**WOMART SALES PREDICTION**

The entire solution for the project has been divided into three parts

* Cleaning the data
* Extracting Information and applying feature engineering
* Choosing the right model and evaluating its performance

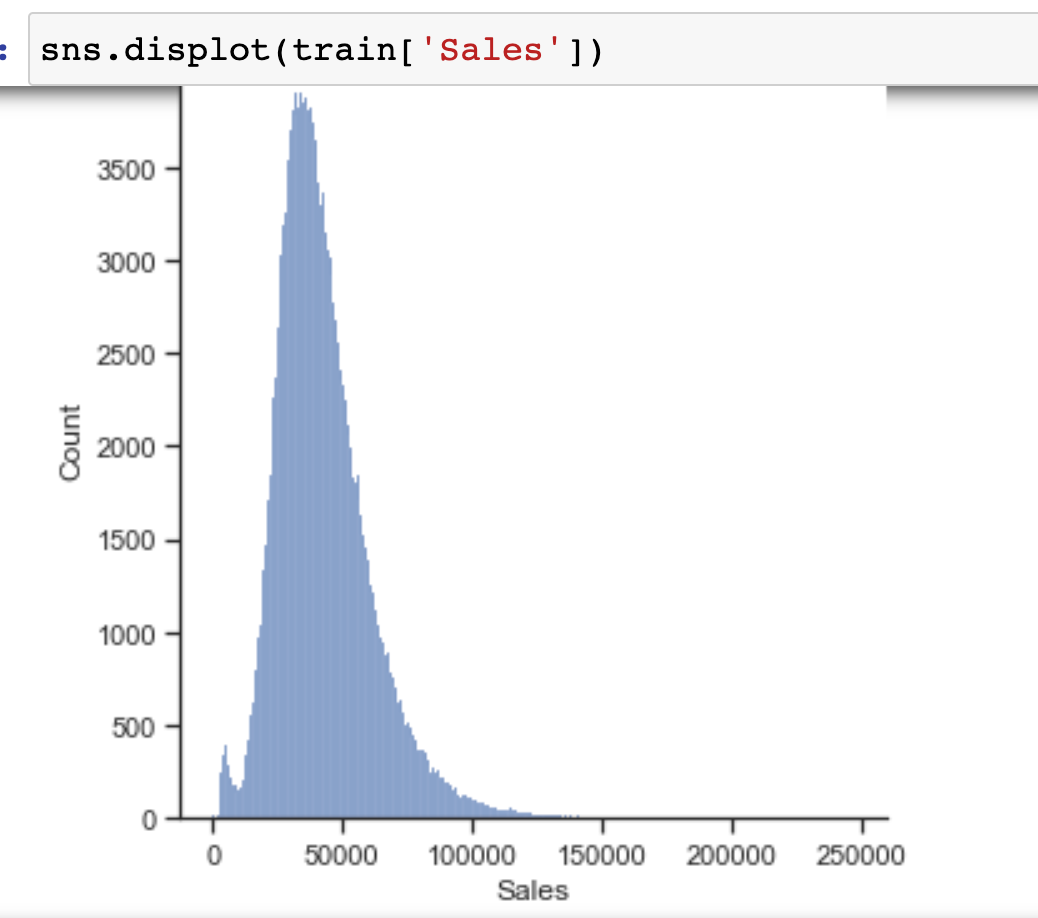
**Cleaning the data**

The training data has 10 columns while test data has 8 columns as expected because we need to predict the sales but the challenge came here is how to find out the ‘#order’ column which is missing in test data.

After observing the data, I have found that order is highly correlated to the sales so it is impossible to the drop the column instead I have trained the model and find out the order for test data.

**Extracting Information and applying feature engineering**

* After finding out the insight from visualisation w.r.t Sales I have observed that Store\_Type', 'Location\_Type', 'Discount' is in ordinal form so applied one hot encoding and ‘region’ is nominal form so applied label encoder (**Note**: This may change upon the circumstances).
* Extracted month and weekend from the date column found out that month sales are more or less same so transformed using label encoding and sale on weekend was high so transformed using one hot encoding.
* Sales columns has 19 missing rows as it is miniscule and doesn’t giving much information, I have dropped the rows although we can replace with a mean but in that case it will overfit the model or unbalanced the columns.
* After visualisation I have found that ‘Sales’ column is slightly rightly skewed



So to deal with the outliers simply discarded the rows with IQR (inter quartile range) around 5k rows has been deleted from the original dataset.

**Model Selection and Performance evaluation**

* Divided the dataset using scikit train test split with 75% for training and 25% for testing but this technique doesn’t give me better accuracy on new test data.

* Now I have trained the model particularly supervised M.L linear regression, lasso regression, Random Forest Regressor, catBoost regressor using cross validation Kfold (because without cv model doesn’t giving better accuracy)
* catBoost outclass every algorithm and give mean accuracy of around 0.97 which is good for any test dataset

**Thanks,**

**Md Tausif**

**Email id =** [**mdtau786@gmail.com**](mailto:mdtau786@gmail.com)

**GitHub :** [**https://github.com/codingf3ver/WoMart-Sales-Prediction**](https://github.com/codingf3ver/WoMart-Sales-Prediction)