Facebook Comment Volume Dataset

Louis Gauthier & Aurore Pistono
DIA1

Objective

We want to create a model to predict the number of comments a Facebook post will receive, taking into account 53 features, both the attributes of the post and the characteristics of the post's Facebook page.

Our objective is to identify the variables that exhibit the strongest correlation with the engagement level of a Facebook post.

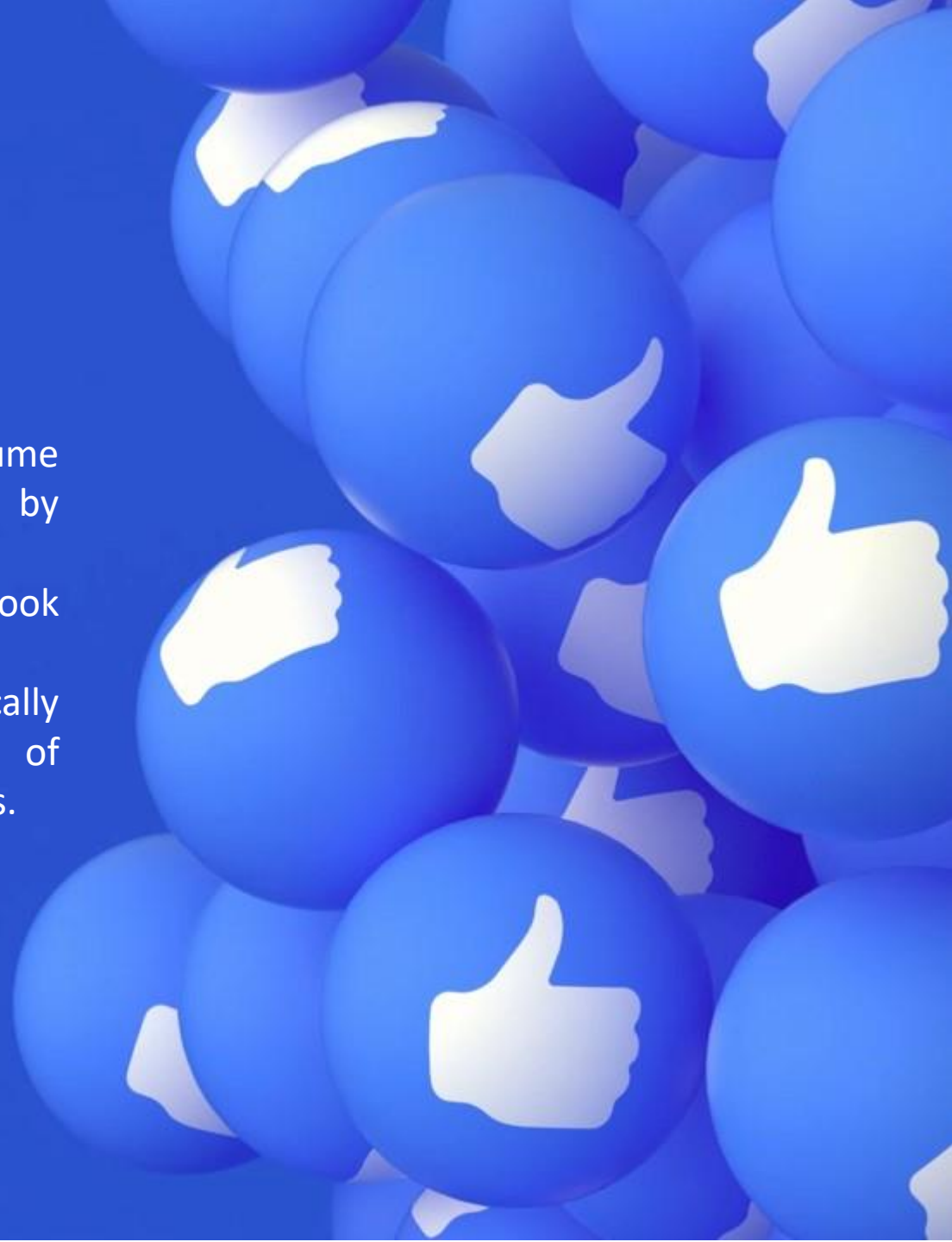
Dataset Presentation

This dataset comes from a research paper exploring "Comment Volume Prediction Using Neural Networks and Decision Trees" authored by Kamaljot Singh, Ranjeet Kaur Sandhu and Dinesh Kumar.

The study delves into modeling user comment patterns on Facebook Pages.

The researchers employed machine learning techniques, specifically Neural Networks and Decision Trees, to predict the number of comments a Facebook post is expected to receive in the next H hours.

The dataset contains 40 949 sample Facebook posts.



Features description :

1 : Page Popularity/likes : number of likes of the page.

2 : Page Checkins : number of people that have physically visited the place.

3 : Page talking about : number of people who come back to the page after liking it.

4 : Page category : category of the document.

5-29 : Derived features : These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features

Features description

30 : CC1 : The total number of comments before selected base date/time.

31 : CC2 : The number of comments in last 24 hours, relative to base date/time.

32 : CC3 : The number of comments in last 48 to last 24 hours relative to base date/time.

33 : CC4 : The number of comments in the first 24 hours after the publication of post but before base date/time.

34 : CC5 : The difference between CC2 and CC3.

35 : Base time : Selected time in order to simulate the scenario.

Features description

36 : Post length : Character count in the post.

37 : Post Share Count : This counts the number of shares of the post

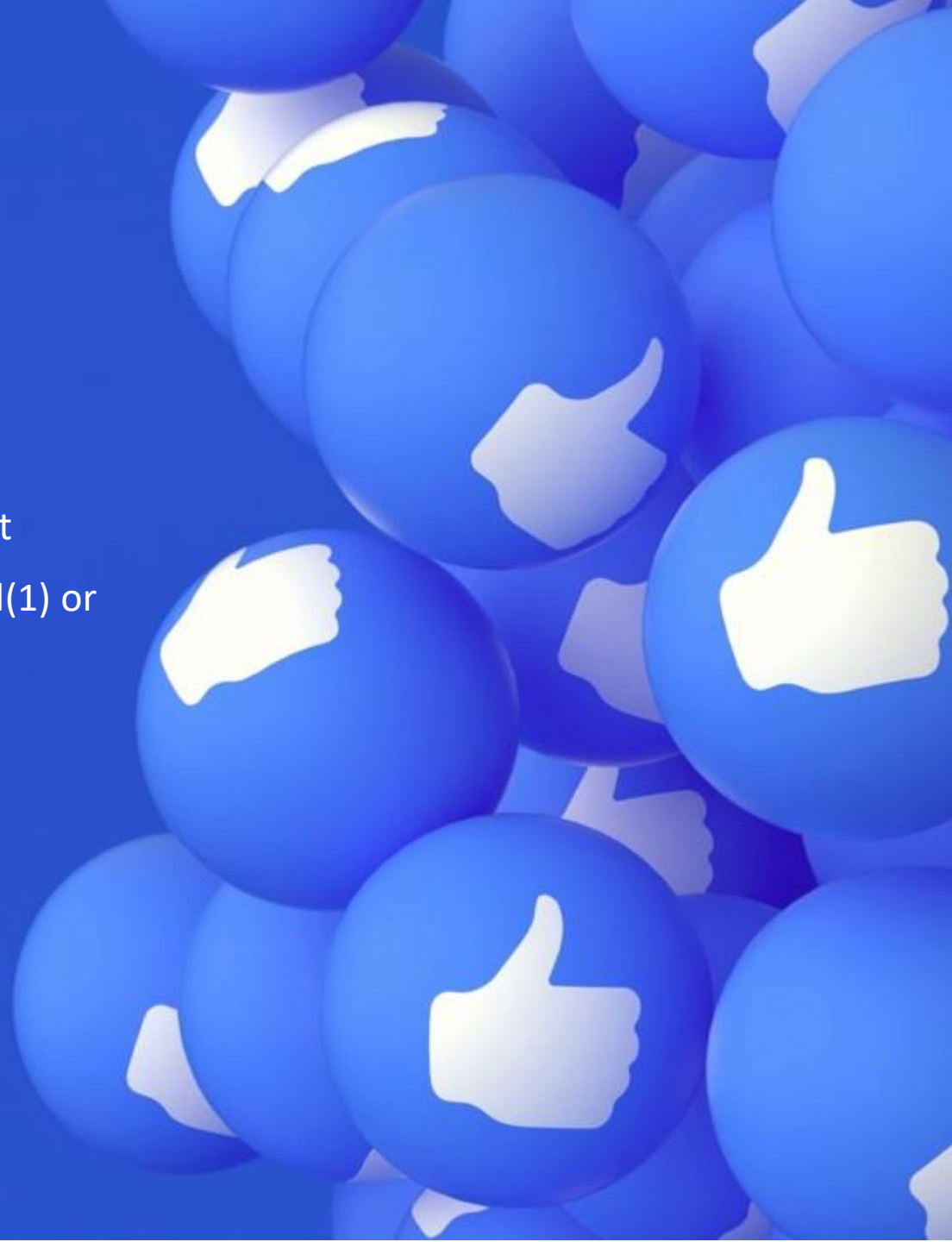
38 : Post Promotion Status : This tells whether the post is promoted(1) or not(0).

39 : H Local : This describes the H hrs, for which we have the target variable/ comments received.

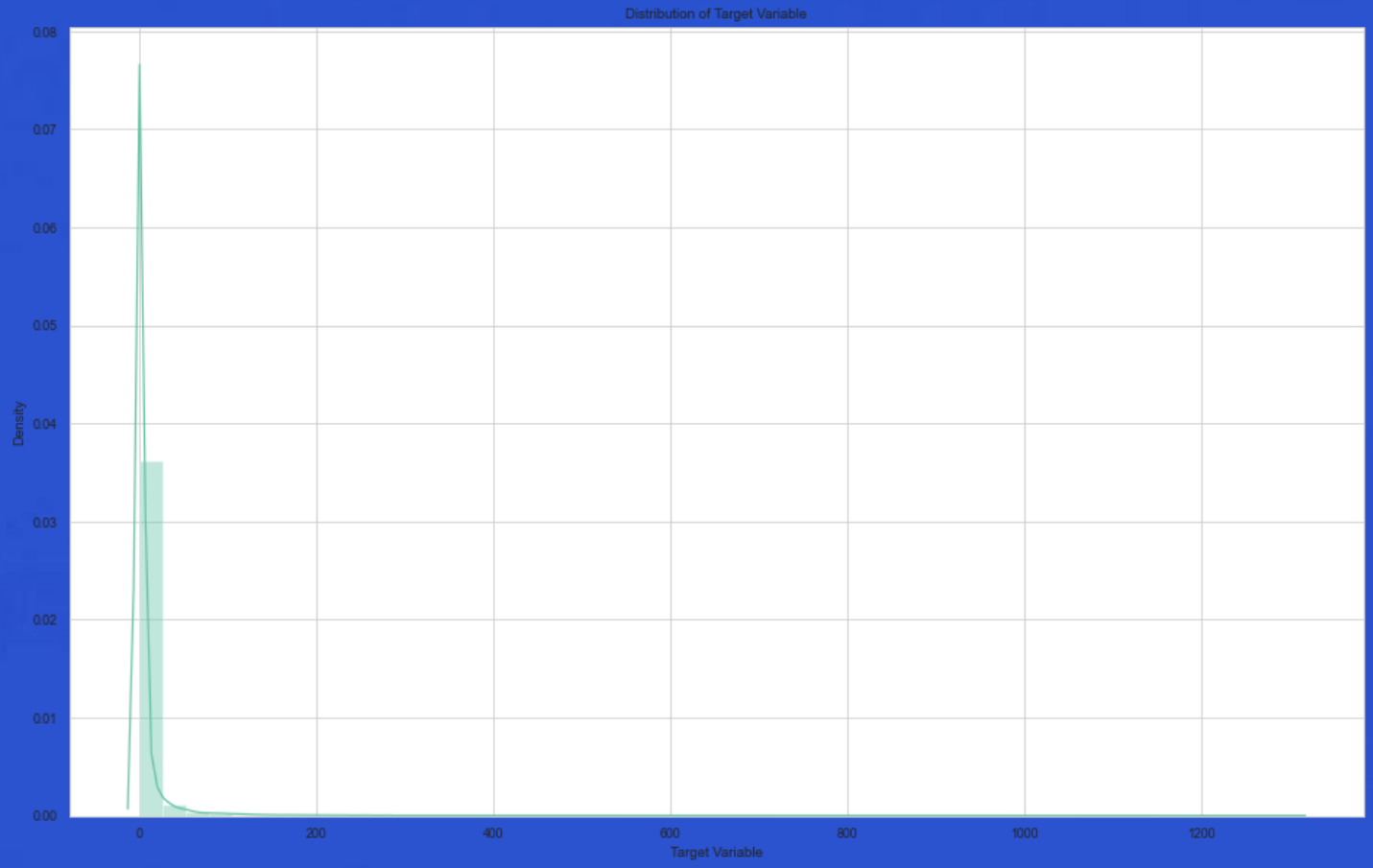
40-46 : Post published weekday : This represents the day(Sunday...Saturday) on which the post was published.

47-53 : Base DateTime weekday : This represents the day(Sunday...Saturday) on selected base Date/Time.

54 : Target Variable : The number of comments in next H hrs

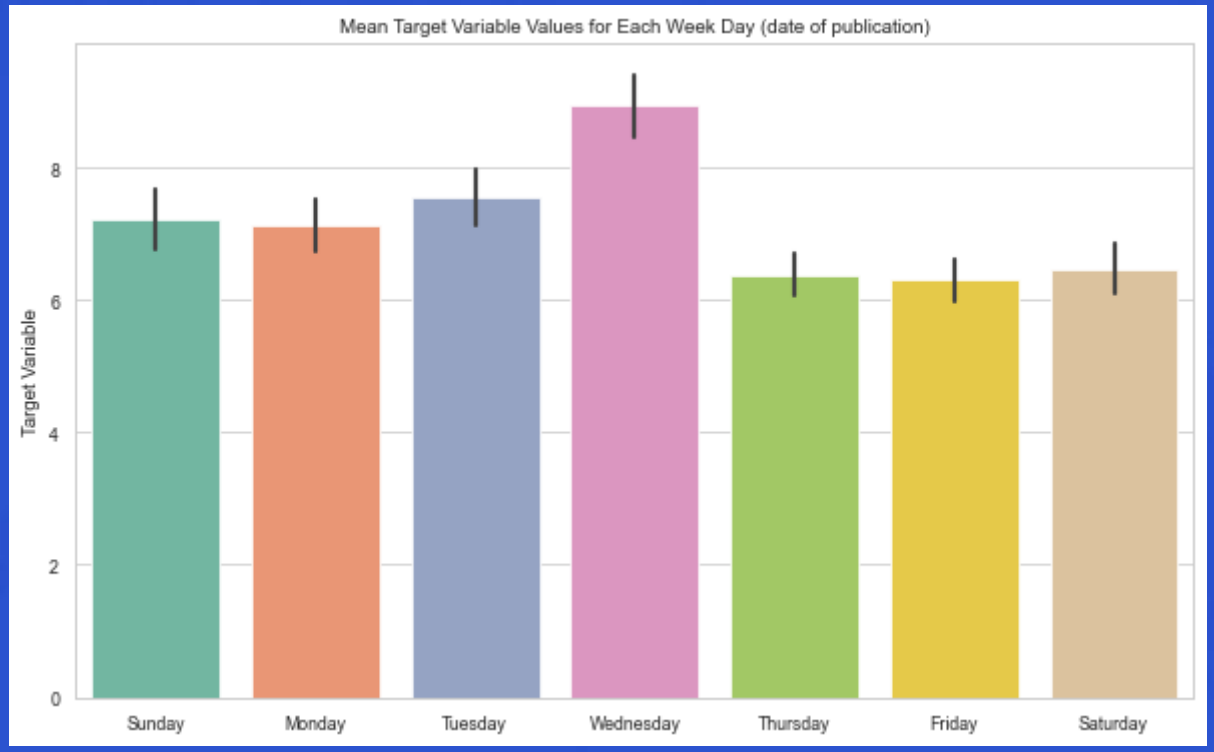


Observations



The plot shows the distribution of the number of comments across the dataset. It indicates that the majority of Facebook posts have zero comments, as evidenced by the peak at 0. However, there are also a few posts that received a higher number of comments, as indicated by the values around 1300.

Observations



This bar plot displays the mean number of comments for each weekday, with the x-axis representing the week days and the y-axis representing the mean number of comments. The highest bar, corresponding to Thursday, indicates that posts published on that day receive the highest average number of comments.

Prediction models

- Linear Regression
- Random Forest
- SVM (Support Vector Machine)
- Gradient Boosting

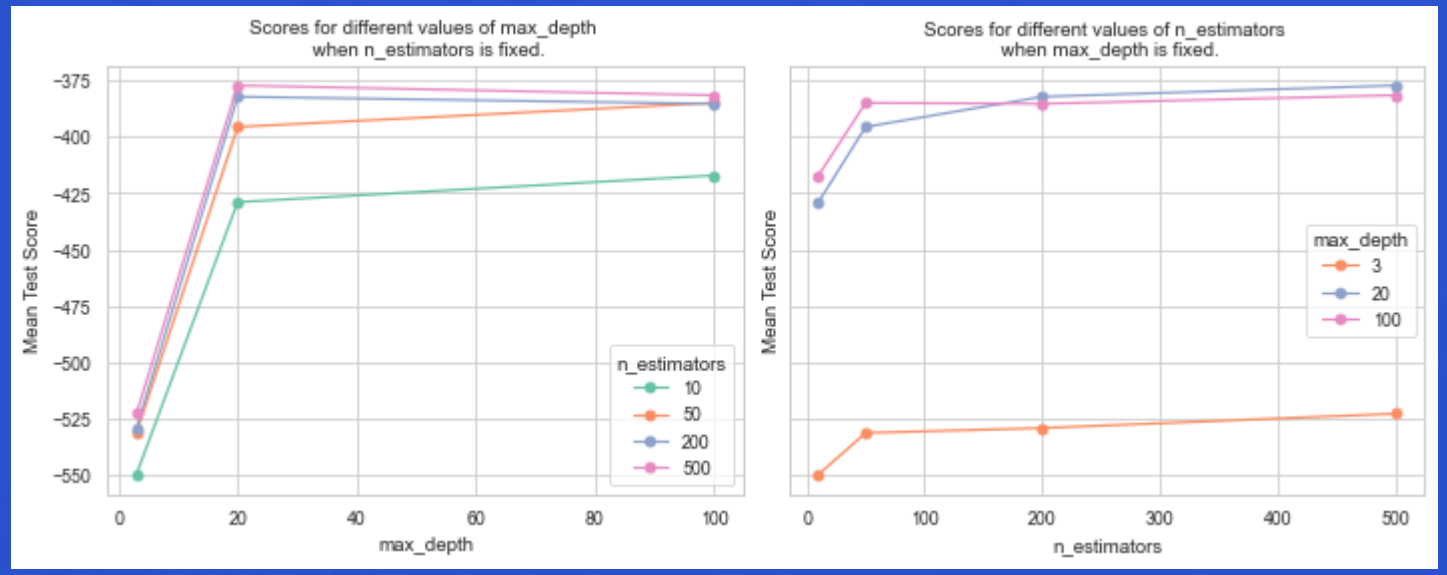
The Scikit-Learn learn library is used for training the models. We perform a grid search with cross-validation to find good hyper-parameters.

Linear Regression

```
Best parameters for Linear Regression: {}
Best score for Linear Regression: 815.7790095725477
RMSE for Linear Regression: 25.085468842114025
```

Linear Regression has no hyperparameters. We achieved a root-mean-square error of 25. In comparison, the mean model performs with an RMSE of 39.68.

Random Forest

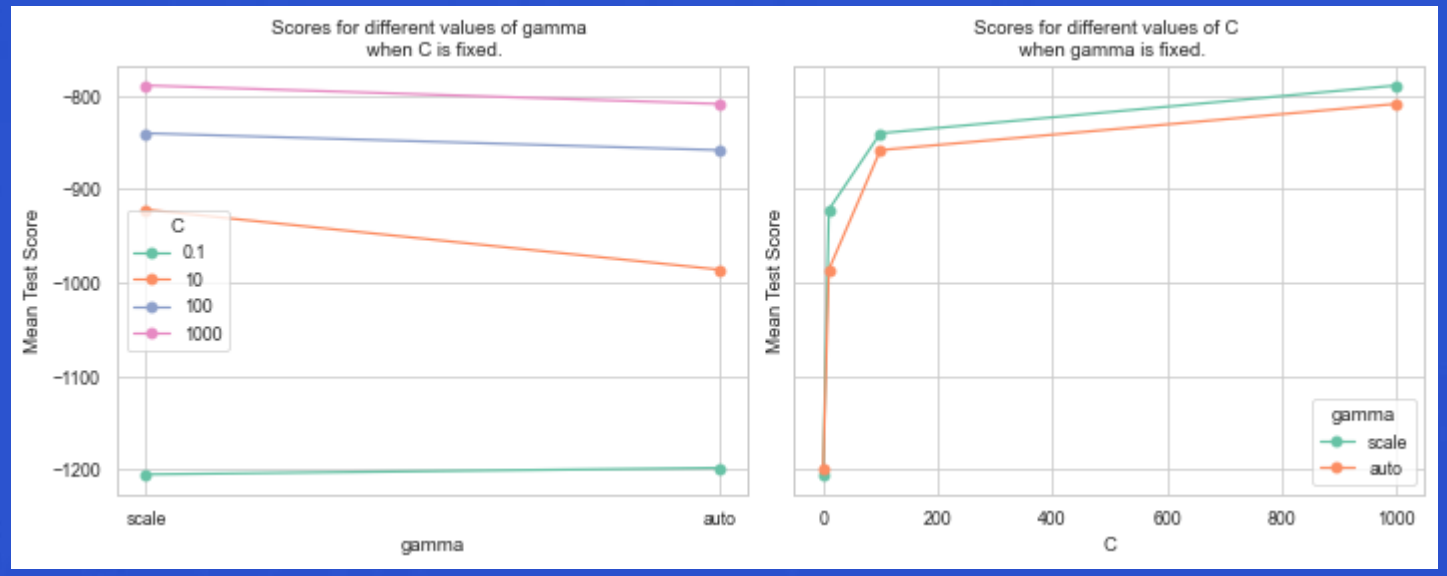


Best parameters for Random Forest: {'max_depth': None, 'n_estimators': 50}

Best score for Random Forest: 384.2732128195555

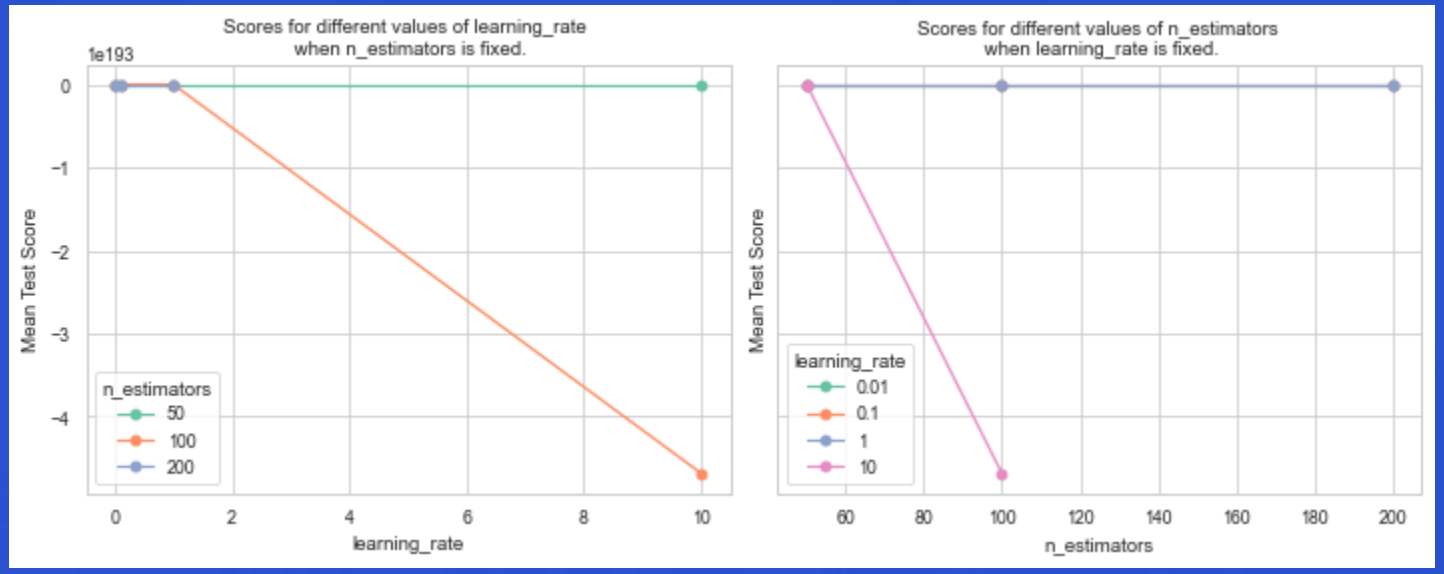
RMSE for Random Forest: 15.71030864703452 (full data set)

SVM




```
Best parameters for SVM: {'C': 1000, 'gamma': 'scale'}
Best score for SVM: 789.1821373099403
RMSE for SVM: 26.736784566136873
```

Gradient Boosting




```
Best parameters for Gradient Boosting: {'learning_rate': 0.1, 'n_estimators': 100}
Best score for Gradient Boosting: 389.7536562833442
RMSE for Gradient Boosting: 18.14342657501043
```

Results

```
RMSE for Linear Regression: 25.085468842114025
RMSE for Random Forest: 15.561348763097282
RMSE for SVM: 26.736784566136873
RMSE for Gradient Boosting: 16.5676917858317
```

After comparing the performance of all the models (trained on the full dataset), we found that the Random Forest model achieved the lowest RMSE of 15.86. This indicates that the Random Forest model provides the most accurate predictions compared to the other models tested.

However, we decided to use the Gradient Boosting model for the API since Random Forest was taking more than 100MB of storage for a very small difference of prediction precision.

Conclusion

In this project, our objective was to predict the number of comments on a Facebook post. We explored various regression models, including Linear Regression, Random Forest, SVM, and Gradient Boosting.

After evaluating the performance of these models, we found that the Random Forest model is well-suited for predicting the number of comments on Facebook posts.

Also, our analysis showed a peak in the mean number of comments on Thursdays, suggesting that publishing posts on this day may result in higher engagement.

Understanding the factors influencing user engagement, such as the day of the week, can be valuable for content creators and marketers looking to increase audience interaction.