Big Mart Sales Prediction

Objective

Predict sales (Item\_Outlet\_Sales) at the item-outlet level using structured data provided for Big Mart. The goal is to minimize RMSE on the test set and gain meaningful insights from model performance across different outlet types and item categories.

Data Understanding & Preprocessing

* Loaded the dataset containing 8,523 rows (train) and 5,681 rows (test).
* Key variables: Item\_Weight, Item\_Visibility, Item\_MRP, Outlet\_Size, Outlet\_Location\_Type, Outlet\_Type, etc.
* Engineered new features:
  + Item\_Category derived from Item\_Identifier
  + Outlet\_Age calculated from Outlet\_Establishment\_Year
  + High\_Performing\_Outlet flag based on outlet performance
* Applied label encoding to categorical features.
* Performed log transformation on Item\_Outlet\_Sales to normalize the target distribution.
* Handled outliers by capping sales at the 99th percentile.

Model Building & Experiments

**Baseline Model: XGBoost Regressor**

* Trained on combined numeric + encoded categorical features.
* Evaluated using RMSE and MAPE on actual (inverse-log-transformed) sales.
* Feature importance plotted to understand key drivers of sales.

**Hyperparameter Tuning**

* Used RandomizedSearchCV to tune:
  + n\_estimators, max\_depth, learning\_rate, subsample, colsample\_bytree
* Improved test RMSE after tuning.

**Residual Analysis**

* Plotted:
  + Residuals vs. Actual sales : Heteroscedasticity observed at higher sales
  + Actual vs. Predicted sales
* Potential solutions : segment-wise modeling or deep learning with weighted loss.

**Deep Learning Experiment**

* Tested a multi-input Keras model with embedding layers for categorical inputs.

**Segmented Modeling**

* Built separate XGBoost models per Outlet\_Type.
* Calculated RMSE and MAPE for each model to assess segment-specific accuracy.

Evaluation summary

* **XGBoost Model:**
  + Test RMSE: ~1070.75
  + Test MAPE: ~50.32%
* **Segmented XgBoost Model by Outlet Type:** Segmented Modelling slightly improved RMSE & MAPE by capturing outlet specific sales dynamics

|  |  |  |
| --- | --- | --- |
| **Outlet Type** | **RMSE** | **MAPE (%)** |
| Supermarket Type 1 | 1021.4 | 48.2 |
| Supermarket Type 2 | 1055.7 | 52.1 |
| Supermarket Type 3 | 980.0 | 47.8 |
| Grocery Store | 1110.2 | 55.3 |

* **Deep Learning Model :** DL model performed slightly better than XGBoost baseline due to its ability to capture non-linear interactions
  + Test RMSE : ~1063.1
  + Test MAPE : ~49.7%

NOTE: I was unable to to submit my results to the Analytics Vidya Platform due to technical issues with the platform and I am sharing my codes and results via email. I would be happy to get an evaluation done by the platform once it’s issues are fixed.