SHIV NADAR UNIVERSITY
DELHI NCR

ACADEMY OF CONTINUING EDUCATION

Analytics Olympiad 2021

Arpita Saggar

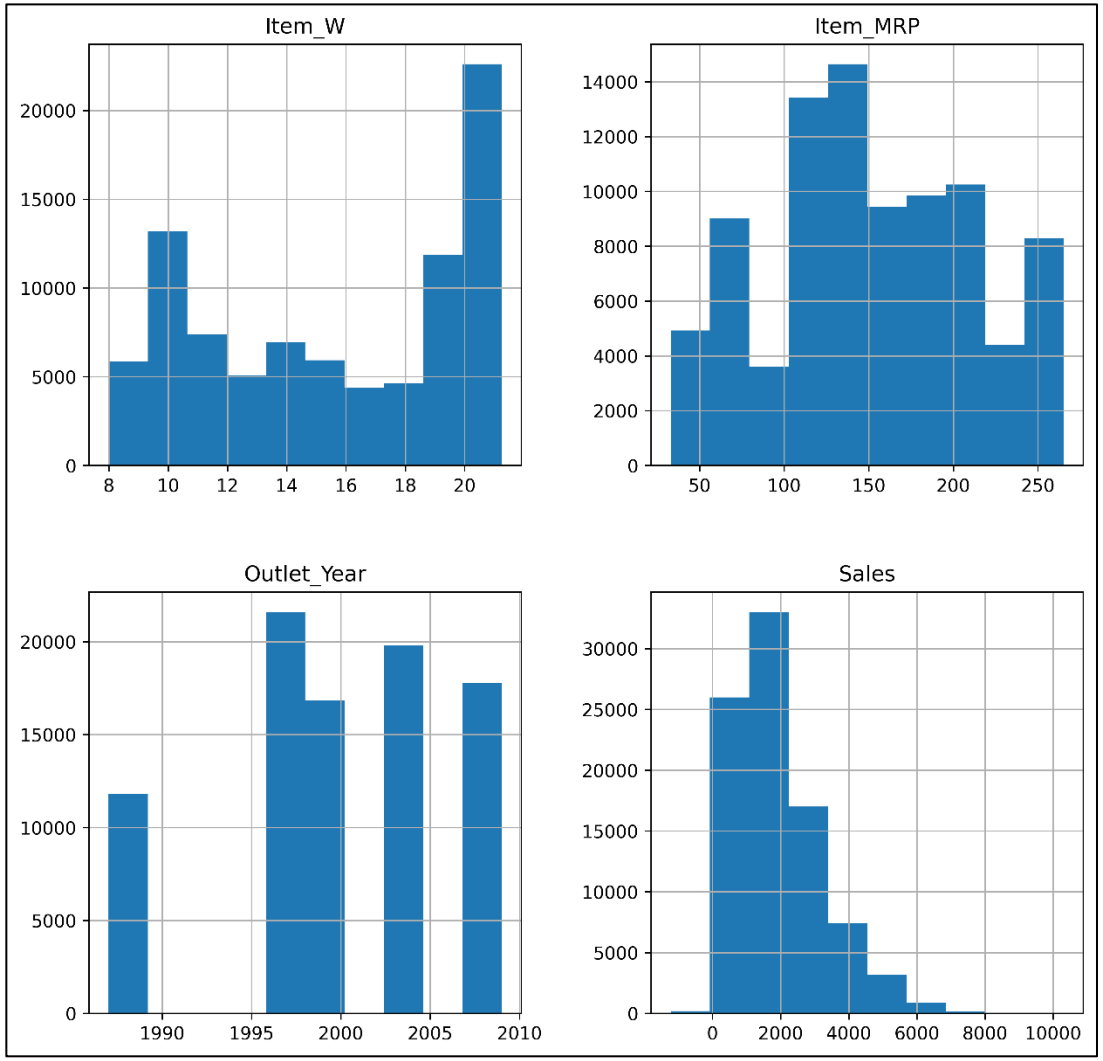
Data Understanding

- Dataset Overview: Sales data from mega marts including details of items and outlets they are sold at, along with total sales output (target variable).
- Data attributes:
  - Item_ID: Item Identification Number
  - Item_W: Item Weight
  - Item_Type: Item
  - Item_MRP: MRP of the Product
  - Outlet_ID: Outlet ID
  - Outlet_Year: Outlet Establishment year
  - Outlet_Size: Size of the outlet
  - Outlet_Type: Type of the outlet
  - Sales: Total sales from the outlet

- Size of Training set: 87864 instances x 9 features
- Size of Test set: 37656 instances x 9 features
- No null values present
- Data types:
  - Item_ID: object
  - Item_W: float64
  - Item_Type: object
  - Item_MRP: float64
  - Outlet_ID: object
  - Outlet_Year: int64
  - Outlet_Size: object
  - Outlet_Location_Type: object
  - Sales: float64

- Outlier removal is not performed since outliers may provide unique insights (due to nature of data), i.e., high sales during festivals/holidays, or low sales for a newly opened outlet, as well as to maximize data available for training.
- Number of unique values for each variable:
  - Item_ID: 895
  - Item_W: 87283
  - Item_Type: 16
  - Item_MRP: 87814
  - Outlet_ID: 5
  - Outlet_Year: 12
  - Outlet_Size: 3
  - Outlet_Location_Type: 3
  - Sales: 87760

- Distribution of Numerical Variables:



- Distribution of Categorical Variables (Item_ID not presented here due to large number of unique values):

	Outlet_ID Values	Count
0	OUT035	24071
1	OUT046	20850
2	OUT018	17156
3	OUT049	13356
4	OUT013	12431

	Outlet_Size Values	Count
0	Small	48614
1	Medium	26683
2	High	12567

	Outlet_Location_Type Values	Count
0	Tier 1	33567
1	Tier 3	29044
2	Tier 2	25253

	Item_Type Values	Count
0	Baking Goods	14666
1	Fruits and Vegetables	14328
2	Meat	8099
3	Snack Foods	7817
4	Household	6018
5	Soft Drinks	5396
6	Frozen Foods	5129
7	Canned	4565
8	Dairy	3807
9	Others	3688
10	Breads	3509
11	Hard Drinks	3129
12	Health and Hygiene	2995
13	Starchy Foods	2974
14	Seafood	1107
15	Breakfast	637

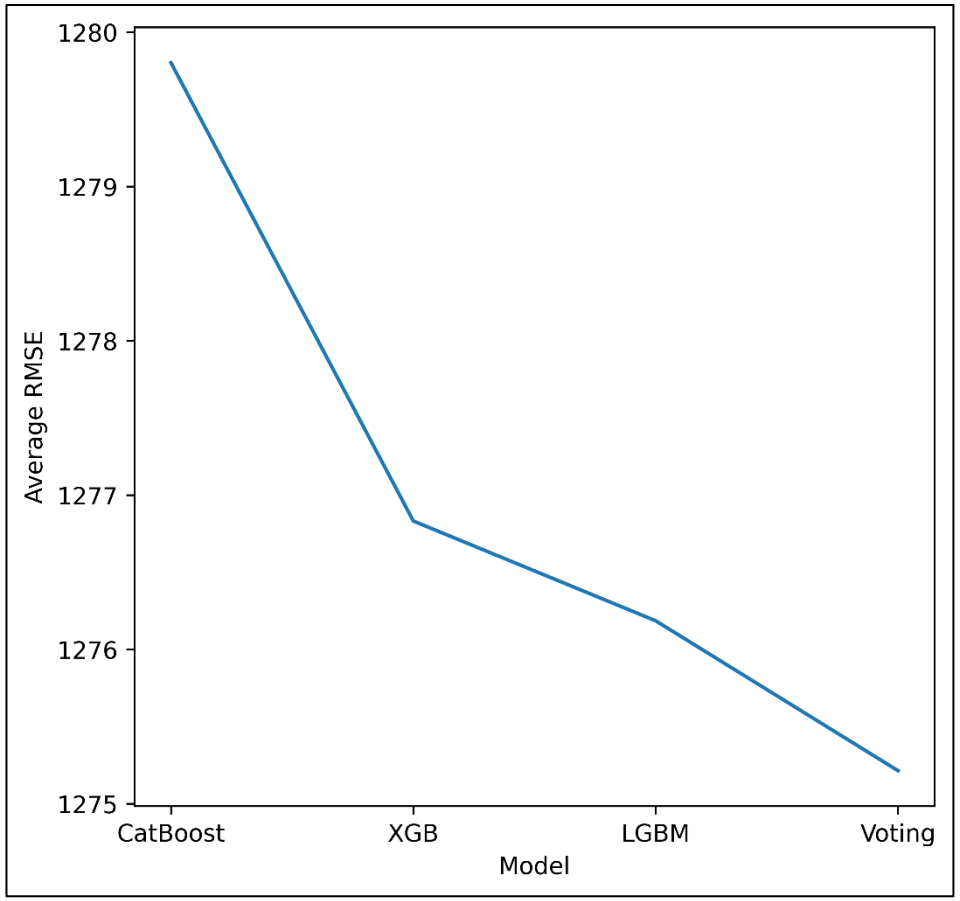
Data Preparation

- No null/missing values present, so no imputation strategy required.
- Outliers not removed for reasons mentioned previously.
- No obvious arithmetic/logical relations exist between columns, so no new features have been derived.
- Initially, all features are selected and model is fit and evaluated. Less relevant features are later removed using permutation importance. Permutation importance evaluates how randomly shuffling a single column of the test data, leaving the target and all other columns in place, affects model performance on shuffled data. Mean feature importance (over repeated calculations) is used to remove less important features.

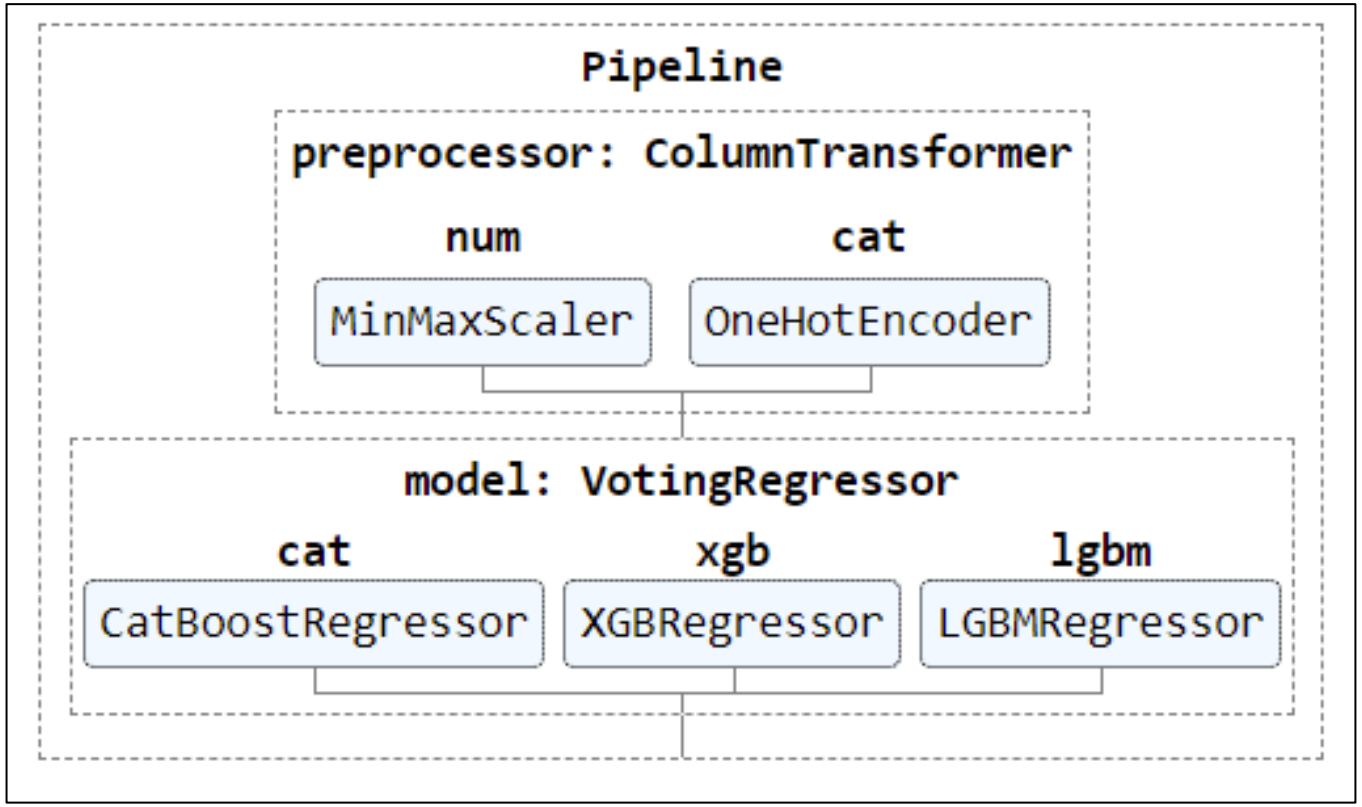
- Binning is not performed. Instead, features are all transformed to range [0, 1].
- Numerical columns are scaled to range [0, 1].
- Categorical columns are encoded as a one-hot numeric array. If an unknown categorical feature is present during transformation, it is ignored.
- Data is partitioned by splitting into 10 consecutive folds (without shuffling). Each fold is then used once as a test set while the 9 remaining folds form the training set.

Model Building & Evaluation

- Boosting methods almost always outperform other regression techniques, so only boosting techniques were tried, namely CatBoost, LightGBM and XGBoost.
- Model finally used is Voting Regressor comprising of CatBoostRegressor, LGMBRegressor and XGBRegressor, since combined model outperforms individual models (lowest RMSE over 10 folds of training data).



- Hyperparameter Tuning is performed, and only hyperparameter set is iterations for CatBoostRegressor, set to 100. For LGBMRegressor and XGBRegressor, default models are used.



- Permutation Importance (PI) is used to evaluate feature importance and remove features with low/negligible impact on model performance.
- Initially, all 8 features are used to train the model. Column Item_ID has lowest PI, so it is removed.

	cols	scores
0	Item_W	0.001720
1	Item_MRP	0.001103
2	Outlet_Year	0.016243
3	Item_Type	0.003028
4	Outlet_ID	0.010537
5	Outlet_Size	0.003784
6	Outlet_Location_Type	0.001473
7	Item_ID	0.000872

- Permutation Importance is again calculated after training model using 7 features this time. Item_W has lowest PI score, so it is removed.
- Next, PI scores are calculated after training using 6 features. Outlet_Location_Type has lowest PI, so it is removed.
- Model is then evaluated using 5 features, however, RMSE increases, so 6 features are finalised, indicated in the table on bottom right.

	cols	scores
0	Item_W	0.000933
1	Item_MRP	0.001533
2	Outlet_Year	0.017038
3	Item_Type	0.004382
4	Outlet_ID	0.009705
5	Outlet_Size	0.005468
6	Outlet_Location_Type	0.002682

	cols	scores
0	Item_MRP	0.002782
1	Outlet_Year	0.018811
2	Item_Type	0.005020
3	Outlet_ID	0.011064
4	Outlet_Size	0.004795
5	Outlet_Location_Type	0.002544

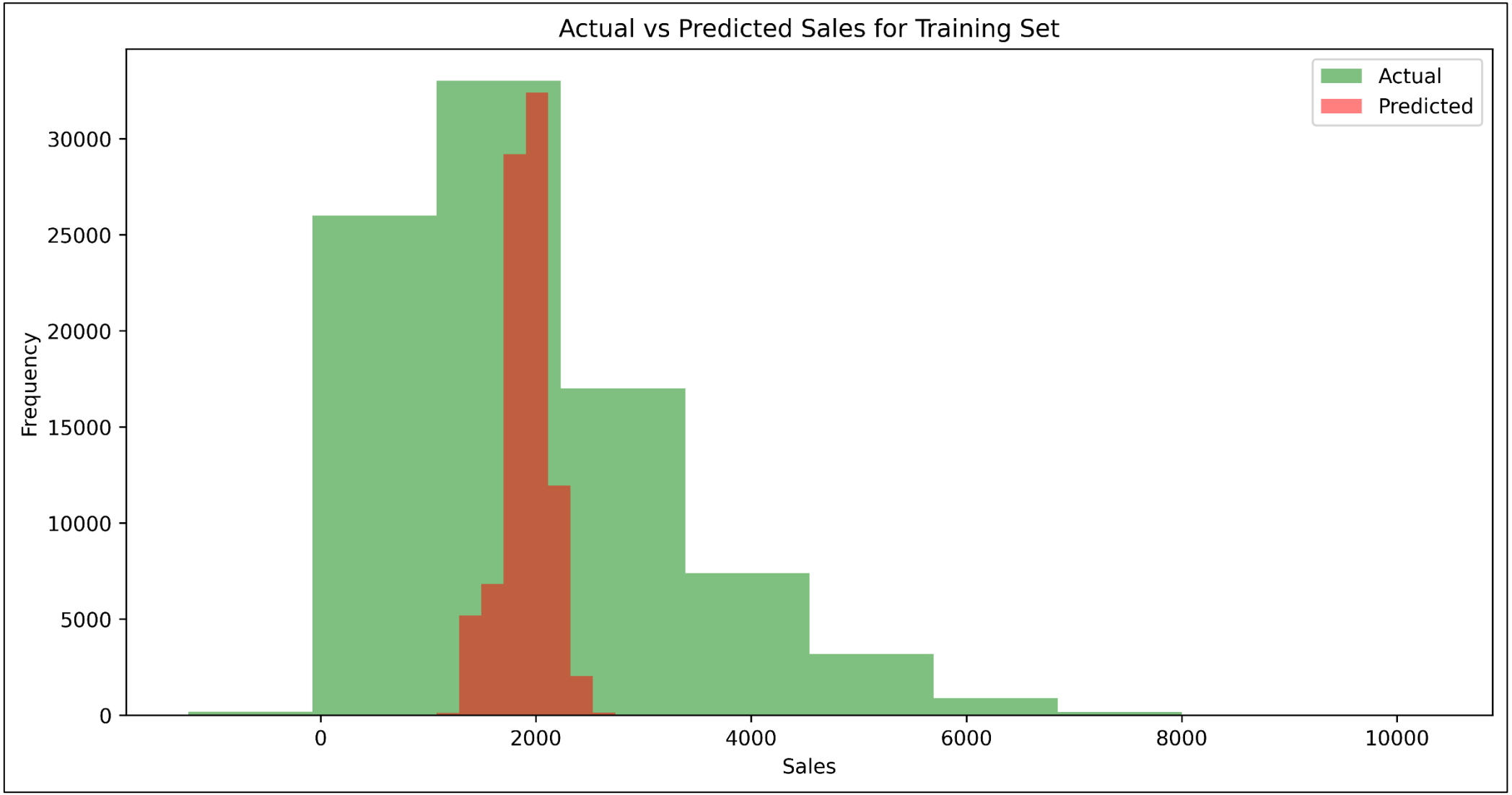
Results

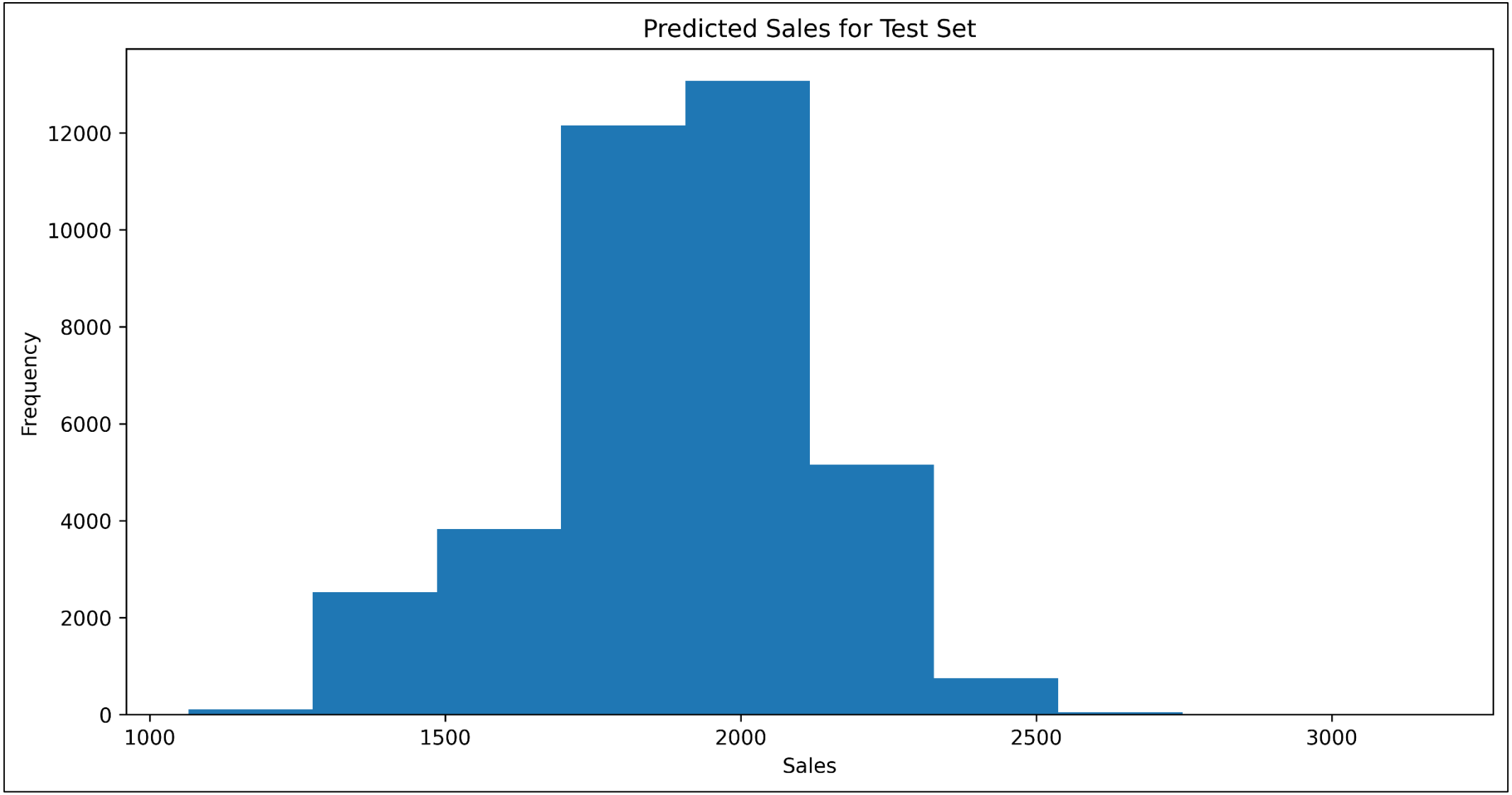
- Average RMSE over 10 folds is 1275.2134. RMSE for each fold is given in the figure.
- RMSE for the entire training set is 1257.478
- Feature importance is as given in the table on bottom right, in previous slide.
- RMSE for test set (Final Score on MachineHack leaderboard) is 1270.27982.

```
Fold 0 RMSE:  1278.275069011596
Fold 1 RMSE:  1268.8580042059336
Fold 2 RMSE:  1259.1690621977698
Fold 3 RMSE:  1297.7764392728168
Fold 4 RMSE:  1270.8664491155937
Fold 5 RMSE:  1268.3107132353502
Fold 6 RMSE:  1249.0513997980543
Fold 7 RMSE:  1295.5000000533314
Fold 8 RMSE:  1302.4432499457123
Fold 9 RMSE:  1261.8838794345063
Average RMSE:  1275.2134266270664
```

Model Insights

- Skewed nature of target variable is reflected in the model predictions for both training set and test set (next 2 slides). Model is biased towards sales values close to 2000. This may also explain variation in RMSE across folds, which is approximately 50.
- Since the model does not produce any predictions less than 0 or greater than 3000 for either training or test set (next 2 slides), it may not be very reliable in extreme cases (very high sales, or failing/negative sales). This situation can be rectified by availability of more/balanced data.


Actual vs Predicted Sales for Training Set
Actual
Predicted
Frequency
0
5000
10000
15000
20000
25000
30000
0
2000
4000
6000
8000
10000
Sales


Predicted Sales for Test Set
Frequency
0
2000
4000
6000
8000
10000
12000
1000
1500
2000
2500
3000
Sales

Business Insights

- In general, permutation importance scores are higher for outlet-based attributes (highest for columns Outlet_ID and Outlet_Year), as compared to item-based attributes (Item_ID and Item_W have lowest scores and are not considered in final model). This suggests that an outlet is a better indicator for estimating sales than an item, so future data collection should focus on including more outlet-based attributes, such as user experience, outlet locality, ease of access etc.

	cols	scores
0	Item_MRP	0.002782
1	Outlet_Year	0.018811
2	Item_Type	0.005020
3	Outlet_ID	0.011064
4	Outlet_Size	0.004795
5	Outlet_Location_Type	0.002544

- Since outlet-based attributes are better indicators of sales, further investment into user experience at an outlet should be prioritised over marketing, research or development of new products.
- Since outlet-based attributes are deemed most important, outlets with large sales output may be used as blueprints to design new outlets/refurbish old/failing ones.

Results and Recommendations

- As noted earlier, since VotingRegressor outperforms individual models, it is most suited for deployment. However, skewed nature of target variable should be kept in mind while deploying.
- Further insights into the model and its features may be obtained using XAI frameworks like LIME and SHAP. However, their drawbacks must be noted while being used for analysis.
- Focus for further data collection should be on greater detail and balance.