Twitter Impact Analysis

Predicting the impact of a tweet

Submitted by:
Aditya Kalkeri

Problem Statement

Jahanna Chronicle is a technology company based in Lunakick. The task is to predict the impact that a tweet has . Impact can be defined as a value that could help Jahanna Chronicle decide if the tweet could go viral.

Chronicle's data team has prepared a dataset for this task. Chronicle's engineering team wants to explore modeling using a decision tree, neural network and linear regression but give the model that fits best.

Goal : Build a model using the dataset provided to predict the impact a tweet can have.

Introduction

Machine learning is a continuously developing field. Machine Learning is considered to be a subfield of Artificial Intelligence (AI), and although it falls under the Computer Science field, its methods differ from traditional programming. The goal of Machine Learning is to understand the structure of data, find patterns, and build models that can then be utilised to predict future 'Target Values' based on input data.

All the modern technologies use Machine Learning in some way. Facial recognition technology, Optical character recognition (OCR) technology, Recommendation engines, Self-driving cars that rely on machine learning to navigate. All major industries use Machine learning and related methods to gain insights on data respective to their domain. This helps them make the best decision possible and makes the process easier.

Regression

Regression Models are used to predict continuous target values, which makes them important in many fields of Science and Engineering, as well as in Finance, Stock markets, sales prediction, etc. It also helps in understanding relation between variables

1. Linear Regression

   The model assumes Linear Relation between Input variable and target variable. There are again two types of Linear Regression:

   Simple Linear Regression

In Simple Linear regression, only a single feature is used to predict the target values. There exists a linear relation between target and the feature, and the relation is represented with an equation.

Thus in Simple linear regression, algorithm finds the best fitting line through the points

Now we generalize the Simple Linear equation line fitting process, to accommodate multiple features. This process is called multiple Linear Regression.

The equation for the model will be

$$Y = a + B1X1 + B2X2 + \ldots + BnXn$$

n = no of features

X1, X2,....Xn = features

2. **Decision Tree Regression**
   The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

   In Decision Trees, for predicting a target value for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Dataset

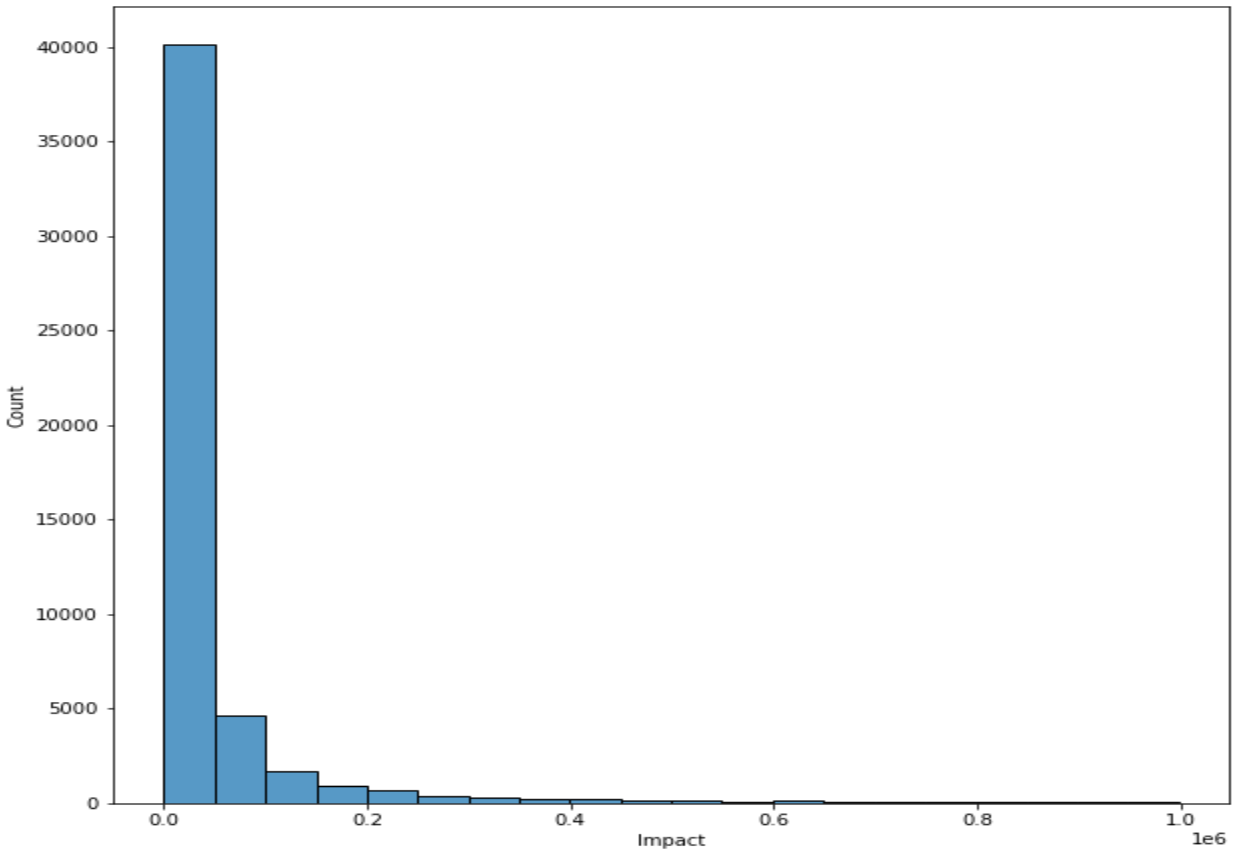
Feature Information:

Post Content - The text in the tweet
Sentiment score - Ranges from -20 to +20 (0 - neutral)
Post Length - The length of the tweet
Hashtag Count - The number of hashtags used in the tweet
Content URL Count - The number of URLs mentioned in the tweet
Tweet Count - The total number of tweets posted by the author of the tweet
Followers Count - The number of followers of the author of the post
Listed Count - the number of lists the post author is a part of

Media Type - The media type of the post (Text, image, video)
Published Datetime - The published time of the tweet
Mentions Count - The number of user mentions in the tweet
Post Author Verified - 1 if author is a verified user
Likes - Likes received for the tweet
Shares - Retweets received for the tweet
Comments - Number of comments for the tweet

Target Variable: Impact

Description:
Numeric variable
Min value = 0
Max value = 997980
Mean = 40038
Median = 2100



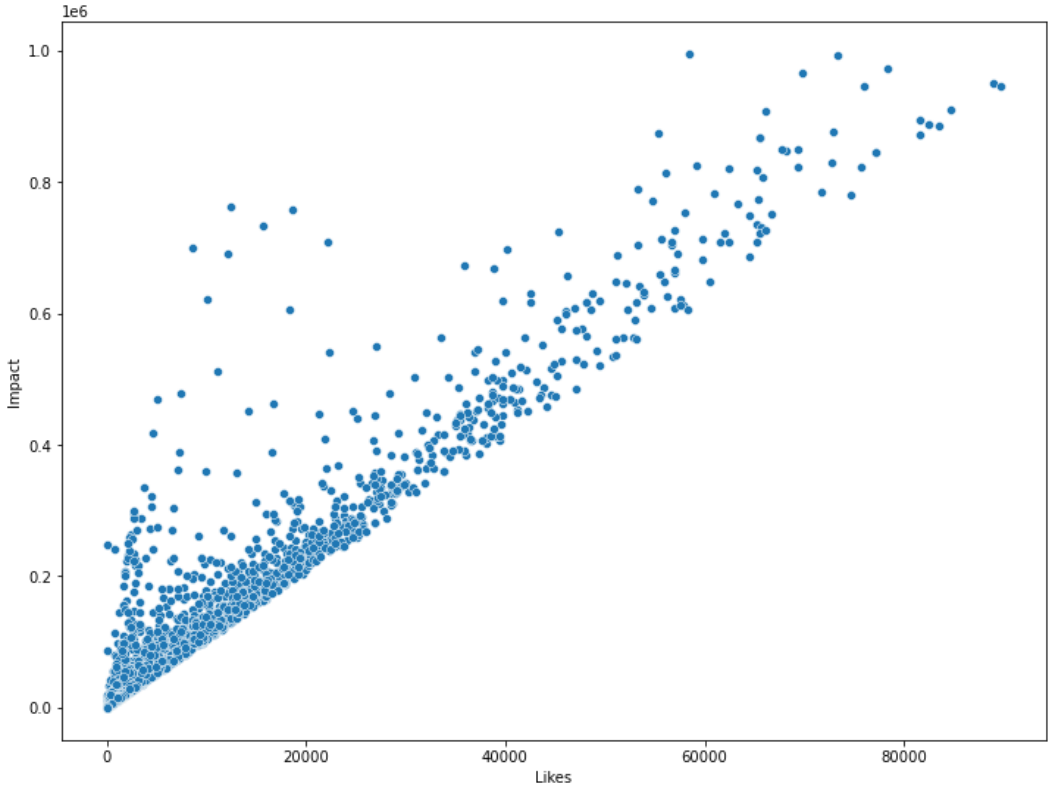
Most of the values are 0. Only a few of the tweets have a high impact. (This type of distribution which has a right skew, is typical in many scenarios : Likes on Instagrams, Salary of employees in any company, etc)

We will divide the dataset into 3 parts:

1. High impact = Upper Limit (75th percentile + IQR) (> 54990)
2. Mid impact = 75th percentile to upper limit (54990 - 27500)
3. Low impact = Below Mid impact (< 27500)

Likes

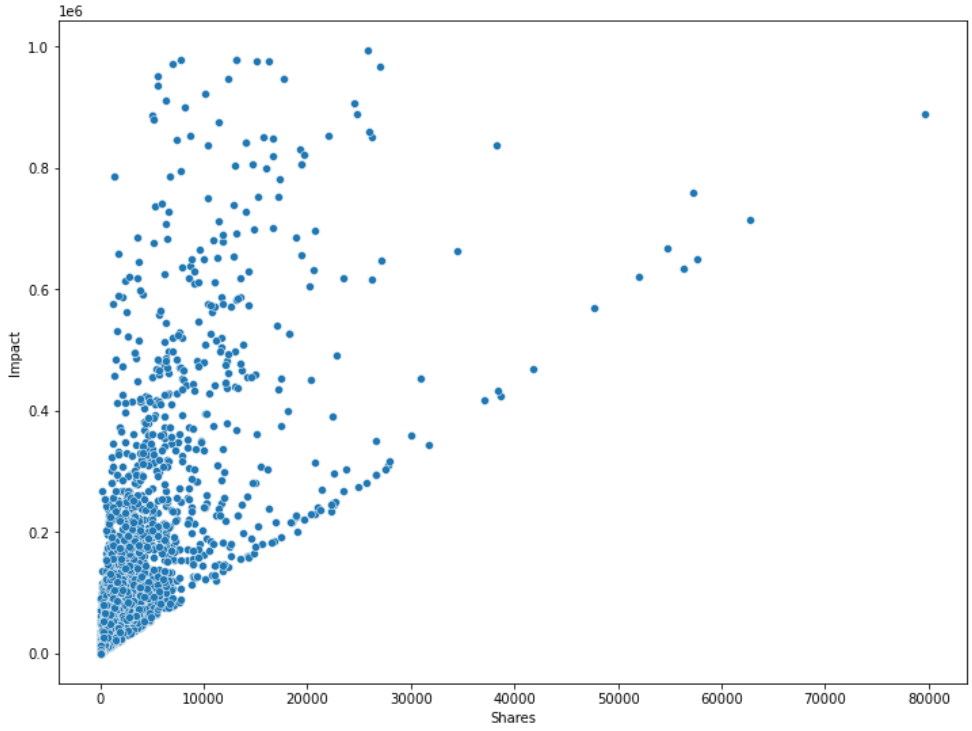
Description
Numeric variable
Min value = 0
Max value = 90919
mean = 3020
Median = 153
Correlation with Impact = 0.93



Highly correlated with target variable, this feature will prove good for prediction.

Shares

Numeric variable
Min value = 0
Max value = 79671
mean = 4508
Median = 2405
Correlation with Impact = 0.73



Good correlation with Impact

Modelling

Data preprocessing

Three steps were performed for data preprocessing

1. Dropping irrelevant features

   The following features were removed from the dataset
   Id, Post Content, Published DateTime, Media Type

2. Scaling the numeric features

   Next we scaled the features by applying standardscaler from sklearn library. This transforms the features into a distribution with mean = 0 and standard deviation = 1

3. Train Test Split

   The dataset was split into two sets: train and test. Train will be used to train the model, test set will be used for evaluation the model

Metrics

Three metrics were used:

1. Root Mean Squared Error
   Calculates the root of the mean of squares of all the residuals from predictions.
2. Mean Absolute Error
   Calculates the mean of absolute error
3. R2 - score
   Gives the ratio of variance of target explained by the model and total variance of the target variable.

Training Models

Testing metrics for all the models

Name	RMSE	MAE	R2
Linear Regression	3.19	0.30	0.99
Ridge	4.00	1.20	0.99
Lasso	3.88	1.05	0.99
Decision Tree	4573	819	0.99
Random Forest	3608	736	0.99
XGBoost	3174	637	0.99

As we can see, all the linear models perform extremely well on the data. Tree based models failed to predict the data. Linear regression performs the best. So we will evaluate the model further.

Evaluating Linear Regression

We did a cross validation test on Linear regression with cv = 10. The score was consistent throughout the dataset and the variation was low (MAE for cv 10 = 0.322 +- 0.06)
The time taken to predict 12500 instances was 0.97 ms

Conclusion

- The model performs well for all instances of data.
- The model is simple and lightning fast.
- we used only 12 features to predict the impact
- Likes, Shares, comments decide the Impact
- It will also help if you are verified and have a high follower count.