Project Proposal

Facebook Advertisement Analysis

INFO 7390

Advance Data Science & Architecture

Professor: Srikanth Krishnamurthy

By: Team 5

Akash Jagtap

Jerin Rajan

Nitin Prince Reuben

Contents

Abstract

Data in the social networking services is increasing day by day. So, there is heavy requirement to study the highly dynamic behavior of the users towards these services. This work is a preliminary work to study and model the user activity patterns. We had targeted the most active social networking service ‘Facebook’ importantly the ‘Facebook Pages’ for analysis. The task here is to estimate the comment count that a post is expected to receive in next few hours. The analysis is done by modeling the comment patterns using variety of regressive modeling techniques.

This study presents a research approach using datamining for predicting the performance metrics of posts published by brands in Facebook pages. Twelve posts extracted from a cosmetic company's page including 790 publications were modeled. One of them, the “Lifetime Post Consumers” model, was assessed using sensitivity analysis to understand how each of the seven input features (category, page total likes, type, month, hour, weekday, paid) influenced it.

We would create a web page for hosting our application using Amazon S3 (ec2 instance) through which we will make provisions for our clients to provide us data. Also, we can show them the output in the UI itself. In this process, we will be automate the models for our dataset making things easier and faster.

Introduction

The increasing use of social networking services had drawn the public attention explosively from last 15 years. The merging up of physical things with the social networking services had enabled the conversion of routine objects into information appliances. These services are acting like a multi-tool with daily applications like: advertisement, news, communication, banking, commenting, marketing etc. These services are revolutionizing day by day and many more on the way. These all services have one thing in common that is daily huge content generation, that is more likely to be stored on Hadoop cluster. As in Facebook, 500+ terabytes of new data ingested into the databases every day, 100+ petabytes of disk space in one of FB's 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 research, we targeted the most active social networking service 'Facebook' importantly the 'Facebook Pages' for analysis. Our research is oriented towards the estimation of comment volume that a post is expected to receive in next few hours. Before continuing to the problem of comment volume prediction, some domain specific concepts are discussed below:

- ***Public Group/Facebook Page:*** It is a public profile specifically created for businesses, brands, celebrities etc.
- ***Post/Feed:*** 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 and it is only one measure of popularity/interest towards post is to which extent readers are inspired to leave comments on document/post.

The worldwide dissemination of social media was triggered by the exponential growth of Internet users, leading to a completely unfamiliar environment for customers to exchange ideas and feedback about products and services. According to Statista Dossier (2014), the number of social network users will increase from 0.97 billion to 2.44 billion users in 2018, predicting an increase around 300% in 8 years. Considering

its rapid development, social media may become the most important media channel for brands to reach their clients in the near.

Companies have realized the potential of using Internet-based social networks to influence customers, incorporating social media marketing communication in their strategies for leveraging their businesses. Several studies focused on finding the relationships between online publications on social networks and the impact of such publications measured by users' interactions. However, fewer studies devoted attention to research for implementing predictive systems that can effectively be used to predict the evolution of a post prior to its publication. A system able to predict the impact of individual published posts can provide a valuable advantage when deciding to communicate through social media, tailoring the promotion of products and services. Advertising managers could make judged decisions on the receptiveness of the posts published, thus aligning strategies toward optimizing the impact of posts, benefiting from the predictions made. Therefore, the predictive tool outlined in this paper could leverage managerial decisions to improve brand recognition. We focused on predicting the impact of publishing individual posts on a social media network company's page. The impact is measured through several available metrics related to customer visualizations and interactions. The predictive knowledge found enables to support manager's decisions on whether to publish each post.

Goals

Come up with reliable model which will help companies to invest their time, resources and money wisely on advertisements on social media especially on Facebook.

Dataset

Dataset which we are using in our project can be found in the link mentioned below.

https://archive.ics.uci.edu/ml/datasets/Facebook+metrics#

We must understand how data is present in our dataset.

```
df.shape
```

```
(500, 19)
```

Our Dataset contains 19 columns and 500 rows. We must understand how the dataset is distributed.

Feature / Column Name	Type of Information	Data type
Posted	Identification	Date/time
Permanent link	Identification	Text
Post ID		
Post message	Content	Text
Type	Categorization	Factor: {Link, Photo, Status, Video}
Category	Categorization	Factor: {action, product, inspiration}
Paid	Categorization	Factor: {yes, no}
Page total likes	Performance	Numeric

Lifetime post total reach		
Lifetime post total impressions		
Lifetime engaged users		
Lifetime post consumers		
Lifetime post consumptions		
Lifetime post impressions by people who have liked your page		
Lifetime post reach by people who like your page		
Lifetime people who have liked your page and engaged with your post		
Comments	Performance	Numeric
Likes	Performance	
Shares	Performance	
Total interactions	Performance	Numeric

Process Outline

1. Data Preprocessing
    - Data Cleaning, handling missing values
2. Exploratory Data Analysis
3. Performing Feature Engineering.
4. Analyzing prediction algorithm.
5. Performing feature selection.
6. Design a pipeline and system to implement this approach and discussion on the system's capabilities
7. Deploy the Model on AWS or Google Cloud Computing Platform
8. Build a web application to demonstrate the prediction and recommendation results.

Deployment Details

1) Language: Python
2) Pipeline: Airflow
3) Container: Docker
4) Cloud Tools/Platforms: AWS (Amazon Web Services) EC2
5) Other Considerations: Google Cloud Platform