Act_Report

December 24, 2018

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. I've gathered their account data from different sources and cleaned them to make them tidy and improve the quality. Here, I would like to share with you three insights I've found from the data which you might found interesting.

1. Breed Identification

One of the exciting dataframes I have is the one identifying the breed of dog from the picture(s) posted along with each tweet like the one presented below. This data was obtained by running each image through a neural network. Each identification comes with a confidence level that shows how much confident the algorithm was in identifying the actual breed.

Among 113 breeds identified by the neural network, the following breeds are the ones with highest probability to be identified correctly on average:

- Komondor (97%)
- Clumber (94%)
- Keeshond (84%)

On the other hand, the hardest breeds to be recognized by the algorithm (on average) include:

- Irish_wolfhound (6.3%)
- Bouvier_des_flandres (8.2%)
- Scottish_deerhound (14.3%)

2. Average dog ratings

If you would like to know what breed of dog has the highest and lowest ratings on average then this section is for you. The lowest belongs to Japanese_spaniel with rating of 5 while Clumber has received 27 which is the highest rating.

3. Retweet_count and favorite_count relationship

The number of likes (favorites) and retweets for each post on WeRateDogs twitter account are other available metrics in my dataframe. I utilized these two metrics to understand whether there is a linear relationship between the two. My personal expectation is that the relationship should be almost linear because it is very likely that a user who likes a tweet also retweets it. Here, I've developed a linear model to find out. The following table summarizes the result:

OLS Regression Results

Dep. Variable:	favorite_count	**R-squared:**	0.864
Model:	OLS	**Adj. R-squared:**	0.864
Method:	Least Squares	**F-statistic:**	1.268e+04
Date:	Mon, 24 Dec 2018	**Prob (F-statistic):**	0.00
Time:	10:40:24	**Log-Likelihood:**	-19690.
No. Observations:	1993	**AIC:**	3.938e+04
Df Residuals:	1991	**BIC:**	3.939e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	2056.9193	121.439	16.938	0.000	1818.758	2295.080
retweet_count	2.5159	0.022	112.607	0.000	2.472	2.560

Omnibus:	525.687	**Durbin-Watson:**	0.761
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16975.761
Skew:	0.569	**Prob(JB):**	0.00
Kurtosis:	17.252	**Cond. No.**	6.23e+03

In the above summary, p-value is zero and R-squared is close to 1. This suggests that there is a statistically linear relationship between the x-variable (retweet_count) and the response (favorite_count). The following scatter plot also confirms the existence of such relationship: