2020

GStore Revenue Management

HAO HAO
YUTING LI
XIQIAO HUANG
MANDY GU
BOYA SUN

Table of Contents

1. Business Objective

As a fast-growing business, Google Merchandise Store (GStore) aims to increase its revenue by gaining a better understanding of its customers and making appropriate investments in the promotional strategies. In the sphere of commerce, Pareto Principle dictates that “80 percent of your profits come from 20 percent of your customers” and holds true for many businesses. To improve its revenue, GStore needs to identify its “20 percent” buyers to make sure it retains these customers. In addition, GStore needs to find out what has worked well for its loyal followers to replicate the successful marketing practices and create more avid buyers.

2. Key Actionable Business Initiatives

By gaining a clear understanding of its customer base, GStore will be able to pursue several business initiatives to improve its sales performance. The first is to target its top 20% customers and create marketing strategies to keep satisfying and retaining them. GStore should figure out where the top buyers converge – the demographic region they fall in, the traffic channels they came from, the ads/campaigns on which they converted and the promotional content that they engaged. With this knowledge, GStore can enhance and optimize such mediums and content. On the other side of the spectrum, GStore should also identify patterns and commonalities among dissatisfied customers that made fewer purchases. For this customer segment, GStore can design promotional strategies to promote action and cultivate loyalty in them.

3. Role of Analytics

We used two analytical approaches to address these business problems. First, we used user-level descriptive RFM (Recency, Frequency and Monetary) analysis to segment customers and give corresponding recommendations. Then, on the transaction level, we used predictive models like random forest to find the most important features that contributed to the revenue and then did further descriptive analysis on the traffic-to-sales conversion rate to see which sections within each of them have the most potential to generate more revenue. Moreover, we can use this information to allocate resources such as financial or human resources more effectively. And higher the accuracy of the prediction model, the more possible that it will help the executives make better business decisions.

4. Metric of Success

We use 3 KPIs to measure the degree of success of our initiative to increase the revenue in the long run. To see if our strategies can cultivate loyalty from non buyers, we calculate the *customer conversion rate*. To make sure we are on the right track of maintaining loyal customers, we track the *customer retention rate*. To evaluate the marketing strategies that target customers with the most important features identified by our analysis, we dynamically track the *spending of top buyers* to see if our strategies can increase their loyalty.

As mentioned above, the key to our success is to accurately identify the most important customer features produced by the prediction models. We measure the performance of our models using R^2 which indicates how much variance in the revenue can be captured and explained by our models.

5. Thinking Through the Analytics

5.1 Overall Descriptive Analysis

The data that we used to conduct analysis is GStore's customer traffic information, which is the existing observational data. Before moving to the prediction model, we first did an exploratory data analysis on the GStore data regarding the ratio of customers that generates revenue as well as the comparison between the user and revenue trend.

We grouped the transaction revenue by visitor ID and then did a scatter plot on the index (represents and reorders visitor ID) and the revenue. As you can see in the chart below, only a small percentage of customers produce most of the revenue, which conforms to the 80/20 rule. To be more specific, among all the users, there are only 1.61% of them that actually generated revenue. We should put more effort on converting visitors to buyers and provide better customer experience for those valuable customers.

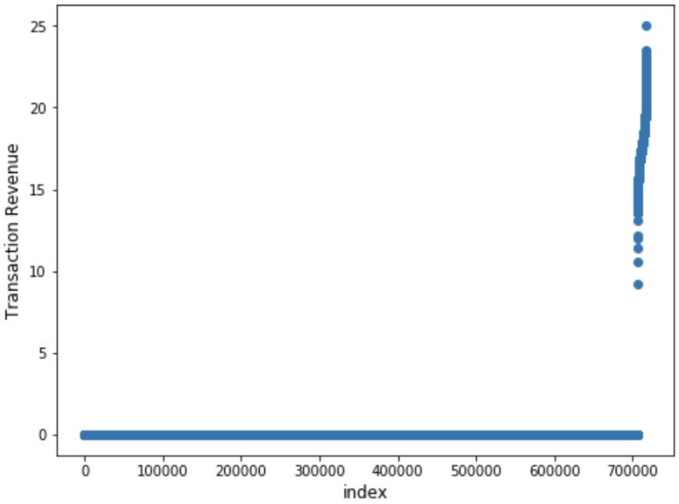


Figure 1. Only a Very Small Percentage (1.61%) of Visitors Contributed to the Transaction Revenue

On the time dimension, we plotted the trend over time for both number of unique visitors and total transaction revenue. As shown below, you can see a sudden increase of visitors from October to December 2016, while the trend of revenue remains stable as before. Therefore, we searched online for possible big events of Gstore that could lead to this huge discrepancy. It turns out that in October 2016, Google opened a temporary pop-up showroom in the SoHo neighborhood of New York City, and in November 2016, a store-within-a-store where Google displayed its main hardware products, named Google Shop[1], was opened. These two events probably largely increased the brand awareness of GStore and brought in lots of visits to it. However, the conversion rate didn't seem to be ideal and the company didn't benefit much from those events during that period. We should reflect on them and try to come up with better campaigns.

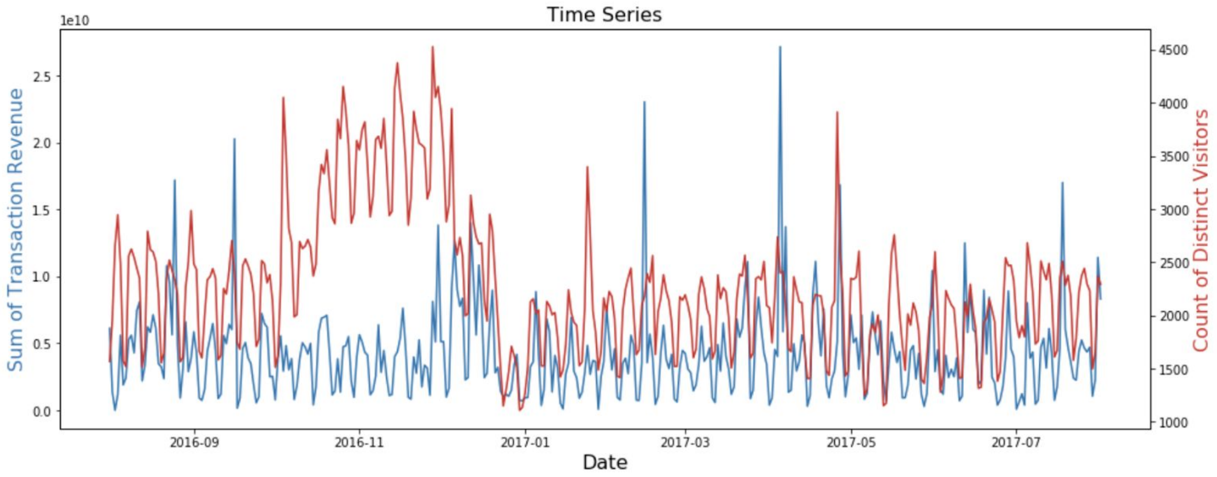


Figure 2. Comparison Between the Trend of Number of Unique Visitors and Total Transaction Revenue

5.2 Data Preparation

We aimed to use the data to find the attributes that affect GStore's revenue, so we chose the "transaction revenue" as our dependent (target) variable. As we mentioned above, only 1.61% of customers in the dataset contributed revenue. To focus our analysis on these groups of customers that contribute revenues, we only kept transactions with nonzero transaction revenue for our modeling. Among the transactions with non-zero revenues, their mean transaction revenue is about $133 and the standard deviation is around $448. To maintain precision, all transaction revenues were multiplied by 1,000,000 in the dataset.

The explanatory variables of our models are the attributes that are related to each transaction. Besides the variables that won't be useful to our model like "visitor id", we dropped variables that contain too many null values or unreadable values, such as "keyword" and "bounces". For numerical variables like "pageviews", "new visits" and "hits", we imputed the NAs with value 0. After the imputation, the variation of these numerical variables are shown as below.

[1] Two events were referred from: Google Store. (2020, April 02). Retrieved June 05, 2020, from https://en.wikipedia.org/wiki/Google_Store

	Pageviews	New visits	Hits
Mean	28	0.38	36
Standard Deviation	21	0.48	30

Table 1. Mean and Standard Deviation of Numerical Variables

We also extracted the month value from "visit start time" to create a feature called "visit month" putting in our model. The majority of our explanatory variables are categorical variables with many categories, for each categorical variable, we aggregated categories with very a small number of transactions under the category named "other".

5.3 Predictive Analysis

We built our prediction model using the machine learning library in Spark, MLlib. Based on data preparation and feature selection in 5.2, we chose the following independent variables as features: ***marketing channel, number of visits, operating system, browser, mobile user*** (boolean)***, city, network domain, page views, number of hits, new visitor*** (boolean)***, medium*** (how the visitor was referred to the site)***, source*** (from where the visitor clicked on a link that led to the site)***, visit month*** (in which month the visitor visited the site).

The target variable is transaction revenue. We first fit a linear regression model which results in poor performance (R^2=0.11). Since most of the features are categorical variables, and we infer that there is a non-linear relationship between the features and the target variable based on the descriptive analysis above, we use the random forest model to predict transaction revenues.

To choose the best random forest prediction model, we first created a pipeline that chains multiple categorical variable transformers such as "StringIndexer" and "OneHotEncoder" and the random forest estimator together as a workflow. Then we parse the pipeline and a grid of parameters such as the number of trees and the depth of trees to the CrossValidator. The CrossValidator then identifies the best parameters by calculating the average evaluation metric R^2. The out of sample R^2 of our best random forest prediction model is 0.25, i.e. 25% of the variance in the target variable is predicted by the model.

The best random forest prediction model allows us to obtain the importance of features. This is because, at each node on each tree, features are used to split samples into multiple buckets. The most important features split samples into distinguished buckets, and such buckets contain samples that are similar among themselves but are different from samples in the other buckets. In our case, the most important features are: ***city, browser, visit month, marketing channel***.

Features	**Importance**[2]
city	4.13%
browser	4.11%
visit month	1.74%
marketing channel	0.46%

Table 2. Feature Importance of the Random Forest Prediction Model

5.4 Further Descriptive Analysis

After building the prediction model, we did further descriptive analysis on the most important four features that contributed to the transaction revenue, which are city, browser, month of visit and channel. As you can see in the four charts below, higher value of total transaction compared to number of unique visitors indicates more efficient conversion of traffic to sales. Namely, New York City was doing great while Mountain View and San Francisco were much less efficient in conversion to purchase, Chrome and Firefox were not bad while Safari needed to be looked into, all months except for October and November were doing great, referral channel was the best, followed by direct channel, while original search and social channel needed extra attention to improve.

[2] Importance: The higher, the more important the feature. The importance of a feature is computed as the (normalized) total reduction of the prediction error brought by that feature.

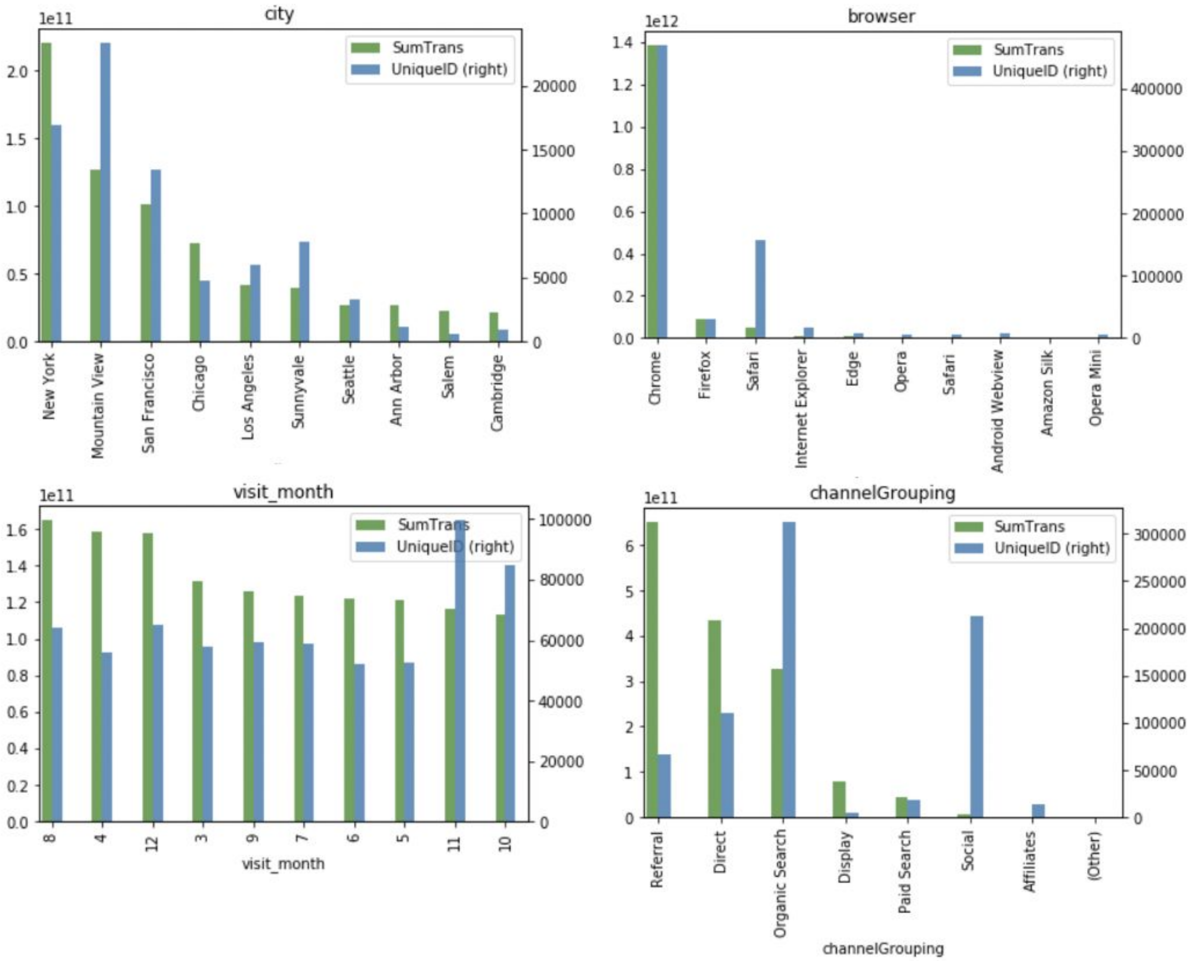


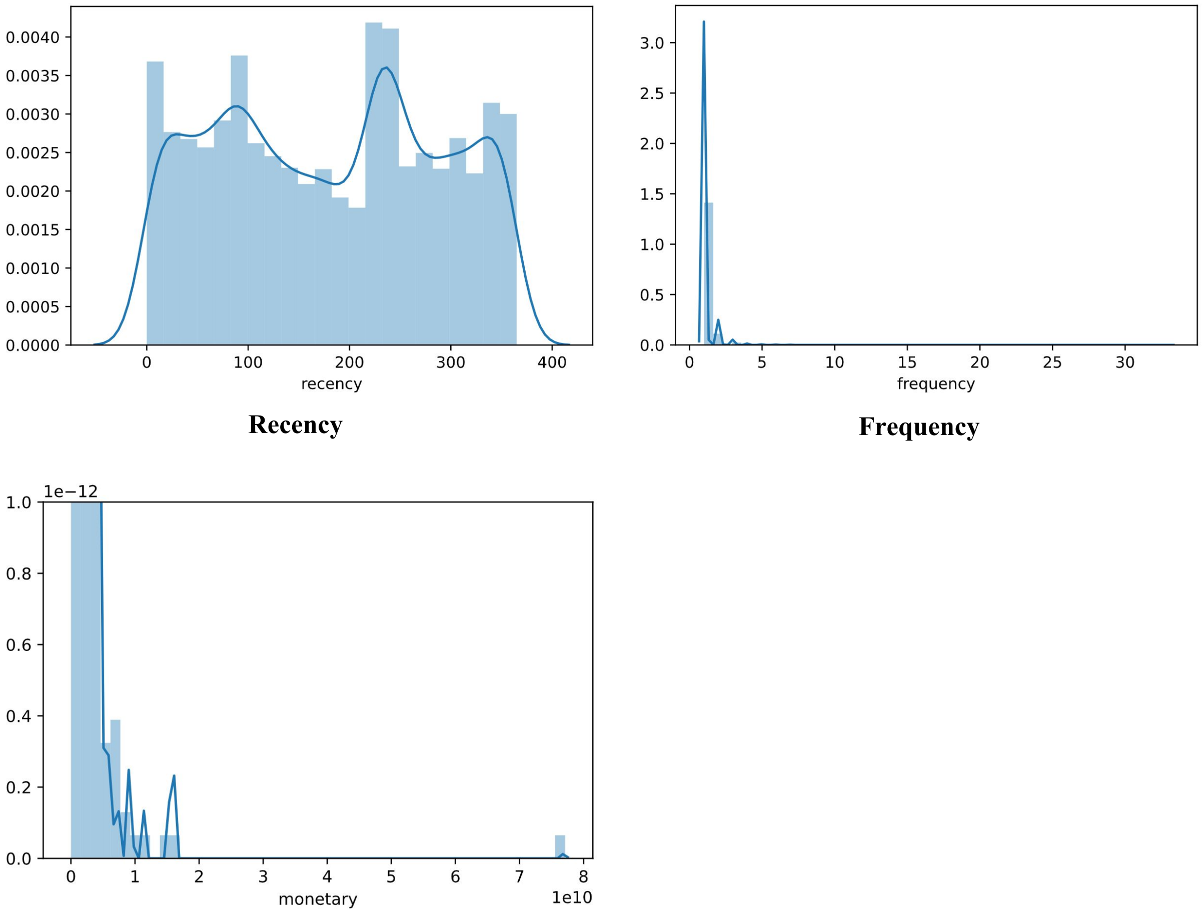
Figure 3. Number of Unique Visitors VS. Total Transaction Revenue for the Most Important Four Variables (city, browser, visit_month, channelGrouping)

5.5 RFM Analysis

To better understand the customers, we also perform *recency, frequency, monetary* (RFM) analysis to segment the customers for marketing insights. RFM captures the three crucial factors of customer spending behavior. Recency stands for the freshness of the customer activity. Frequency refers to the frequency of the customer's transactions. Monetary encapsulates the purchasing power and willingness to pay of customers. We define the metrics as:

Recency = last transaction date - user's last transaction date
Frequency = user's total transactions
Monetary = user's total transaction value

As the first step of our analysis, we create the recency, frequency and monetary metrics based on user-level transaction data. The calculated RFM metrics have the following distribution:



Recency

Frequency

Monetary

Figure 4. Distributions of Recency, Frequency and Monetary Values

As shown above, the customers' recency values are approximately uniformly distributed. The majority of the customers have only purchased once (frequency = 1) at GStore. Finally, the monetary values are right skewed.

As the next step, we rank the customers using each RFM metrics in an ascending order. While we bucket the recency and monetary metrics into 4 groups, following the standard approach, we bucket the frequency metric into 2 groups as most of the customers have only frequency value of 1. The following table summarizes the method with which we rank the customers RFM values:

Recency	**Frequency**	**Monetary**
R-Tier-4 (most recent)	F-Tier-2 (most frequent)	M-Tier-4 (highest spend)
R-Tier-3		M-Tier-3
R-Tier-2		M-Tier-2
R-Tier-1 (least recent)	F-Tier-1 (only one transaction)	M-Tier-1 (lowest spend)

Table 3. FRM Tier Summary

With the RFM rankings ready, we create a criteria of segmenting customers using RFM rankings. We create a total of 7 customer segments: best, loyal, highest paying, promising, at risk, newest and churned. The following summarizes the criteria with which we create the customer groups (note that X indicates any feasible value):

1. **Core - Best Customers**	RFM Score: 424
Highly engaged customers who have bought the most recent, the most often, and generated the most revenue.	
2. **Loyal - Most Loyal Customers**	RFM Score: X2X
Customers who buy the most often from your store.	
3. **Whales - Highest Paying Customers**	RFM Score: XX4
Customers who have generated the most revenue for your store.	
4. **Promising - Faithful Customers**	RFM Score: X22, X21
Customers who return often, but do not spend a lot.	
5. **Potential – At Risk Customers**	RFM Score: 124, 224
Customers who purchased often and big but not recently.	
6. **Rookies - Newest Customers**	RFM Score: 41X
First time buyers on your site.	

7. **Churned -Once Loyal, Now Gone**	RFM Score: 11X
Great past customers who haven't bought in a while.	

Table 4. Customer Groups by RFM segmentation

Using these criteria, we segment the customers into 7 groups and generate the following tree map that visualizes the customer groups:

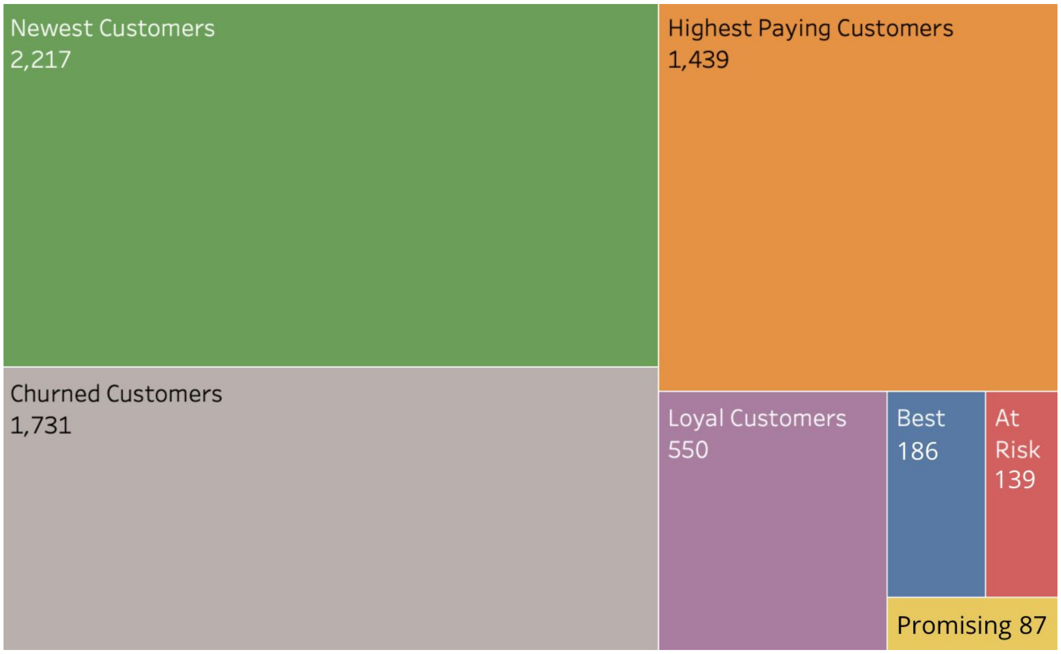


Figure 5. Treemap of Customer Groups

These segment outcomes can promote actionable changes in GStore's marketing strategy. Gstore can provide the newest customers with welcome offers and onboard support to increase their interest and encourage them to become loyal buyers. For its best customers, loyal customers and highest paying customers, GStore can roll out loyalty programs, subscription tiers and cross-selling bundles to increase monetization of these strong buyers. There are also two interesting segments to which GStore should allocate special attention. The first is the at-risk customers, who purchased often and spent big amounts but haven't purchased recently. GStore should send these customers personalized reactivation campaigns to try reconnecting with these once-valuable customers. The other group is the promising customers, who return often but do not spend a lot. In this case, GStore has successfully cultivated loyalty among these customers, GStore needs to engage their interest and increase monetization by sending them product recommendations.

6. Executing the Analytics

The dataset contains both primary data (e.g. date, device and totals) and second-hand information such as “channelGrouping” and “socialEngagementType”, which can be obtained by the marketing team in Gstore. The primary data can be captured automatically once a customer browses the website while second-hand information is collected by the database about customers’ personal information gathered by the marketing team. In this way, by combining both first-hand and second-hand data, the whole dataset can be organized at the transaction level (including customers’ behavior of only browsing the website).

With the data ready, the analytics team can further take the responsibility to clean the data and execute the analytics as well as create a model to predict revenue. To evaluate the success of the model, performance metrics such as Root Mean Square Error will be applied to verify if the model has predictive power. Apart from the model, both Exploratory Data Analysis and Recency, Frequency and Momentary (RFM) analysis are also critical for the analytics. EDA verifies the 80/20 rule that 20% of customers generate 80% of revenue while RFM analysis categorizes customers into eight groups by which marketing team can decide where they should lay more emphasis. Given the model and all the analysis, the result can be passed back to the marketing team to make a strategic plan either to run the campaign or try to improve target customers’ satisfaction.

However, we encountered both technical and organizational impediments during the analytics process including big data issues and having the model less ideal than we expected. The data is collected on a transaction level, indicating a huge amount of data we need to process. Thus, it is time consuming and computational for us to analyze data. To tackle the big data issue, we introduce spark and will further use cloud to help us process big data so that it will be more efficient to run the model. Another big issue comes from the model performance. The random forest model has comparatively low predictive power, making it hard for the analytics team to present a convincing predicted revenue given by a new customer. The problem may lie in data collection such as having too much missing data, which lower the data quality and further impair the model performance. The organizational impediment comes from the potential silo between marketing and analytics team. The model itself cannot be automatically converted to actionable insights without domain knowledge, which the analytics team needs to communicate with the marketing team. As a result, we need to improve the communication between these two teams.

7. Implementing the Analytics

Based on the insights we generated regarding “city”, “browser”, “visit month” and “channel grouping” through our analysis, we provide GStore with the following recommendations to help them increase revenues. Among the channels, most of the visitors came to GStore through organic search and social, however, most of the revenues contributed by visitors came through referral or directly typing the website. Referral has the highest rate of converting traffic to sales, therefore, we recommend GStore to promote referrals especially in the segment that has low traffic to sales conversion rate to help increase the revenues.

The majority of the visitors who contributed most revenues browse GStore through Chrome, however, most of them came to GStore through organic search. Since Chrome is also Google's product, we recommend Google to add features in Chrome to make it more convenient for these high-value customers to access GStore through Chrome and make purchases.

In terms of the cities, we found that Mountain View has much more unique visitors than New York even though the population of Mountain View is much less than New York, while Mountain View's revenues are much less than New York. Our hypothesis is that Google's headquarters are located in Mountain View, it's likely that a lot of the unique visitors are Google employees. GStore should further investigate to find out why so many unique visitors are from Mountain View and why traffic to sales conversion rate is so low in Mountain View, and then improve the conversion rate based on the findings.

When we looked at the visit month, we noticed that GStore has the most revenues in every four months (April, August and December), whereas October and November have the most traffic. Due to the lack of information from the marketing team, we can only come up with the hypothesis that GStore might have certain promotions in April, August and December to drive revenues, and in October and November around Thanksgiving shopping season, customers might check the Google product in GStore, but eventually purchase the products from elsewhere with more discounts. We recommend GStore to do further research and test our hypothesis. If our hypothesis is confirmed to be true, GStore can increase promotions or discounts in October and November instead of other months to generate more revenues.

As you can see here, we are unable to convert some of our analytics results into actionable insights due to the lack of GStore's marketing information such as marketing strategies and promotion and referral policies. In order to ensure the adoption, we recommend GStore to embed its data analytics team within the marketing team to break the silos, so that the data analytics team can have more clear understanding of the market, data source, marketing strategies, promotions, etc. These marketing information along with the domain knowledge from the marketing team will help the analysts to drive meaningful and actionable insights. Without silos, analysts can also better assist the marketing team in implementing these insights, which will eventually lead to GStore's revenues increase.

8. Scaling Up

With the current dataset, what we have done is to create a proof of concept to predict the total revenue, which is a static model. However, after having the access to real-time data, we will no longer be able to run the model offline, thus we can alternatively deploy the model on cloud to map each transaction (or user) to a predicted revenue. As the cloud can handle big data, it is a better way for us to overcome the big data issue at the same time. Moreover, with the business initiative to increase sales and improve customer loyalty, a real-time model can provide the marketing team with the latest information for each customer so that they can reach out to customers more precisely and timely.

To better scaling up the analytics initiative, the way to design the database can be improved. Since it takes great efforts to do data cleaning because of a large amount of information embedded in several columns, it will be more efficient for the analytics team to carry out the model if the database is well structured and it will take less time to run the model on cloud, which is also cost-efficient. Moreover, we should break the silo between marketing and analytics teams. Once the communication is efficient, more insightful actions can be taken to serve our customers.