MS&E 226 - Small Data

Mini Project Report - Part I

SUNet ID: [dhruvj, tromero1]
Name: [Dhruv Joshi, Tyler Romero]

1 Introduction

1.1 About the Dataset

The dataset selected is from the publicly available Kaggle Dataset for the "Outbrain Click Prediction" competition, launched on October 5^{th}, 2016. It contains a sample of users page views and clicks, as observed on multiple publisher sites in the United States between 14-June-2016 and 28-June-2016. At a high level, the data tells us about users (represented uniquely by **uuid**) who visited a content-rich page (represented uniquely by **document_id**), and whether or not they clicked one (key **ad_id**) of a set of ads in a group (key **display_id**), present on the page. The objective of the competition is to predict **which pageview instance would lead to a clickthrough on an advertisement**. In addition, we will be predicting **time on page in the event that a user clicks on an add.** We are provided rich metadata about each of these members -

a About the user:

i timestamp (milliseconds since 1970-01-01 - 1465876799998 - "Epoch time")

ii platform (desktop = 1, mobile = 2, tablet =3)

iii geo_location (country $>$ state $>$ DMA)

iv traffic_source (internal = 1, search = 2, social = 3)

b About the host Webpage

i source_id (the part of the site on which the document is displayed, e.g. edition.cnn.com)

ii publisher_id (anonymized numeric identifier)

iii publish_time (time from epoch when the content was published)

iv topic_id (an integer indicating the topic the document content is grouped into).

c About the Advertisement: We have campaign_id and advertiser_id giving us information about the advertiser and their specific campaign.

1.2 Preliminary Analysis and Concerns

There are no NULL/NA values. We suspect that all or a part of the dataset will be useful to build the covariate matrix - and there would be several interaction terms which revolve around interest in the document content as well (such as the interaction between a unique user OR country and the particular document/webpage/publisher). Some covariates like 'publish_time' may not be required for this analysis and would be discarded. The details about the user and the document are given in one table (*page_views.csv*) and the information about the events that take place, i.e. a click on an ad in another table (*events.csv*). We would need to match a particular instance in both these by looking at specific parameters defining an instance (i.e. the uuid, the geo_location and the platform), which are common to both tables. The actual dataset provided in this competition has information on 2 Billion pageviews, which we deemed far beyond the scope of this course project. Hence, we would be analyzing a smaller subset of it (1 million pageviews). One of our primary concerns is how the unique user ID was fixed and whether it was consistent across platforms/devices used to access these websites. However, since we will not be 'personalizing' predictions per user, this information is should not affect our analysis.

Since the number of ads on the display in a document is not fixed, this may affect the user actually clicking on the ad and would require further investigation. Also, we are given very little metadata and contextual information about the host page (besides *topic_id* and *document_category*) and the advertisements themselves, which means we would need to use other clustering strategies to find correlations.

2 Understanding the Users

The users were overwhelmingly from the US (followed by Canada, Australia and the UK - see Fig 1), and mostly used Desktop and mobile devices, with tablets used least of all. An overwhelming majority of users have very few pageviews (i.e. they show up very few times in the dataset) - 99.9% users show up less than 10 times in our training set (See Fig 3).

3 Understanding the Host webpages

The document topics have been categorized by an automated system of outbrain or another provider (not mentioned), we receive the following metadata about a page:

document_id	topic_id	confidence_level

The confidence_level is the confidence with which the document content falls into that category. We would need to set a threshold for the confidence levels if they are to be considered in the models we build. The mean value of confidence is 4.888786% and only 0.013% documents have a confidence level greater than 70% of the topic (See Fig 5). Hence this metadata may not be very useful to us.

4 Understanding the Events

There are 1.037% events/sessions which actually led to an advertisement click (This is normal as per industry standards). There are more than 60% pages which had less than 10% views. It might be helpful (while training our prediction model) to only look at the events which had a out-click traffic higher than a threshold (since we can find useful correlations in this). In Fig 2, we can see that there is a positive correlation between number of clicks and number of views (as expected) but a large majority of pages had extremely (1 or less) few clicks. In this figure it is worth noting that each time a page was loaded, the groups of ads would be different, hence our model should favour pages (i.e. documents) which had a higher click-through rate. This would be done through a weighted model.

5 Continuous Response Variable

We will use the ”time on page” - this is created from the difference between the timestamp for when the user entered the document and when they left the page (i.e. clicked on an ad). This is a crucial factor in deciding the relevance and quality of an advertisement and predicting the confidence of the click-through by the user on the ad. Is should be noted that quantity only exists for those sessions which resulted in a click. See Fig 4.

6 Binary Response Variable

This would be an indicator of whether the session resulted in the user clicking an ad or not. This is a 'discrete' transformation of the continuous response variable.

Appendix

Fig 1: Country breakdown with platform

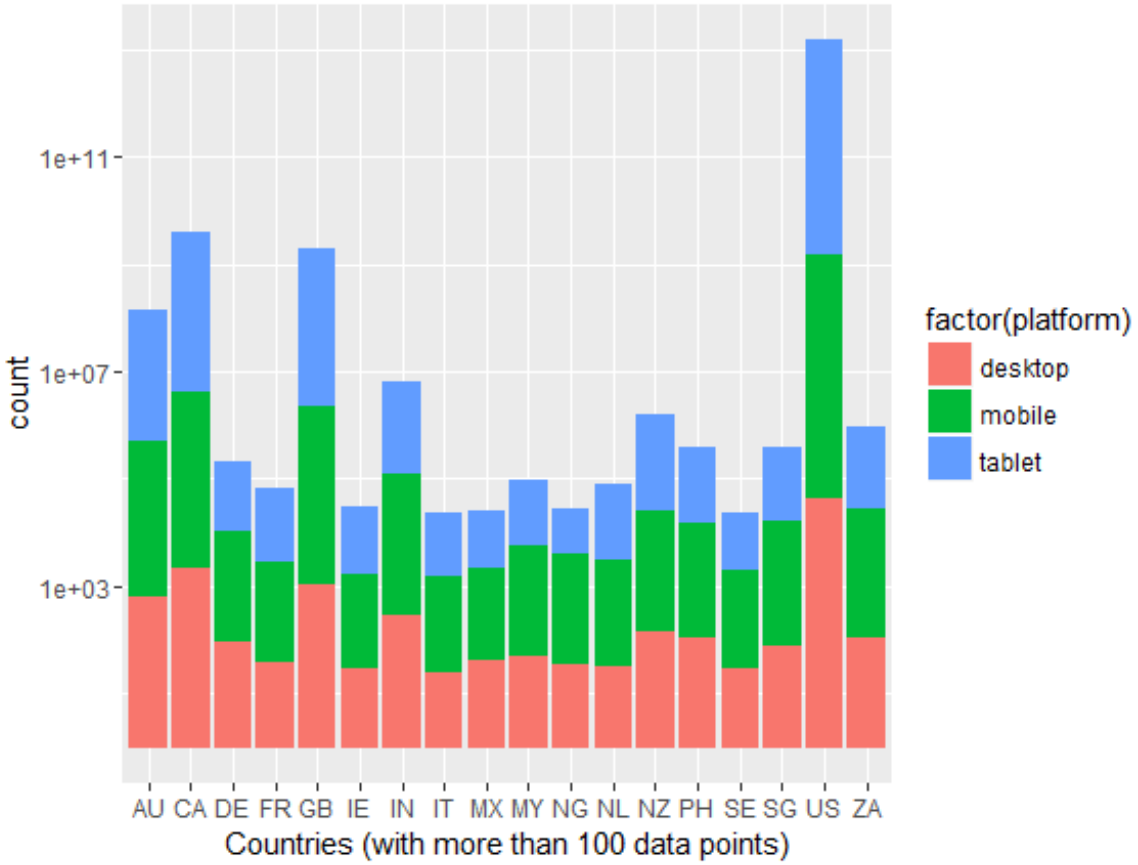


Fig 2: No of clicks vs No of views (logarithmic axes)

Fig 3: Number of times each user appears in the dataset

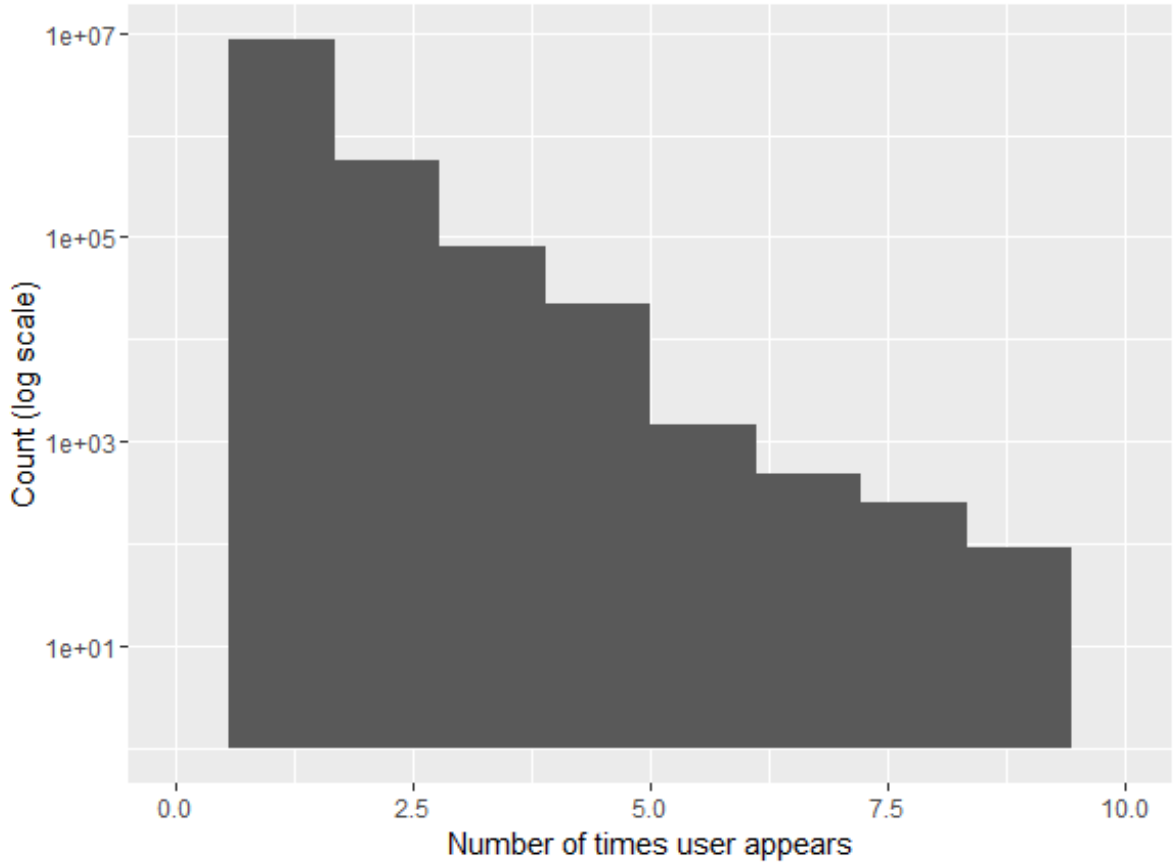


Fig 4: Continuous Response Variable - time on page

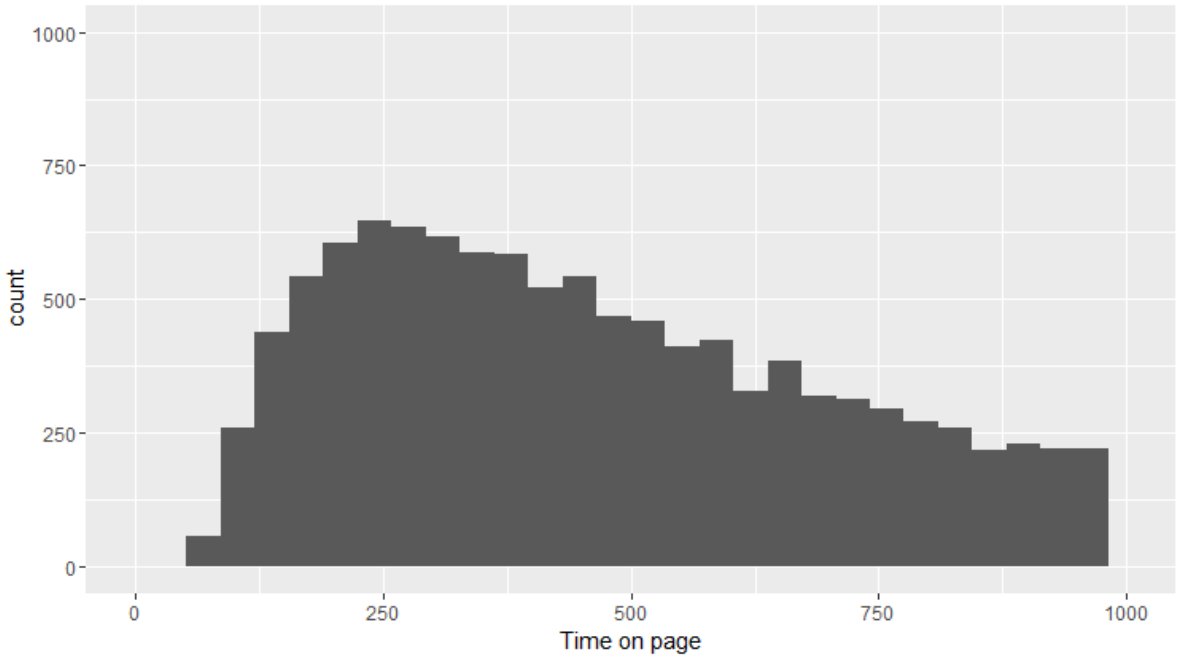
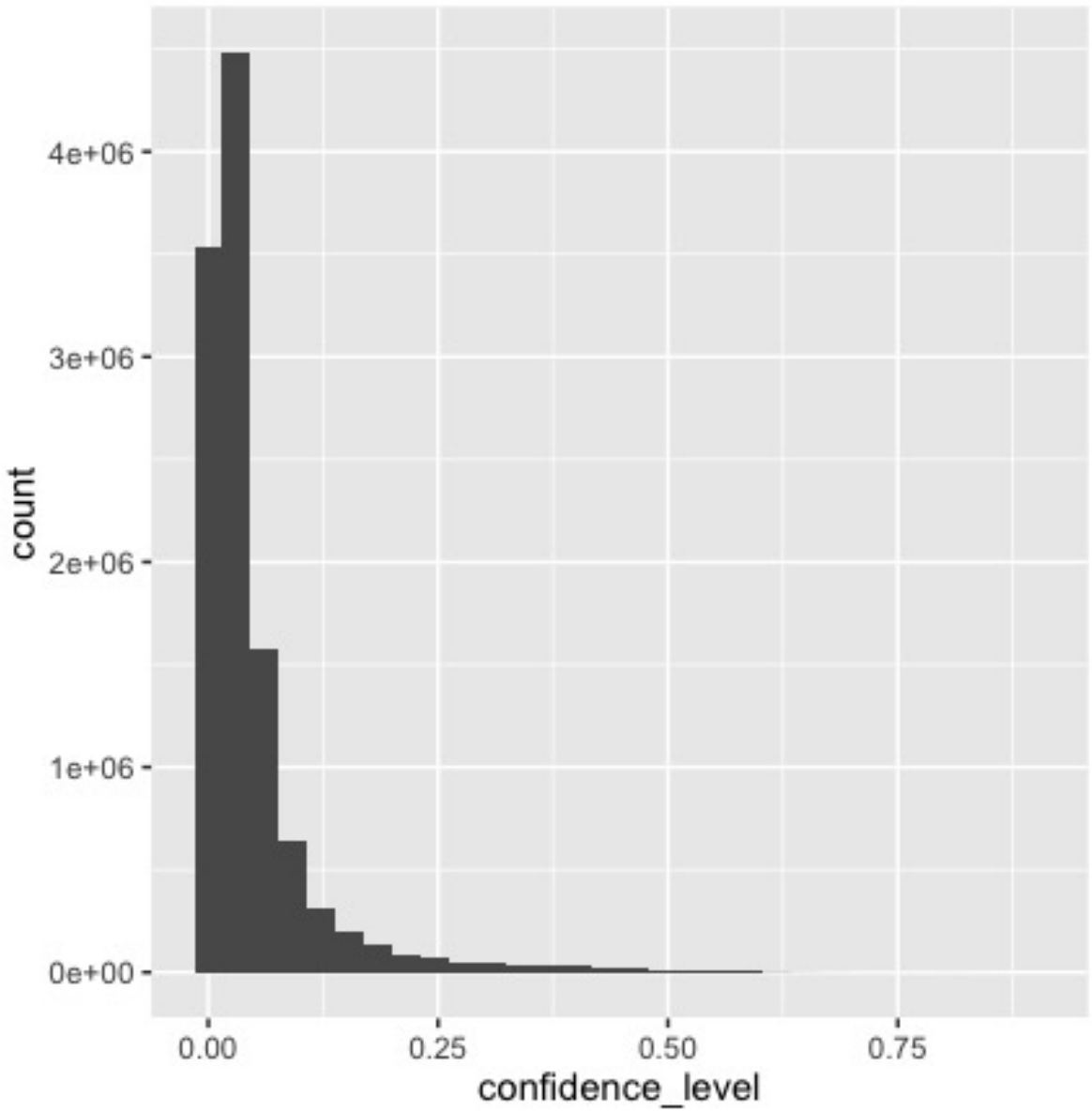


Fig 5: Confidence levels in document_topic



References and Acknowledgments

1. Outbrain competition Exploratory Data Analysis by user *anokas*, https://www.kaggle.com/anokas/outbrain-click-prediction/outbrain-eda, Accessed October 27th, 2016