Student: Akshay Kumar Student ID: 202095271

Monday, June 13, 2022

Stores Sales Analysis and Prediction

CMSC-6950-001 (Comp Based Tools&Applications

The project is about performing statistical analysis on stores sales data and make prediction on future sales based on some information. The dataset was published on Kaggle.

Dataset:

	Store ID	Store_Area	Items_Available	Daily_Customer_Count	Store_Sales
0	1	1659	1961	530	66490
1	2	1461	1752	210	39820
2	3	1340	1609	720	54010
3	4	1451	1748	620	53730
4	5	1770	2111	450	46620
...	...	...	...	...	...
891	892	1582	1910	1080	66390
892	893	1387	1663	850	82080
893	894	1200	1436	1060	76440
894	895	1299	1560	770	96610
895	896	1174	1429	1110	54340

The dataset contains 896 rows and 5 columns.
Store ID: Contains the unique IDs for each store. The data type is int

Store_Area: It has the size of each store mentioned in sq feet. The data type is int

Items_Available: Contains the count of items each store possess. The data type is int

Daily_Customer_Count: Shows the number of customers each store gets every day. The data type is int

Store_Sales: Contains the sale of the store. The data type is int.

Tools:
The project has been implemented using Jupyter Notebook

Programming Language:
Python v3 has been used as the preferred programming language

Libraries used:

1. **Pandas:** Used to read dataset which is in csv format and convert it into dataframe
2. **Numpy:** It was used to convert dataframe generated by pandas into array to perform statistical operations. There were some functions like mean, median, standard deviation and percentile which were used.
3. **Scipy:** This library was used to find corelation among two columns such as Store_Area and any other column.
4. **Matplotlib:** It is one of main library which was used to create visualizations to show statistical information.
5. **Seaborn:** Like Matplotlib, a Histogram was made using this library
6. **Scikit Learn:** The main task of using this library was to predict future sales based on some information.

Machine Learning:
The project uses Linear Regression and K-Nearest Regressor to predict future store sales. The model is trained using 25% of the data. After, training is completed, we perform the testing of the model using the remaining 75% of the

data. Both these models had the accuracy of close to 90%. After the models are tested, there comes a time for actual prediction. To predict store sales, we pass two parameters in form of 2D array. The parameters are Store_Area and Items_Available. The format looks like [[1202,8523]] where 1202 represents the size of store and 8523 is the number of items available in the store. The model predicts the store sale based on these arguments.

Object Oriented Programming:

All the code has been written keeping OOP concepts in mind. Classes, Objects and Functions has been used to increase readability.

Implementation:

The code which performs all the computation and create visualizations is written in Project.ipynb file. There is another file called Main.ipynb which imports the former file and run it. In Main.ipynb we create objects and using those objects call functions to perform computation and make graphs.

Version Controlling:

Github has been used as the platform and tool for version controlling. The repository has been uploaded on Github under main branch.