Tweet Impact prediction
(Jahanna Chronicle)

A report by
W.L.N. Laavanya

PROBLEM STATEMENT

Jahanna Chronicle wants to decide if a particular tweet could go viral or not and for this reason they want to predict the 'impact' of a tweet with the help of its features, like the number of likes, number of shares, number of followers of the user who posted the tweet, etc.

DATA ANALYSIS AND FEATURE SELECTION

The data that has been given includes **18 features and 1 target variable**. There are a total of **1,30,443** data points (tweets) given in the dataset. The different features include:

- EsId
- URL
- Created Date
- Sentiment
- Post
- Post length
- Hashtag count
- Content URL count
- Like count
- Share count
- Comment count
- Followers count
- Following count
- Tweet count
- Gender
- Peak velocity
- Estimated Reach
- Impact

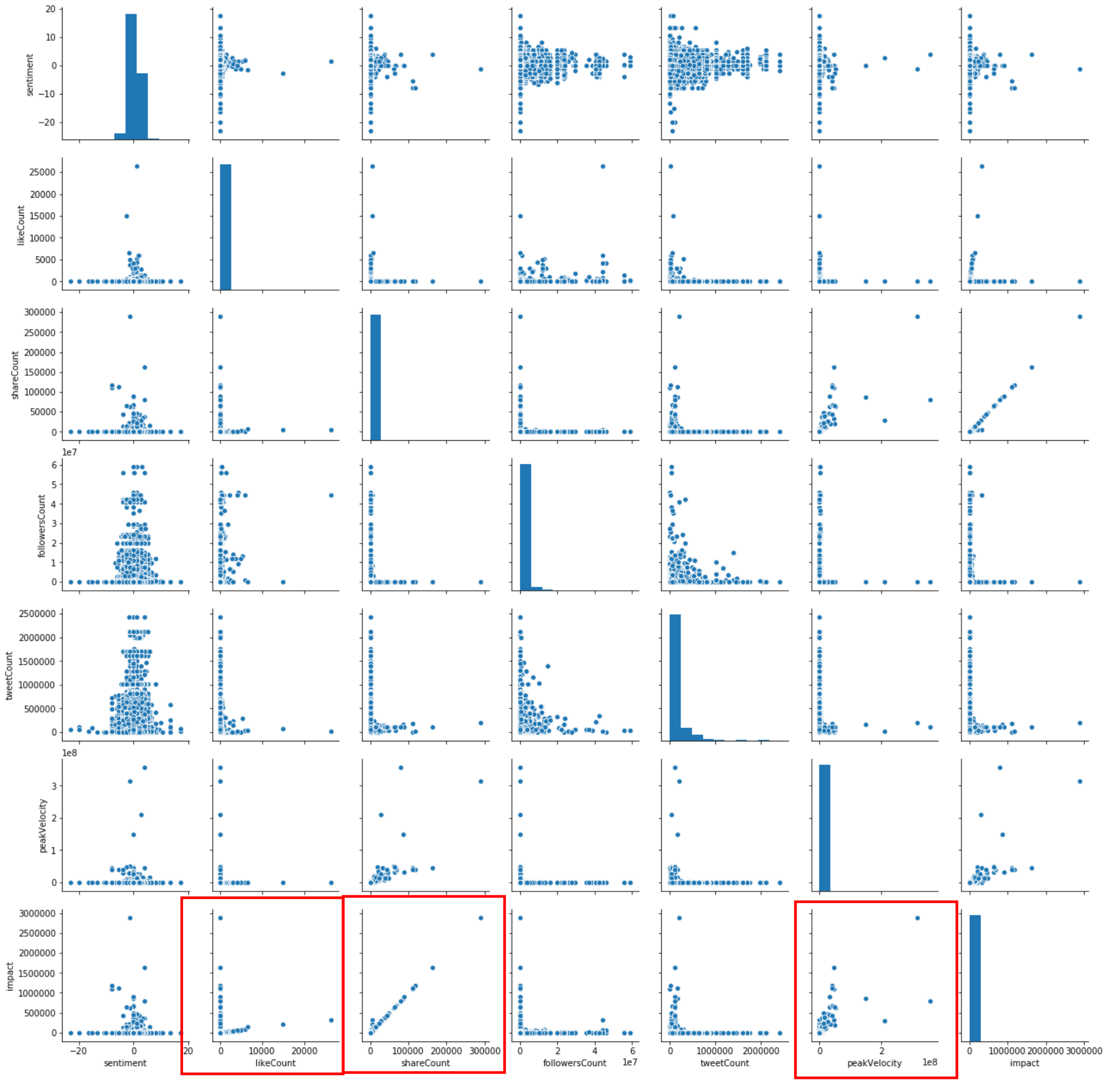


Fig. 1 Pair Plots of top 5 attributes

Attributes like EsId, URL, createdDate do not add any specific detail to the tweet's impact and so have been removed from the analysis. It has been observed that the commentCount is 0 for all tweets, which does not provide any value, and so, that attribute has also been removed from the data analysis. It was also observed that Gender, secondsElapsed, following count, contentURLcount, postLength, post, hashtagcount features. This was found by trying various combination of features for each regression model. **The best features were – 'share count' and 'like count'.**

In order to get an idea which attributes are largely correlated with the impact of a tweet, we take a look at the pair plots of the features. Due to space constraint, we only show the pair plots (Fig. 1) of the top 5 important features – 'shareCount', 'likeCount', 'peakVelocity' 'followersCount', 'sentiment', 'tweetCount'.We can see that the share count and like count are positively correlated with the impact of the tweet and so it is clear that these two features play a vital role in deciding the impact.

To look at it more clearly, we can see the plots (Fig. 2a, 2b, 2c) of 'share count', 'peakVelocity' and 'like count' with

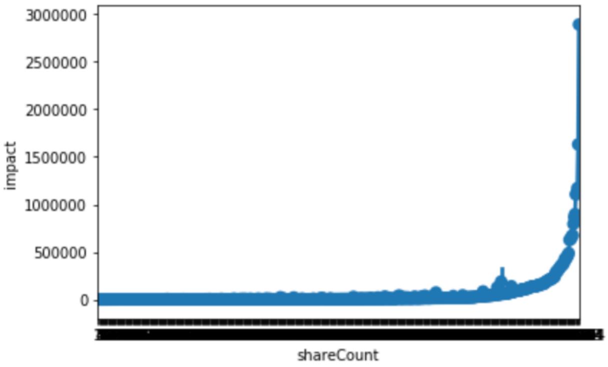


Fig. 2a Share count vs Impact

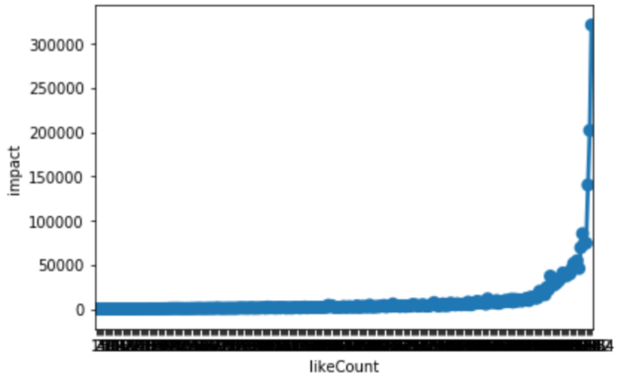


Fig. 2b Like count vs Impact

'Impact' and observe the positive correlation.

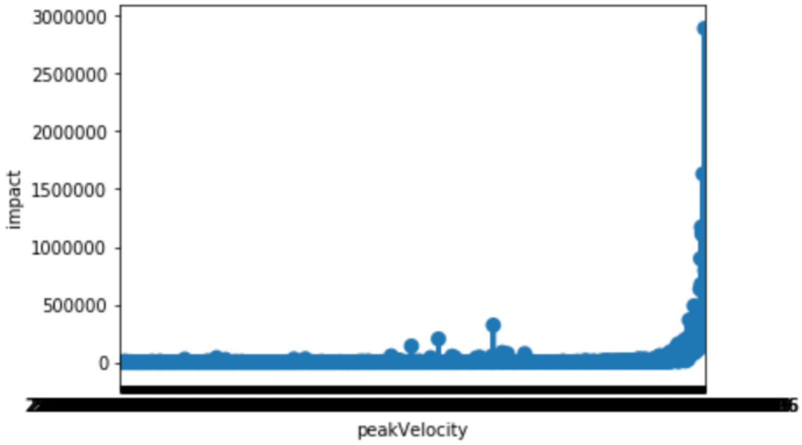


Fig. 2c Peak Velocity vs Impact

If we take ‘sentiment’ and ‘followers count’ (Fig. 2d, 2e), they don’t add much value in telling something about the impact of the tweet.

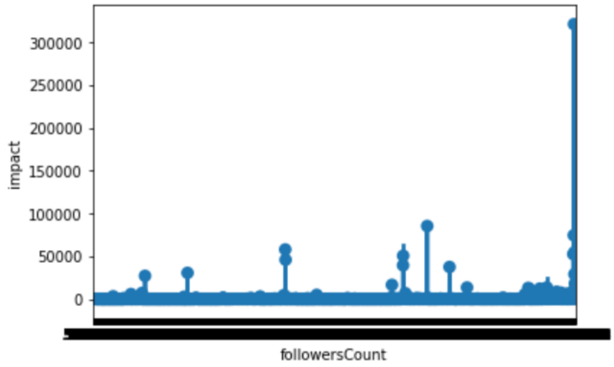


Fig. 2d Followers count vs Impact

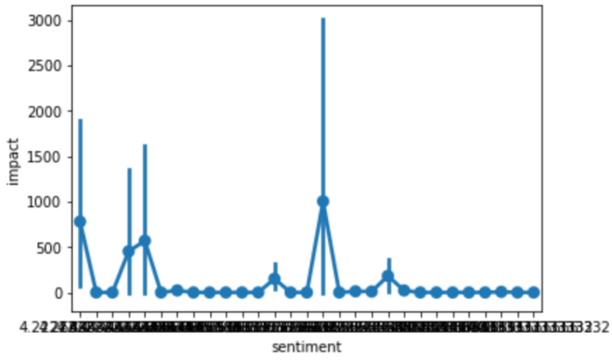


Fig. 2e Sentiment vs Impact

PREDICTION MODELS

We divide the data into 80% training and 20% test set initially and also check performance by doing 10-fold cross-validation of the entire dataset.

ERROR COMPARISON BETWEEN MODELS

	LINEAR REGRESSION	DECISION TREE (Random Forest)	NEURAL NETWORK
RMSE	50.34	656.01	50.45
MAE	1.23	11.30	1.07

- **LINEAR REGRESSION**

The **best prediction models** were found to be the linear regression model and the MLP Regressor (using just ‘share count’ and ‘like count’ features). By best, we mean the models with the least RMSE value as the R2 score is almost the same for all models. The main reason for this is that, the impact score is observed to be a result of a linear combination of the share count and like count of the tweet (Approximately, (like count + share count) *10 = impact, as observed from the data manually).

Features	share count, like count
Model parameters	**Intercept** : -0.69 **Coefficients** : [10.00018299 9.99971358]
Model Evaluation	***Before 10-fold Cross-Validation*** **MAE (Mean Absolute Error)** – 1.237 **R2 score** – 0.99 **RMSE (Root mean square error)** – 50.34 **Training time** – 0.02s

	After 10-fold Cross-Validation **MAE (Mean Absolute Error)** – 1.234 **R2 score** – 0.99 **RMSE (Root mean square error)** – 69.92

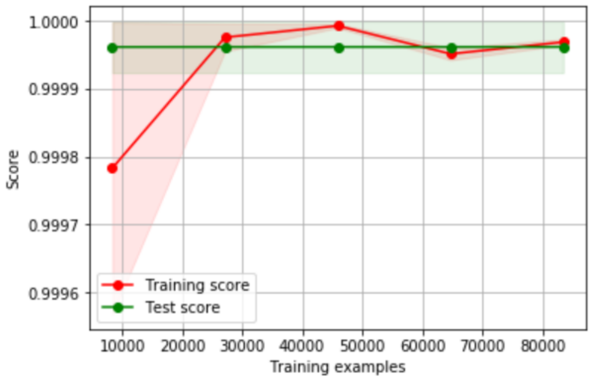


Fig. 3 Performance Plot

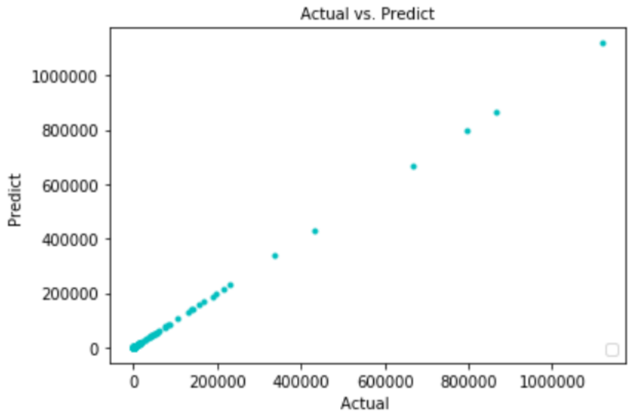
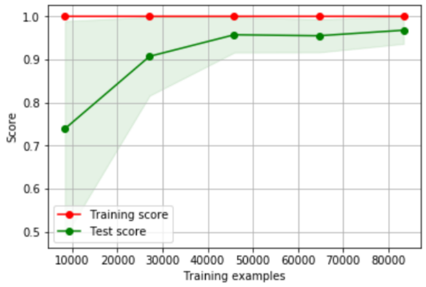


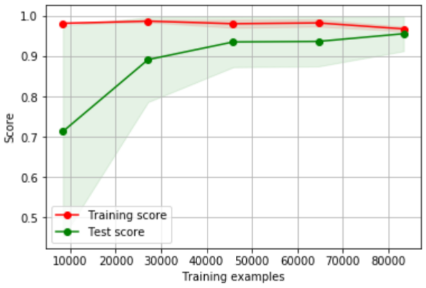
Fig.4 Actual vs Prediction

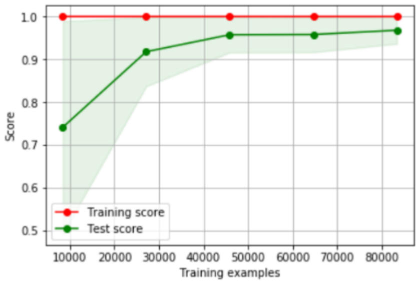
- **DECISION TREE REGRESSION**

Decision Tree Regressor	**Random Forest Regressor with 76 estimators**	**Adaboost with Decision Tree with 100 estimators**
MAE (Mean Absolute Error) – 13.003 **RMSE (Root mean square error)** – 812.10 **R2 Score** – 0.995 **Training time** – 0.052s	**MAE (Mean Absolute Error)** – 11.303 **RMSE (Root mean square error)** – 656.01 **R2 Score** – 0.996 **Training time** – 3.385s	**MAE (Mean Absolute Error)** – 12.7 **RMSE (Root mean square error)** – 804.99 **R2 Score** – 0.995 **Training time** – 7.552s

Decision Trees also get an R2 score of 0.995 (close to linear regression) but the RMSE is very high in this case as compared to Linear Regression. Even with ensemble variations like Random Forest and Boosting of decision trees, the RMSE does not improve much.







- **MULTI LAYER PERCEPTRON REGRESSION**

The MLP regressor’s performance is also similar to the linear regressor, however, the MLP regressor takes more time to train than the linear regressor.

Features – share count and like count

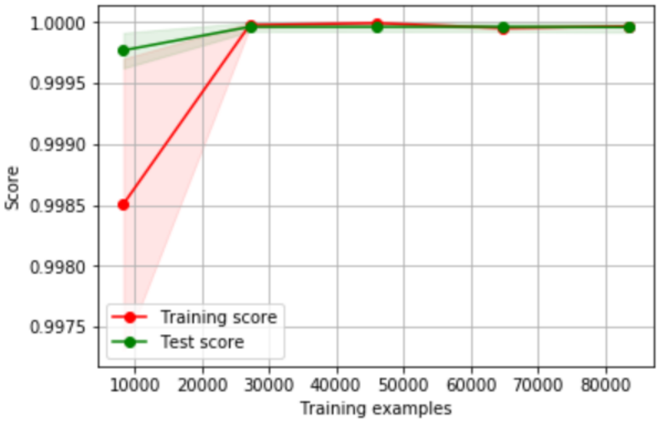
Model parameters – 1 hidden layer with 5 neurons

Model Evaluation – MAE (Mean Absolute Error) – 1.13

R2 score – 0.99

RMSE (Root mean square error) – 50.84

Training time – 15.49s



SUMMARY

In conclusion, the best model that can predict the impact a tweet in terms of performance and time is the Linear Regression model. This model has the best performance using just the features – share count and like count.