Facebook Comment Prediction

Goal of the task

Predict the number of comment a Facebook post will receive based on 53 features on the page, the post and other related factors.

Since social media has a large influence on individuals and society, there is a massive demand to study dynamic behavior in these social networking services.

Dataset

The raw data is crawled, cleaned, preprocessed and 5 variants was generated. Each variant has the same post but with different time at random. The dataset contains 53 input attributes and one target value. Input attributes come from page features, essential features, weekday features and other basic features. It includes but not limited to

- Page likes
- Page category
- Comment received in last 24 hours
- Comment received between 24 and 48 hours
- Weekday or weekend
- Time of the day
- Length of the post

Approach

I use mse to measure accuracy. I experiment with a portion of data with linear regression, decision tree, random forest, boosting and neural network. After evaluation, I proceed to the whole dataset with decision tree, random forest and neural network. I got the most accurate prediction with random forest.

model	mse
decision tree	60.52922905258934
random forest	57.8040005639679
neural network	111.21034538248148

- Decision tree: max depth 12, random state 42
- Random forest: max depth 8 random state 0 n estimators 100
- Neural network: Multi-Layer Perceptron, used MLP regressor from scikit learn with early stopping

Analysis

Top 6 features

name	Type of feature	description
CC3 Min	Derived feature	These features are aggregated by page, by calculating min, max,

		average, median and standard
Post Promotion Status	Other feature	To reach more people with posts in News Feed, individual promote their post and this features tells that whether the post is promoted(1) or not(0).
CC3	Essential feature	Essential feature The number of comments in last 48 to last 24 hours relative to base date/time
Post Published Weekday 40	Weekdays feature	This represents the day(Sunday...Saturday) on which the post was published
Base DateTime Weekday 47	Weekdays feature	This represents the day(Sunday...Saturday) on selected base Date/Time.,
Base DateTime Weekday 52	Weekdays feature	This represents the day(Sunday...Saturday) on selected base Date/Time
CC4 Min	Derived feature	These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features
H Local	Other feature	This describes the H hrs, for which we have the target variable/ comments received
Post Published Weekday 45	Weekdays feature	This represents the day(Sunday...Saturday) on which the post was published
CC2 Min	Derived feature	These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features

Conclusion

Time of post, page feature and post promotion have the most impact on comment volume. Post itself have relatively lower influence on comment.

Continuing:

- Use Hits@10 to measure accuracy (introduced in the original published paper)
- Improve model
- Study effect of promotion on comment volume