**Objective:** To predict item outlet sales

**Data Filter:** Considered only records in training data which contains outlets specified in test records

**Outlier Analysis:** Item\_Outlet\_Sales attribute in training dataset capped at 1%/99%

**Approach-1:**

Attributes Item\_Weight and Outlet\_Size had missing values. Following methods were experimented for missing value imputation: KNN Imputer, with adversarial validation to evaluate if imputed value is causing a different behavior of the variables

Data Transformations:

* Converted ‘Outlet\_Size’ from (1/2/3) as its ordinal variable
* Derived variable ‘Outlet\_Age’ by subtracting ‘Outlet\_Establishment\_Year’ from current year
* Aggregating at minimum, maximum, variance, standard deviation (measure of dispersion) and average (measure of central tendency) for item\_outlet\_sales for Item\_Identifier, Item\_Outlet\_Sales and Outlet\_Identifier as these 3 attributes had higher cardinality

Feature Engineering/Model Training

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Prediction Model Technique | | | | | |
| Random Forest | LightGBM | XGBoost | CatBoost | KNN | SVM |
| Feature Engineering Technique | Lasso | **P** | **P** | **P** | **P** | **P** | **P** |
| Ridge | **P** | **P** | **P** | **P** | **P** | **P** |
| ElasticNet | **P** | **P** | **P** | **P** | **P** | **P** |
| Random Forest | **P** | **P** | **P** | **P** | **P** | **P** |
| LightGBM | **P** | **P** | **P** | **P** | **P** | **P** |
| XGBoost | **P** | **P** | **P** | **P** | **P** | **P** |
| CatBoost | **P** | **P** | **P** | **P** | **P** | **P** |

From the above approach tested, Lasso Regularisation + CatBoost Regression model gave the best RMSE of around 1200.

**Approach-2**

Attributes Item\_Weight and Outlet\_Size had missing values. Following methods were experimented for missing value imputation: Outlet\_Size imputed to dummy value “Not Available” with adversarial validation to evaluate if imputed value is causing a different behavior of variable. Did not treat missing value of ‘Item\_Weight’ attribute as CatBoostRegressor can handle missing values.

Feature Engineering: CatBoostRegressor feature importance method, with categorical attributes specified under ‘cat\_features’ parameter

Model Training: CatBoostRegressor model training gave the RMSE value of 1148, being the best approach as compared to models in Approach-1