Final Project: Facebook Comment Prediction

Introduction

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.

Objectives

The goal of this project is to predict the number of comment a Facebook post will receive based on 53 features on the page, the post and other related factors.

Dataset Introduction

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, derived 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

In particular, derived features are calculated by the aggregated properties of all the post belongs to one page.

Procedure

I performed preliminary analysis on the attributes, I found that some features have a positive linear relation with label value, such as amount of post received before selected base time. I selected 7 columns that have most significant weight from my preliminary analysis with decision trees and plot each attribute against each other in order to find correlation. I found that for each page, the average of comment received for all its post before selected base time, last 24h, last 24h to 48h, first 24h after publication but before base time have a positive correlation. It indicates that the speed for amount of posts received over time is constant.

In the data preprocessing stage, I delete one column because it has exactly the same values in all the training examples, thus it doesn't provide useful information.

I didn't make modifications to categorical attributes because of the following reasons:

- I didn't find any particular categorical attribute that have strong linear relation with label value
- Any modification such as encoding doesn't affect tree based model
- the number of categorical features in the dataset is huge, which can lead to high memory consumption. This situation can be solved by combining PCA, but it might lose data and same as the above reason, categorical attributes don't play a significant role in my preliminary analysis, so it's not worth performing one hot encoding and PCA.

I didn't perform any feature scaling on numerical attributes because all of them has been decimal encoded.

To measure the model accuracy, I use mean square error and grid search with 10 folds.

I used linear regression, decision tree, random forest and neural network on each variant and whole dataset. I adjust the parameter for each model manually. I got the most accurate prediction with random forest.

model	mse
Linear regression	69.00928537125812
decision tree	60.52922905258934
random forest	57.8040005639679
neural network	111.21034538248148

Details on the machine learning models:

- 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

Findings

Since I got the most accurate prediction with random forest, the top 10 most weighted features are examined. 6 of these features are time related, 2 of them are related to page and only 1 are related to post itself. Top 10 features are the following:

name	Type of feature	description
CC2	Essential feature	The number of comments in last 24 hours, relative to base date/time
Base Time	Other feature	Selected time in order to simulate the scenario
Post Share Count	Other feature	This features counts the no of shares of the post, that how many peoples had shared this post on to their timeline.
CC1 Avg	Derived feature	These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features(cc1 is The total number of comments before selected base date/time)
CC5	Weekdays feature	post published on weekday orweekend
CC1	Essential feature	The total number of comments before selected base date/time.
CC4 Avg	Derived feature	Page aggregated features(cc4 is The number of comments in the first 24 hours after the publication of post

		but before base date/time.,Ä†)
CC4	Essential feature	The number of comments in the first 24 hours after the publication of post but before base date/time
CC3 Avg	Derived feature	These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features(cc3 is The number of comments in last 48 to last 24 hours relative to base date/time)
CC5 Median	Derived feature	These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features(cc5 is The difference between CC2 and CC3)

The top 10 features match the result from preliminary analysis

Conclusion

Post features and page features both contributes to comment volume equally. Time of post published also has influence on comment volume. Post promotion has low influence on comment.

Appendices

Kamaljot Singh(2015), Comment Volume Prediction using Neural Networks and Decision Trees. *Department of Computer Science* DAV University, Jalandhar Punjab, INDIA Retrieved April , 2019

http://uksim.info/uksim2015/data/8713a015.pdf

Facebook Comment Volume Dataset Data Set, Kamaljot Singh, Assistant Professor, Lovely Professional University, Jalandhar. University of California Irene. machine learning repository
https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset