**Myntra Fashion Report**

**Problem Statement:**

But as is the nature of the business, the fashion industry moves fast and no one trend remains

for long, this renders your best trained forecasting models useless after some time. To resolve

Thus, the team needs to constantly update and retrain their models all the time. The data you are

working with has the following columns: -

1. URL - URL of the Product

2. Product\_id - ID of the Product

3. BrandName - Product Brand Name

4. Category - Category of product

5. Individual\_category - Sub category of Each product

6. category\_by\_Gender - Category based on gender

7. Description - Description about the project

8. DiscountPrice (in Rs) - Price after Discount

9. OriginalPrice (in Rs) - Original Price of the Product

10. DiscountOffer - Offer on the product

11. SizeOption - Different Size Options

12. Ratings - Customer Ratings

13. Reviews - Total Customer Reviews

You have to create a model that can predict user ratings.

**Data Description:**

This dataset contains **526564** clothes records and has **13** columns. The columns are website address, the id of the product, the name of the brand, small category, large category, for which gender, description, the price after discounting, the original price, the discount rate, all sizes that can be chosen for each product, ratings, how many reviews for each product.

Most discount prices are less than **1469** and most original prices are less than **2899**. Most ratings are higher than **3.9**, while most products have less than **52** reviews. So, many products are relatively cheap with less discounts. The ratings are relatively high but not many products are so attractive that customers want to give it a comment.

There are some empty values in discount price, discount rate, ratings and reviews. There are no duplicate values present in the dataset.

**Data Preprocessing:**

1. There are some empty values in discount price, discount rate, the empty in discount price and discount rate probably means that there isn't any discount, so it can be represented by 0, but the empty in discount rate are more than that in discount price, which means they can be calculated. So, the record with empty discount rate are filled with 0 and the record with empty discount price but with clear discount rate are calculated by the original price and the discount rate.
2. Renaming some of the columns - "Individual\_category" : "IndividualCategory",

"Product\_id" : "ProductId",

"category\_by\_Gender" : "Gender",

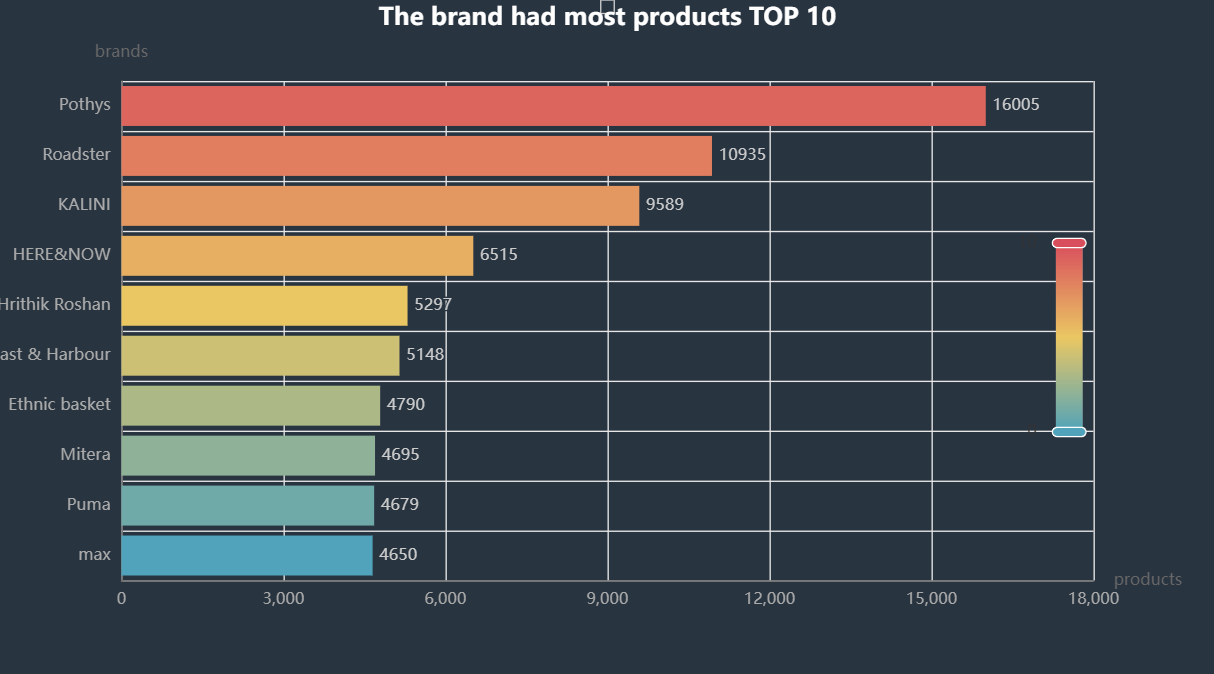
"DiscountPrice (in Rs)" : "DiscountPrice",

"OriginalPrice (in Rs)" : "OriginalPrice"

1. For DiscountOffer. First fill empty values with "0". Since there are "...% OFF", "... OFF" AND "0" in this column, I change "... OFF" to "... % OFF", change "... % OFF" to float and change others to 0. Then calculate the price after discount by OriginalPrice\*(1-DiscountOffer)
2. For Reviews. Fill empty values with “0”
3. Drop the URL (it can be gotten by merge the final dataset with df dataset or use df2 dataset directly) and the description.
4. For Ratings. Fill empty values using KNN Imputer with n\_neighbors = 5. The use of a KNN model to predict or fill missing values is referred to as “*Nearest Neighbor Imputation*” or “*KNN imputation*.” KNN Imputer is used as it utilize the similarity between users or items to predict missing ratings. If a user has not rated an item but has similar preferences to other users who have rated that item, KNN can predict the missing rating based on the ratings of similar users for that item.

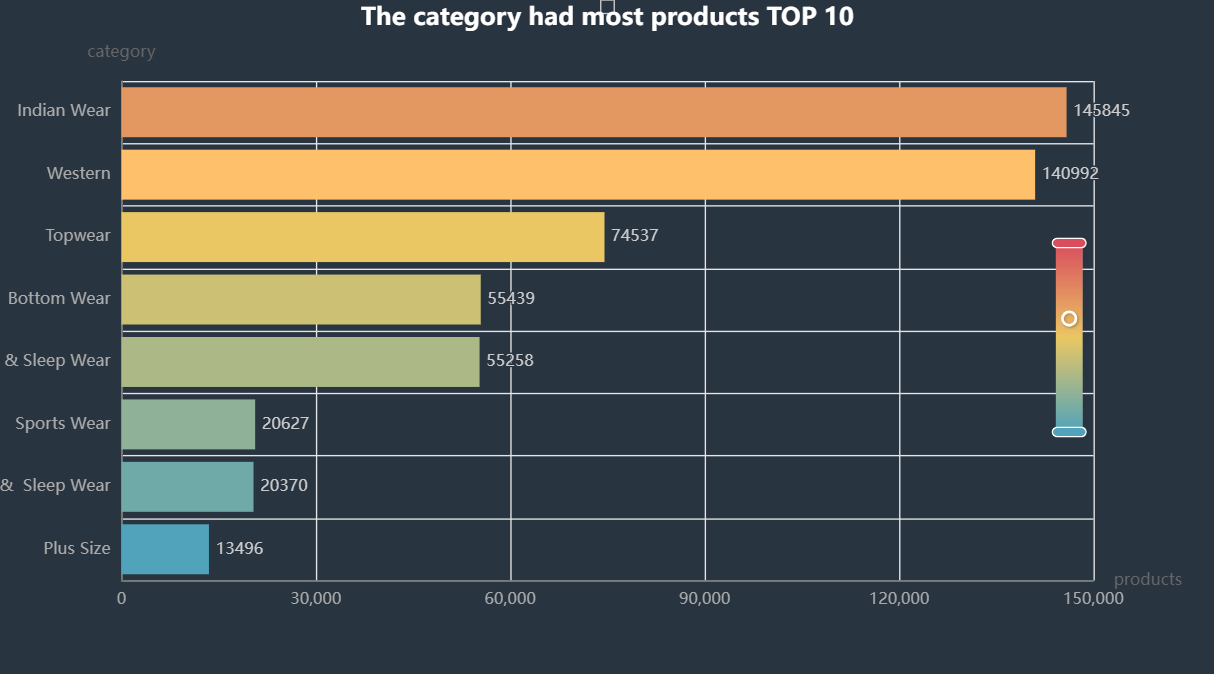
**EDA:**

1. Which brand had most products on myntra?

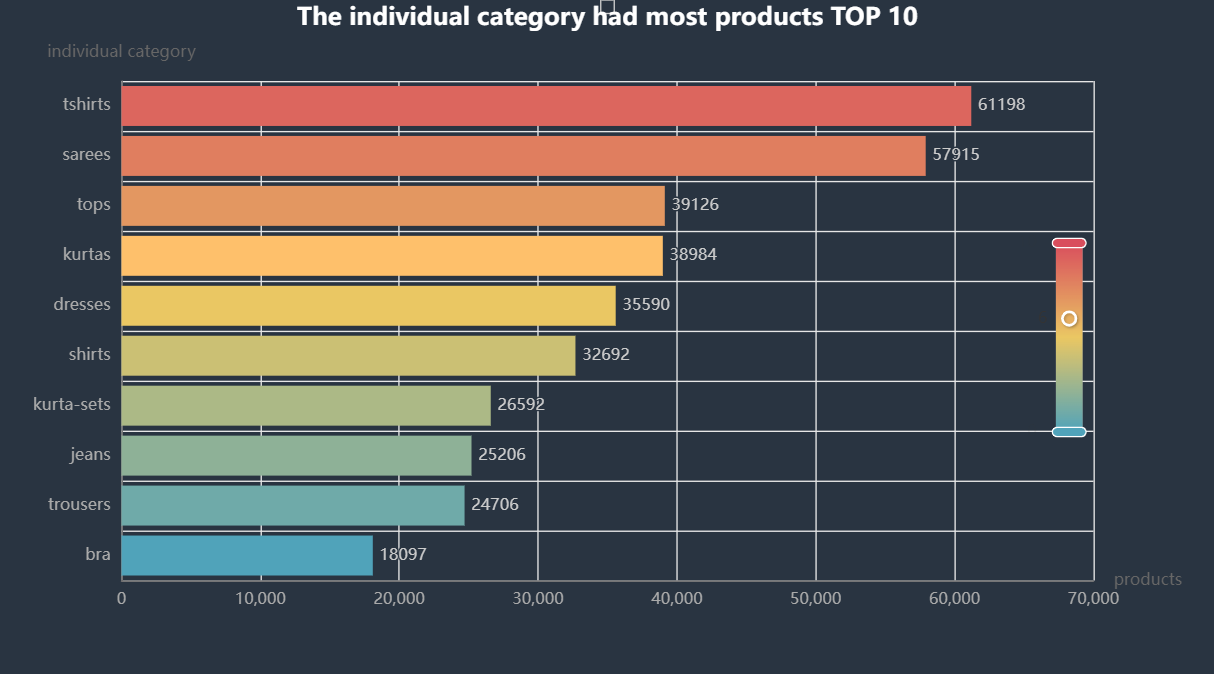


Pothys, Roadster and KALINI are the top3 brands who had most products on myntra. Pothys had more than 16K products, which was much more than other brands.

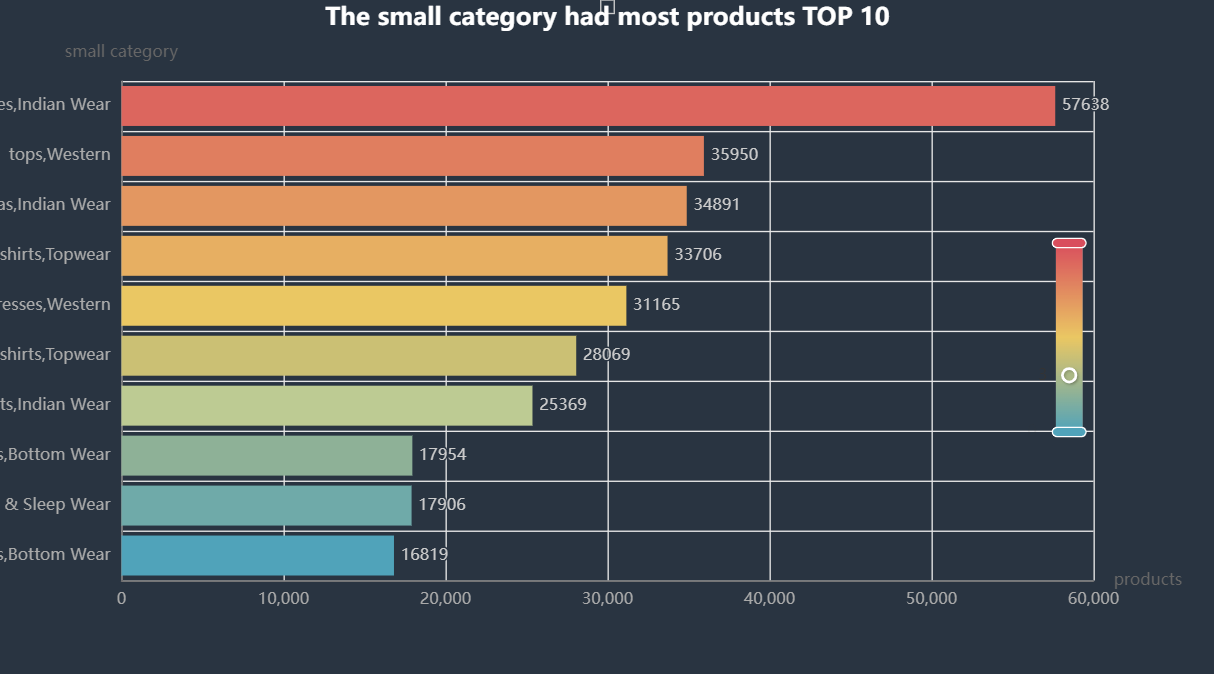
1. what kinds of clothes were the most on myntra?



There were more than 140K Indian wear and Western clothes on myntra, while only 13K plus size clothes were sold on myntra.

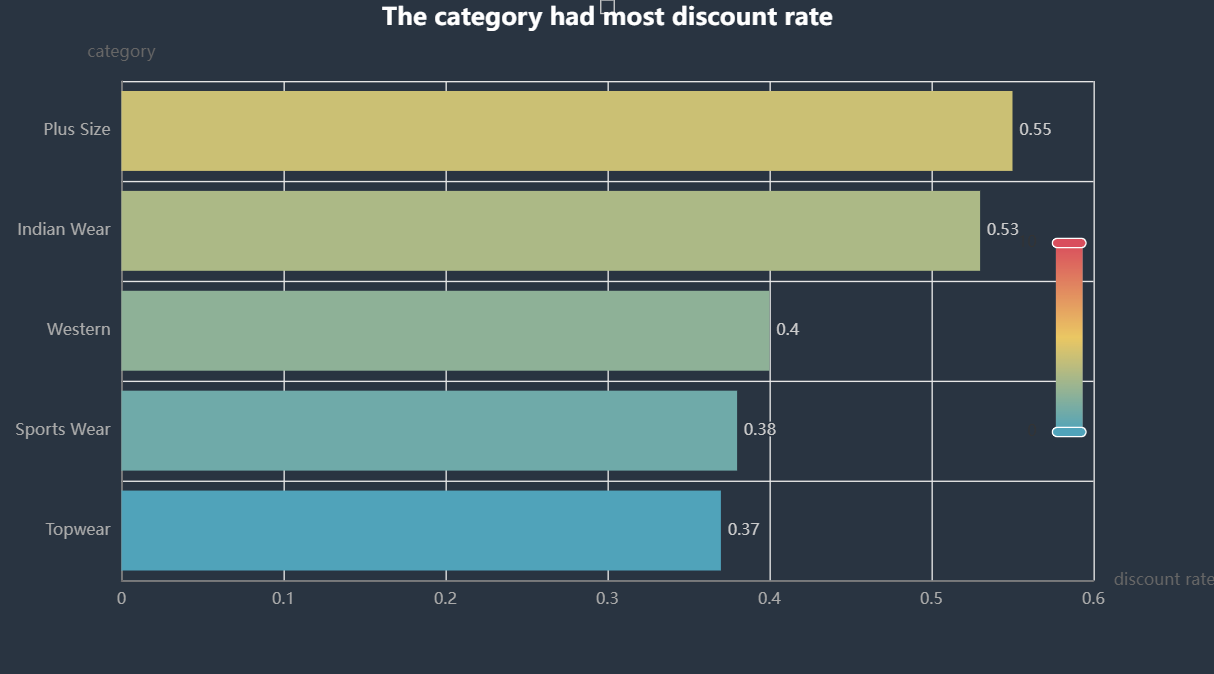


Most clothes on myntra were t-shirts or sarees, about 61K and 58K respectively, which were much higher than that of others.



The indian wear sarees had a dominant role on myntra, with about 57K products. And we can see that almost all the sarees are indian wear.

1. what kinds of clothes have the highest discount rate?

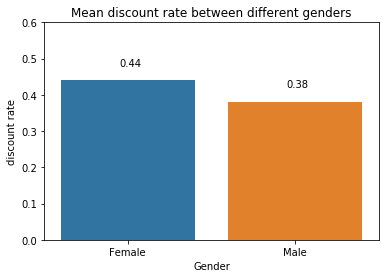


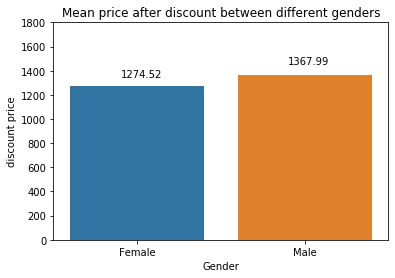
Plus size category has the highest discount rate of 0.55



As for discount rate, there is nothing special for small category or individual category, so only category and brand are shown here. Plus size clothes and Indian wear have high discount rate, which is near half of the price. This may because Indian wear is the most popular one and high discount rate can encourage the consumption, while plus size clothes are not so popular, so high discount rate can draw consumers' attention. TOP5 brands have a relatively high mean discount rate, which are all higher than 80%. Especially HRITIKA, which have a high mean discount rate at 89%.

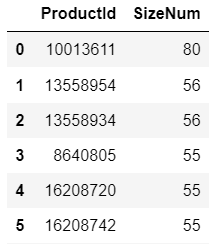
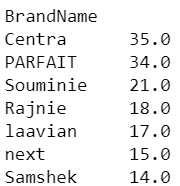
1. Whose clothes are cheaper, and whose clothes have large discount rate, Men or Women?





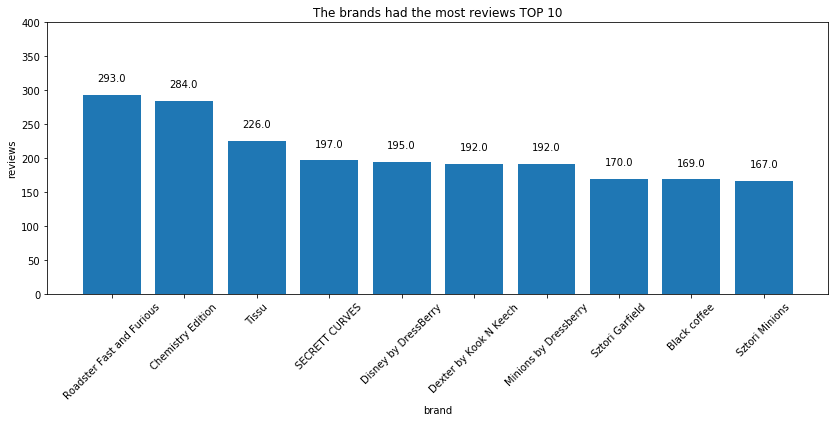
From the bar plots above, we can see that average discount rate on female clothes was higher than that of male clothes and the average price after discount of female clothes was lower than that of male clothes.

1. Which brand offered more size choice? (average)

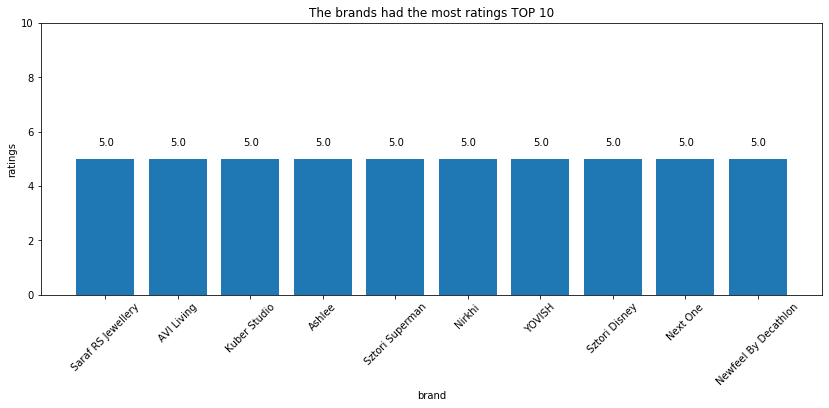
Many items have lots of size options available, some items offer more than 50 sizes. The product whose productid is “10013611” has most available size, about 80 size options. The information of it is shown above. As for the brand, Centra, PARFAIT and Souminie are the top 3 brand, who offers most size options, the numbers are 35,34 and 21 respectively.

1. Which brand had most reviews?



Roadster Fast and Furious had the most reviews, at 293 reviews per item, while Chemistry Edition had the second place. Tissu is the third popular brand with 226 reviews per item. They all had more than 200 reviews per item. Then came SECRETT CURVES, Disney by DressBerry, Dexter by Kook N Keech and Minions by Dressberry, all had about 190 reviews per item.

1. Which brand had highest rating?



When looking at the ratings, the full mark is 5.0 and only 7 brands get full mean ratings. However, we have to take the number of reviews into account. For the highest-rating-brand, only Minions by Dressberry had 192 reviews while others had much less reviews, while for the most-review-brands, only Minions by Dressberry had full ratings. So Minions by Dressberry was the most popular brand for fashion clothes on myntra probably.

**Feature Engineering:**

1. ProductId has few meaning, just discard it
2. SizeOption has few meaning, SizeNum is extracted from it and will be one of the features
3. DiscountPrice and OriginalPrice can used to calculate DiscountOffer directly, so DiscountPrice will be wiped out of the model
4. BrandName, Category, IndividualCategory and Gender will be turned to many 0-1 variables by get\_dummy method. However, there are too many brandname, individualcategory, so some of them will be turned to "others". brandtop118 have more than 1000 items. individualcategory top17 have more than 8000 items
5. Drop brandname, category, individual category, gender changed to IsMan
6. Standarized the "OriginalPrice","Reviews" and "SizeNum" columns using StandardScaler so that dataset can be used to train models conveniently

**Model Building:**

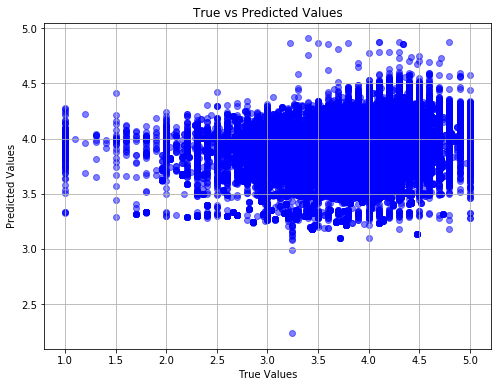
After dropping “Ratings” column all the columns from the cleaned dataset is saved as X and “Ratings” column is saved as Y. X has shape - (526564, 150) and Y has shape - (526564, 1)

Split the dataset into 3 equal parts – (X\_test, y\_test) which will be used as new dataset for retraining the model. (X\_train\_2, y\_train\_2) which is used for training and (X\_test, y\_test) which is used for testing.

**SGDRegressor -**

Trained SGDRegressor model with hyperparameters as

(*alpha=0.0001, average=False, early\_stopping=False, epsilon=0.1,eta0=0.01, fit\_intercept=True, l1\_ratio=0.15, learning\_rate='invscaling', loss='squared\_loss',max\_iter=1000,n\_iter\_no\_change=5, penalty='l2',power\_t=0.25,random\_state=42,shuffle=True,tol=0.001,validation\_fraction=0.1, verbose=0,warm\_start=False*)



RMSE = 0.4385 R2 = 0.1335

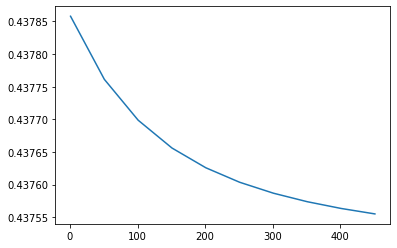
Time for Retraining the model and check for new dataset how model performs after retraining with new dataset.

After first training on top -

RMSE = 0.438

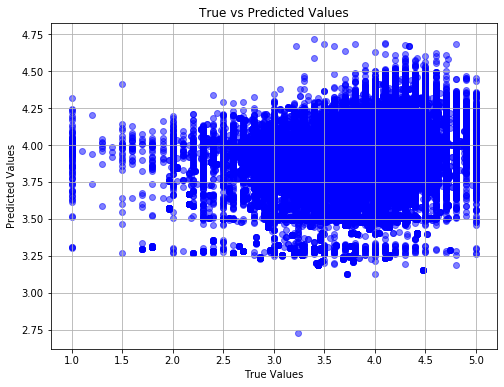
R2 = 0.1335

For multiple training on top -



From the plot we can see as the number of epochs is increasing RMSE value keeps on decreasing.

For 400 training runs on top -

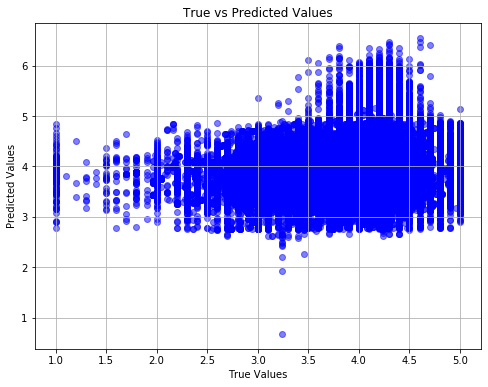


RMSE = 0.4375 R2 = 0.1373

**PassiveAggressiveRegressor :**

Trained PassiveAggressiveRegressor model with hyperparameters as

*(C=1.0, average=False, early\_stopping=False, epsilon=0.1, fit\_intercept=True,*  
 *loss='epsilon\_insensitive', max\_iter=1000, n\_iter\_no\_change=5, random\_state=21, shuffle=True, tol=0.001, validation\_fraction=0.1, verbose=0, warm\_start=False)*



RMSE = 0.5806 R2 = -0.5187

After first training on top -

RMSE = 0.5948

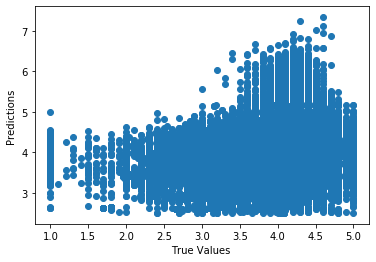
R2 = -0.5942

For multiple training on top -



From the above plot it can be seen as Retraining has no effect on the PassiveAggressiveRegressor. As we increase number of epochs for this model RMSE score is almost constant.

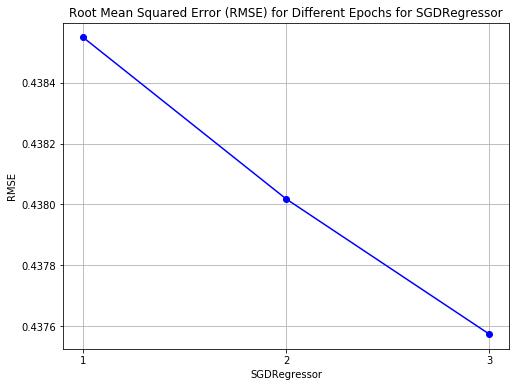
For 400 training runs on top -

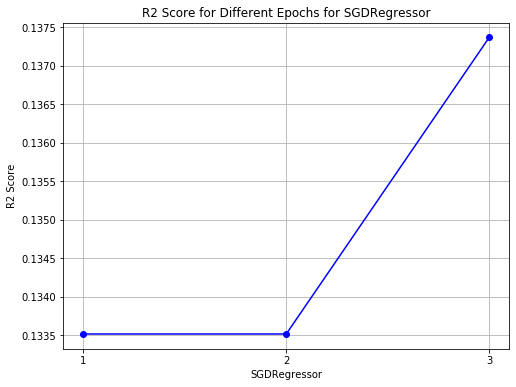


RMSE = 0.5948 R2 = -0.5942

**Model Comparison:**

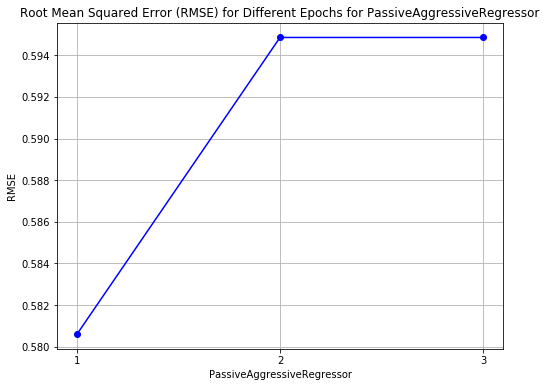
**For SGDRegressor**

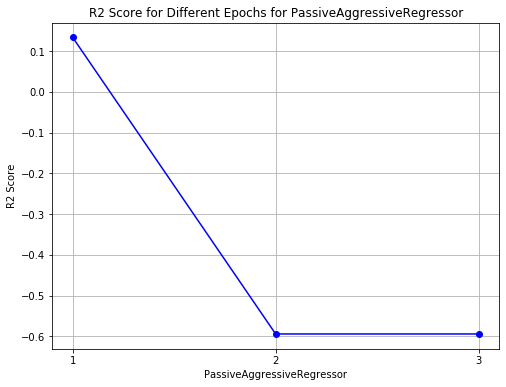




With Retraining the Root Mean Score value is decreasing and R2 score is increasing. Hence with retraining model is able to make good predictions.

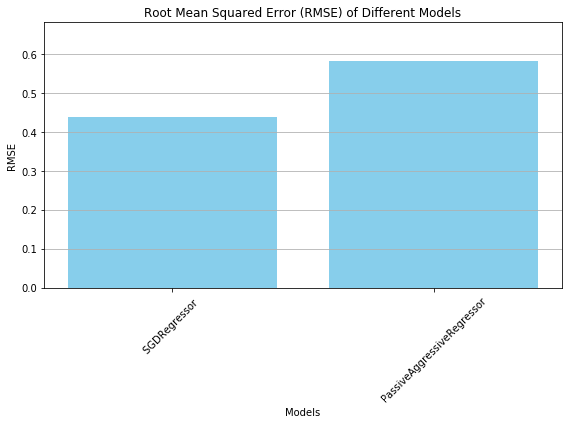
**For PassiveAggressiveRegressor**

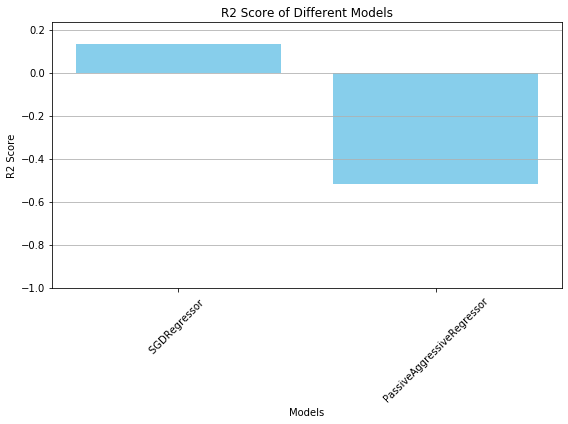




With Retraining and increasing number of epochs the RMSE score and R2 Score are constant. Hence Retraining has no effect on PassiveAggressiveRegressor.

**SGDRegressor** **vs PassiveAggressiveRegressor**





Hence SGDRegressor is good model as compare to PassiveAggressiveRegressor for our dataset as it keeps on getting better after Retraining and after increasing number of epochs whereas PassiveAggressiveRegressor gives negative r2 score value and after training on top it gives same result with no improvement.

**Input and recommendations for management:**

For regression models also I feel there should may be better variables that should be available. Since, most of the columns were not useful and hence dropped, leaving very few columns to make the predictions. There should be columns for positive and negative reviews so that positive can be used for promotional activities and negative can be taken care for understanding improvement areas.

**User perspective:**

As a user I prefer weighted positive and negative customer reviews along with how many people have given reviews. There were a lot of missing values in target variable itself and that needs more of the research to select which imputation is better to be done.