ENTEVYUV 10.0

COMMENT2LIKES: ESTIMATING VIDEO LIKES USING COMMENT DATA

Presented by Vinodhini Rajamanickam

Index

Topic Highlights

Problem Statement
Project aim
Tools Used
Project Processes (step wise)
Approaches
Exploratory Data Analysis (EDA) Insights
Model Performance Evaluation
Conclusion

PROBLEM STATEMENT .

Content creators, marketers, and platform administrators lack a reliable tool to estimate the potential popularity of their videos based on comment information.

Project's Aim

to develop a machine learning model that predicts video popularity using comments, empowering decision-making, content optimization, and marketing strategies on video-sharing platforms.

Tools Used

PROGRAMMING LANGUAGE

Python

IDE

Jupiter Notebook

DATA LOADING AND MANIPULATION

Pandas

DATA VISUALIZATION

Matplotlib, Seaborn

MACHINE LEARNING

scikit-learn

NATURAL LANGUAGE PROCESSING

NLTK

Tools Used

REGULAR EXPRESSION

re

TEXT VECTORIZATION AND FEATURE EXTRACTION

Count Vectorizer

DATA SCALING

Standard Scaler

ML ALGORITHM

Decision Tree Regressor

Random Forest Regressor

Gradient Boosting Regressor

EVALUATION METRICS

Mean Squared Error(MSE)

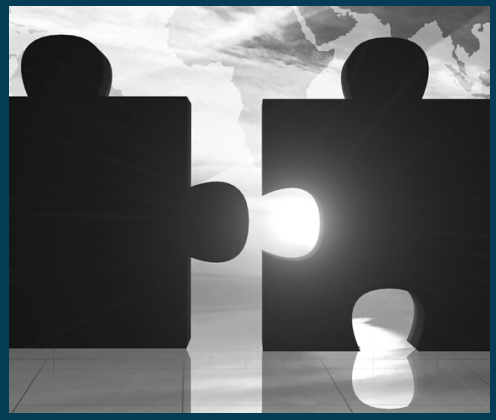
Mean Absolute Error(MAE)

R2 score

PROJECT PROCESS

PHASE 1

- Importing necesssary libraries
- Loading Data
- View Data
- Getting to know the data



PHASE 2

Data Merging

SInce there are two datasets, merging is necessary

PHASE 3

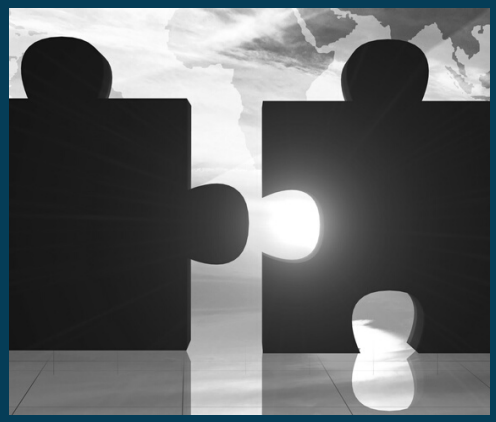
Data Cleaning

- Remove unnecessary features
- Rename columns
- handle null values
- Handle regular expressions

PHASE 4

Exploratory Data Analysis

Exploring the data and getting insights from them



PHASE 5

Define X and y

separating data into X and y (Target Variable)

PHASE 6

Preprocessing

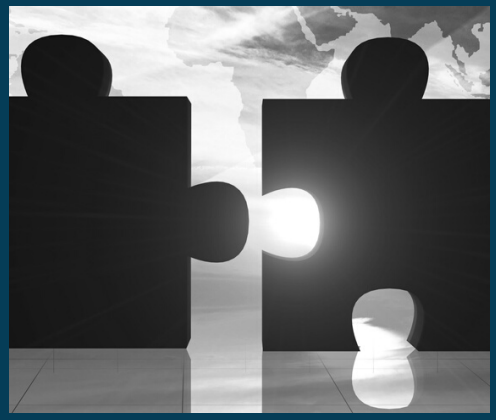
- Remove unnecessary space, symbols, numbers etc.
- Remove stopwords
- Lemmatize the text

PROJECT PROCESS

PHASE 7

Vectorization and Feature Extarction

Turn the text into numerical form and extract a range of features



PHASE 8

Train Test Split

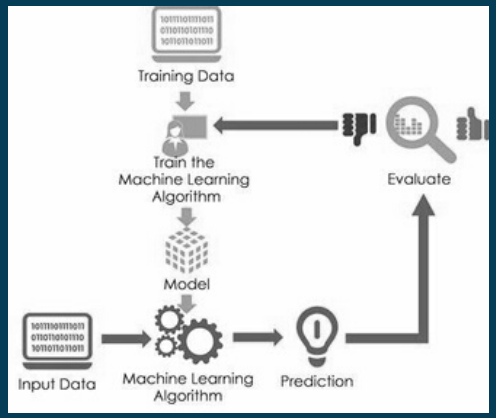
Split the data into 80% train and 20% test dataset

PHASE 9

Scaling

As the data shows a range of variations Scaling is a option to normalize the data

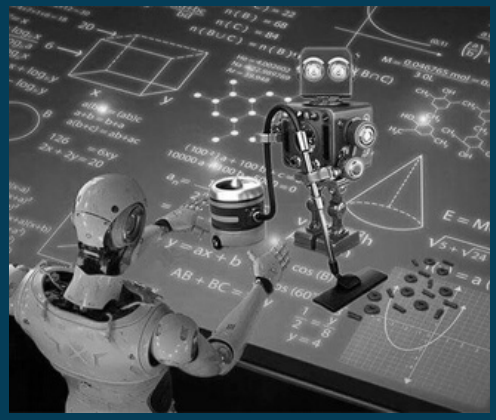
PROJECT PROCESS



PHASE 10

Model Training and Evaluation

train the model on different algorithms and evaluate the model using MSE, MAE and R2 score



PHASE 11

Model Comparison

Compare the models according to their performaces

PHASE 12

Final Model

Finalize the best performing model.

PROJECT PROCESS

Approaches

DATA CLEANING

cleaning data is a necessary part in any ML project.

- **Remove Features**

(Unnamed: 0_x, Unnamed: 0_y, Published At)

- **Rename Columns**

(Likes_x to Video likes, Likes_y to comment likes).

- **Handle null values**

few columns cannot be filled using any methods(eg. Comment) and target variable(Video Likes), it is better to remove them.

PREPROCESSING

- **Remove unnecessary space**
- **Remove stopwords.**

(I, a, an, is,the ,and, etc.)

- **Lemmatize the text**

(smiling, smiles, smiled---> smile)

Approaches

VECTORIZATION AND FEATURE EXTRACTION

applied and compared two different vectorization techniques :

- Count vectorizer
- TF-IDF Vectorizer

selected **count vectorizer**, as this showed better performance after comparison from (96% to 99%)

for feature extraction used **max_features** method.

SCALING

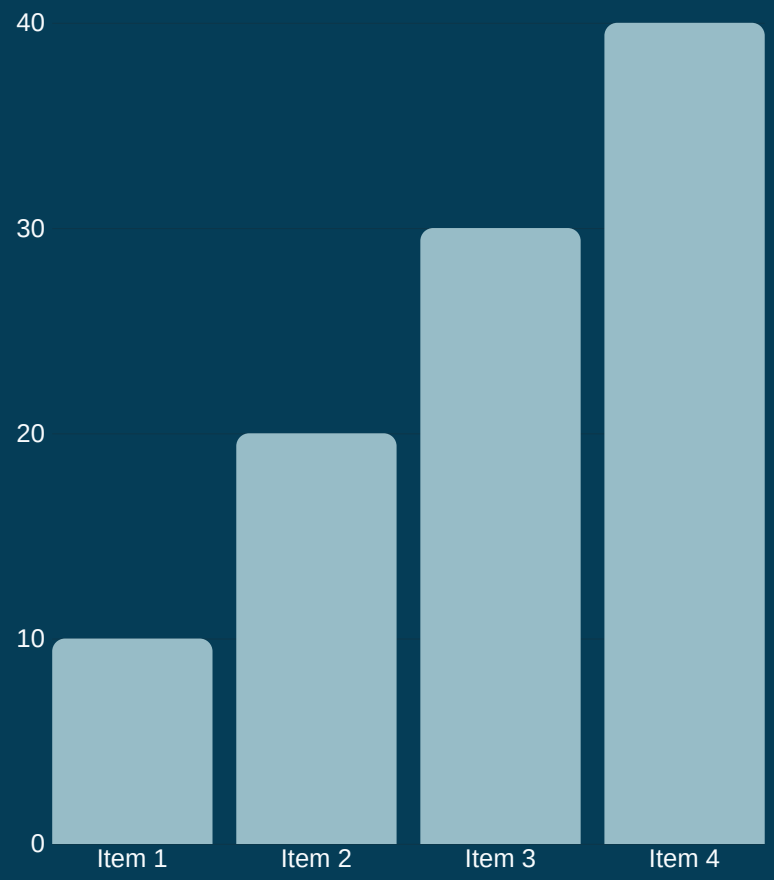
the features showed a range of variations. which needed scaling. Compared two different techniques:

- StandardScaler
- MinMaxScaler

selected **standard Scaler** for scaling.

Exploratory Data Analysis

Insights



Title with highest number of comments

El Chombo | Dame Tu Cosita feat. Cutty Ranks | Official Video | Ultra Music

Title with lowest number of comments

BEST Auditions Of Songs From Movies Amazing Auditions

Title with most number of likes

El Chombo | Dame Tu Cosita feat. Cutty Ranks | Official Video | Ultra Music

Title with least number of likes

How To Build A Business That Works Brian Tracy | GENIUS

Title with most number of views

l Chombo | Dame Tu Cosita feat. Cutty Ranks | Official Video | Ultra Music.

Title with least number of views

Mathematics and Chemistry MathChemistry.com | Masters Degree in Math

Title with most

Positive sentiments

Champions Chess Tour FTX Crypto Cup Day Commentary by David Jovanka Kaja | amp Simon

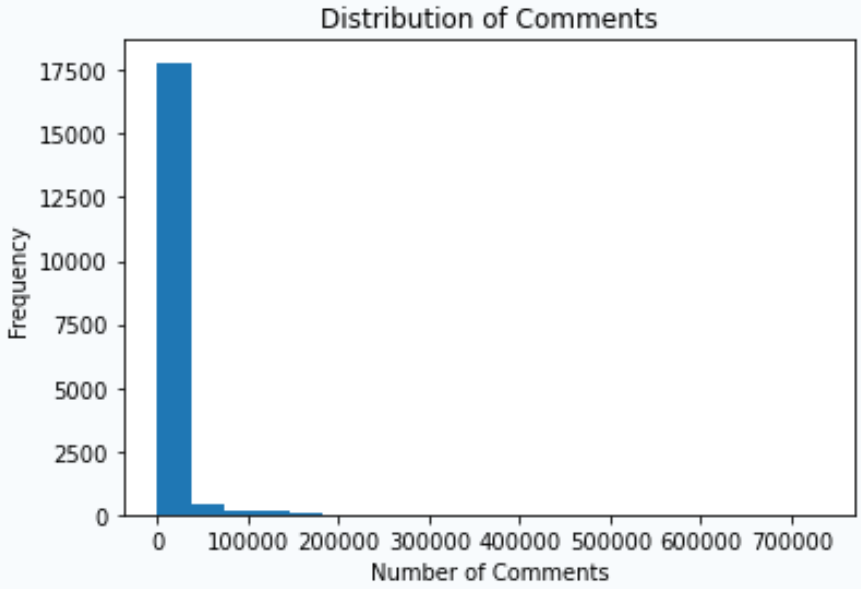
neutral sentiments

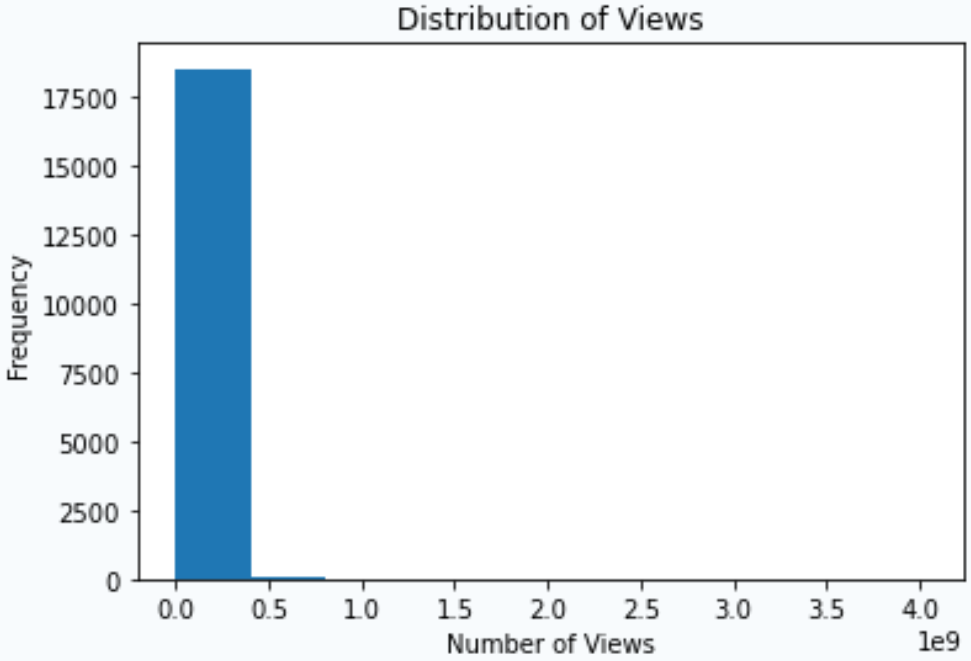
D P CHEZ VOUS Ces PROMOS disparaissent bient t du PS Store Xbox Store Nintendo eShop

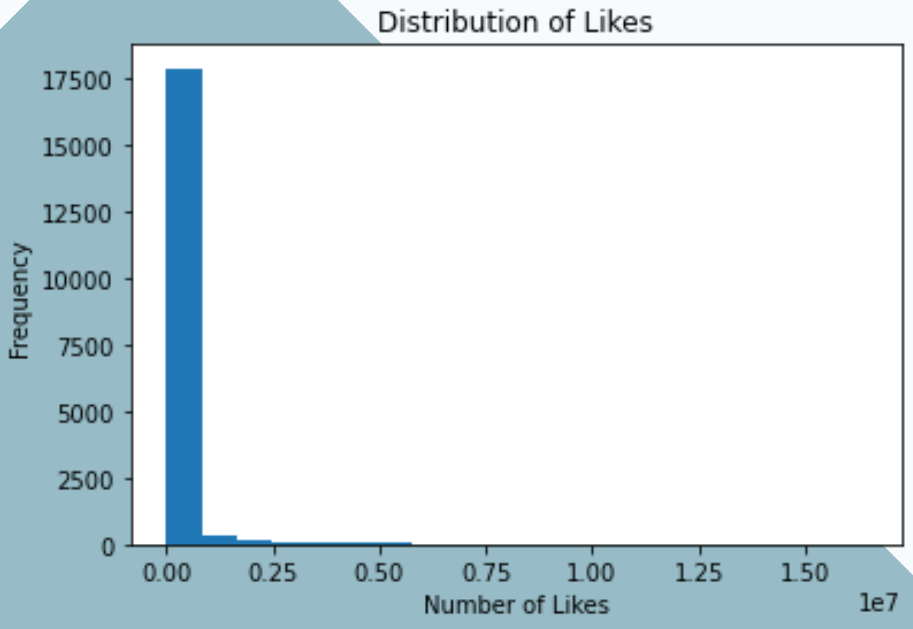
Negative sentiments

Nightly News Full Broadcast Aug

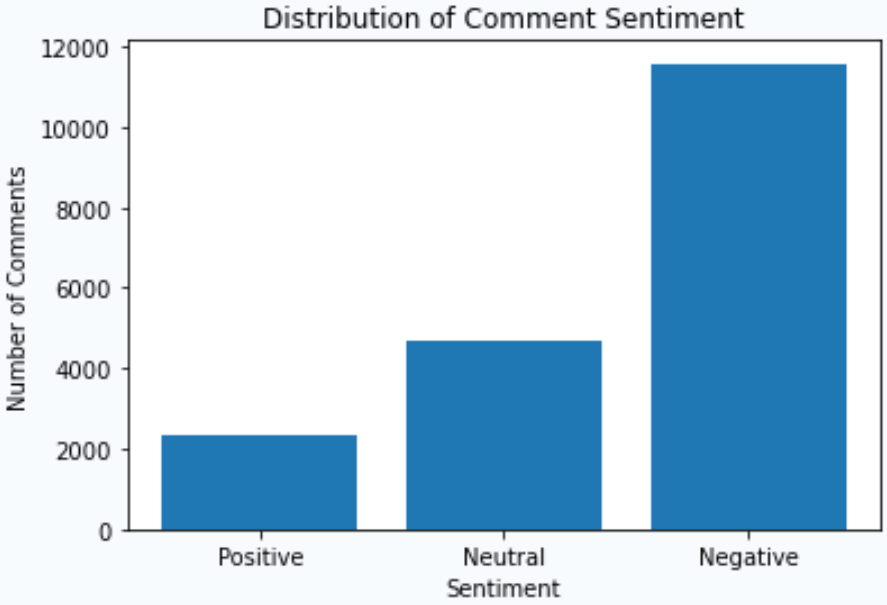
Visualizations

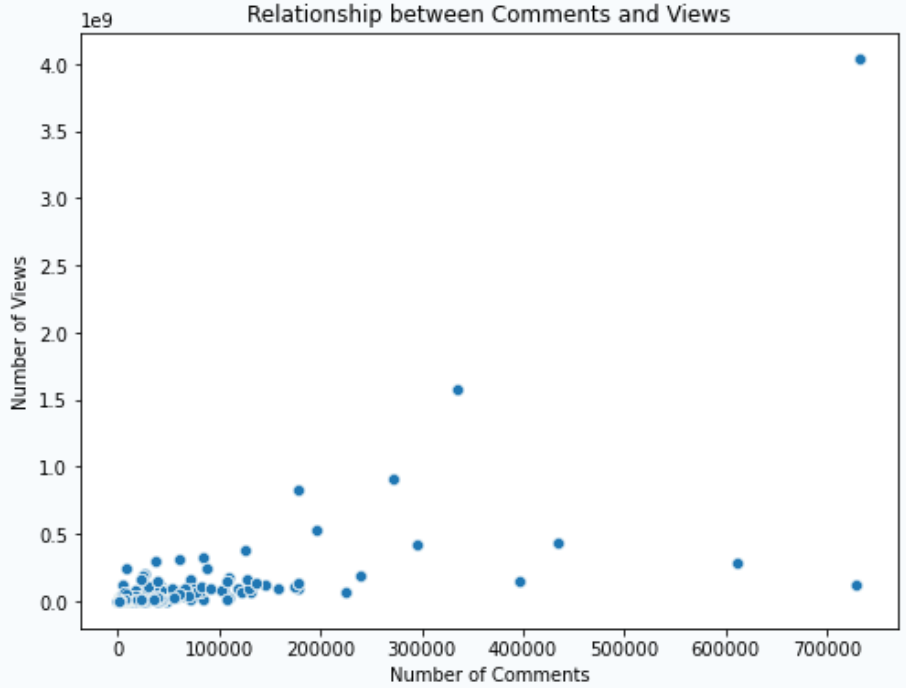


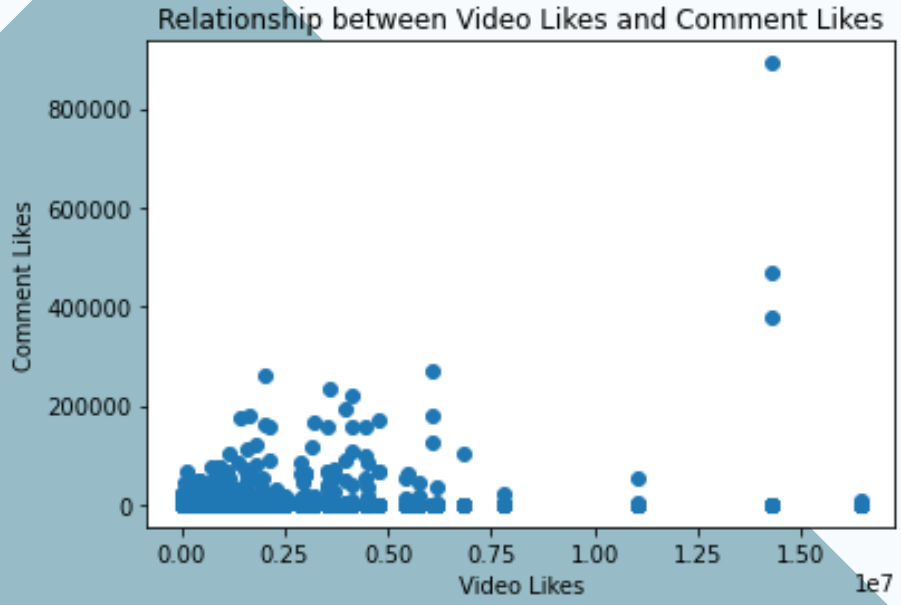




Visualizations

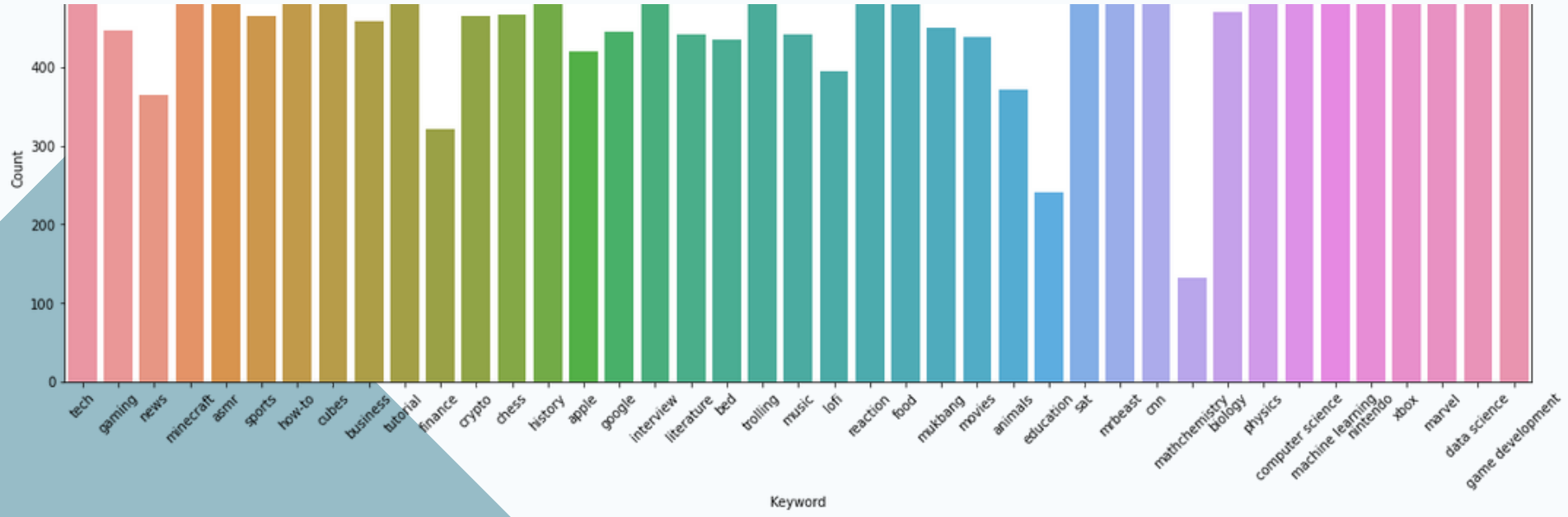






Visualizations

Keywords



MODEL PERFORMANCES AND EVALUATION

Comparing Different models

DECISION TREE

MAE:
80026413.019835

MSE :
2546.195794183445

R2 score :
0.9988

RANDOM FOREST

MAE:
1124419410.13664

MSE :
7154.4537125500

R2 score :
0.9982

GRADIENT BOOSTING

MAE:
7387141277.98380

MSE :
33273.009188345066

R2 score :
0.9887

99.8%

Accuracy of the model

RANDOM FOREST REGRESSOR

Thank you

NAME

Vinodhini Rajamanickam

COURSE

Master Data Science

BATCH

D50