SHIV NADAR UNIVERSITY
DELHI NCR

ACADEMY OF CONTINUING EDUCATION

# Analytics Olympiad 2021

Arpita Saggar

# Data Understanding

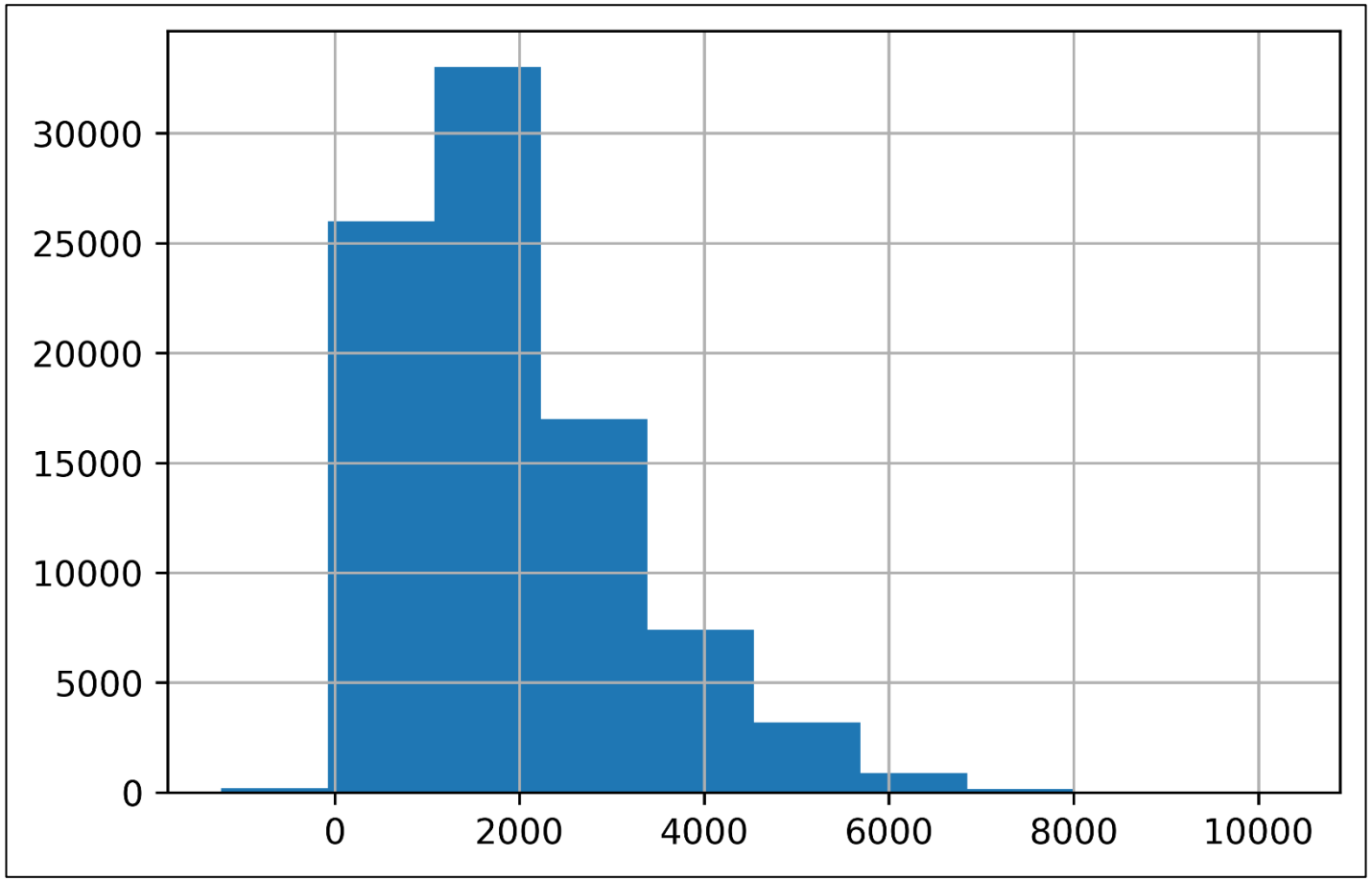
- Dataset Overview: Sales data from mega marts, including details of items and outlets they are sold at, along with total sales output).
- Size of Training set: 87864 instances x 12 features
- Size of Test set: 37656 instances x 12 features
- No null values present

- Data attributes:
  - Item_ID: Item Identification Number
  - Item_W: Item Weight
  - Item_Type: Item
  - Item_FC: Fat content in Item
  - Item_Vis: Item Visibility
  - Item_MRP: MRP of the Product
  - Outlet_ID: Outlet ID
  - Outlet_Year: Outlet Establishment year
  - Outlet_Size: Size of the outlet
  - Outlet_Location_Type: Type of the outlet location
  - Outlet_Type: Type of the outlet
  - Sales: Total sales from the outlet

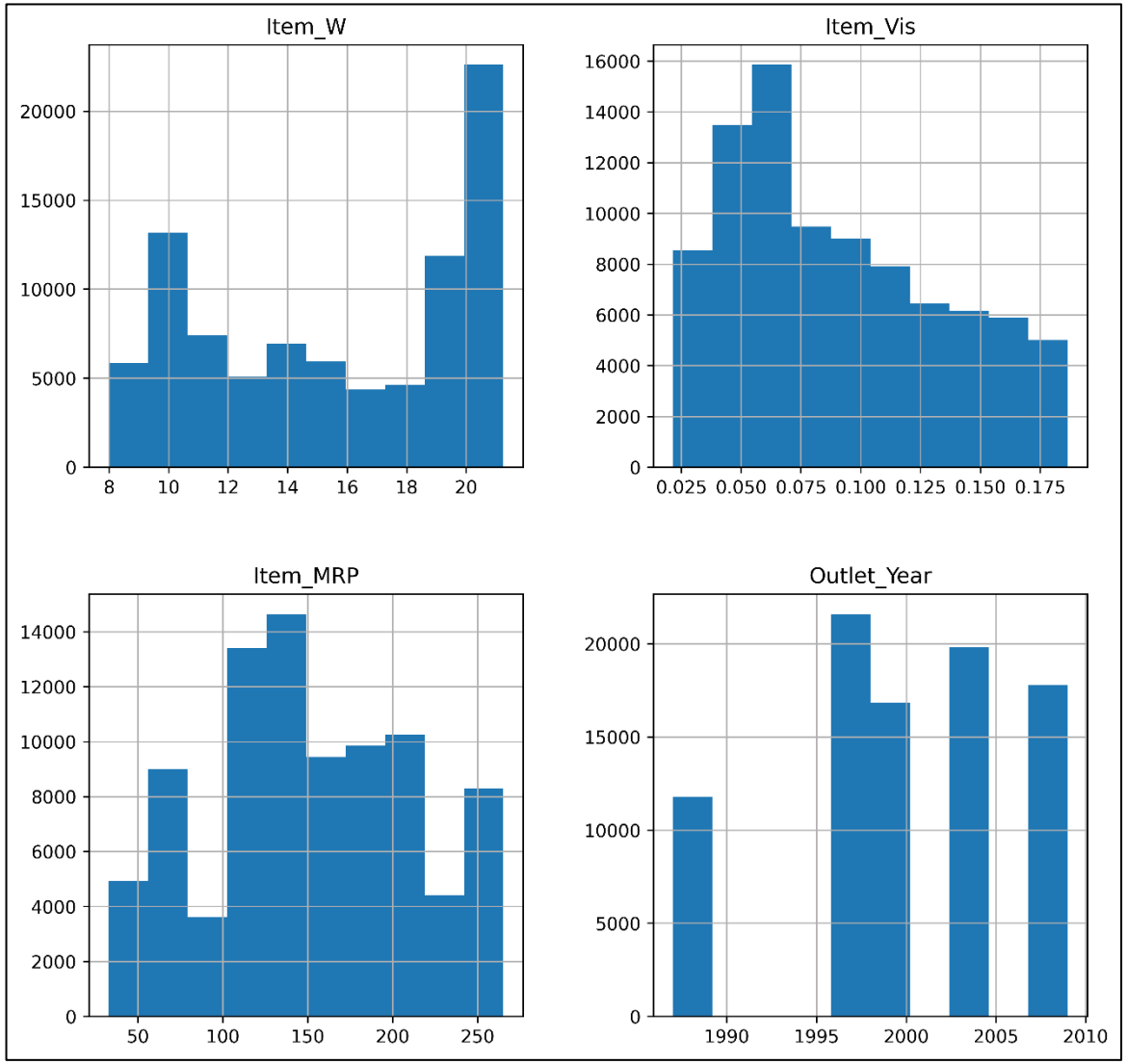
- Data types:
  - Item_ID: object
  - Item_W: float64
  - Item_FC: object
  - Item_Vis: float64
  - Item_Type: object
  - Item_MRP: float64
  - Outlet_ID: object
  - Outlet_Year: int64
  - Outlet_Size: object
  - Outlet_Location_Type: object
  - Outlet_Type: object
  - Sales: float64

- 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_FC: 5
  - Item_Vis: 87826
  - Item_Type: 16
  - Item_MRP: 87814
  - Outlet_ID: 5
  - Outlet_Year: 12
  - Outlet_Size: 3
  - Outlet_Location_Type: 3
  - Outlet_Type: 2
  - Sales: 87760

- Distribution of Target Variable(Sales):



- 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 |

| | Outlet_Type Values | Count |
|---|---|---|
| 0 | Supermarket Type1 | 71621 |
| 1 | Supermarket Type2 | 16243 |

| | Item_FC Values | Count |
|---|---|---|
| 0 | Low Fat | 46761 |
| 1 | Regular | 32214 |
| 2 | LF | 5858 |
| 3 | low fat | 2098 |
| 4 | reg | 933 |

| | 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.
- Outlier removal has not been performed.
- 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.

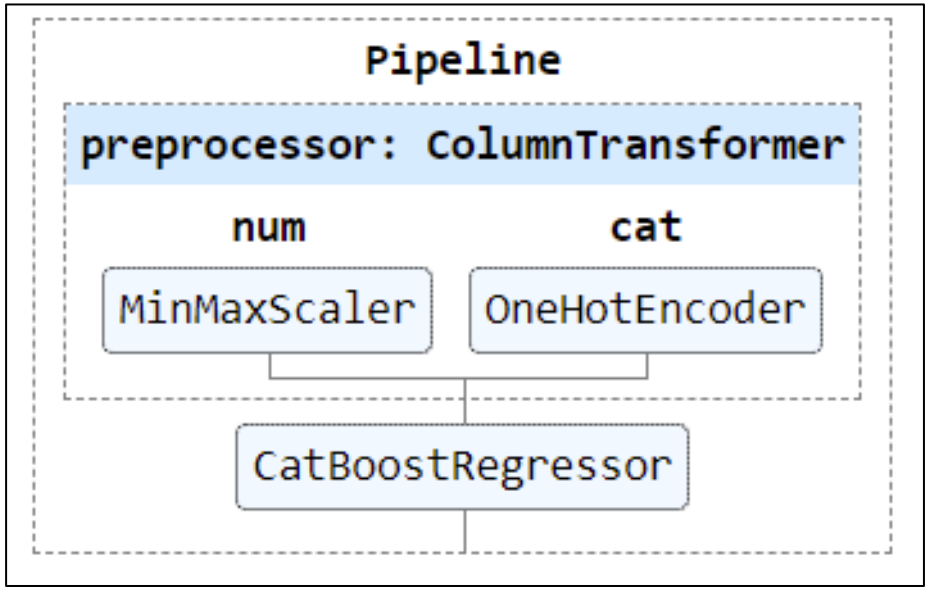
- Binning is not performed. Instead, features are all transformed to range [0, 1].
- Numerical columns are scaled to range [0, 1].
- Categorical columns are encoded as a one-hot numeric array. If an unknown categorical feature is present during transformation, it is ignored.
- Data is partitioned by 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. VotingRegressor consisting of all 3 models was also evaluated.
- Model finally used is CatBoostRegressor, LGMBRegressor and XGBRegressor, since it outperforms the other models (lowest RMSE over 10 folds of training data).

| Model | Average RMSE over 10 folds |
|---|---|
| LGMBRegressor | 1274.314 |
| XGBRegressor | 1274.400 |
| VotingRegressor | 1273.747 |
| CatBoostRegressor | 1272.993 |

- Hyperparameter Tuning is performed, and iterations are set to 500 with learning_rate=0.04.
- Permutation Importance (PI) is used to evaluate feature importance and remove features with low/negligible impact on model performance.
- Initially, all 11 features are used to train the model. These are recursively reduced by eliminating the feature with lowest PI at each step, till RMSE stops decreasing.
- Item_W, Item_Vis and Item_ID are removed in this manner.
- Final model is trained on the entire training set, using 8 features.



# Results

- Average RMSE over 10 folds is 1272.99. RMSE for each fold is given in the figure.
- RMSE for the entire training set is 1260.25.
- RMSE for test set is 1268.84 (from MachineHack platform).

```
Fold 0 RMSE:  1271.6469883121435

Fold 1 RMSE:  1269.7460058259733

Fold 2 RMSE:  1269.540823160822

Fold 3 RMSE:  1270.631445824456

Fold 4 RMSE:  1263.9201114672865

Fold 5 RMSE:  1273.8158401503713

Fold 6 RMSE:  1271.5772724979722

Fold 7 RMSE:  1288.443992726281

Fold 8 RMSE:  1274.7788468146039

Fold 9 RMSE:  1275.833250837431

Average RMSE:  1272.9934577617341
```

# Model Insights

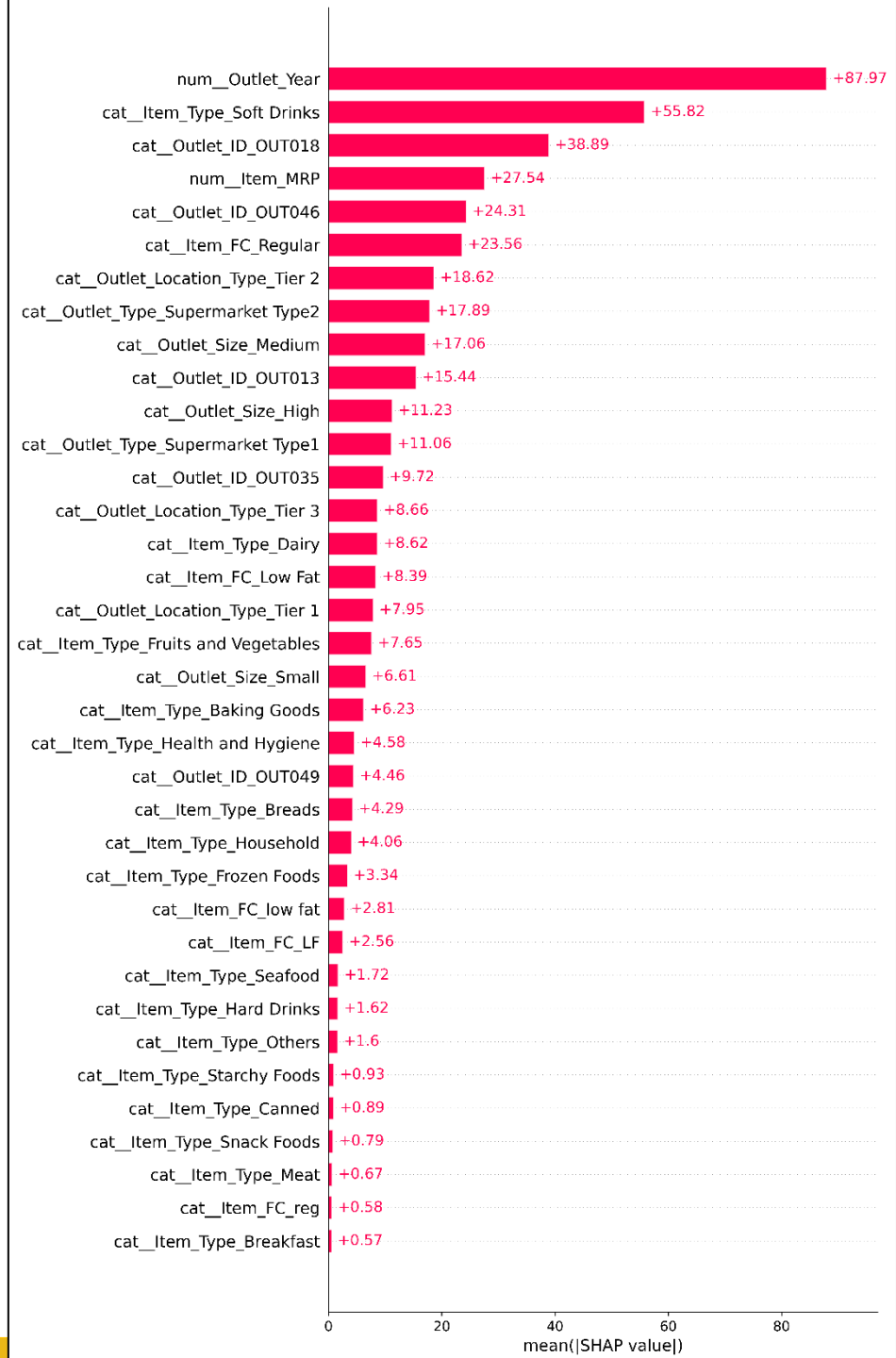
- Table shows PI scores for the final model.
- All attributes eliminated from the model are item attributes. Additionally, outlet based attributes have higher relevance scores than item based attributes, which suggests that sales are more reliant on outlets rather than inventory.
- Future data collection should focus on including more outlet-based attributes, such as user experience, outlet locality, ease of access etc.
- Outlet_Year is the most relevant feature, while Item_MRP is least relevant.

| | cols | scores |
|---|---|---|
| 0 | Item_MRP | 0.002222 |
| 1 | Outlet_Year | 0.018928 |
| 2 | Item_Type | 0.003635 |
| 3 | Outlet_ID | 0.007938 |
| 4 | Item_FC | 0.002733 |
| 5 | Outlet_Type | 0.003994 |
| 6 | Outlet_Size | 0.002587 |
| 7 | Outlet_Location_Type | 0.002807 |

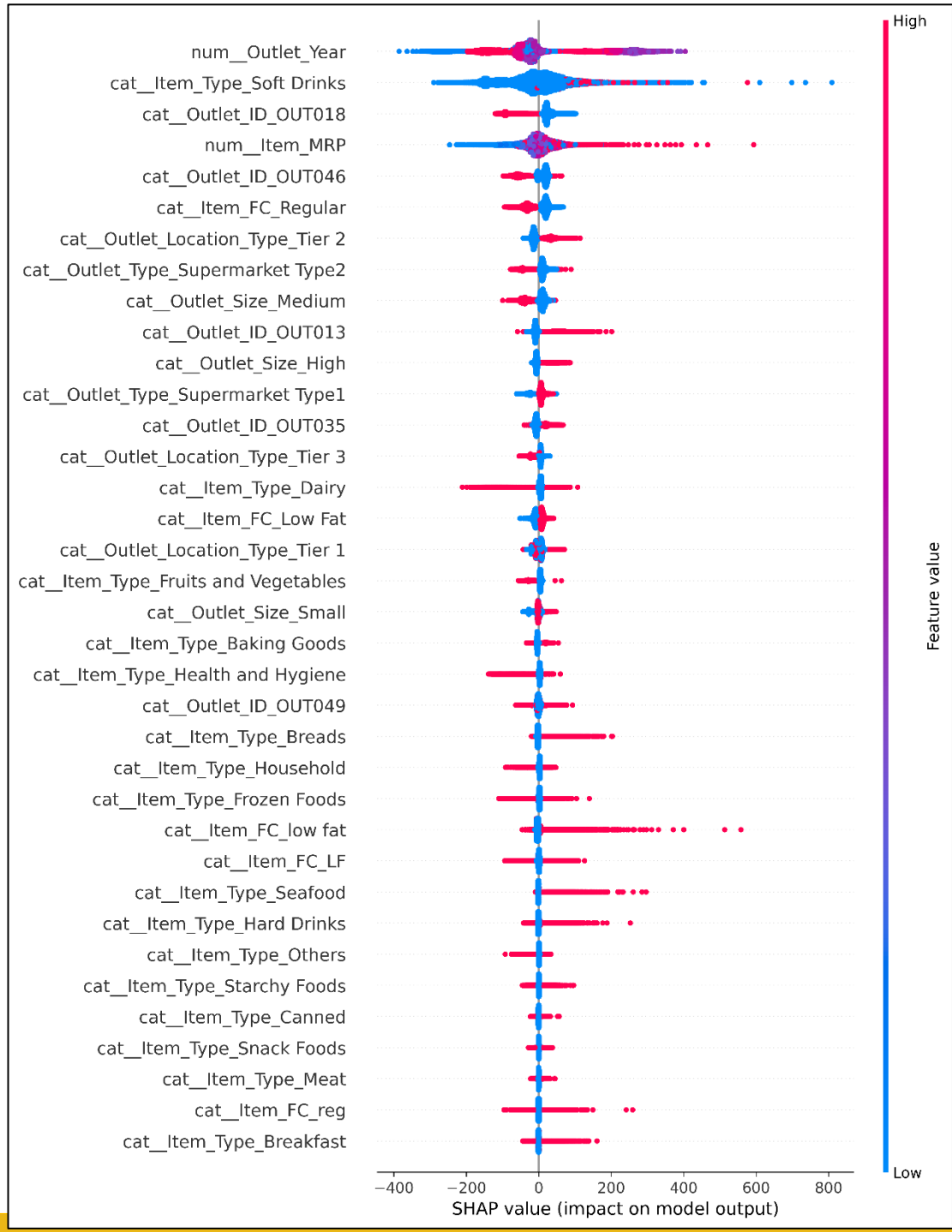
# Business Insights & Recommendations

- Permutation importance scores are useful in determining which features are important, but not how they affect model performance.
- For this purpose SHAP Values (an acronym from SHapley Additive exPlanations) have been used. They break down a prediction to show the impact of each feature.
- Shap values show how much a given feature changed our prediction (compared to if we made that prediction at some baseline value of that feature).
- SHAP has multiple Explainer modules. However, since CatBoostRegressor is a tree-base method, SHAP's TreeExplainer is used to generate explanations.

- A global feature importance plot shows the mean absolute value for that feature over all the given samples. The feature space here is modified due to use of One Hot Encoding.
- While permutation importance gave an overview of which features are importance, the modified feature space in SHAP also specifies values of a particular feature which are relevant.
- For instance, whether or not an item is a soft drink is also a strong indicator for predicting Sales, more so than other Item_Type(s) such as Meat or Seafood.

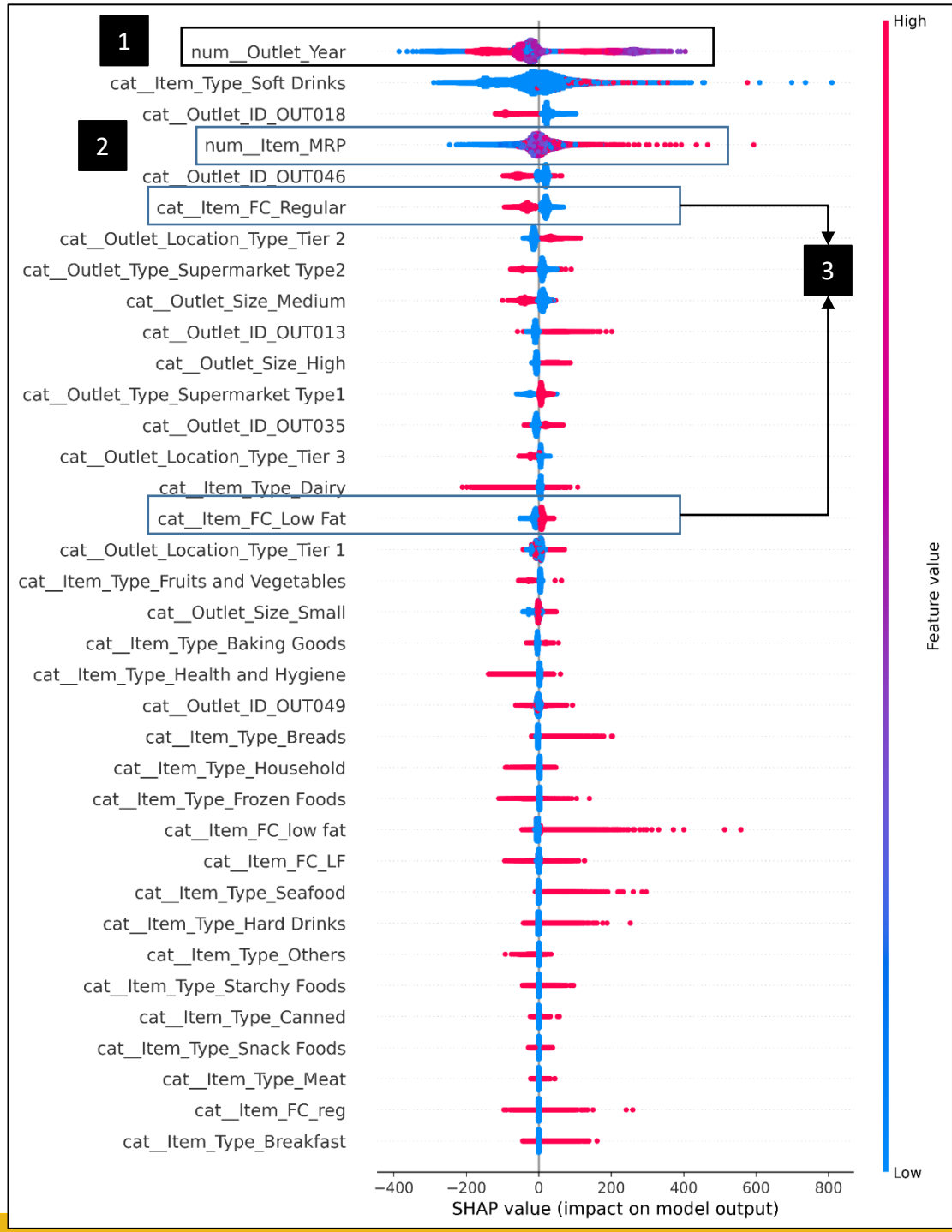


- The beeswarm plot is designed to display an information-dense summary of how the top features in a dataset impact the model's output. Each instance the given explanation is represented by a single dot on each feature fow.
- Color is used to display the original value of a feature. Purple indicates overlap of positive and negative values.
- This plot is made of many dots. Each dot has three characteristics:
  - Vertical location shows what feature it is depicting
  - Color shows whether that feature was high or low for that row of the dataset
  - Horizontal location shows whether the effect of that value caused a higher or lower prediction.



1. Generally, low values of Outlet_Year causes lower predictions, while higher values cause higher predictions. Thus newer outlets attract more business. This information may be used to decide which outlets to close/upgrade.
2. Products with high MRP cause higher sales while cheaper products cause lower sales.
3. Products with Regular fat content lower sales while products with Low fat content increase sales.

Such insights may be used to drive business decisions.



# The End