CAPSTONE PROJECT ON SUPERVISED MACHINE LEARNING

Website Behavior Analysis

KARAN K. KARLE

PCAI, FEB 2021 BATCH

Contents

Introduction

- Machine learning is a subfield of Data Science,
- Machine learning is an approach to data analysis that involves building and adapting models, which allow programs to "learn" through experience
- Machine Learning has two types i.e. Supervised and Unsupervised Learning,
- Here we have to build Supervised regression model for predict the probability of a user buying a product
- Decision Tree , Random Forest , K-Nearest Neighbor , Linear Regression , Support Vector machine ,etc. Algorithm are used for creating model
- Final model selection is based on accuracy score like R2 score of all models

Objectives

- To build a predictive regression model which predict the probability of a user buying a product, based on the characteristics of user observed from the website browsing history data.
  - It will be benefiting for making better merchandising decisions, understand and plan marketing efforts.
  - Understanding customers relations with company
  - Avoid stockouts and Identify opportunities to boost revenue

Python Packages

- Pandas, Numpy, Seaborn, Matplotlib
- StandardScaler, mean_squared_error, r2_score
- Algorithms
  - Linear Regression
  - Decision Tree
  - Random Forest
  - KNN
  - SVM
- GridSearchCV and all other needed dependencies

The Solution Approach

The Data

DATA 1

Browsing Data

- Timestamp
- UserID
- Website_Section_Visited

	Timestamp	UserID	Website_section_visited
0	2017-07-26 00:03:18.448	0	product
1	2017-07-26 00:36:59.028	0	default
2	2017-07-26 00:41:17.273	0	product-listing-category
3	2017-07-26 00:45:39.197	0	content
4	2017-07-26 00:45:48.487	0	home
...	...	...	...
5535918	2017-07-26 23:18:53.789	9221827579306644828	iroa
5535919	2017-07-26 23:19:03.394	9221827579306644828	iroa
5535920	2017-07-26 23:19:11.569	9221827579306644828	product
5535921	2017-07-26 23:21:56.085	9221827579306644828	product
5535922	2017-07-26 23:16:32.835	9223103337073924884	product

5535923 rows × 3 columns

5535923 Browsing entries are there

Data

DATA 2

Final Conversion Data

- Timestamp
- UserID
- Product_purchased
- OverAllCartValue

```
In [26]: purchase_data.columns=["Timestamp","UserID","Products_purchased","OverAllCartValue"]
         purchase_data.head()
```

Out[26]:

	Timestamp	UserID	Products_purchased	OverAllCartValue
0	2017-07-26 00:00:12.301	0	H209597	31.50
1	2017-07-26 00:00:12.388	0	H211370	30.48
2	2017-07-26 00:00:14.389	0	A282331	51.00
3	2017-07-26 00:00:16.837	0	H211410	16.74
4	2017-07-26 00:00:19.625	0	H211801	34.35

```
In [27]: purchase_data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 79794 entries, 0 to 79793
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Timestamp           79794 non-null  object
 1   UserID              79794 non-null  int64
 2   Products_purchased  79794 non-null  object
 3   OverAllCartValue    79794 non-null  float64
dtypes: float64(1), int64(1), object(2)
memory usage: 2.4+ MB
```

```
In [28]: purchase_data.shape
Out[28]: (79794, 4)
```

79794 Final Conversion entries are there

Data

Final Data

- By merging Browsing and final conversion Data
- Data Cleaning:
  - Duplicate values are removed
  - Missing values are imputed

```
In [42]: final_data
```

Out[42]:

	Timestamp_x	UserID_x	Products_purchased	OverAllCartValue	isGuestUser	Timestamp_y	Website_section_visited
0	2017-07-26 00:00:12.301	0	H209597	31.50	True	2017-07-26 00:03:18.448	product
1	2017-07-26 00:00:12.388	0	H211370	30.48	True	2017-07-26 00:36:59.028	default
2	2017-07-26 00:00:14.389	0	A282331	51.00	True	2017-07-26 00:41:17.273	product-listing-category
3	2017-07-26 00:00:16.837	0	H211410	16.74	True	2017-07-26 00:45:39.197	content
4	2017-07-26 00:00:19.625	0	H211801	34.35	True	2017-07-26 00:45:48.487	home
...	...	...	...	...	...	...	...
79789	2017-07-26 23:09:08.202	9174973170462435039	K45766	89.96	False	2017-07-26 00:39:14.899	home
79790	2017-07-26 23:44:19.505	9179943977593655876	V34738	24.66	False	2017-07-26 00:39:14.909	content
79791	2017-07-26 23:53:15.661	9179943977593655876	H210000	21.64	False	2017-07-26 00:39:14.944	home
79792	2017-07-26 23:13:02.55	9211905364441411643	A209343	73.00	False	2017-07-26 00:39:15.161	product
79793	2017-07-26 23:21:05.221	9221827579306644828	V34417	33.50	False	2017-07-26 00:39:15.223	product

79794 rows × 7 columns

79794 Final entries are there

Exploratory Data Analysis

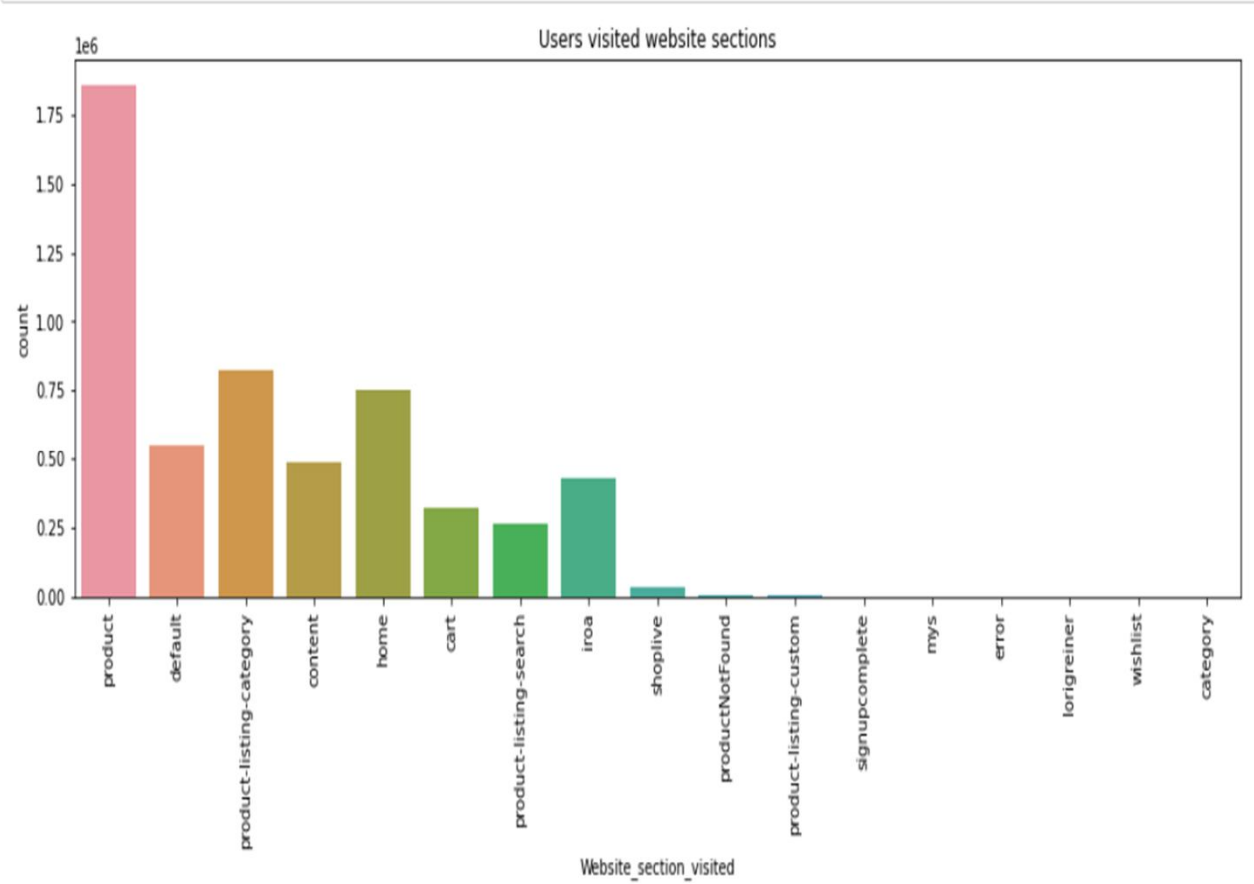
- From Browsing Data most visited Section of Website is “ PRODUCT ”
- UserID ‘0’ is denoted for Guest Users

Out[33]:

	Timestamp	UserID	Website_section_visited	isGuestUser
0	2017-07-26 00:03:18.448	0	product	True
1	2017-07-26 00:36:59.028	0	default	True
2	2017-07-26 00:41:17.273	0	product-listing-category	True
3	2017-07-26 00:45:39.197	0	content	True
4	2017-07-26 00:45:48.487	0	home	True

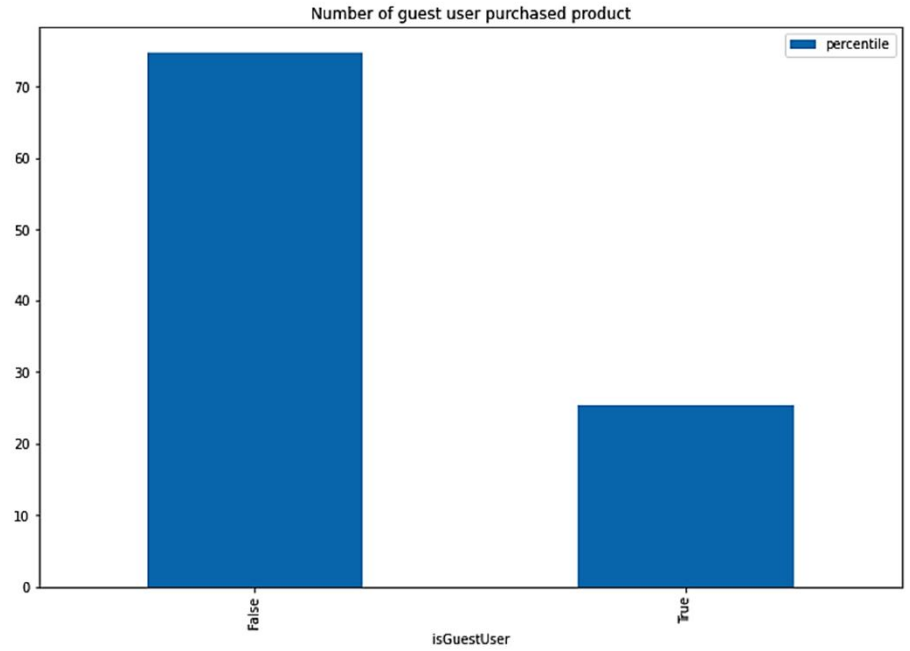
In [34]:
```
brows_data['isGuestUser'].value_counts()
```

Out[34]:
```
False    4128045
True     1407878
Name: isGuestUser, dtype: int64
```



Guest User Analysis

	isGuestUser	userCount	percentile
0	False	59598	74.689826
1	True	20196	25.310174



Final Conversion of Guest users

❖ **Here**

- ‘False’ denoted for registered user
- ‘True’ denoted for Guest user

❖ **Out of all Final Conversion(Product Purchasing) Customers 25% of are Guest Users**

Feature Engineering

Features added :-

I. Total_minutes_spent

II. Number Of Time Visited

III. Total Cart Value

IV. Total Products Purchased

V. Buy Probability

VI. User Rank (Score)

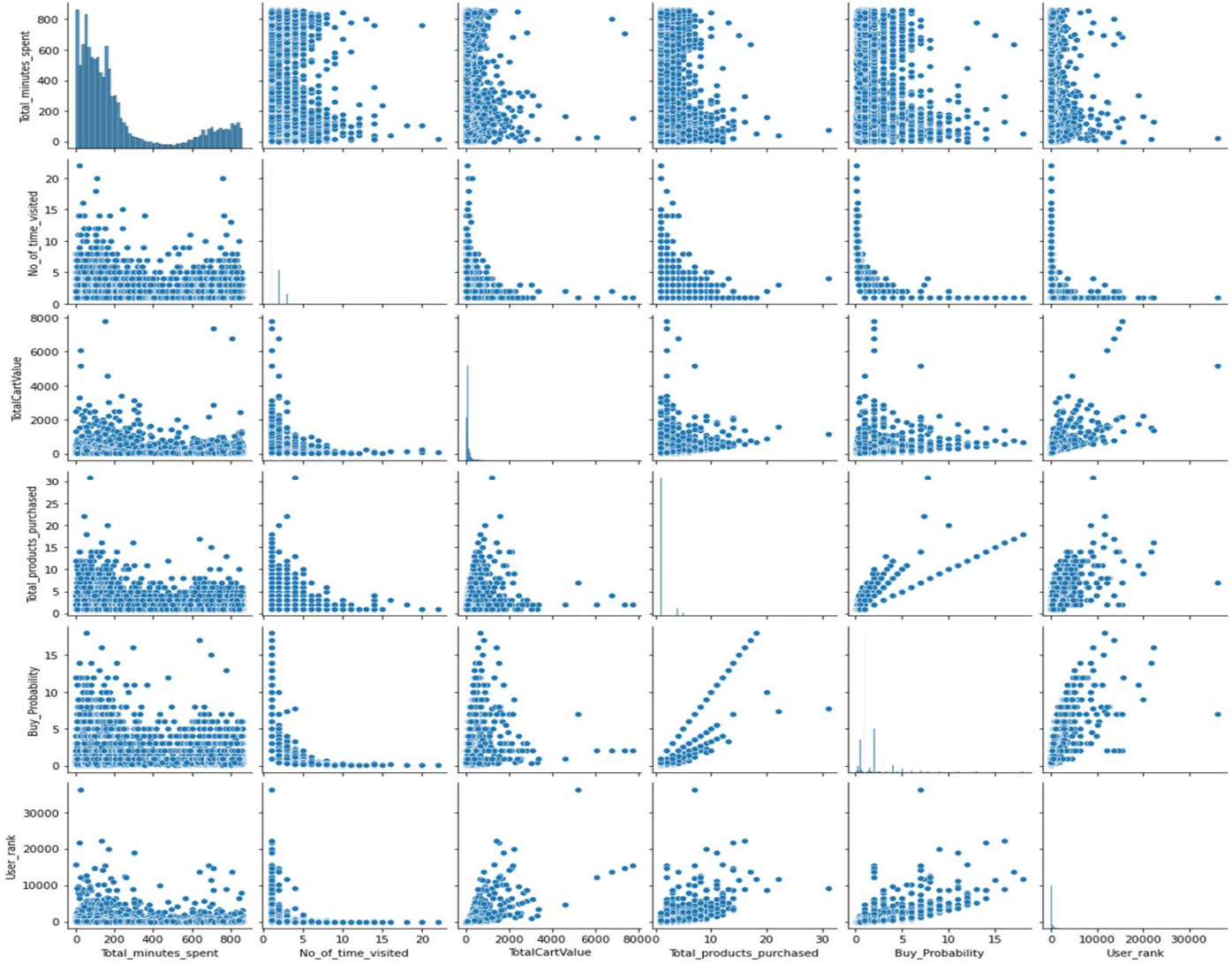
	Total_minutes_spent	No_of_time_visited	TotalCartValue	Total_products_purchased	Buy_Probability	User_rank
547	0.449517	1	157.95	1	1.000000	157.95
548	26.025000	3	60.96	2	0.666667	40.64
550	54.915333	1	216.12	5	5.000000	1080.60
551	10.063067	1	53.24	1	1.000000	53.24
552	14.423817	1	52.48	2	2.000000	104.96
...	...	...	...	...	...	...
41003	856.985867	1	135.82	2	2.000000	271.64
41004	843.631617	1	89.95	1	1.000000	89.95
41005	816.368967	1	79.14	2	2.000000	158.28
41006	835.334100	1	78.00	2	2.000000	156.00
41008	843.922600	1	368.90	1	1.000000	368.90

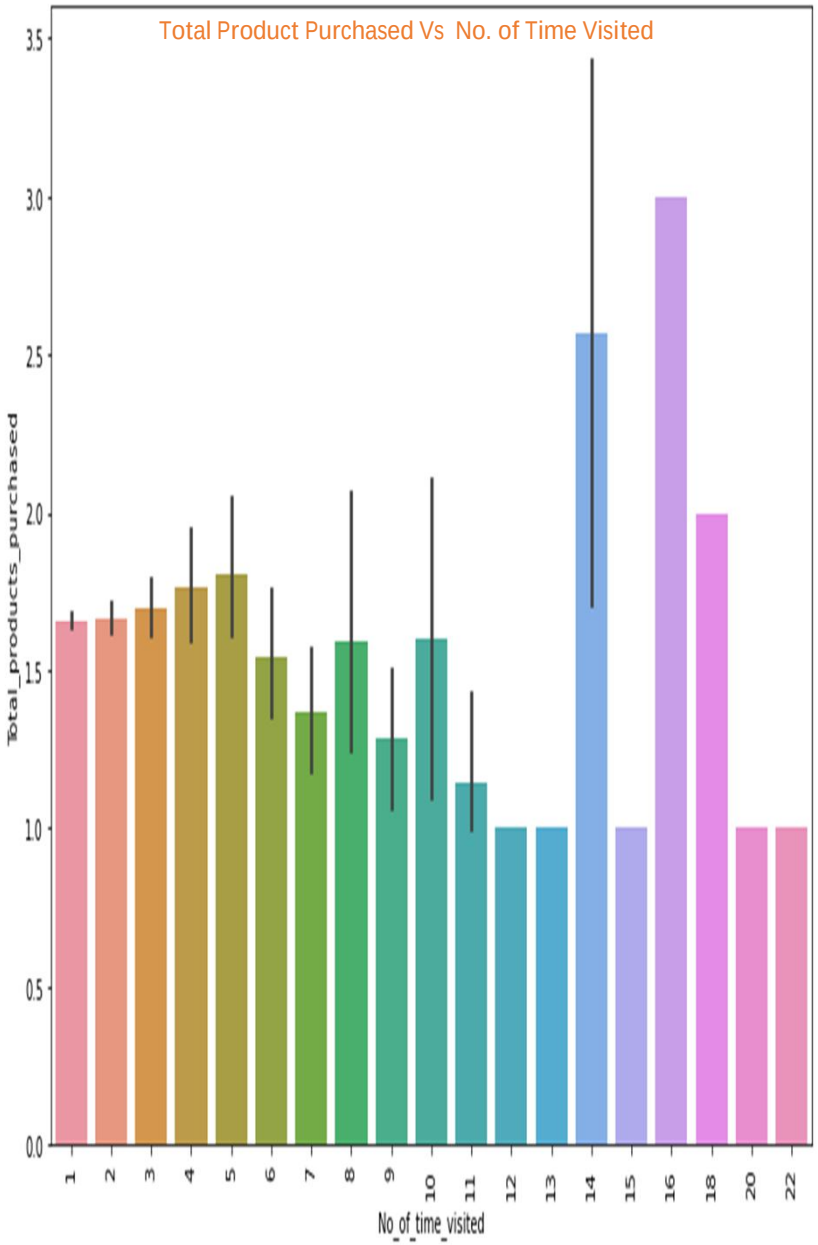
19829 rows × 6 columns

19829 Final entries are there

Data Visualization

Pair Plot :-





Model Building and Performance Evaluation

Data trained is by different Regression Algorithms

- Linear Regression
- Decision Tree
- Random Forest
- KNN
- SVM

Model	RMSE	MAE	R2_score
LinearRegession_train	4.295553e-01	1.845177e-01	0.874388
LinearRegession_test	4.163728e-01	1.733663e-01	0.875851
DecisionTree_train	3.750756e-17	1.406817e-33	1.000000
DecisionTree_test	4.122151e-02	1.699213e-03	0.998783
SVR_train	2.917852e-01	8.513860e-02	0.942041
SVR_test	2.633424e-01	6.934922e-02	0.950338

Model Building – K-Nearest Neighbor

K-Nearest Neighbor Regressor is trained with different hyper parameters selected by Grid Search

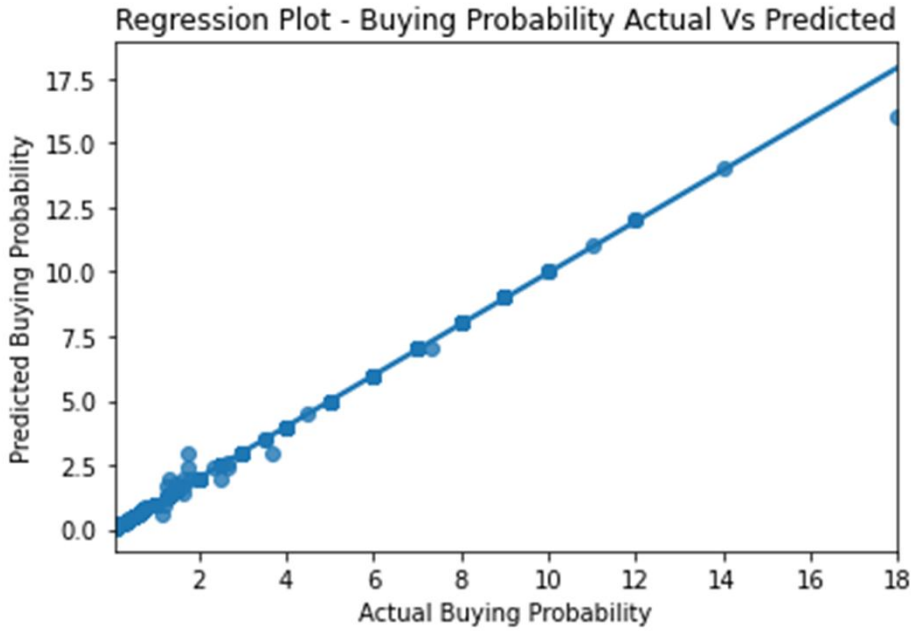
KNN_train	6.076236e-02	3.692065e-03	0.997487
KNN_test	1.096742e-01	1.202843e-02	0.991386
KNN_tuned_train	0.000000e+00	0.000000e+00	1.000000
KNN_tuned_test	7.523394e-02	5.660145e-03	0.995947

Model Building – Random Forest

RandomForest_train	1.897698e-02	3.601260e-04	0.999755
RandomForest_test	4.228119e-02	1.787699e-03	0.998720
RandomForest_tuned_train	1.414871e-02	2.001860e-04	0.999864
RandomForest_tuned_test	3.708925e-02	1.375613e-03	0.999015

Random forest Regressor is trained with different hyper parameters with 5 fold grid search.

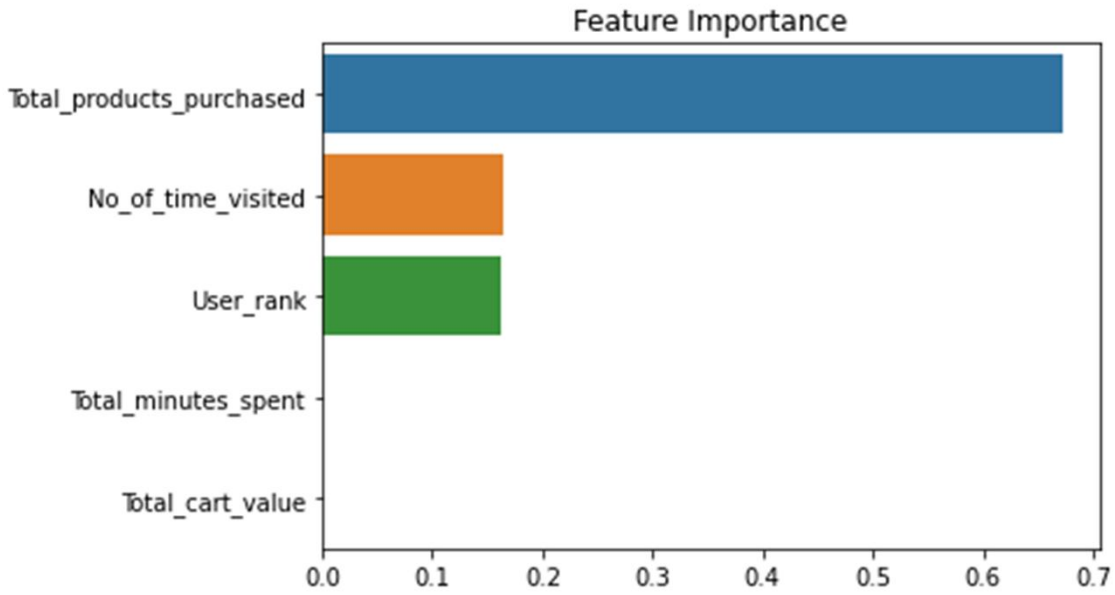
- Accuracy for both Train and Test is above 99%



Feature selection

Feature Importance :

- Total_products_purchased
- No_of_time_visited
- User_rank
- Total_minutes_spent
- Total_cart_value



Model Comparison and Conclusion

- ❖ Data is trained with all Regression Models.
  - All models have best accuracy
  - Accuracy achieved with most models
    - 99% with training data
    - 99% with test data
- ❖ Random Forest or KNN can be best for Building a Good Model.

	Model	RMSE	MAE	R2_score
10	KNN_tuned_train	0.000000e+00	0.000000e+00	1.000000
2	DecisionTree_train	3.750756e-17	1.406817e-33	1.000000
6	RandomForest_tuned_train	1.414871e-02	2.001860e-04	0.999864
4	RandomForest_train	1.897698e-02	3.601260e-04	0.999755
7	RandomForest_tuned_test	3.708925e-02	1.375613e-03	0.999015
3	DecisionTree_test	4.122151e-02	1.699213e-03	0.998783
5	RandomForest_test	4.228119e-02	1.787699e-03	0.998720
8	KNN_train	6.076236e-02	3.692065e-03	0.997487
11	KNN_tuned_test	7.523394e-02	5.660145e-03	0.995947
9	KNN_test	1.096742e-01	1.202843e-02	0.991386
13	SVR_test	2.633424e-01	6.934922e-02	0.950338
12	SVR_train	2.917852e-01	8.513860e-02	0.942041
1	LinearRegession_test	4.163728e-01	1.733663e-01	0.875851
0	LinearRegession_train	4.295553e-01	1.845177e-01	0.874388

Future Improvements

❖ Scope for Improvements

- With more Features
- With Gathering more Data
- Try with different Hyperparameters

Thank You