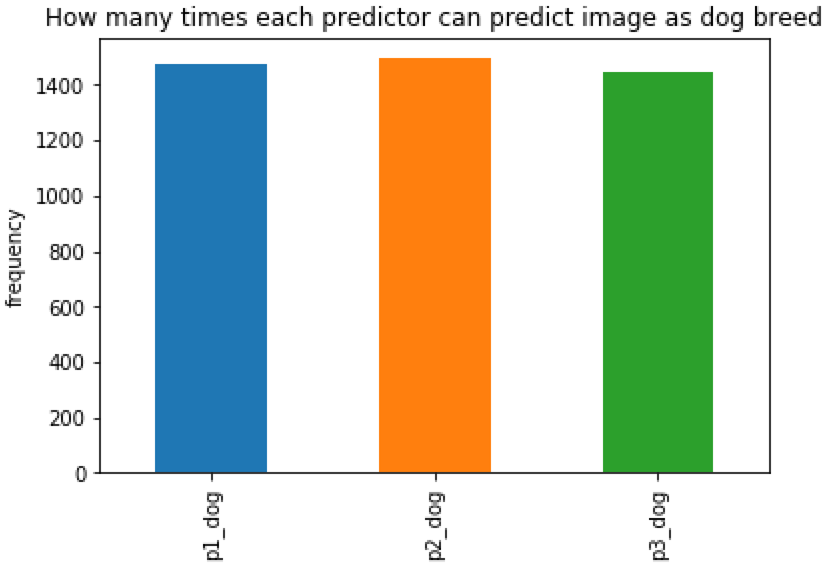
Act Report

By : Irfan Septiyana Putra

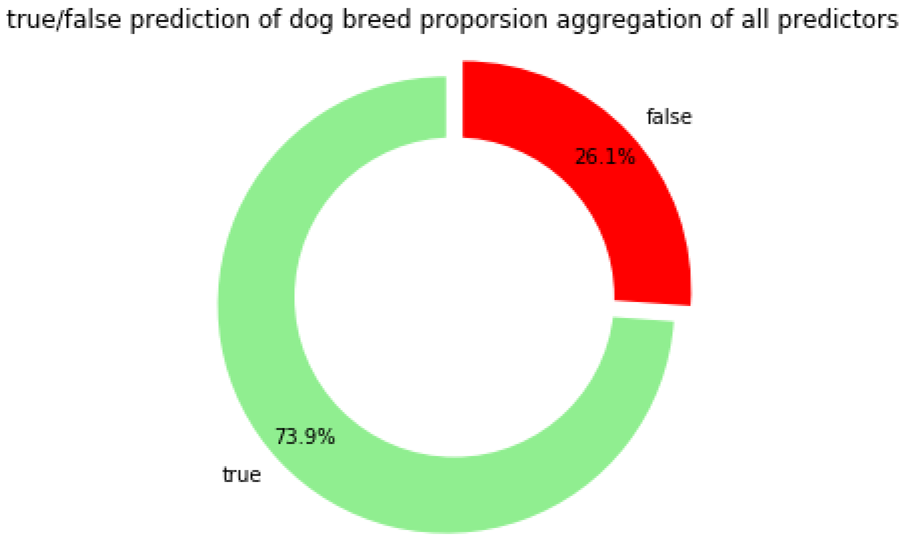
There are 3 data file gotten from data wrangle process : tweeter_archive_master, image_prediction_filtered and favorite_retweeted_filtered. From those data, we can visualize interesting information like performance of predictors, tweet rating distribution and correlation between retweet count and favorite count. Using visualize intertretation, all people can get information about those items quicker with more interesting way than seeing data manually.

Seeing data in image_prediction_filtered, there are 3 perdictors used to predict the image in tweet. Interesting question about this issue is all predictors have same performance to predict dog breed or not. We can get the information by counting true value in p1_dog, p2_dog, and p3_dog.



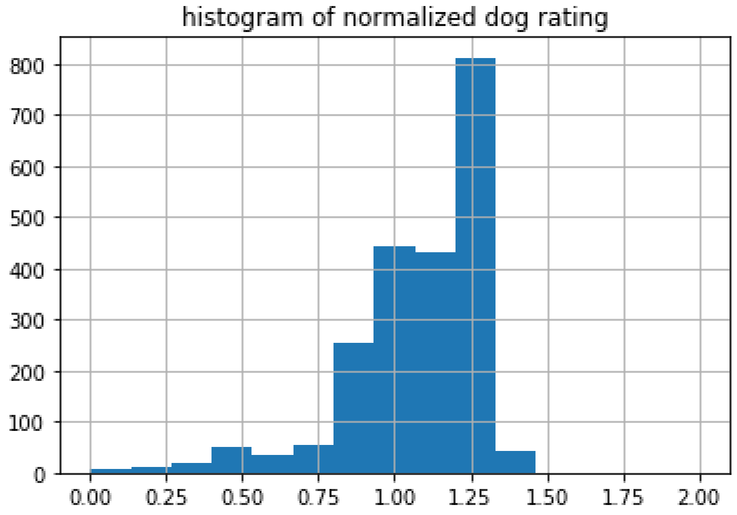
Seeing bar plot above, all three predictors has only slight performance differencies. All predictors can predict image as dog_breed in around 1400 times.This data is taken as observed data comparison.

From information before, I continue to gain information of true prediction proporsion vs false prediction proportion aggregated from all predictors. Aggregation is done to see true prediction proporsion in whole predictors which is shown in pie chart below.



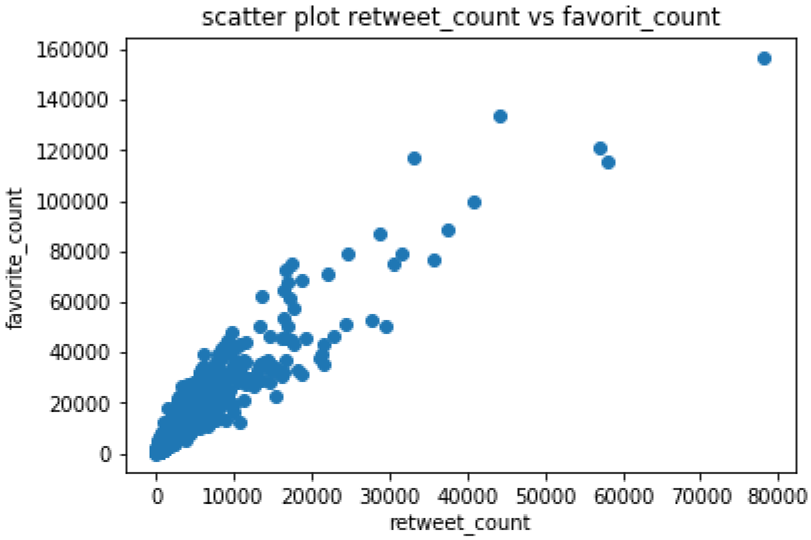
Meaned from all predictors, true prediction is around 73.9% or 0.739 in proporsion. This data can be used to evaluate dog breed predictors performance. 26.1% false prediction is high value for error percentage. Based on data, suggested to improve predictors model to gain smaller error prediction percentage.

Moving to tweeter archive master data, interesting data to explore is dog rating. Due to differences of scale raring, it will be easier to compare dog rating using normalize_rating. Distribution shown below.

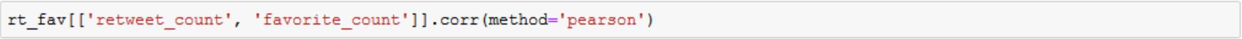


Distribution of dog rating is almost in normal distribution with mean in 1.22. From this data, tweeter user often give rate more than the scale rating. For example, if scale is 10, people give score 12. This issue is based on mean normalize rating that exceeds 1.0 .

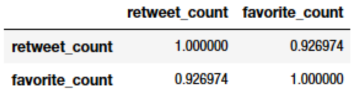
In the last data, it seems interesting to see that is retweet count and favorite count correlate one another or not. Used scatter plot on retweet and favorite count and graph is shown below.



Retweet count and favorite count like having positive correlation. That assumption is based on scatter plot, the more big retweet count the more count in favorite count. For assuring this assumption, used corr() method to calculate correlation between 2 variables.



```
rt_fav[['retweet_count', 'favorite_count']].corr(method='pearson')
```



	retweet_count	favorite_count
retweet_count	1.000000	0.926974
favorite_count	0.926974	1.000000

Calculation result of favorite and retweet count correlation is 0.926. This value is very high for correlation value. It can be concluded that favorite can and retweet count have positive correlation in 0.926.

From all data visualization, I can say that we can get insight by visualizing data and from visual data we can make our first assumption about data. This conclusion is based on activity visualizing distribution of dog rating and scatter plot favorite and retweet count. Furthermore, with visual data we can share information with all people even they are not know about data analyst process in deep, like I did in first data visualization.