In-Depth Analysis

Traditional Approach

The categorical variables were encoded using one-hot encoding:
Gender as Gender_enc, Age as Age_*, City_Category as City_*, Stay_In_Current_City_Years as stay, Occupation as occu, ProdCombo as pc_*, ProdCat1 as pc1_*, ProdCat2 as pc2_*, ProdCat3 as pc3_*. The continuous variable was chosen as target and various models were tried out.

Linear Regression: There are 122 features that give the below results
R^2 with Prod Categories: 0.6376223400425602
Root Mean Squared Error with Prod Categories: 2989.4402029238413
Cross validation score is with Prod Categories: 0.6389821282926864

Decision Tree: A configuration of (max_depth=15, min_samples_leaf=100) gave the below results.
R^2: 0.635121867047957
Root Mean Squared Error: 2999.736317822526

Elastic CV: A configuration of (cv=5, alphas=np.linspace(0.001,1,50)) gave the below results
R^2: 0.6400716117954796
Root Mean Squared Error: 2993.6060388057804

Random Forest: A configuration of (max_depth=16, n_estimators=90) gave the below result
R^2: 0.6424529101374172
Root Mean Squared Error: 2983.686705716457

In general we are not getting past the 0.65 accuracy mark. Let's try some feature engineering to improve the scores.

Feature Engineering

Four functions were created that return

1. Average Purchase Value per column.
2. Average Count of column
3. Total count of column
4. Median purchase value of column

New columns were created after applying the above functions for each of the categorical variables such as Age, Occupation, Stay_In_Current_City_Years, ProdCat1, ProdCat2, ProdCat3, Gender, City_Category, ProdCombo, User_ID and Product_ID.

After Feature engineering results: The models were again tried to see if there was any improvement.

Linear Regression:
R^2: 0.7176789707438745
Root Mean Squared Error: 2638.6438593275593
Cross validation score with new features is: 0.7191447714961551

Decision Tree: Configuration of **(**max_depth=400, min_samples_leaf=112, min_samples_split=40) gave
R^2 Test: 0.7373382707106608
Root Mean Squared Error Test: 2552.277662635797

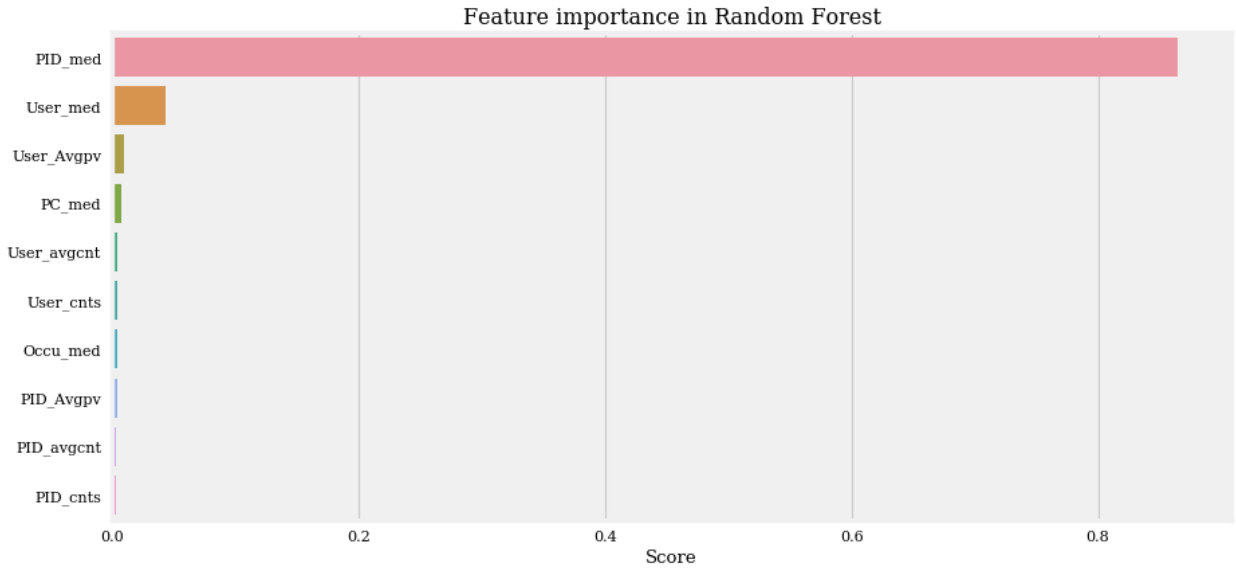
Elastic CV: A configuration of (cv=5, alphas=np.linspace(0.001,1,50)) gave the below results
R^2: 0.7197624634394806
Root Mean Squared Error: 2641.494731114654

Random Forest: A configuration of (max_depth=16, n_estimators=90) gave the below result
R^2: 0.7467830310391308
Root Mean Squared Error: 2510.7525519630262

Conclusion

It can be concluded that feature engineering helped boost the overall accuracy score past 70% to 74.6. The model of choice would be Random Forest.

Here is the feature importance chart.



As you can see, the engineered feature of median price of Product_ID is the most important one.