Python for Data Analysis

Facebook Comment Volume Prediction

Pierre-Antoine DELEU

DIA 2

Introduction

My datatset was the Facebook Comment Volume Dataset.

Facebook was created in 2006 by Mark Zuckerberg in his room in Harvard. It is nowadays the most popular social network on internet with more than 2 billions users since June 2017.

So it’s normal to see a massive demand to study dynamic behaviour on this social network.

And that’s where the dataset is useful.

The goal of the dataset is to predict the number of comment a Facebook post will receive in the next H hours (H being a feature of the dataset) based on features on the page, the post and other related factors.

The dataset contains 54 features:

Page features:

- The popularity/likes of the page
- The category of the page
- The number of people who visited the place
- The count of users that are engaged with the page

Post related features:

- The lenght of the post (character count)
- How many people shared this post
- If the post is promoted or not
- Selected time in order to simulate the scenario
- The number of hours after the selected time where the target is recorder

Weekday features:

- 7 features which defines on which day the post was published
- 7 features which defines on which day the post was used to make a prediction
- They are indicated with binary indicators (0,1)

The target feature:

- How many comment the post will have in H hours.

The derived features:

- These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features.

And the essential features

- C1: Total comment count before selected base date/time
- C2: Comment count in the last 24 hours with respect to selected base date/time
- C3: Comment count in the last 48 hours to last 24 hours with respect to base date/time
- C4: Comment count in first 24 hours after publishing the document, but before the selected base date/time
- C5: The difference between C2 and C3

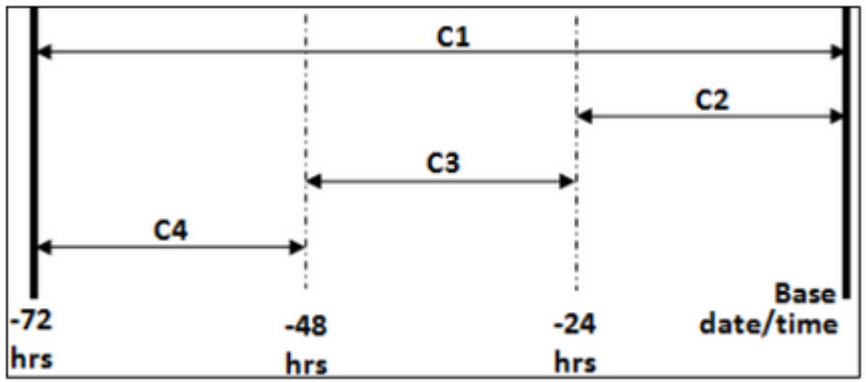

Figure 2. Demonstrating the essential feature details.

Figure from Kamaliot Singh's report, the creator of this dataset

Work with the initial dataset

There was no title for any of the columns in the dataset, so I modified myself the dataset I was using and add a row at the beginning with the name for all the features.

While doing that, I couldn’t figure out how I will be able to use the derived features because we could not know which one was exactly what and furthermore I thought there was enough variable to make a good work so I decided to not name them and after I erased them from the dataFrame in the beginning of the notebook.

I also removed the ‘promote’ feature since no one of all the post I work with has been promoted so it was just a column full of 0.

And I worked with all the others features, which mean 28 features.

We have in this dataset 40949 entries and here you can see all the features I decided to keep with their name.

We can also see there is no null in this dataset so it’s perfect.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40949 entries, 0 to 40948
Data columns (total 28 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Page_likes  40949 non-null  int64
 1   Checkins    40949 non-null  int64
 2   Returns     40949 non-null  int64
 3   Category    40949 non-null  int64
 4   C1          40949 non-null  int64
 5   C2          40949 non-null  int64
 6   C3          40949 non-null  int64
 7   C4          40949 non-null  int64
 8   C5          40949 non-null  int64
 9   baseTime    40949 non-null  int64
 10  length      40949 non-null  int64
 11  shares      40949 non-null  int64
 12  hrs         40949 non-null  int64
 13  sun_pub     40949 non-null  int64
 14  mon_pub     40949 non-null  int64
 15  tue_pub     40949 non-null  int64
 16  wed_pub     40949 non-null  int64
 17  thu_pub     40949 non-null  int64
 18  fri_pub     40949 non-null  int64
 19  sat_pub     40949 non-null  int64
 20  sun_base    40949 non-null  int64
 21  mon_base    40949 non-null  int64
 22  tue_base    40949 non-null  int64
 23  wed_base    40949 non-null  int64
 24  thu_base    40949 non-null  int64
 25  fri_base    40949 non-null  int64
 26  sat_base    40949 non-null  int64
 27  output      40949 non-null  int64
dtypes: int64(28)
memory usage: 8.7 MB
```

Data Visualization

I begin with normalizing the data set to be able to scatter a little since it was unligible without it, and after I came back to the original dataset.

I’ve done many different visualization to be sure to understand better the dataset and the importance of some of the features in this work.

Some of them are what day people comment the most the posts on facebook or which category received the most comments.

Models

I used 4 different model to try to estimate the test set.
I used linear regression, but the result were really bad. So I tried Random Forest and found good result. After I used the Decision Tree and found worse result than the random Forest but still good. And I finished by using my grid with 34 estimators and there I found also a great score.

So at the end, it's clearly the Random Forest which is the best model to choose.

score	model
0.295337	Linear Regression
0.600592	Random Forest
0.541255	Decision Tree
0.585967	Grid 34 estimators