Google Gstore Revenue Prediction

Abstract

"Half the money I spend on advertising is wasted, the trouble is I don't know which half" -John Wanamaker

One of the largest problems in running a business is identifying who is likely to generate revenue. Online business' have access to far more data on who is visiting their 'store' than a brick and mortar business, who might only have access to a small amount of data on people who browsed but bought nothing. To gain insight into this problem, in this paper we will analyze a dataset from the Google Merchandise Store (also known as GStore, where Google swag is sold) customer database. The raw dataset contains 12 columns to predict the transaction revenue per customer. The outcome from this analysis will aid in better use of marketing budgets. This paper will model the customer data, and create predictive models to estimate revenue generated from a holdout set.

Introduction and background

Identifying customers who will purchase a product after visitng a website is one of the most important factors in online commerce. The Pareto Principle, a business principle that asserts 80% of outcomes result from 20% of causes . In our case this means that we would expect 80% of the revenue to be generated by 20% of the customers [1]. The data from the Gstore contains 55 variables for around 900,000 site visits. A rigorous investigation of these determinants is needed to identify if this rule of thumb holds, and which customers marketing resources should be put towards.

The Pareto Principle was developed by Italian economist Vilfredo Pareto, who discovered that 80% of the land in Italy was owned by 20% of the population. This is a description of a power law distribution, known as the Pareto distribution. It has been found that natural phenomena exhibit this distribution, particulary in commerce. This was translated to business administration by consultant Joseph M. Juran, who urges decision makers to find "their golden 20%".[1] Where they can make actionable statements like "These are my top transaction locations" or "these are my top traffic channels for revenue generation."

A lay approach to identifying the golden 20% is to graph a single categorical variable and identify which category most of

the revenue was generated from. (chart 1) This however makes it more difficult to address variable interaction.

The purpose of this paper is to generate a multivariate model of transaction behavior of the Gstore visitors so that we can identify customers most likely to generate revenue using data from the Google Gstore. In particular, it develops a Random Forest model to predict the revenue generated by a specific transaction. Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. We will compare this with a generalized linear model that generalizes linear regression by allowing the linear model to be related to the response variable via a link function (gaussian) and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Each model is evaluated using the Root Mean Squared Error (RMSE) function. The model with the lowest RMSE is chosen.

This paper reports which factors we deem to be the most important, providing insights for a better decision making model for how to target specific customers to an online store.

The organization of this paper will be: -description of the methodological framework -description of the Gstore data - empirical results of the modeling -conculsions drawn from our findings

Methodology

Modeling transaction revenue can be done with a generalized linear model a randomForest regression model

We expect Random Forest modeling to work better if: 1. The underlying function is not truly linear 2. We end up selecting a high number of predictor variables 3. There is high covariate shift (the distributions shift between the train and the test set) [6]

While the Random Forest Regression seems like the best fit for our data, we will employ the 'no free lunch' theorem [7] and generate linear models as well for comparison.

We will be predicting the natural log of the sum of all transactions per user (further discussed in the data section)

$$y_{user} = \sum_{i=1}^{n} transaction_{user_i}$$

$$target_{user} = \ln(y_{user} + 1)$$

Data

Data analyzed in this paper as collected between 2016 and 2017 for assessment of the Gstore. The data provided is a random sample for all visitors to the Google Gstore during this period.

There are almost 1 million rows and 55 columns once unpacked, The following is a description of the most consequential columns from the data.

fullVisitorId: - A unique identifier for each user of the Google Merchandise Store. This will be important for summing each transactors behavior since it is constant between a users unique visits over time.

channelGrouping: - The channel via which the user came to the Store. This is where the visitors to the Gstore "came from" on the internet. Some examples of categoreis are 'Social' (social media links), Referral or Organic Search.

date: - The date on which the user visited the Store.

device: - The specifications for the device used to access the Store, such as tablet, phone or desktop.

geoNetwork: - This section contains information about the geography of the user, such as the country, state and city from which they accessed the store.

socialEngagementType: -Engagement type, either "Socially Engaged" or "Not Socially Engaged", referring to social media usage

totals: -This section contains aggregate values across the session.

trafficSource: -This section contains information about the Traffic Source from which the session originated.

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.

visitNumber: - The session number for this user. If this is the first session, then this is set to 1.

visitStartTime: - The timestamp (expressed as POSIX time).

hits: - This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.

customDimensions: - This section contains any user-level or session-level custom dimensions that are set for a session. This is a repeated field and has an entry for each dimension that is set.
totals - This set of columns mostly includes high-level aggregate data

The most notable transformation that is employed is the natural log of the target revenue. As shown in in chart 2, the transactions are on the whole all zero, causing a severe right skew. Taking the log of revenue makes the data more normally distributed, allowing for more precise analysis. This also gives us our first insights into finding the golden 20%, since most of the visitors to the Gstore are not revenue generators. As shown in chart 2 (pageviews vs revenue) only about 1% of visitors to the site produced renvenue, however the 80 20 rule applies to pageviews where 80% of the revenue came from 20% of the users.

Another notable feature of the data is that 40% of the variables are either sparse or only contain one value, which can be excluded from our analysis for containing no row-wise information. For example, the predictor adwordsClickInfo.isVideoAd is NA for 98% of the data and False() for the remainder. There is no correlation between the positive class and the response variable so it is excluded. Similar situations occur in 17 of the 55 rows and were dropped before analysis.

Columns that are duplicated. SessionID and visitID are about the same, no more info. However fullvisitorID is unique.

One other unique preprocessing decision was made involving the predictors SessionID and visitID. The data in the columns were very similar and therefore did not carry unique information. The usefulness of these columns would be to identify when a user returned to the site, but this is captured in the variable fullvisitorID, so SessionID and visitID were dropped.

Preprocessing

Certain features are selected based on the exploratory data analysis. Some of those features are then thresholded based on frequency below 1%. Other features are imputed with knn. Step order [12]. We thresholded columns with many levels, such as country. where US and Canada were the two most frequent countries, while there were many other countries that only appeared a few dozen times, so they were aggregated into an ‘other’ category.

Empirical Results

A randomForest model was used to optimize the RMSE and select the optimal number of predictors and node size for our decision trees. Random Forest is a bagging technique that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting. [11] Using parallel processing 10 models are trained across a grid of 20 to find the best hyperparameters, mtry and minN. Mtry is the number of variables randomly sampled at each split for the random forest model and min_n is the minimum number of data points in a node that are required to be split further [13]. The optimal Random Forest regression obtained a .913 RMSE, while the GLM RMSE is 1.044

In Chart 3, it is clear that the random forest model is more narrow in its distribution of predicted values, leading to a lower RMSE. Table 1 is the correlation matrix of select variables included in the modeling. A correlation matrix will give decision makers an intuitive understanding of the different factors on purchasing behavior, as well as indication of possible predictors for revenue generating transactors. For example, device category (which hardware people use to access the Gstore) is highly correlated with revenue, as the majority of revenue has come from desktop users, while a predictor like GeoNetwork does not show high correlation with the target variable, and is likely to be eliminated in the model building process

Table 2 is the feature importance output from our Random Forest regression modeling. In terms of decision tree modeling, the higher up the branch network, the more impact the variable has to selecting which leaf the data belongs to. The ten most important variables for classification are, in order: -hits -pageviews -source -channelGrouping -medium -metro -operating system -deviceCategory -region -isMobile

In Table 2, we show the differences in in estimation between a GLM and a Random Forest Regression. In this table we report the estimation results of both models. THE RMSE metric reports the differences between values predicted by a model and the values observed. In our case this is total revenue per customer. RMSE is a measure of accuracy that is useful in comparing different models

RMSE can be calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{\hat{y}_i - y_i}{\sigma_i}\right)^2}$$

The predictor with the most impact was hits. This intuitively makes sense as this is the most direct measure for interest in the Gstore that we have in the dataset. This also goes for pageviews, and could be combined into an engagement metric for decision makers. Because most pageviews are low, the log of these predictors allowed us to generate more detail in our model. Approximately 80% of the revenue came from the top 20% of pageviewers.

The next group of predictors are TrafficSource and channelGrouping. Both of these variables tackle where the visitor came from on the internet. The largest share of revenue came from referral. Out of the possible places a visitor could have come from, referral conveys the most interest in the product before the visit even starts. Recent studies from Neilsen show that "…referrals are the most trusted form of advertising, with 92% of consumers reporting they completely or somewhat trust referrals from people that they know."[8]

The predictor metro is next up in terms of importance. In a recent CBS news report, more than 90% of tech jobs are located in just 5 US cities. It would make sense that someone who wished to purchase Google products would be in some way involved in technology. [9] This predictor falls in the same importance band as operatingSystem, deviceCategory, region and isMobile. As is to be expected, most of the revenue comes from American Google Chrome users. Chrome is a product of google, which is a country located in America.

Surprisingly, even though Macintosh operating systems are second to Windows OS users, Mac OS users generate far more revenue than almost every other category combined. This might be due to the high brand identification that Apple has generated for its products. Marketer Marc Globe, author of **Emotional Branding** had this to say about Apple users. This high rate of revenue from Apple users could be indicative that the brand loyalty is transferred to Google products.

"Apple's brand is the key to its survival. It's got nothing to do with innovative products like the iMac or the iPod[10]."

Summary and Conclusions

In this paper, we compare the results of a GLM and Random Forest Regression model to accurately predict the revenue generated by a visitor to the Google Gstore. Several important

factors are found to influence a Gstore visitor's decision to make a purchase. The most important predictors point us towards visitors who are cosmopolitan Americans emotionally invested in technology. They were also very likely to be referred to the Gstore by someone they trust. Our predictors line up with business intuition, particuarly the 80 20 rule. As seen in Table 3, Mac OS users accounted for about 60% of the revenue, but only made up 28% of the visitors. While this isn't exactly in line with the 80 20 rule, it still an actionable item to target advertising. The 80-20 rule is a rule of thumb highlighting that an entreprenuer should be looking for small segments of its users that drive profit.

This study generates useful insights for a better understanding of customer behavior on the Google Gstore. For decision makers, this could lead to more targeted marketing, and even what data is important to collect from a visitor to an online market.

References

[1] Tardi, Carla .**80-20 Rule**, 2020, https://www.investopedia.com/terms/1/80-20-rule.asp

[6] Trotter, Matt. **What is Covariate Shift?**, July 08 2021, https://www.seldon.io/what-is-covariate-shift/#:~:text=Covariate%20shift%20is%20a%20specific,training%20environment%20and%20live%20environment.&

[7] Brownlee, Jason.**No Free Lunch Theorem for Machine Learning** October 21, 2021, https://machinelearningmastery.com/no-free-lunch-theorem-for-machine-learning/

[8] Kunis, Leigh. **Referral Marketing: What it is and Why it Works**, August 23 2018, https://www.springboard.com/blog/business-and-marketing/referral-marketing-what-it-is-and-why-it-works/#:~:text=A%20recent%20Nielsen%20study%20shows,referrals%20from%20people%20they%20know.&text=

[9] Min, Sarah, **Most US Tech Jobs Clustered i in Just 5 Major Cities**, December 10 2019, https://www.cbsnews.com/news/90-percent-of-tech-jobs-growth-concentrated-in-just-5-cities-according-to-brookings-institute-report/

[10] Kahney, Leander **Apple: It's All About the Brand**, December 04 2002, https://www.wired.com/2002/12/apple-its-all-about-the-brand/

[11] Gupta, Aman **XGBoost versus Random Forest**, April 26 2021, https://medium.com/geekculture/xgboost-versus-

random-forest-898e42870f30

[12] **Ordering of Steps** , tidy Recipes, https://recipes.tidymodels.org/articles/Ordering.html

[13] **Random Forest**, tidy models, https://parsnip.tidymodels.org/reference/rand_forest.html

Charts

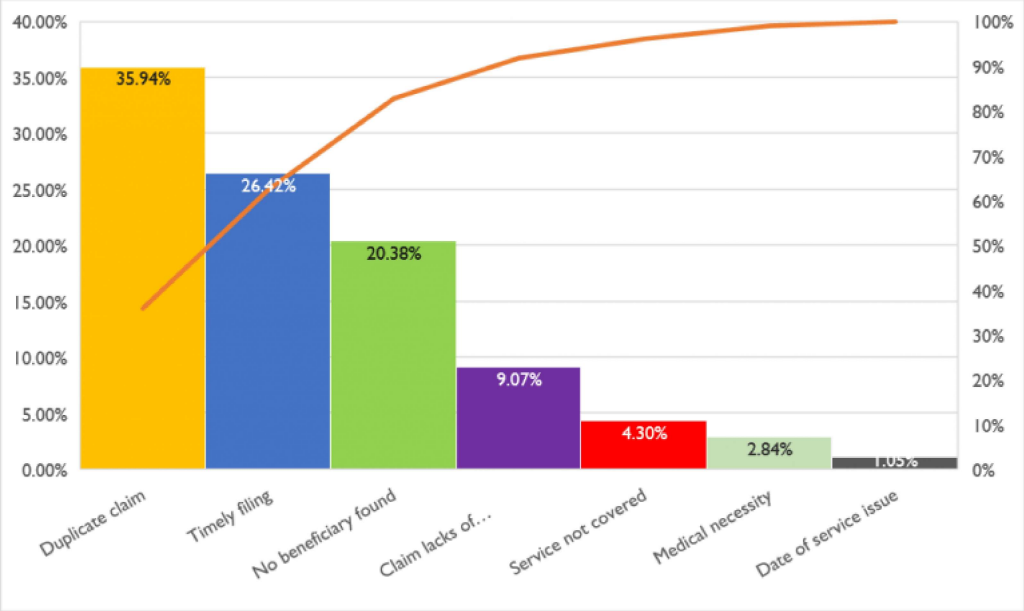

Chart 1

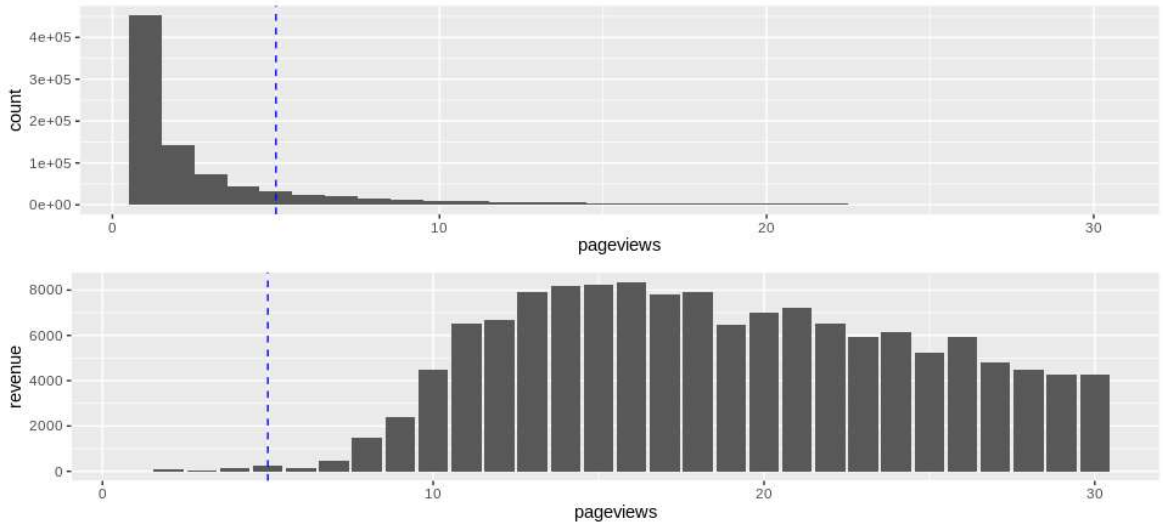

Chart 2

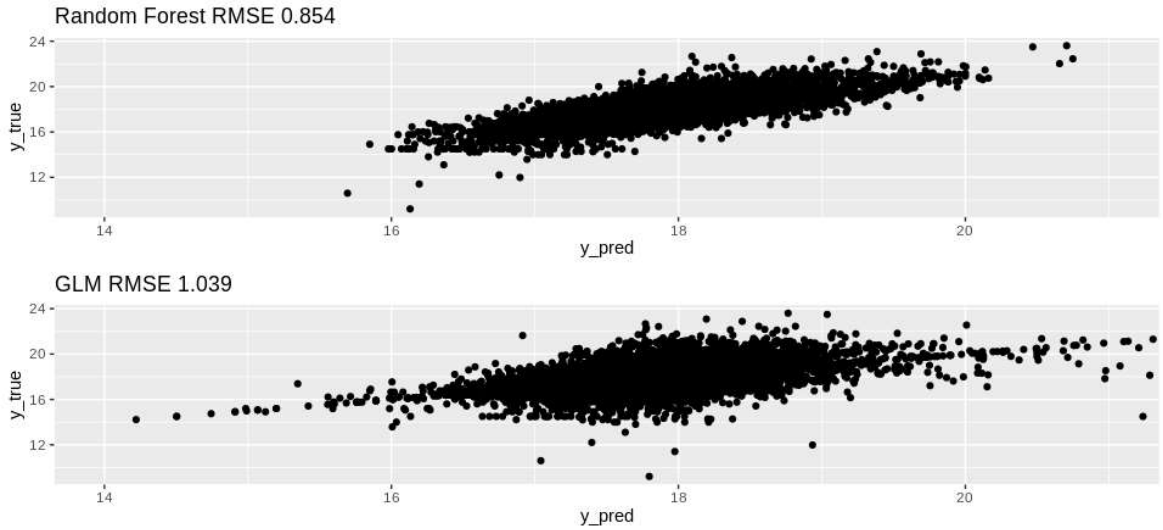

Chart 3

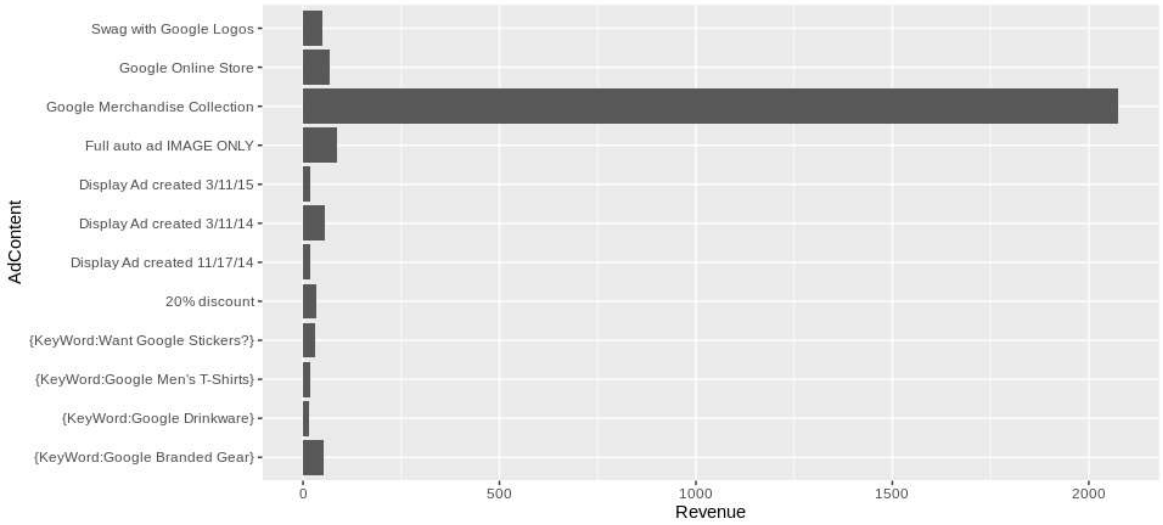

Table 2

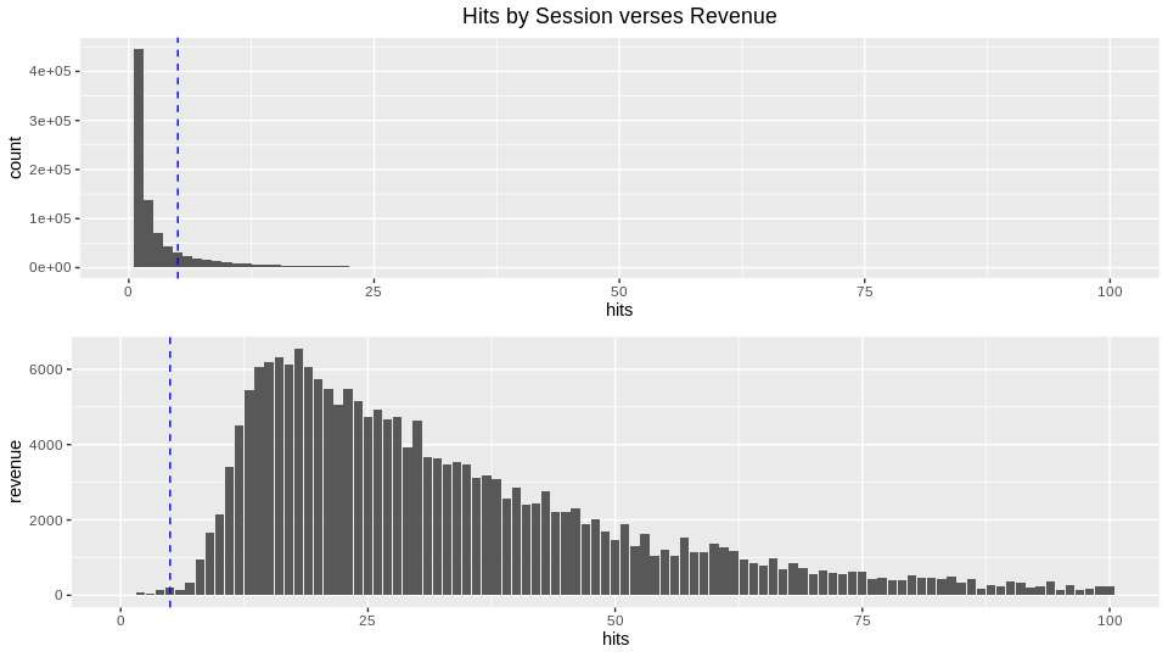
Hits by Session verses Revenue
count
4e+05
3e+05
2e+05
1e+05
0e+00
0
25
50
75
100
hits
revenue
6000
4000
2000
0
0
25
50
75
100
hits

Table 3