Project 1: How popular are your social network posts?

Chen Nan

Due date: October 01, 2017

1 Introduction

The leading trends toward social networking services had drawn massive public attention from "one and half" decade. The merging up of computing with the physical things had enabled the conversion of everyday objects into information appliances. These services are revolutionizing day by day, and much more are on the way. As in Facebook, 500+ terabytes of new data ingested into the databases every day, 100+ petabytes of disk space in one of FBs largest Hadoop (HDFS) clusters and there are 2.5 billion content items shared per day (status updates + wall posts + photos + videos + comments). The Twitter went from 5,000 tweets per day in 2007 to 500,000,000 tweets per day in 2013. Flickr features 5.5 billion images as that of January 31, 2011 and around 3k- 5k images are adding up per minute.

In this project, we focused on the leading social networking service, in particularly "Facebook Pages", for automatic analysis of trends and patterns. In particular, we are interested in the comment volume prediction (CVP) that a document is expected to receive in next H hours. To make the context, here are the definitions of the key terms used.

- Page: It is a public profile specifically created for businesses, brands, celebrities etc. Each page can contain multiple posts.
- Post: These are basically the individual stories published on page by administrators of page.
- Comment: It is an important activity in social sites, that gives potential to become a discussion forum. The comment is for each Post.

We focused on fine grained predictive modeling techniques. Given some posts that appeared in past, whose target values (comments received) are already known, we simulated the scenario. For this, Facebook pages are crawled for raw data. The raw data are then pre-processed, and transformed into vectors for each post. For each post, 41 features and 1 target value have been identified. The features can be classified into following categories, with a complete list shown in the appendix.

1. *Page Features*: There are 4 features in this category. Page likes: defines users support for pages; Page Category: defines the category of source of documents: Local business or place, brand or product, company or institution, artist etc; (see Appendix for the complete list). Page Checkin's: an act of showing presence at particular place and under the category of place, institution pages only. Active Users: This is the actual count of users that were 'engaged' and interacting with that Facebook Page. The users who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares by visitors to the page.

2. *Essential Features*: This includes the pattern of comment on the post in various time intervals w.r.t to the randomly selected base date/time, named as C1 to C5. C1: Total comment count before selected base date/time. C2: Comment count in last 24 hrs with respect to selected base date/time. C3: Comment count in last 48 hrs to last 24 hrs with respect to base date/time. C4: Comment count in first 24 hrs after publishing the document, but before the selected base date/time. C5: The difference between C2 and C3. Furthermore, we aggregated these features by source and developed some derived features by calculating min, max, average, median and standard deviation of 5 above mentioned features within a page. So, adding up the 5 essential features and 25 derived essential features, we got 30 features in this category.

3. *Weekday Features*: are used to represent the day on which the post was published and the day of selected base date/time.

4. *Other Basic Features*: 5 features of this category are identified. They include length of document, time gap between selected base date/time and document published date/time ranging from (0,71), document promotion status (with values 0 or 1), post share count, and the number of hours H, for which the comments amount is received.

2 Dataset

The project has two datasets. The training dataset is for model estimation and model selection. It includes 40949 samples. Each sample has one output value (i.e., the number of comments received within next H hours since the base time), and 41 features defined in the previous section. It is noted that H could be different for each sample, and its value is given as one of the features.

The testing dataset has the same format, except that the output values are not provided. You are expected to predict the output values for each sample in the test dataset. The accuracies of your prediction will be evaluated based on the numbers you provided.

3 Project Assignment

Your task is to build a regression model to predict the number of comments within certain period from baseline.

3.1 Step 1: Simple Regression Model

In this part, you are required to develop a simple model that can be used for predicting the comment volume. To reduce the difficulty, you are allowed only limited manipulations of the original data set. You are allowed to take power transformations of the original variables (square roots, logs, inverses, squares, etc), but you are *NOT* allowed to create interaction variables. Your model should include *NO* more than 6 predictors/covariates, but should explain as much variability as possible.

After obtaining the model with aforementioned features, you are required to analyze the model and provide meaningful interpretations. Please focus your attention on the interpretation of the model. A strong analysis should include the interpretation of various coefficients, statistics, and plots associated with their model and the verification of any necessary assumptions.

3.2 Step 2: Complex Regression Model

In this part, you are free to construct the "best" regression model for predicting comment volumes. You are encouraged to experiment with any of the methods that were discussed during the semester for finding a suitable model. You are allowed to create any new variables you desire (such as quadratic, interaction, or indicator variables). Your model needs to be estimated based on the training data, and provides prediction on the testing data. Forecast errors will be evaluated as a component of your project score.

Note: You are allowed to construct multiple regression models to make the forecasting. Only the final forecasting results should be submitted for evaluation.

3.3 Step 3: (Optional) Free Form Model

You can choose any arbitrary model, including but not limited to regression models, for prediction purpose. If you choose to do this part, you need to summarize the method you choose, report the results, and compare the results with regression models in your report. The forecasting accuracy from this model will be evaluated. If your accuracy is better than that of the best regression model in the class, you will be awarded *2 bonus points*.

Attention: Make sure your results are replicable by the codes you submitted. Unreproducible results are considered cheating/plagiarism.

Appendix

Description of the variables

The data has 42 columns which include 39 continuous variables, and 3 attribute variables.

0 Target Variable Decimal Target
The no of comments in next H hrs (H is given in Feature no 39).

1 Page Popularity/likes Decimal Encoding Page feature
Defines the popularity or support for the source of the document.

2 Page Checkins Decimal Encoding Page feature
Describes how many individuals so far visited this place. This feature is only associated with the places eg:some institution, place, theater etc.

3 Page talking about Decimal Encoding Page feature
Defines the daily interest of individuals towards source of the document/ Post. The people who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares, etc by visitors to the page.

4 Page Category Value Encoding Page feature
Defines the category of the source of the document eg: place, institution, brand etc.

5 - 29 Derived Decimal Encoding Derived (page) feature
These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features.

30 C1 Decimal Encoding Essential feature
The total number of comments before selected base date/time.

31 C2 Decimal Encoding Essential feature
The number of comments in last 24 hours, relative to base date/time.

32 C3 Decimal Encoding Essential feature
The number of comments in last 48 to last 24 hours relative to base date/time.

33 C4 Decimal Encoding Essential feature

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

34 C5 Decimal Encoding Essential feature
The difference between CC2 and CC3.

35 Base time Decimal(0-71) Encoding Other feature
The time gap (in hours) between publish time and base time

36 Post length Decimal Encoding Other feature
Character count in the post.

37 Post Share Count Decimal Encoding Other feature
This features counts the no of shares of the post, that how many peoples had shared this post on to their timeline.

38 Post Promotion Status Binary Encoding 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).

39 H Decimal(0-23) Encoding Other feature
This describes the H hrs, for which we have the target variable/ comments received.

40 Post published weekday String Weekdays feature
This represents the day (Sunday...Saturday) on which the post was published.

41 Base DateTime weekday String Weekdays feature
This represents the day (Sunday...Saturday) on selected base Date/Time.

List of Page Categories

Seq	Category	Seq	Category	Seq	Category
1	Product, service	41	Automobiles and parts	81	Producer
2	Public figure	42	Tv channel	82	Landmark
3	Retail and consumer merchandise	43	Telecommunication	83	Cause
4	Athlete	44	Entertainment website	84	Organization
5	Education website	45	Shopping, retail	85	Tv, movie award
6	Arts, entertainment, nightlife	46	Personal blog	86	Hotel

7	Aerospace, defense	47	App page	87	Health, medical, pharmaceuticals
8	Actor, director	48	Vitamins, supplements	88	Transportation
9	Professional sports team	49	Professional services	89	Local, travel website
10	Travel, leisure	50	Movie theater	90	Musical instrument
11	Arts, humanities website	51	Software	91	Radio station
12	Food, beverages	52	Magazine	92	Other
13	Record label	53	Electronics	93	Computers
14	Movie	54	School	94	Phone, tablet
15	Song	55	Just for fun	95	Coach
16	Community	56	Club	96	Tools, equipment
17	Company	57	Comedian	97	Internet, software
18	Artist	58	Sports venue	98	Bank, financial institution
19	Non-governmental organization (ngo)	59	Sports, recreation, activities	99	Society, culture website
20	Media, news, publishing	60	Publisher	100	Small business
21	Cars	61	Tv network	101	News personality
22	Clothing	62	Health, medical, pharmacy	102	Teens, kids website
23	Local business	63	Studio	103	Government official
24	Musician, band	64	Home decor	104	Photographer
25	Politician	65	Jewelry, watches	105	Spas, beauty, personal care
26	News, media website	66	Writer	106	Video game
27	Education	67	Health, beauty		
28	Author	68	Music video		
29	Sports event	69	Appliances		
30	Restaurant, cafe	70	Computers, technology		
31	School sports team	71	Insurance company		
32	University	72	Music award		
33	Tv show	73	Recreation, sports website		
34	Website	74	Reference website		
35	Outdoor gear, sporting goods	75	Business, economy website		
36	Political party	76	Bar		
37	Sports league	77	Album		
38	Entertainer	78	Games, toys		
39	Church, religious organization	79	Camera, photo		
40	Non-profit organization	80	Book		