PR-Milestone 1 Report

Table of Contents

\

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

2- Replacing the following in item fat amount column using pandas rename:

a. LF to Low Fat
b. Low fat to Low Fat
c. Reg Regular

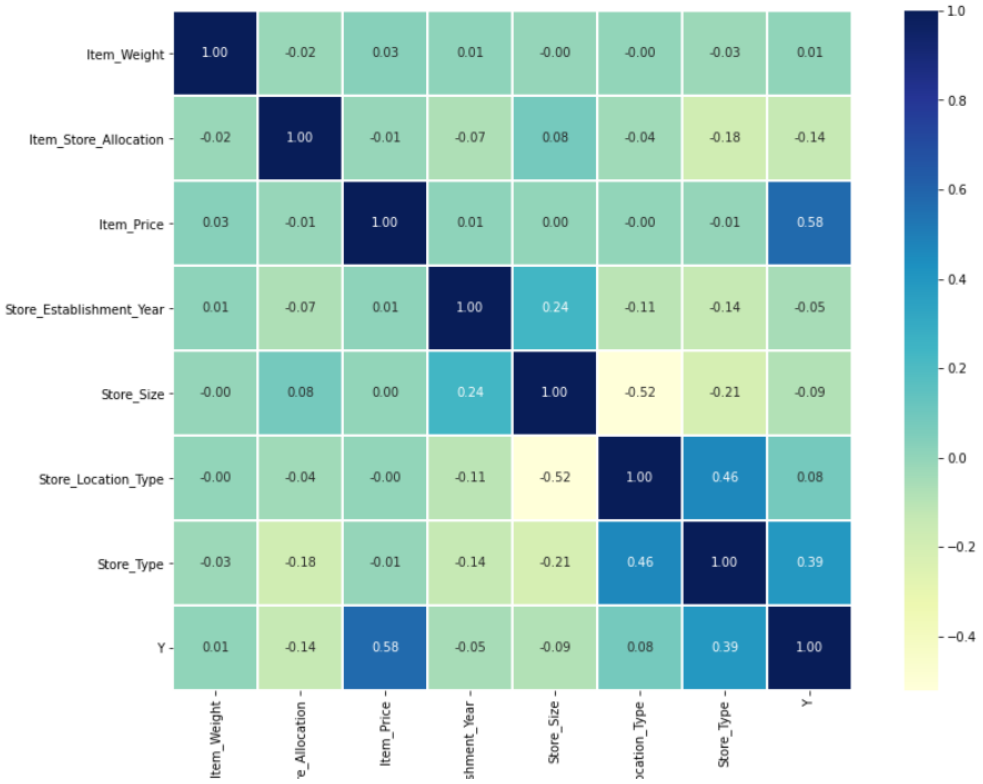
3- Filling the NaN's in item weight using with backward fill– using fillna function found in pandas with a method parameter='bfill'

4- Filling the 0's in item store allocation with backward fill -using fillna function found in pandas with a method parameter='bfill'

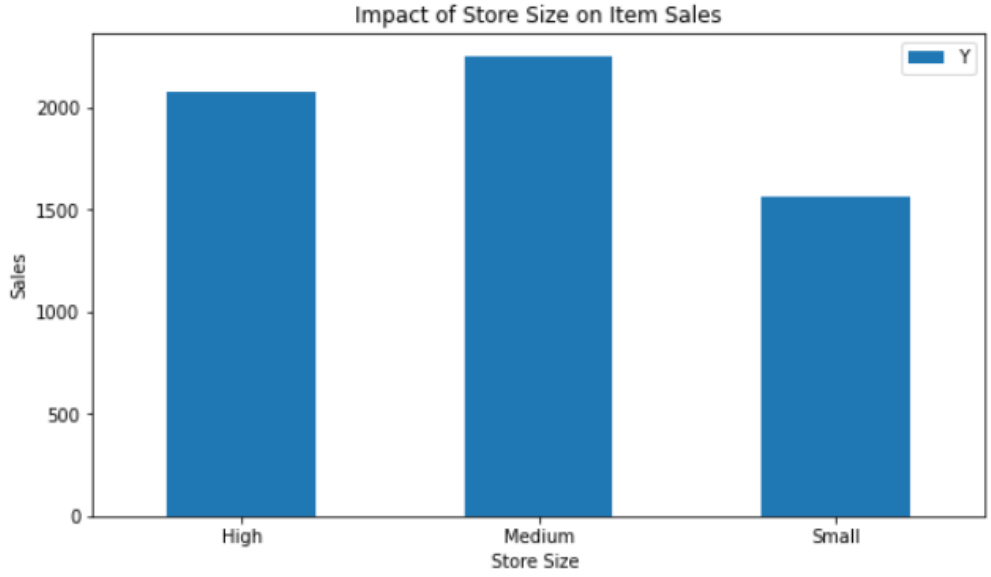
5- Filling the NaN's in store size with backward fill - using fillna function found in pandas with a method parameter='bfill'

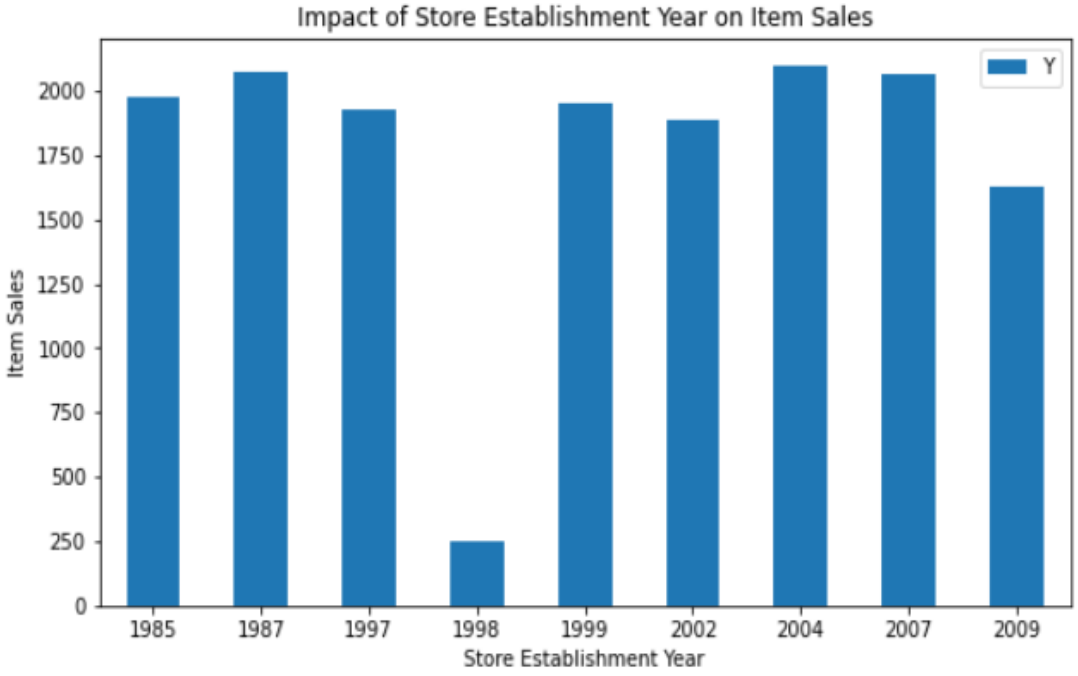
6- We performed label encoding on Store_Size , Store_Location_type, store_type

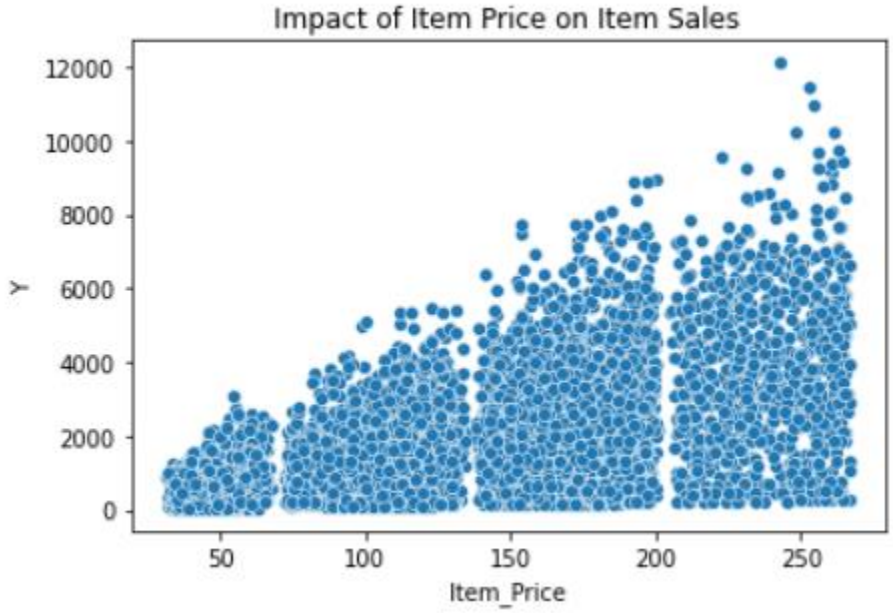
Data Analysis

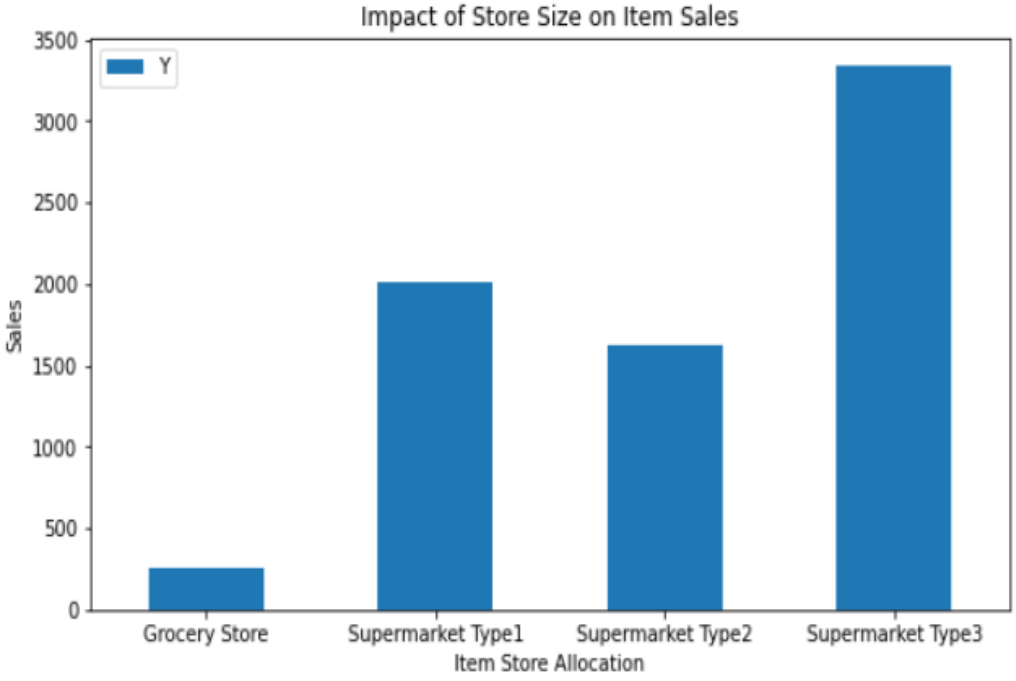


Visually, as we can see above. This is the correlation between each feature with the other. In our case, we will mainly focus on the relation between all features with our target which is “Y”.


Impact of Store Size on Item Sales
Y
2000
1500
1000
500
0
Sales
High
Medium
Small
Store Size


Impact of Store Establishment Year on Item Sales
Y
2000
1750
1500
1250
1000
750
500
250
0
Item Sales
1985
1987
1997
1998
1999
2002
2004
2007
2009
Store Establishment Year


Impact of Item Price on Item Sales
12000
10000
8000
6000
4000
2000
0
Y
50
100
150
200
250
Item_Price


Impact of Store Size on Item Sales
Y
3500
3000
2500
2000
1500
1000
500
0
Sales
Grocery Store
Supermarket Type1
Supermarket Type2
Supermarket Type3
Item Store Allocation

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 Acquired Results

Model	Description	Advantages	Mean absolute error
Random Forest	**Random forest builds multiple decision tree s 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**

Features Used/Discarded

- **Features Used:**
  - Item_Store_Allocation
  - Item_Price
  - Store_Size
  - Store_Location_Type
  - Store_Type

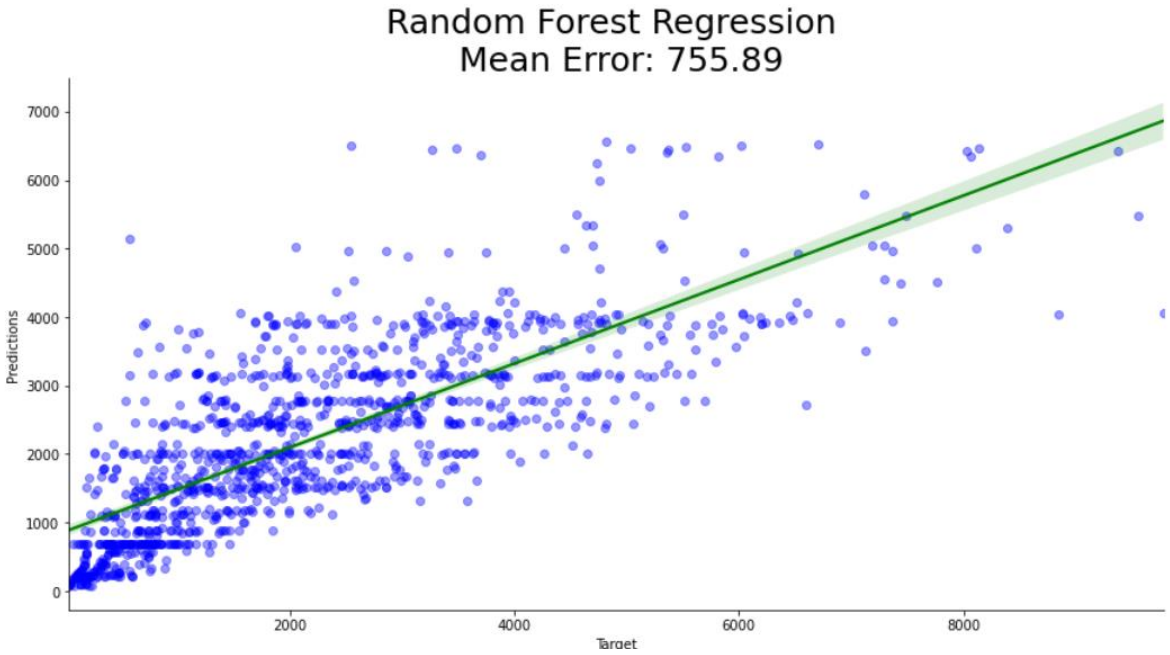
- **Features Discarded:**
  - Item ID
  - Item Weight
  - Item fat amount
  - Item category
  - Store ID
  - Establishment year

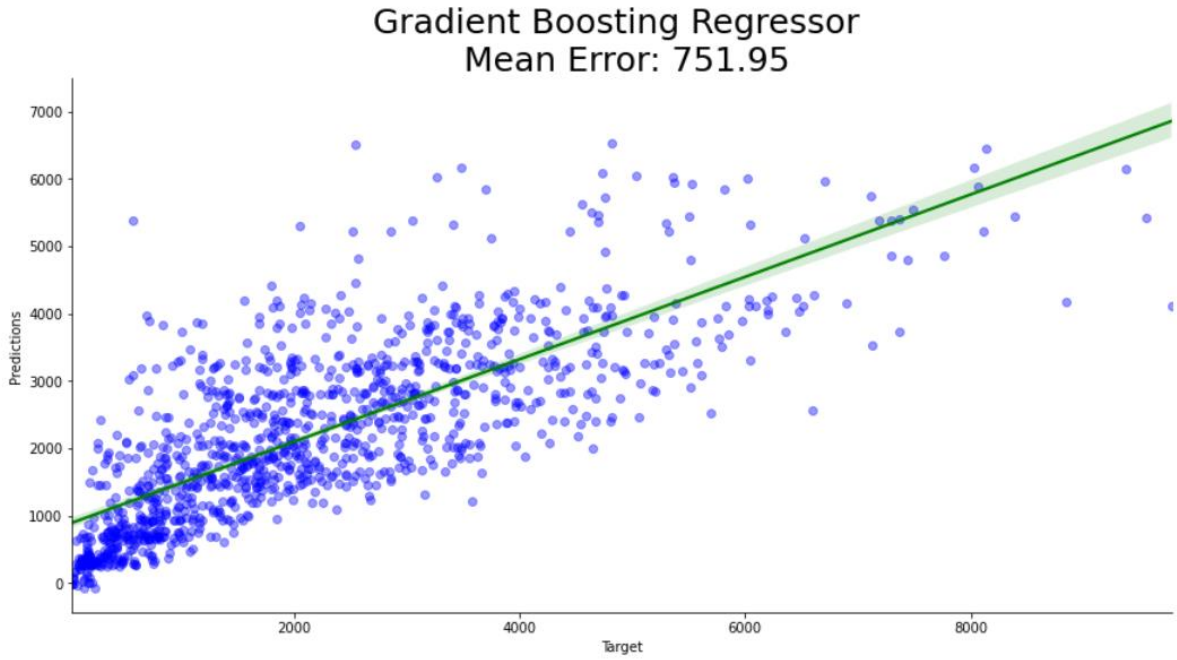
Data Size

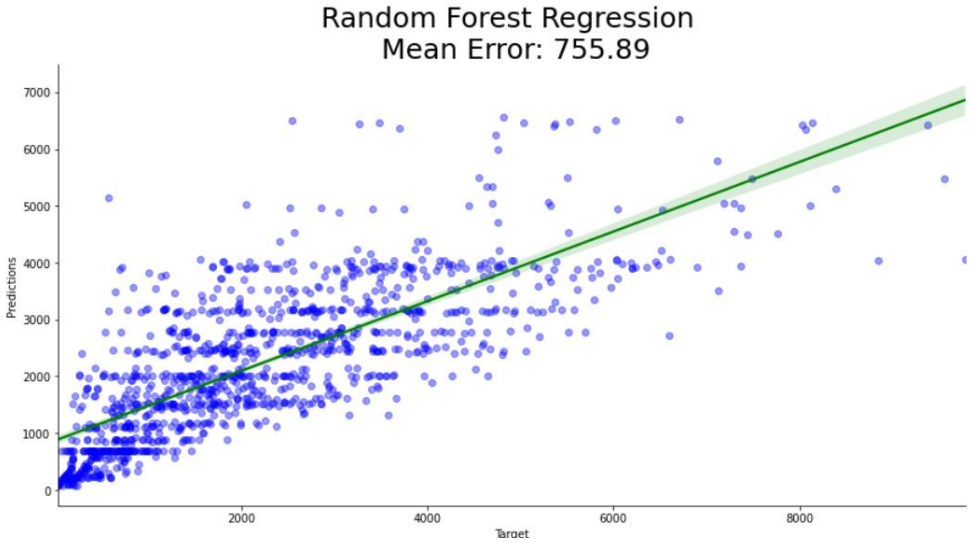
Training Data Size: 80% -> 4800

Testing Data Size: 20% -> 1200

Acquired Results






Random Forest Regression
Mean Error: 755.89
Predictions
Target
0
1000
2000
3000
4000
5000
6000
7000
2000
4000
6000
8000

Conclusion

In this phase, we saw the effect of pre-processing on the accuracy & erro r. In addition, we saw the impact of preprocessing on highly correlated fe atures.

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