Google Analytics Customer Revenue Prediction

German E. Novakovskiy, Vibudh Agrawal
Bioinformatics graduate program
University of British Columbia
Vancouver, BC Canada V6T 1Z4
gnovakovsky@cmmt.ubc.ca, vagrawal@prostatecentre.com

Abstract

Using dataset provided by Google and Kaggle we tried to predict customers, who bought some items in Google Store (GStore). For this purporse we used supervised learning algorithms such as Random Forest, Gradient Boosting and Logistic Regression for predicting the natural log of customer's total revenue. Data is higly inbalanced (there are only ~1.5% who actually bought something).

1 Introduction

The problem with current market is that it follows well-known 80/20 rule true: 80% of revenue comes from 20% of customers. Thus, If we can predict these 20% customers it can help to change company's marketing strategy, focusing the investment in marketing promotions more towards these customers.
The customer data for this problem comes from Google's GStore and we are trying to predict ***revenue from each customer.*** The data set is collected from Kaggle challenge, which has the same name is our project (reference).
The potential outcome and contribution of this project will be more actionable operational changes and a better use of marketing budgets for different companies.

2 Data exploration

For training we have 903,653 examples and for testing we have 800,000 examples.

The customer data has the following features associated with it:

1. ***fullVisitorId*** - A unique identifier for each user of the Google Merchandise Store.
2. ***channelGrouping*** - The channel via which the user came to the Store.
3. ***date*** - The date on which the user visited the Store.
4. ***device*** - The specifications for the device used to access the Store.
5. ***geoNetwork*** - This section contains information about the geography of the user.
6. ***sessionId*** - A unique identifier for this visit to the store.
7. ***socialEngagementType*** - Engagement type, either "Socially Engaged" or "Not Socially Engaged".
8. ***totals*** - This section contains aggregate values across the session.
9. ***trafficSource*** - This section contains information about the Traffic Source from which the session originated.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

10. ***visitId*** - An identifier for this session. This is part of the value usually stored as the utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
11. ***visitNumber*** - The session number for this user. If this is the first session, then this is set to 1.
12. ***visitStartTime*** - The timestamp (expressed as POSIX time).

The problem here is that several columns are acutally in Json format, and we need to parse them in order to retreive relevant features. If we do this we end up with 59 columns. One of them, transactionRevenue, part of ***totals***, log of which we want to predict.

First step in each data mining task is data exploration. First of all we should remove features that have the same values across all samples since we can't learn anything from them: there are 21 such columns (for example longitude from geoNetwork Json dictionary). So it the end we have 38 features.

If we simply explore the main column, with value we want to predict, transactionRevenue, we can see that there are 11,515 samples with non-zero value, and 892,138 with zero. So overall we have only 1.3% of people who purchased something on the website. It makes this problem highly imbalanced.

The distribution of log revenue values for users with non-zero revenues is depicted on figures 1 and 2:

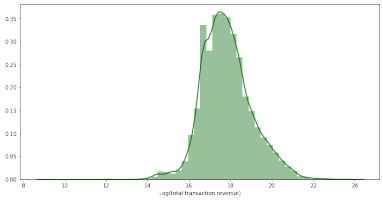

Figure 1: Logarithmic distribution of total transaction revenue (users with non-zero revenue only).

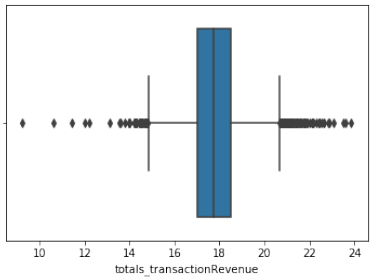

Figure 2: Box plot of logarithmic distribution of total transaction revenue (users with non-zero revenue only).

During data exploration we identified that certain features might be very informative for prediction. For example, according to the distribution plot of users with non-zero revenue (Figure 3) we see that more than 95% of them are from USA (country feature from geoNetwork).

But we need to emphasize that among all users there are ~350,000 from US, which is 38% of all people (Figure 4). Also, we noticed that in the data set we have a lot of irrelevant features. For example, continet feature form geoNetwork has America entry for ~400,000 users, but we already see that majority of them are from US, so this is quite redundant. Moreover, features such as city or metro have around 500,000 users with value Not set.

Almost all of our features are categorical. In order to deal with them in classification and regression problems we decided to convert them to binary features using one-hot encoding. For majority of categorical features there were more than 2000 unique values, but distribution of counts was very skewed: only several terms were present in the data with reasonable counts. Good example is country feature (Figure 4). We used only the most abundant terms for one-hot encoding: by setting threshold

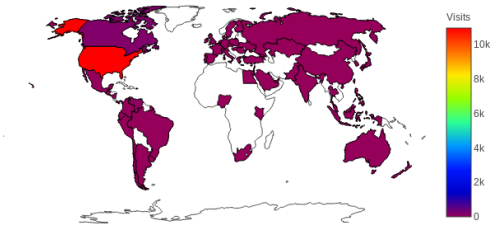

Figure 3: Number of users with non-zero transactions across countries.

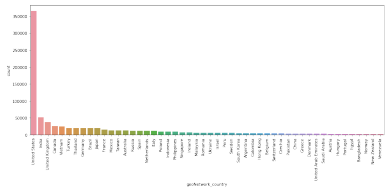

Figure 4: Histogram that shows distribution of number of users across countries.

we used only those categorical values that were present in at least 0.5% of users. Doing this we removed all potentially irrelevant features and reduced the feature set in order to avoid overfitting.

At the end we had train data set with 242 features (instead of 59 as was initially) including variable of interest, transactionRevenue.

3 Results

We tried to solve this problem using different approaches. First of all, it is a regression problem, we chose different regression models to directly predict log of transactionRevenue for each user (just predict revenue and THEN CONVERT TO LOG???). Second, we have highly inbalanced problem, only slightly more than 1% of users have non-zero revenue. This can be a huge problem for regression. Thus, we decided to divide the problem in two parts: classification (where we predict users with zero and non-zero transactions) and regression (where we run regression only on those users that were predicted to have non-zero transactions).

3.1 Regression

3.1.1 Random Forest

3.1.2 Gradient Boosting

3.2 Two-step model

3.2.1 Classification

Random Forest

Gradient Boosting

Logistic Regression

3.2.2 Regression

Random Forest

Gradient Boosting

4 Discussion

5 Conclusion

6 Methods

7 Reference

7.1 Style

Papers to be submitted to NeurIPS 2018 must be prepared according to the instructions presented here. Papers may only be up to eight pages long, including figures. Additional pages *containing only acknowledgments and/or cited references* are allowed. Papers that exceed eight pages of content (ignoring references) will not be reviewed, or in any other way considered for presentation at the conference.

The margins in 2018 are the same as since 2007, which allow for $\sim 15\%$ more words in the paper compared to earlier years.

Authors are required to use the NeurIPS LaTeX style files obtainable at the NeurIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

7.2 Retrieval of style files

The style files for NeurIPS and other conference information are available on the World Wide Web at

`http://www.neurips.cc/`

The file `neurips_2018.pdf` contains these instructions and illustrates the various formatting requirements your NeurIPS paper must satisfy.

The only supported style file for NeurIPS 2018 is `neurips_2018.sty`, rewritten for LaTeX 2_ε. **Previous style files for LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

The LaTeX style file contains three optional arguments: `final`, which creates a camera-ready copy, `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

New preprint option for 2018 If you wish to post a preprint of your work online, e.g., on arXiv, using the NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text "Preprint. Work in progress." in the footer. This version may be distributed as you see fit. Please **do not** use the `final` option, which should **only** be used for papers accepted to NeurIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please do *not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `neurips_2018.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections 8, 9, and 10 below.

8 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 10 regarding figures, tables, acknowledgments, and references.

9 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

9.1 Headings: second level

Second-level headings should be in 10-point type.

9.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

10 Citations, figures, tables, references

These instructions apply to everyone.

10.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated…

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2018` package:

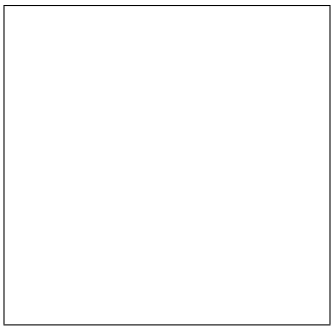

Figure 5: Sample figure caption.

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2018}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

10.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[2]

10.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

10.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

[1] Sample of the first footnote.

[2] As in this example.

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

11 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

12 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NeurIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

  ```
  \newcommand{\RR}{I\!\!R} %real numbers
  \newcommand{\Nat}{I\!\!N} %natural numbers
  \newcommand{\CC}{I\!\!\!\!C} %complex numbers
  ```

  Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

12.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Remember that you can use more than eight pages as long as the additional pages contain *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.