جامعة مصر للمعلوماتية
EGYPT UNIVERSITY
OF INFORMATICS

Social media and its impact

Made by:

Eyad Mohamed:22-101037

Mohamed Salah 22-101093

Nour eldin Mohamed 22-101075

Yassin Elshahbour 22-101006

Youssef Hariry 22-101263

Youssef Gamaleldin 21-101114

FACULTY OF
ENGINEERING

Table of Contents

1)Abstract:

In this project, we sought to analyze data from a study conducted on Facebook posts to explore how different sections of the study affect one another. Our goal is to understand the relationships between various features and how they influence each other and how they would influence a specific output using both regression and classification.

We start by using regression to investigate how the output would change in accordance with the other sections of the data. In addition, we apply classification methods to group a sections output into categories and study how these categories are affected by different features.

To build the relations and be able to predict the outputs, we build models using both approaches regression for predicting continuous outputs and classification for predicting categories. We also visualize the relationships between sections to better understand the structure and connections in the data.

The regression model obtained an R2 of 0.95, which indicates that 95% of the variation in the y values is explained by the x values, our classification model achieved an accuracy of 85% when predicting the correct class. By experimenting with various algorithms and optimizing their parameters for improved performance, these outcomes were attained. Later in the report, specifics regarding the techniques utilized and their results will be given.

2)Dataset:

We used the same dataset for both regression and classification. The data set used was the facebook metrics dataset which studies different interactions in facebook.We used standard scalar normalization on this data for both classification and regression.

Data set feature	Description
Page total likes	Number of people who have liked the company's page.
Type	Type of content (Link, Photo, Status, Video).
Category	User selected Category: action (special offers), product (direct advertisement) and inspiration (non-explicit brand related content).
Post month	What month was the post made
Post weekday	What day is the post made in
Post hour	What hour is the post made
Paid	If the company paid Facebook for advertising.
Lifetime post total reach	How many people did the post reach till now
Lifetime post total impressions	Impressions are the number of times a post from a page is displayed, whether the post is clicked or not.
Lifetime post consumers	The number of people who clicked anywhere in a post

Lifetime engaged users	The number of people who clicked anywhere in a post (unique users).
Lifetime post consumptions	The number of clicks anywhere in a post.
Lifetime post impressions by who have liked a page.	Total number of impressions just from people who have liked the page
Lifetime post reach by people who liked a page	The number of people who saw a page post because they have liked that page
Lifetime people who have liked a page and engaged with a post	The number of people who have liked a Page and clicked anywhere in a post
Comments	Number of comments on the publication.
Likes	Number of "Likes" on the publication.
Shares	Number of times the publication was shared.
Total interactions	The sum of likes comments and shares of a post

Table (1):Features

3) Methods

We tested three different algorithms for both regression and classification tasks.

3.1 Classification

The features used during classifications were:

(Page total likes, Post Hour, Lifetime Post Total Reach, Lifetime Post Total Impressions, Lifetime Engaged Users, Lifetime Post Consumers, Lifetime Post Consumptions, Lifetime Post Impressions by people who have liked your Page, Lifetime Post reach by people who like your Page, Lifetime People who have liked your Page and engaged with your post, comment, like, share, Total Interactions)

3.1.1 Random Forest Classification

We implemented Random Forest Classification using different combinations of parameters:

Max depth values: 2, 4, 6, 8

Number of estimators: 10, 50, 100, 200

Each combination's precision was computed, and we chose the combination with the best precision for our model.

3.1.2 K-Nearest Neighbors (KNN) Classification

We tested the KNN algorithm with k-values ranging from 1 to 20 to assess the model's performance across different neighborhood sizes. The value of k that had the highest precision was then chosen for our model.

3.1.3 Naïve Bayes Classification

A Naïve Bayes classifier was implemented using the Gaussian distribution to model feature probabilities.

3.2 Regression

Regression for each type was made with the output(Y) being the total interactions and the features was

(Lifetime Post Total Reach, Lifetime Post Total Impressions,Lifetime Post Consumers,Lifetime Engaged Users,Lifetime Post Consumptions, Lifetime Post Impressions by people who have liked your Page, Lifetime Post reach by people who like your Page, Lifetime People who have liked your Page and engaged with your post, Paid)

3.2.1 Random Forest Regression

Random Forest Regression was tested using the same set of parameters as the classification model:

Max depth values: 2, 4, 6, 8

Number of estimators: 10, 50, 100, 200

Each combination's R2 was computed, and we chose the combination with the best R2 for our model.

3.2.2 Polynomial Regression

Polynomial regression was tested with degrees 1 and 2.
Higher degrees (4 and 8) were also attempted but resulted in long runtimes followed by unrealistic R2 scores likely due to overfitting.

3.2.3 Ridge Regression

We implemented the Ridge regression using the same polynomial degrees (0, 1, 2) and tested a range of alpha values: 0.001, 0.1, 1, 10, 100.

4)Evaluation

4.1 Classification

We evaluate each classification by accuracy.

4.2 Regression

We evaluated each regression by their R2.

5)Results

5.1 Classification

Random forest Classification best results were with Number of estimators = 200 Max depth = 2

KNN Classification bests results were with k = 19

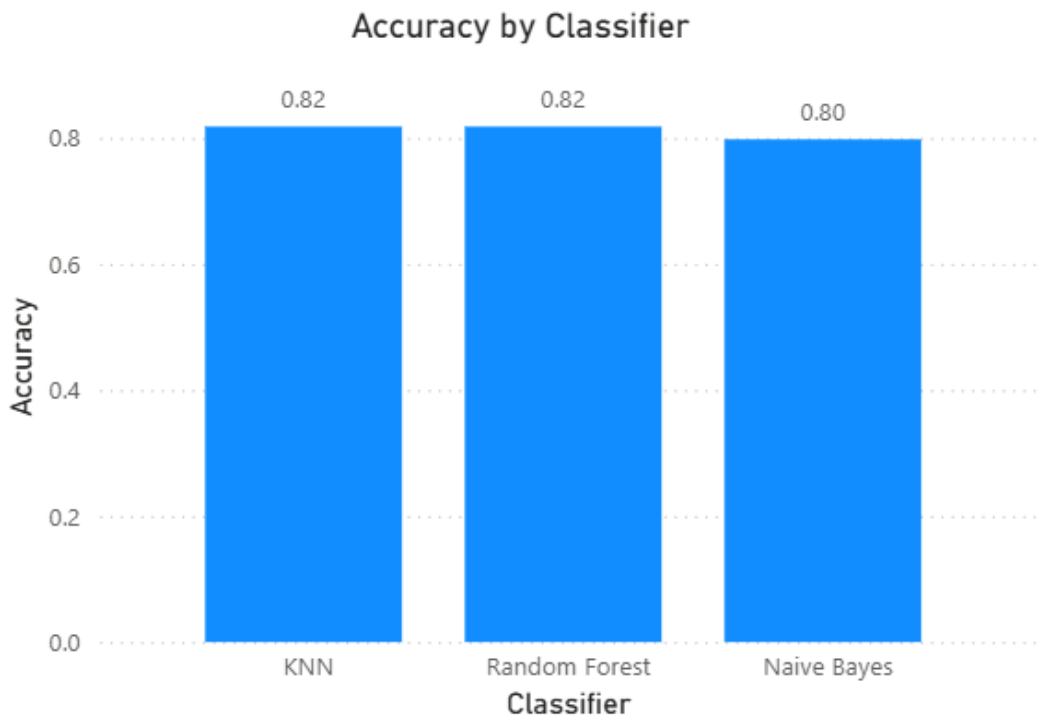
5.2 Regression

Random forest regression best results were with Number of estimators = 50 Max depth = 6

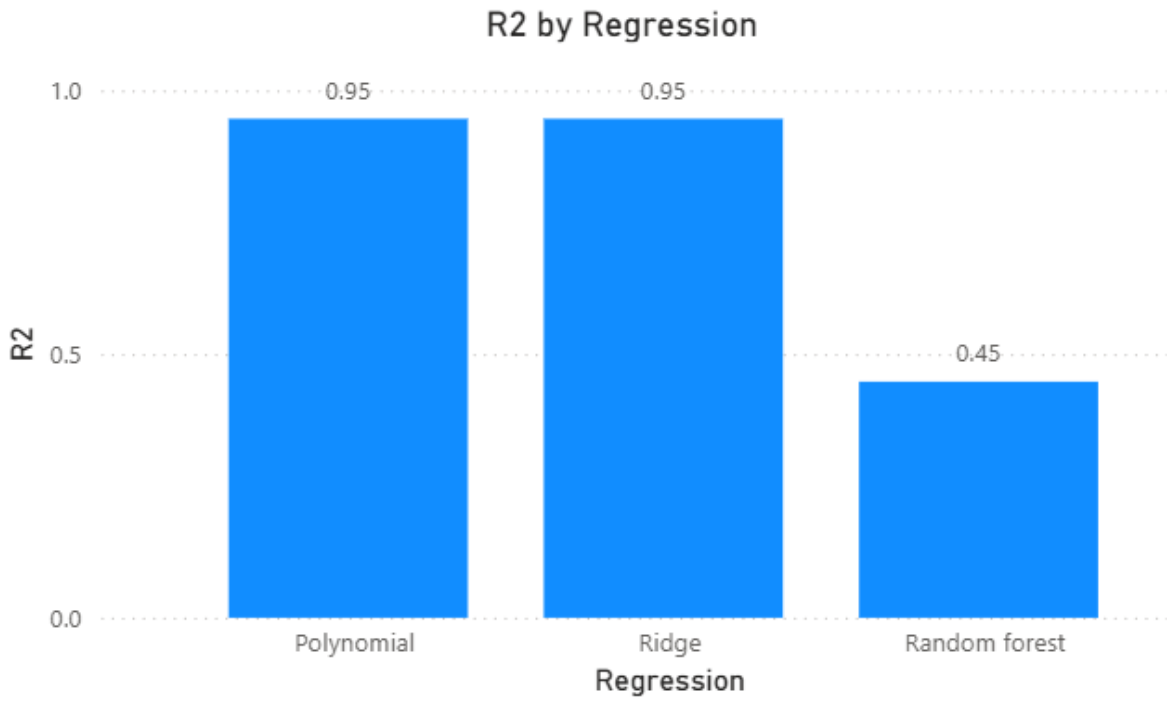
Polynomial regression had the best results with degree = 1

Ridge regression had the best results at degree = 1 and alpha = 0.001

5.3 Visuals



Figure(1):Accuracy according to the classification model



Figure(2):R2 according to the regression model

6)Conclusion

To wrap things up according to our dataset and our parameters the Random Forest Classification had the best precision coming behind it was the KNN in second place and the Naive bayes was the third place. The regression models had the ridge at degree 1 and alpha = 0.001 at the same R2 of polynomial and following them was the random forest.

At any other degree the polynomial had a negative R2, it being negative meant the model could have been because of a severe overfitting.

The ridge had improved the R2 to the same degree suggesting that the issue indeed was over fitment.

References:

https://www.youtube.com/watch?v=gkXX4h3qYm4&pp=ygUacmFuZG9tIGZvcmVzdCBhbGdvcml0bWFzxLE%3D

UCI Machine Learning Repository

https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html