**Project: Twitter Impact Prediction**

Created on Friday, July 18, 07:10:2020

|  |  |
| --- | --- |
| Author | Shivam Kolhe |
| Dataset | Twitter Impact Prediction |
| File | 1 |
| Type | Regression |
| Minimum Test RMSE Achieved | 0.57 |
| Maximum R2 Achieved | 0.988 |
| Best Algorithm (Lowest RMSE) | Light Gradient Boosting |
| Best Balanced Algorithm | Linear Regression |
| Variable With Highest Impact | LIKES |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Train RMSE** | **Test RMSE** | **R2 Score** | **Time Taken** |
| **Linear Regression** | 0.802 | 0.839 | 0.966 | 46.2 ms |
| **Ridge Regression** | 0.802 | 0.839 | 0.966 | 14 ms |
| **Decision Tree (Pruned)** | 0.609 | 0.657 | 0.98 | 41.8 ms, |
| **Random Forest** | 0.502 | 0.586 | 0.987 | 12.8 s,  38.7 s for randomized tuning |
| **Light Gradient**  **Boosting** | 0.477 | 0.57 | 0.988 | 508 ms,  53.2 s for Grid Tuning |

|  |  |
| --- | --- |
| **Model** | **Parameters** |
| **Ridge Regression** | alpha = 1, max\_iter=500 |
| **Decision Tree (Pruned)** | max\_depth=4 |
| **Random Forest** | n\_estimators= 250,  min\_samples\_split= 2,  min\_samples\_leaf= 4,  max\_features= 'auto',  max\_depth= 9 |
| **Light Gradient Boosting** | learning\_rate= 0.1,  n\_estimators= 850,  num\_leaves= 9 |

* The reason for this high score is the Independent Variable “LIKES”. It has the highest correlation with the target variable “Impact”.
* Final chosen variables for model building are:

**Numeric:**

'Sentiment score',

'Post Length',

'Hashtag count',

'Content URL count',

'Listed Count',

'Likes'

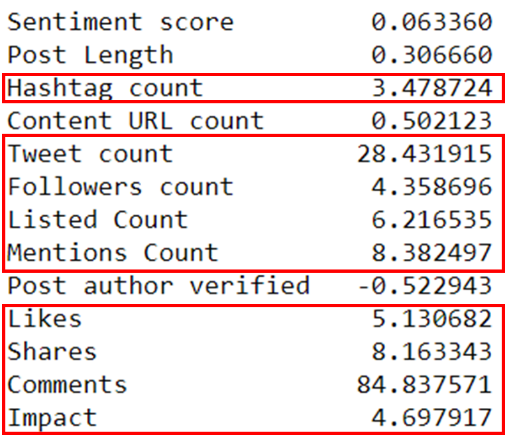
**Categorical:**

'Media Type\_TEXT',

'Media Type\_VIDEO',

'Post author verified\_1.0'

* Skewness

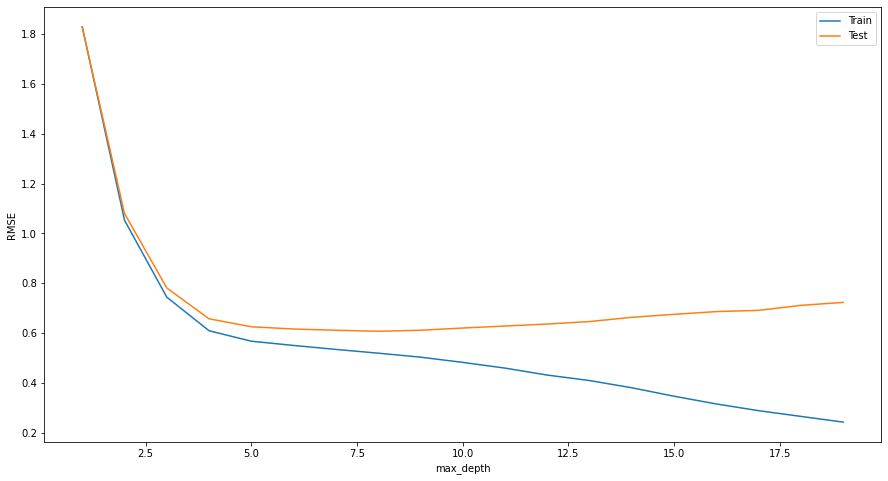


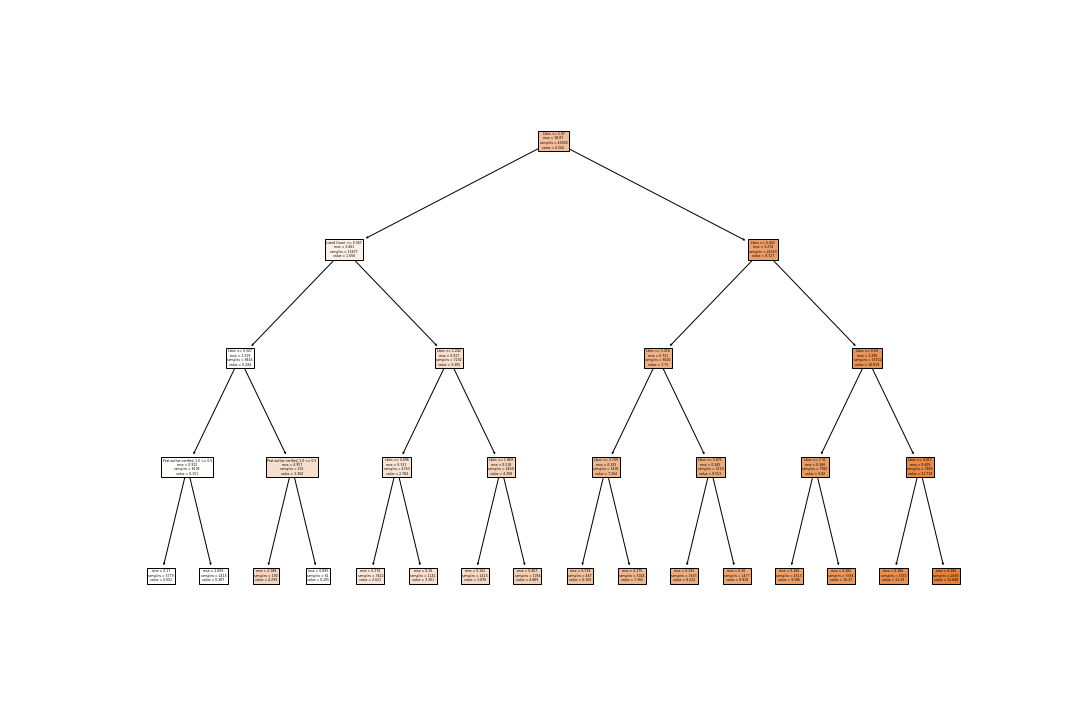
* From the analysis it is seen that no matter how we treat the columns Mentions Count and Comments, the outliers are not removed and the skewness is present.
* One more thing is that these columns contains more than 87% zero values. Thus, these are dropped.
* *'Tweet count', 'Followers count', 'Listed Count', 'Media Type', 'Mentions Count', 'Likes', 'Shares', 'Comments', 'Impact'*;

These columns are log transformed to balance the skewness. The feature Hashtag count is exponentially transformed.



* It can be seen that Shares and Likes are highly correlated with each other. Thus, remove Shares variable
* After performing the Variation Inflation Factor with threshold 10, the remaining multicollinear variable were treated.
* Decision tree was overfitting. Thus, pruning was needed. Pruning was performed and the max\_depth is selected as 4.





*Pruned Decision Tree*