Big Mart Sales Prediction

EDA:

- There are 8523 rows and 12 columns in the dataset. Item_Outlet_Sales is the target variable and Item_Identifier & Outlet_Identifier are the identifier columns.
- Mainly we have 9 columns to work with.
- Categorical cols: ['Item_Fat_Content', 'Item_Type', 'Outlet_Size', 'Outlet_Location_Type', 'Outlet_Type']
- Numerical cols: ['Item_Weight', 'Item_Visibility', 'Item_MRP', 'Outlet_Establishment_Year']
- Checked the columns for missing values. The next section describes that process.
- Plotted histogram graphs of all numerical columns to see if there are any natural groupings, outliers, skewness in the data.
- Found that Item_MRP had 4 distinct groups. Used that to create an Item_MRP_level feature.
- Plotted correlation heat map for all the features and the target column. Excluded highly correlated features from training.

Missing value imputation:

- 'Item_Weight' and ‘Outlet_Size’ columns have missing values.
- For filling 'Item_Weight' missing values I found the mean value of 'Item_Weight' per ‘Item_Identifier’ and replaced the missing values with it.
- ‘Outlet_Size’ takes value from {Small, Medium, High}. The ‘Outlet_Size’ missing values had ('Outlet_Type','Outlet_Location_Type') values as (Supermarket 1, Tier2) and (Grocery store, Tier 3). I compared the 'Outlet_Type' and 'Outlet_Location_Type' columns for non missing values and found that they had the ‘Outlet_Size’ as Small. Hence the missing values in ‘Outlet_Size’ were filled with Small.

Feature Engineering:

- 'Item_Fat_Content' had values ['Low Fat', 'Regular', 'low fat', 'LF', 'reg']. Since similar values are repeated with different names, I grouped them into just 2 categories: LF and reg.
- Created age column as: df['age'] = 2013 - df['Outlet_Establishment_Year'].
- Created price per weight column: df['price_per_weight'] = df['Item_MRP']/df['Item_Weight']
- Created grouped columns for Item_Weight, Item_MRP and Item_Visibility. The groupings had values of Small, Medium and High. Tried several cutoff values (based on percentile values and the histogram plots) for grouping each of these columns.
- Created Item_Identifier_cat column which is the first two characters of Item_Identifier column. FD, NC, DR were the values. These might indicate information like non-consumable, food, etc which will be helpful in modelling.
- Created columns like is_perishable, is_processed, is_supermarket.
- Grouped Item_Type into 3 broad categories: food, drinks, non-edible.
- Created interaction features like Visibility_Perishable, visibility_mrp,

- Created log_visibility as visibility column had values skewed towards left.
- Overall experimented with about 25 features in the modelling:
  - 'Item_Weight',
  - 'Item_Fat_Content',
  - 'Item_Type',
  - 'Item_MRP',
  - 'age',
  - 'Outlet_Size',
  - 'Outlet_Location_Type',
  - 'Outlet_Type',
  - 'MRP_level',
  - 'weigh_level',
  - 'log_visibility_level',
  - 'price_per_weight',
  - 'Item_Visibility',
  - 'log_visibility',
  - 'Outlet_Identifier',
  - 'Item_Identifier_cat',
  - 'is_perishable',
  - 'visibility_mrp',
  - 'is_processed',
  - 'is_supermarket',
  - 'item_type_cat',
  - 'log_visibility_mrp',
  - 'Outlet_Tier',
  - 'item_avg_sales',
  - 'outlet_avg_sales'

Model:

- Since I had several categorical features, I used the CatBoostRegressor model as it handles the categorical features quite well without us having to do processing like label encoding, etc.
- I split the training dataset into 4:1 ratio for train and validation datasets.
- For hyperparameter tuning I used GridSearchCV to experiment with 'depth', 'learning_rate' and 'iterations'.
- This doc contains the RMSE obtained on different set of features: https://docs.google.com/spreadsheets/d/1oxNGw3buiNa7S1qHEBC3tMECsGN7vmUl8oJgI7NIKYA/edit?usp=sharing

Results:

Achieved best RMSE of 1146.69 on test data on this feature set: age, Item_Fat_Content, Item_MRP, Item_Type, Item_Weight, MRP_level, Outlet_Location_Type, Outlet_Size, Outlet_Type