Big Mart Sales Prediction – Approach

Overview of Final Submission

- Final model: Stacked Regressor Model

- The Base Learners used are
    - Lasso Regression
    - XGBoost Regressor model
    - CatBoost Regressor model

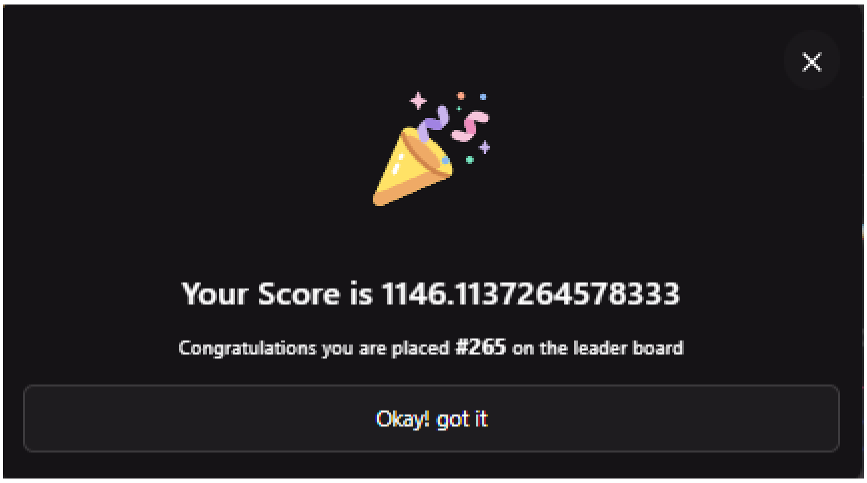
- The Meta Learner used is
    - Polynomial Ridge Regression

Error Metric

RMSE on private test data: 1146.1137

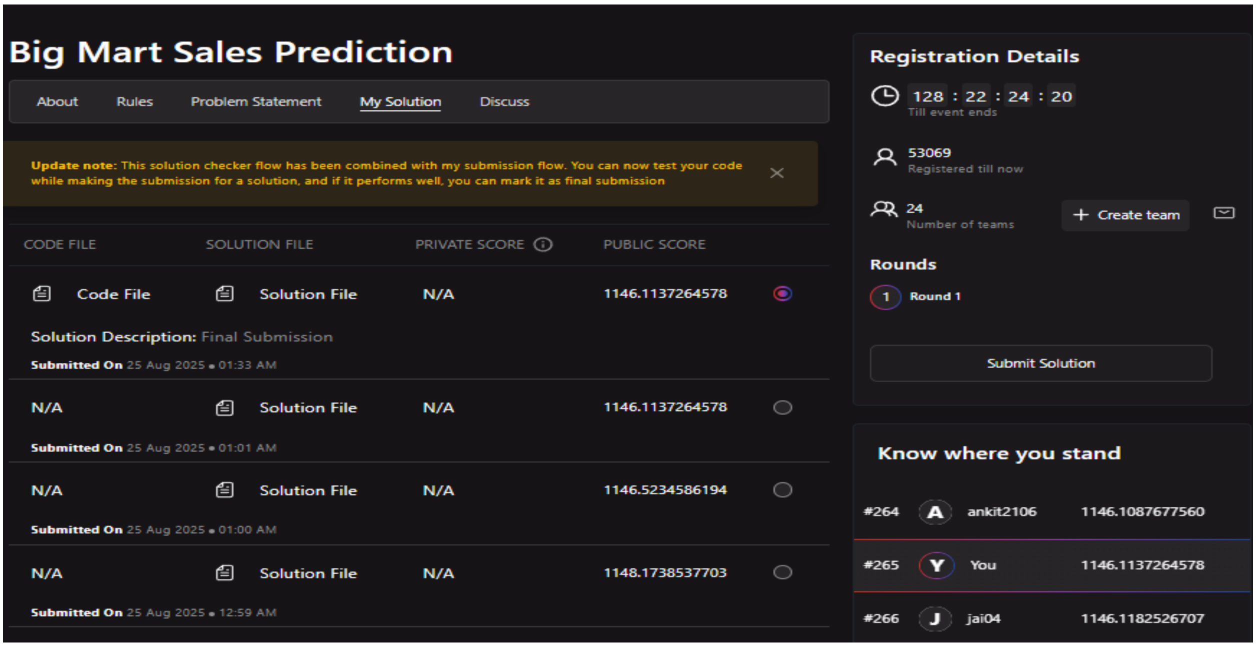
Rank: 265

Screenshot of Submission



Your Score is 1146.1137264578333

Congratulations you are placed #265 on the leader board



Big Mart Sales Prediction

Update note: This solution checker flow has been combined with my submission flow. You can now test your code while making the submission for a solution, and if it performs well, you can mark it as final submission

CODE FILE	SOLUTION FILE	PRIVATE SCORE	PUBLIC SCORE	
Code File	Solution File	N/A	1146.1137264578	●
Solution Description: Final Submission **Submitted On** 25 Aug 2025 • 01:33 AM				
N/A	Solution File	N/A	1146.1137264578	○
Submitted On 25 Aug 2025 • 01:01 AM				
N/A	Solution File	N/A	1146.5234586194	○
Submitted On 25 Aug 2025 • 01:00 AM				
N/A	Solution File	N/A	1148.1738537703	○
Submitted On 25 Aug 2025 • 12:59 AM				

Registration Details

128 : 22 : 24 : 20
Till event ends

53069
Registered till now

24
Number of teams

Rounds

1 Round 1

Know where you stand

#264	A	ankit2106	1146.1087677560
#265	Y	You	1146.1137264578
#266	J	jai04	1146.1182526707

Data Exploration

The detailed functions used for Data explorations can be found in the src/utils module in github

Step 1: Understanding the item and outlet data in detail.

- Count of items and outlets.
- Check for duplicate entries or wrong entries.
- Check for null values.

Step 2: Feature Engineering and Data Cleaning

Item Weight (Method 1)

- It was observed that there were null values in the case of item weight. Also, the same item which were having missing weight had their weight mentioned in other outlets.
- Parse through both train and test data to gather all item details and store as a metadata json file.
- All item details were identified with this method.
- Output file: Item_mapping.json in Data/Output folder.

Item Weight (Method 2)

- Missing values were imputed with the mean of item weight

Outlet details (Method 1)

- No direct method was identified to fill in the missing details in case of outlet type and outlet location type.
- In this case the details were filled using information from outlets having similar sales output.
- A similar metadata json file was created for outlet details.
- Output file: Outlet_mapping.json in Data/Output folder.

Outlet details (Method 2)

- For outlet_size, missing values were imputed with the mode of outlet_size

Item_Fat_Content

- Standardized all the categories into two. For example, low fat was mentioned as LF, low fat and Low Fat.
- Standardized the categories to: Low Fat and Regular.
- For non-consumables, fat content was changed to non-edible.

Item Visibility

- Few items had visibility mentioned as 0 which is suspicious.
- In such cases the visibility is taken as the mean item visibility.

All the data exploration steps are consolidated in the following folder: Link

Relevant plots and figures can be found here : Link

Years (New Feature)

- This is the number of years of operations of the outlet which is formulated as

  Years = reference year – year of establishment of outlet

  Where reference year = 2025

MRP Bands

- A categorical tagging of the item into 4 bands based on the quantile in which they were priced at.
  - Q1: 0 – 25th quantile
  - Q2: 25-50th quantile
  - Q3: 50-75th quantile
  - Q4: 75-100th quantile

Item Category

- The Item identifier started with a pair of strings categorizing each item into 3. This was used as new item category feature
  - Category 1: FD (Food)
  - Category2: NC (Non-Consumables)

        - Category3: DR (Drinks)

MRP Squared

- Square of the MRP price was added as a new feature.
- This was done to generalize the model to the higher MRP value items.

Dropped columns

- Item identifier
    - high cardinality
    - replaced by other item related features like item category etc.
- Outlet_established_year
    - Replaced by years
- Item_type
    - Replaced by combined item type

These were the major newly added features

Feature Transformations

- One hot encoding
    - The following columns were one hot encoded as they were categorical and didn't have an inherit hierarchical order of categories
        - Item_Fat_Content
        - Outlet_Type
        - Item_Category
- Label Encoding
    - The following columns were ordinal encoded as the categories involved are much larger and there is an inherit hierarchy for categories.
        - MRP Band
        - Outlet Identifier
        - Outlet Size
        - Outlet location type
- Scaling
    - The following numerical features were scaled using Standard Scalar
        - Item_weight
        - Item_visibility
        - Years
        - MRP squared
        - Item mrp

All the steps are consolidated in the following transformer function and model pipeline

Transformer functions that were build can be found in the utils module : Transformers

Model Developed

Baselines

- Linear regression with regularization (Lasso and Ridge) :
- Random Forest Regression:
- XGB Regression:

Baselines models were able to perform well and provided an RMSE of 1188.2 in the competition test data. This took the score up to a rank of 4000 in the leaderboard.

Details of the model implementation can be found in the links provided along with them.

All attempted models can be found here : Models

Ensemble model

Further improvements were achieved using ensemble model which took the model up to a rank of 256 in leaderboard with a RMSE of 1046.11 on test data.

The final ensemble model is achieved after finetuning the base learners parameters via trial and error. **The initial models were overfitting significantly and the regularization parameters were increased as well as constraining the branches and depth of the tree model to reduce complexity.**

The final model link can be found here: MultiEnsembleModel

Final submission file can be found here: Final Submission