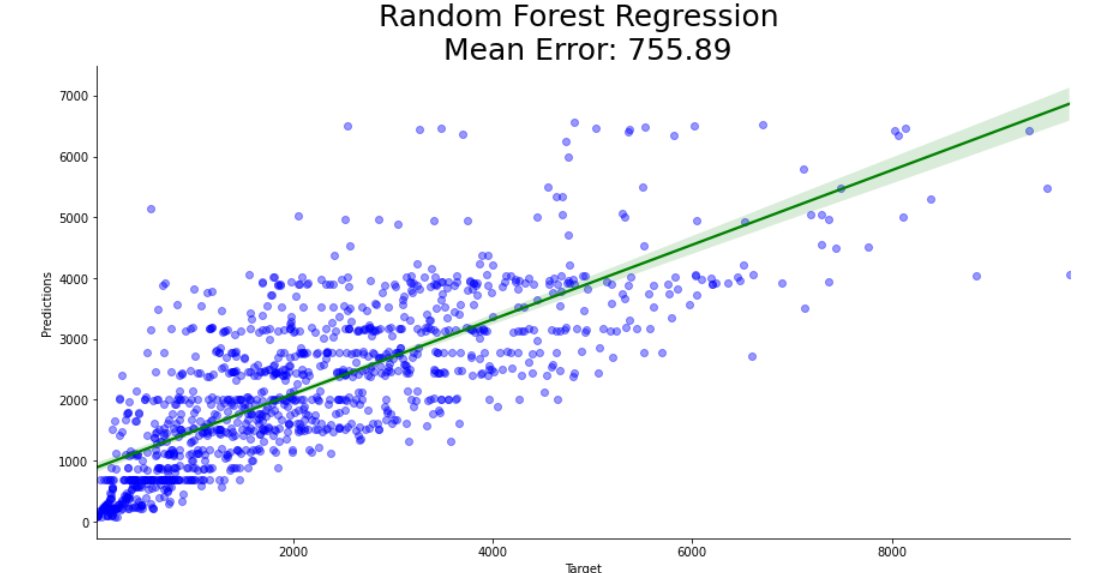
Training Data Size: 80% -> 4800

Testing Data Size: 20% -> 1200

# **Acquired Results**







# Conclusion

**In this phase, we saw the effect of pre-processing on the accuracy & error. In addition, we saw the impact of preprocessing on highly correlated features.**

**My initial intuition was adding new columns that holds extra information or simplifying the categorical data may have a positive effect on the output but, unfortunately it did not.**