**Work Sample – Campaign Targeting Audiences**

Shiqi Bao

**Descriptive Analyses (Part A)**

1. What percent of customers returned after the first visit?

**11.82%** of customers returned after the first visit. This is computed by number of customers who paid a second visit divided by number of customers.

2. What are the top three best-performing stories in each section, by pageviews?

I use headline to define story. The top three best-performing stories in each section are listed here in Table 1.

Graphical user interface, text, application

Description automatically generated

Table 1: Best Three Performing Stories in Each Section

3. Based on this data, would you choose to promote a Tech story or a Markets story on social media? Why?

With a closer look at this data, in Table 2, Markets story performs better than Tech story in terms of pageviews of the top three best performing stories.

Text

Description automatically generated

Table 2: Best Three Performing Stories in Tech and Markets

As shown in Table 3, in terms of the overall pageviews and return rate, which is calculated by dividing the number of second visits by the number of first visits, the Markets section performs better than the Tech section. The two-sample t-test of the null hypothesis of equal return rates of Markets and Tech has a p-value of 0.3714, which is greater than 0.05. This shows that there is no statistically significant difference between these return rates. We need more data to accurately measure the return rates of different sections. However, based on the realized pageviews and return rates, I will **choose to promote a Markets** story to social media, as it is slightly better than Tech.

Table

Description automatically generated

Table 3: Visits and Return Rate in Each Section

4. Create a visualization exploring the relationship between any of the content characteristics (such as section, author, keywords, etc...) and returning visitors. This is an open-ended task. Briefly describe the visualization and the insight.

To study the relationship between **section and returning visitors**, I find that Politics has the highest return rate at 13.53%, followed by Business at 12.13%, and Markets at 11.89%. Life has the lowest return rate at 9.26%. The lines in Figure 1 show the range of one standard error above and below the estimates. Again, the exact ranking of return rates is subject to large estimation errors.

Chart, bar chart

Description automatically generated

Figure 1: Return rate in Each Section

To study the relationship between **author and returning visitors**, I compute the return rates of frequently read authors (pageviews above 200). Figure 2 shows the frequently read authors weighted by return rate. The font size is proportional to return rates. *Khadeeja Safdar* ranked in the top with 565 pageviews and 14.15% return rate. *Michael C. Bender* ranked in the second place with 1717 pageviews and 13.8% return rate.

Text

Description automatically generated

Figure 2: Word Cloud for Popular Authors by Return Rate

To study the relationship between **keywords and returning visitors**, I compute the return rates of frequently read keywords (pageview above 100). Figure 3 shows the frequently viewed keywords weighted by return rate. The font size is proportional to return rates. *Leder* viewed 266 times with a return rate of 16.16% ranks at the top. And *lawyers*, *interview*, *obstruction of justice* and *2016 presidential election* followed, with return rates around 14%.

Text

Description automatically generated

Figure 3: Word Cloud for Popular Keywords by Return Rate

5. What other interesting stories can you tell with this data?

Frequent readers generate more revenues than others. It is worth exploring what frequent readers’ first read is about. I use total visits as a proxy for how frequent customers view and divide customers into different blocks:

* Block 1: totalVisits between 2-5, which contains 1104 observations;
* Block 2: totalVisits between 6-9, which contains 112 observations;
* Block 3: totalVisits between 10-20, which contains 54 observations;
* Block 4: totalVisits above 20, which contains 23 observations.

Then I examine the distribution of first-reads’ sections within each customer block. The result is shown in Figure 4. The length of stacked bars represents the probability of first-reads’ sections. During their first visits, frequent readers tend to read more **Business** and **Opinion** sections, and less **Markets** sections relative to infrequent readers.

Chart

Description automatically generated

Figure 4: Returned Visitors’ First Visit Distribution across Sections

**Predictive Modeling (Part B)**

1. Create a simple model predicting the likelihood of a first-time visitor returning.

I choose *wordCount*, *videoCount*, *author*, *topicKeywords* as predictors. To reduce the number of insignificant predictors, I only include authors of more than 200 first-time pageviews and keywords of more than 500 first-time pageviews. Categorical variables are transformed into dummy variables, and numerical variables are divided by the sample maximum to reduce to ranges of zero and one. We end up with a total of 32 predictors.

I use the Random Forest model to predict return probabilities. Logistic regression regularized by elastic net serves as the benchmark. The sample is divided into a training set of 8,000 observations and a test of 2,000 observations. Table 4 shows the models’ out-of-sample Area under Receiver Operating Characteristics (AUROC)[[1]](#footnote-1). Random Forest yields AUROC of 0.53 which is lower than the benchmark’s AUROC of 0.57.

Figure 5 shows the variable importance of random forest. The number of words dominates other variables’ importance. Figure 6 shows slope coefficients of non-zero predictors in the logistic model. Only two predictors matter in the logistic model: whether the article is in Politics section and the number of videos.

Based on the result, a few variables, number of words, number of videos and the political section drives most of the model outcomes. To interpret their impacts on the model, I build a decision tree with the three predictors. Figure 7 shows the decision tree. The nodes of the tree show Politics section without video and longer articles with video tend to yield higher return probabilities of 13% relative to other circumstances (10% and 7%).

|  |  |
| --- | --- |
| Model | AUC |
| Random Forest | 0.5318 |
| Logistic Regression | 0.5674 |
| Decision Tree with Selected Features | 0.5609 |

Table 4: Model Performance Comparison

Timeline

Description automatically generated

Figure 5: Importance of Predictors in Random Forest

Chart, bar chart

Description automatically generated

Figure 6: Coefficient of Features in Logistic Regression

A picture containing text, receipt, screenshot

Description automatically generated

Figure 7: Decision Tree

2. What additional data would you add to improve the model?

I will add additional data describing visitor profiles and behaviors. The first visitors are not subscribers, it’s hard to know their demographic information, but we can certainly take advantage of the traffic information to evaluate their engagement, such as the number of visited pages, page view duration, and events describing multi-tasking, click to other articles, click to videos and watching time. We can also take advantage of visitor profile information such as geographics, languages, and platform information.

1. ROC curve plots all combinations of true positive rates and false positive rates. AUROC measures the area between the ROC curve and 45-degree line. A higher AUROC indicates better performance. [↑](#footnote-ref-1)