Project Report
CS 412 - Introduction to Machine Learning

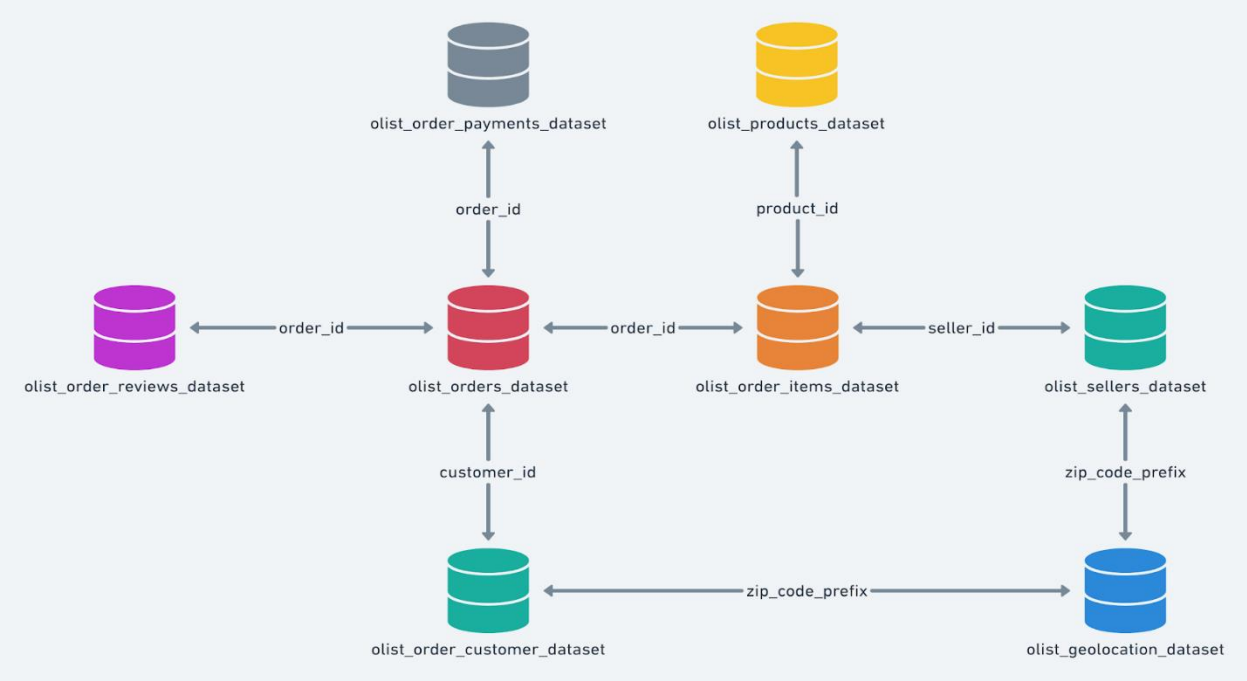
Machine Learning Tasks on Brazilian E-Commerce Dataset

Summary:

Dataset:

The dataset that we used for our machine learning tasks is Brazilian E-commerce public dataset by Olist[1]. Olist is an online store in Brazil using which merchants are able to sell their products and ship them directly to the customers using Olist's logistic partners.

The dataset has transactional details of 990,000 orders placed through the website. The dataset has details about customers, sellers, products, order delivery, order payments and reviews of the products by the customers in Portuguese. There are 9 csv files in total, each representing a specific table. The data schema is shown as below:



Regression:

Task : Predict sales price for a product category in a particular state in a given month.

Pre-processing :

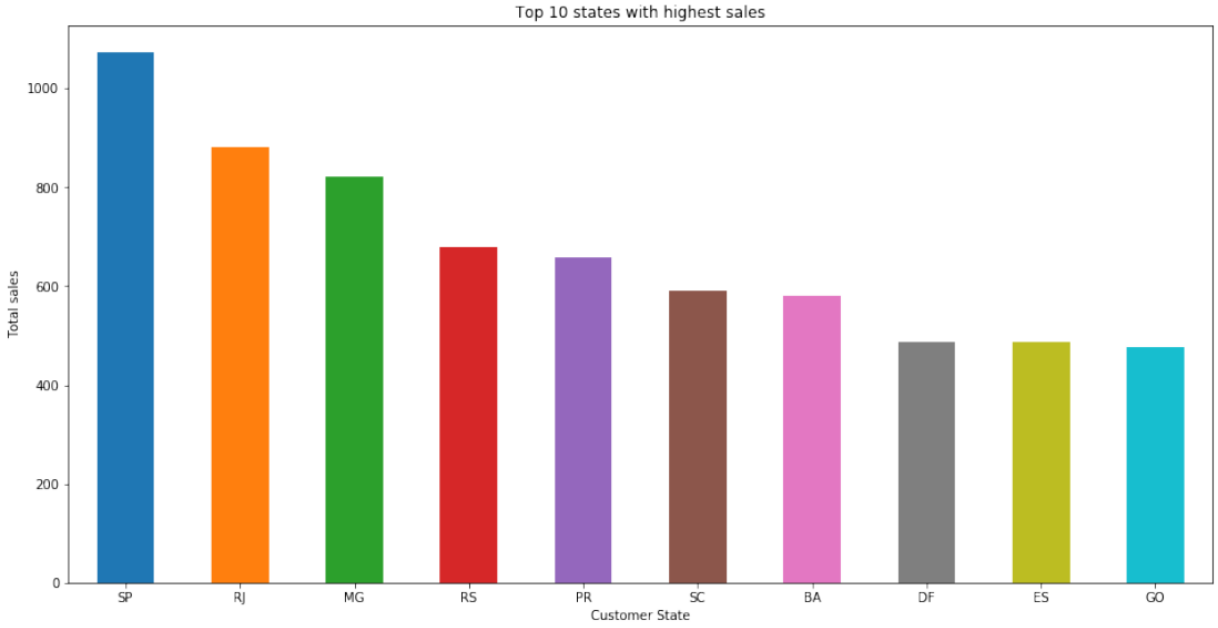
To perform the regression task, we needed overall sales price of previous transaction records based on product category for each state filtered by the month. To get the data in the required format, we had to combine the records from multiple tables. Below is the list of tables from which data was extracted for specific attributes as part of the pre-processing step:

- Orders - order id, customer id and order purchase time
- Products - product id and product category
- Customer - customer id and customer state
- Order Items - order id, product id and price
- Product Category Translation - product category in english

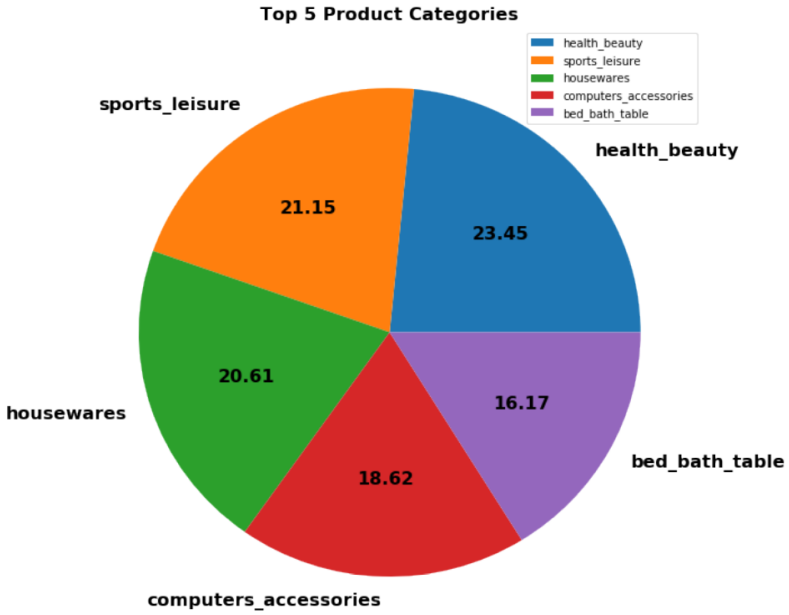
Groupby is done using year, month, state, product category on price and label encoding is done on the columns state and product category.

Data Visualization :

To find out states with the highest sales -



To find out top product categories :



Models used for the machine learning task:

- Decision Tree
- AdaBoost with Decision Tree
- Random Forest

Approach :

- 30-Fold cross-validation was performed to tune the hyperparameters of the different classifiers.
- The preprocessed dataset was divided into 80% training set and 20% test set.
- The hyperparameters of the classifiers were tuned based on their performance on the validation set.
- The classifiers were again trained using the entire 80 % training set with the best hyperparameters found.

Evaluation metrics and hyperparameters used (Hyperparameter tuning) :

Results of the train data for Decision Tree Regressor

Maximum depth	Mean Absolute Error	Root Mean Squared Error	R2 Score
5	1092.1505880055329	2387.452384331795	0.2471057019604561
10	750.9615215363106	1585.1622207098558	0.6375643616474612
15	632.8068940119049	1410.1097557562869	0.712912764000986
20	631.5873044208989	1415.1079686093794	0.7098189790235698

Results of the train data for AdaBoost Regressor with Decision Tree

Learning rate	Mean Absolute Error	Root Mean Squared Error	R2 Score
0.001	551.1588153027816	1257.2213610796541	0.770629513182874
0.01	551.2618506065185	1246.0219846700834	0.7733823891465103
0.1	555.2040367156006	1282.6121940563507	0.7415096411057142
1	588.2326780061336	1304.1987414921098	0.7440626737476563

Results of the train data for Random Forest Regressor

No. of estimators	Mean Absolute Error	Root Mean Squared Error	R2 Score
25	583.5497669844525	1301.3779597660023	0.7572830780351757
50	578.9325566017578	1292.5144573768569	0.7598764699725129
75	575.4414755753714	1288.1483537969848	0.7620529594995923
100	576.047737988877	1287.4969971975797	0.7628303625361482

Resulting performances of the 3 classifiers :

Results of the test data

Model	Mean Absolute Error	Root Mean Squared Error	R2 Score
AdaBoost Regressor with Decision Tree	579.5692997088307	1367.984558910878	0.810747067547194
Decision Tree Regressor	668.7982882046557	1533.9859702253607	0.762029566301638
Random Forest Regressor	598.6570582067619	1339.5212003084177	0.818540625342986

Challenges faced :

- Data was skewed. Very high sales in some states and very low in the others.
- The variance in sales was very high ranging from $4 to $40000 leading to higher values while evaluating the mean squared error metric.
- Hence, metrics mean absolute error and R2 (for evaluating variance) were also used for evaluation.