Google Analytics User Revenue Prediction

Blaine Murphy

September 2021

Executive Summary

The Google merchandise store has a problem that is familiar to most retail stores, a majority of their revenue comes from a minority of their visitors. Naturally the marketing team wishes to spend limited add dollars to make the most positive impact on revenue. I analyzed the data of user visits to the GStore over the course of the year and found several factors that are correlated with whether a user will make a purchase and the size of that purchase. In addition, I created a neural-network model to predict total user revenue that significantly improved upon the base model. Finally, I identified the most impactful features on model prediction which include the deviation of pageviews, the median of hits, the internet penetration in the users country, and the number of visits to the store.

Problem

The Google merchandise store gets a large portion of its revenue form a fairly small portion of its visitors/users. This poses a problem for the marketing team, which is how best to allocate limited marketing dollars to make the most positive impact on revenue. The task then is to explore the data, find features that correlate to transaction revenue, and build a predictive model that will help identify the factors that are most impactful for revenue prediction.

Data and Wrangling

The data available for this project is browsing information for visitors to the GStore website between August 1st, 2016 and August 1st, 2017. Each row in these data is one visit to the store and has a unique value that is the concatenation of the user id and the time of the visit. A small number of columns contain JSON blobs that vary in depth and need to be unpacked. The target revenue for each visit is in one such column.

After inspecting these data, I quickly noticed that this is an international data set. To supplement these data, I merged in country GDP, population, life expectancy, and the percent of the population who use the internet data from the World Bank. Some GStore visitors were from countries not in the World bank data set and their values were filled in with median values from their sub-continent.

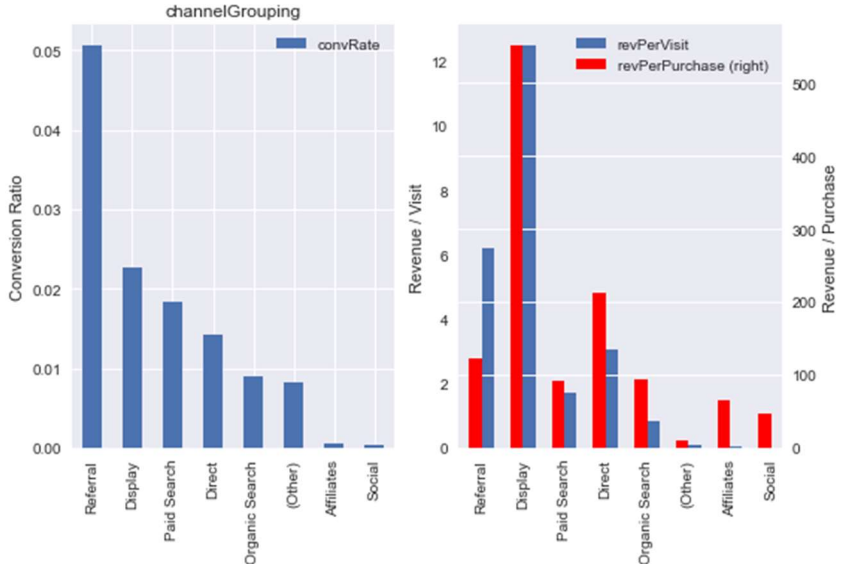

Figure 1. The figure shows two bar plots demonstrating the effect of the 'channelGrouping' feature. On the left is the conversion ratio, the fraction of visits that result in a purchase. On the right is the average revenue per purchase and per visit.

Exploratory Analysis

I profiled all of the features in the data, identifying how they relate to the conversion ratio for each visit and the average amount spent for each visit. During this process I also checked how highly some features correlate to others in the data. As an example, the 'channelGrouping' feature is shown in Figure 1. The conversion ratio varies from .05 for the 'Referral' category to approximately 0 for the 'Social' category. However, the average size of the transaction is highest for the 'Display' category. A χ^2 test between 'channelGrouping' and whether a purchase was made or not demonstrates a p-value of ~0.0 but a Cramer's V value of 0.13 indicating it is a statistically significant feature, but has a low correlation to a transaction. In a few instances there were redundant features in the data, such as the 'medium column being highly correlated with 'channelGrouping' with a Cramer's V of 1. 'channelGrouping' has a higher correlation to whether a purchase was made, so I dropped the 'medium' column to aid in interpretability of the final model.

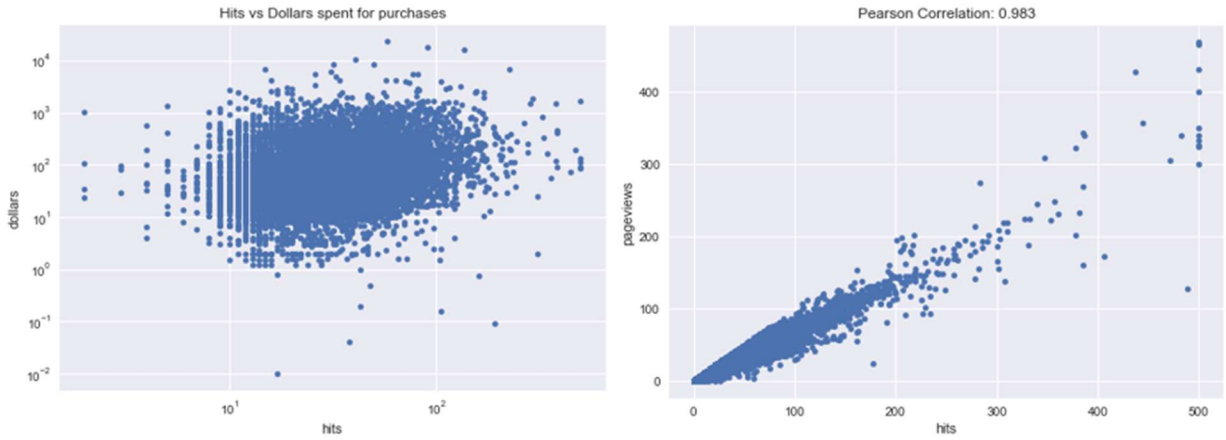

Figure 2. The figure on the left is the cross plot of column 'hits' and 'dollars' for all visits that resulted in a purchase and on the right is the cross plot of 'hits' and 'pageviews' for all visits.

There are only two numeric columns in the data, 'hits' and 'pageviews', and they are highly correlated to each other (Figure 2). For the unfamiliar, 'pageviews' represents the number of unique pages seen by that user on the website and 'hits' is the number of files downloaded during that visit, so it makes sense that they are both positively correlated with the transaction and highly correlated with each other. The dollars spent relates to 'hits' with a Pearson correlation of 0.15 with a p-value of 0 and to 'pageviews' with a Pearson correlation of 0.19 with a p-value of 0. 'hits' and 'pageviews are highly correlated (Pearson 0.98) and I dropped several of the summary statistics using VIF before modelling.

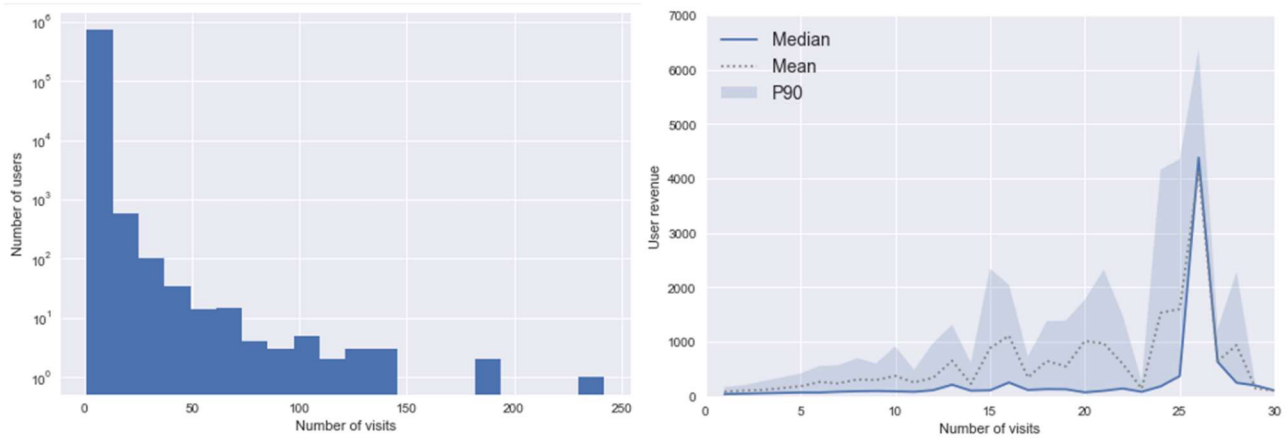

Figure 3. Two plots describe the number of visits among users. On the left is a histogram of users' visits. On the right is a plot of the median, mean and 90th percentile of purchases for users grouped by the number of times they visited the GStore.

Finally, I analyzed the effect of repeated user visits which is shown in Figure 3. 631,804 of the 722,668 unique visitors to the GStore only visited the store once, but repeated users account for most of the total revenue. My first thought was to use repeated visits to the store like a time series, but since the majority of the data I don't believe this is possible.

Instead, I will aggregate the visit-level data to a user-level data with summaries of individual visit characteristics. The number of visits is positively correlated with user revenue as seen in Figure 3.

Feature Engineering

I created a few new features in hopes of improving the predictive power of the model. First, I created a column that recorded the difference in time between a user's visits to the store. Next, I created columns that represented when in the time window they visited the GStore. Using the users visit time, I created columns for time of the day, day of the week, and month of the year. The geographic 'region' column for each user was mostly null, but contained the state for visitors from the United States. Because the probability of a purchase varied greatly across the United States, I created a new column from region that contained the user's state or if international a single value representing not from the United States. I greatly reduced the number of unique values in the 'city' column by lumping all cities that had no record of a purchase into one. Finally for keyword searches, I calculated the most common words in searches that resulted in a purchase. Then I created binary columns from the keywords and dropped the 'keyword' column, which is mostly null. Finally, I grouped the data by the user id and aggregated the columns by appropriate means. The result is a table of 722,668 unique users and 358 features.

Next to remove collinear features I calculated variance inflation factor and iteratively removed those features that scored higher than 2. This action resulted in dropping 68 predictors leaving 290 features for modelling. Finally, I split the data into 70% training, 20% validation and 10% test sets in preparation for modelling. Using only the training set I trained a standard scaler to normalize all of the predictors and applied the scaling function to the training and validation sets.

Modelling

To predict the revenue generated from each user I decided to use a Keras neural network due to their high accuracy, the large number of predictive features, and their customizability. In addition, it is straightforward to understand feature impacts with a NN using the shap library. I started with a 3-layer NN with one dropout layer for regularization. In the training data there should be a purchaser every ~70 users, so the batch size was set at 500 to ensure purchasers are in every model update batch. To prevent overfitting to the training set, a callback was employed for stopping network updates when there was no additional improvement in the error of the hold out validation set (Figure 4). This model produced a training RMSE of 1.629, a validation RMSE of 1.639, and a testing RMSE of 1.658. This is an improvement over the base case mean of training prediction which produced an RMSE of 2.08.

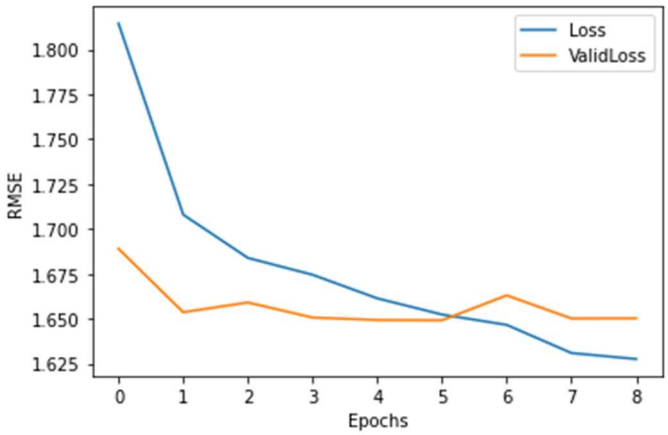

Figure 4. The training history of the 3-layer neural network. The RMSE decreased quickly and training was stopped early due to no additional improvement in validation error.

Next, because the data is heavily imbalanced (99% of users did not make a purchase) I tried creating a classification NN to predict whether a user will have made a purchase, and then use this as input into the regression NN. Despite a very high classification ROC AUC score the regression prediction RMSE using the classifier probability as a predictor did not improve the error over the base NN. I also tried creating complex NN's that performed classification and regression in the same network, but again these did not perform improve the RMSE score of the regression to the hold out sets.

Finally, I tried creating segmentation models to improve the prediction for user revenue. I created pairs of Keras models for the following situations: mobile vs. not-mobile users, US vs. international users, weekend vs. weekday users, and single visitors and repeat visitors. In each instance the prediction error increased for the model that was predicting most of the revenue. For example, the validation RMSE score for the model trained only on users from the USA is 3.09, but

these users account for 94% of the revenue. The prediction for international users improved to an RMSE of 1.17, but overall, the prediction RMSE did not improve. Moving forward I will use the first 3-layer neural network as the best and simplest predictive model.

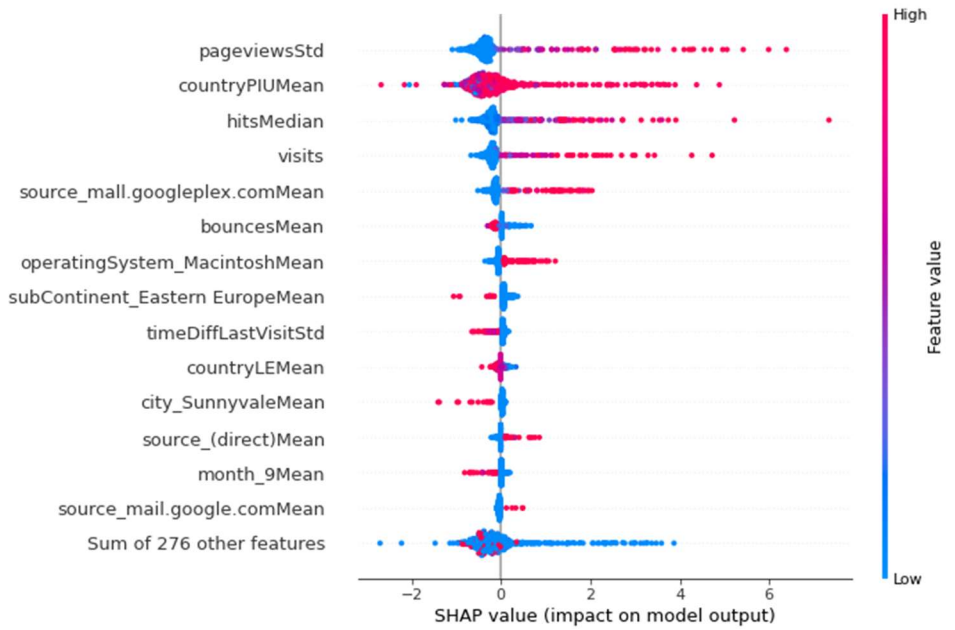

Figure 5. The figure shows the feature importance of the final 3-layer NN model. Features are ranked in order of average effect on the regression output and the horizontal axis represents the magnitude of the impact on the regression prediction. The color represents the relative value of the feature and each point is one permutation of record fed to the model.

I utilized the 'shap' library to visualize the feature impacts on the regression prediction, and to save on computation time I randomly subset the data. I created a 50-sample background set from a random selection of the training data with 10 purchasers and 40 non-purchasers. Next, I randomly sampled 500 users (with the same ratio) from the validation set that is permutated in the feature importance calculation. Figure 5 shows that the features that on average have the greatest impact in predicting a user's total revenue are the standard deviation of the users pageviews, the percent of the country population that use the internet, the median hits from each visit and the number of visits. Unsurprisingly high values of pageviews deviation, median hits, and visits have a positive impact on the regression output. In addition, higher country percent of internet users has a positive effect on the prediction. There are a couple of features in the top 15 that are negatively correlated with the regression output. These include 'bounces', which occurs when a user leaves the site without clicking through, and the deviation of the time between visits. When these values are high, they have a negative impact on the output. These observations make intuitive sense in the context of the problem.

Recommendations and Future Work

The neural network model is an accurate predictor of total user revenue and is ready for deployment. The marketing team can utilize the data and model's insights immediately. The most important features for predicting user revenue are the deviation of pageviews, the percent of the country's population that use the internet, the median hits, the number of user visits, and the source "mall.googleplex.com." I interpret pageviews and hits being key to the prediction as a signal that users who spend more time clicking through the site are more likely to spend more money. Site developers should ensure that the website is easy to use and navigate. That the percent of population that use the internet is important is not a surprise, but it reinforces that international marketing efforts should focus on countries with high internet penetration. The number of user visits to the GStore is also an important feature to the regression prediction, indicating steps should be taken to increase the probability of a user to return.

The project should be revisited periodically to utilize newly acquired visitor data, every 3 or 6 months. It could be useful to integrate more detailed data for this problem, such as data on the purchased items or detailed user data joined from browser history for example. Further work could include updates to the GStore site. Developers could look at improving the site for ease of use, navigation and ability to hold user attention. Additionally, the marketing team could try sending notifications or email marketing materials to encourage previous visitors to return to the site. A/B tests should then be conducted to see if these changes positively impact user revenue or key model metrics, like pageviews and the probability of a user to return.