BUMK776: Action Learning Project

Recommendations on Digital Marketing in Google Merchandise Store

Amy ChiaJung Chan
Yu-Tung Chang
Farid Freyha
Yu-Cheng Lai
Nan Wang
Yi-Hsuan Wu

03/11/2022

Honor Pledge:
We pledge in our honor that we have not given or received any unauthorized assistance on this assignment.

Executive Summary

In this project, we focused on developing the customer value prediction model for the Google Merchandise Store, assisting the Google Merchandise team with identifying new customers with higher long-term values. We found that most of the customers are located in the United States, especially New York and California with 28 millions future revenue and 119 millions future revenue respectively. Besides, customers who are familiar with Google products are most likely to contribute to future revenue. Also, Although most of the customers are from referrals, they have little impact on future revenues. Finally, We recommend the Google Merchandise Store focusing more on chromeOS users and AndroidOS users, allocating more budget to display channel and direct channel grouping but less to affiliates and social channel grouping of referrals.

Introduction and Background

The Google Merchandise Store, an ecommerce site that sells google-branded merchandise, is interested in acquiring new customers through digital advertising. In order to achieve the ultimate goal of increasing revenues for the Google Merchandise Store through launching digital advertising to new customers, it's important for the Google Merchandise team to predict the long-term value of first time purchasers, identifying who will spend more in the long-term. What's more, Google advertising platform's bidding algorithm will be informed by conversion values that are exactly these predicted long-term values. Thus, in this project, our primary objective is to build a model that can predict the long-term value of first time purchasers for the Google Merchandise Store, providing recommendations to the Google Merchandise team.

Descriptive Analysis

In order to have a better understanding of the given dataset and the code provided at the Google Colaboratory (colab), we explored through the BigQuery Export schema. Based on our intuition about customer value, we decided to start with exploring the correlation between the variable '`future revenue`' and other variables. First, we performed exploratory research on the existing variables by running SQL queries at Google Cloud Platform (GCP). We reran all the existing variables, such as '`channelGrouping`', '`device.browser`', '`device.browser`', '`device.operatingSystem`', etc to have a deeper perspective on the sub variables under these nested queries. After finding sub variables under these nested variables and some new variables we thought would impact our initial model, we modified the existing variables by adding these new variables. However, before actually implementing these new variables, we applied a correlation regression function on colab to test the correlations between our chosen variables. We converted the correlation table into a heatmap to have a clearer visualization. Finally, after checking the correlations between the chosen variables and '`future revenue`', we dummy-coded these variables: '`device.deciveCategory`',

'trafficSource.medium','trafficSource.source','trafficSource.keyword' , 'newVisits' by assigning 1 indicating true and 0 indicating false , which have higher correlations and adding them to our model.

Exploratory Data Analysis

To further understand and investigate the target characteristics of the dataset, we installed the Sweetviz package to conduct exploratory data analysis that can quickly generate summarized reports and help create a visualized dashboard.

At first, we set *'future revenue'* as our target value (dependent variable) to conduct the exploratory data analysis. We then dropped *'fullVisitorId'* and *'firstPurchaseSessionTime'* from the datasets since these two variables may distort the relationship between independent and dependent variables.

We dropped the outliers that do not fit the model, including *'future revenue' >=10000*, '*revenue*'*>=10000*, '*visitNumber*'*>=150*, and '*productQuantityPurchases*'*>=1500*. After we re-conduct the exploratory analysis, we found that although *'Display'* only accounts for 1% of the channel grouping, it has nearly 1.3 times higher average future revenue than other channel groups. Also, the average future revenue from traffic source medium *'Referral'* is 1.8 times higher than other mediums.

After investigating the relationship between different variables and revenues, we found that '*Direct Search*' is the top channel that could lead to higher future revenue compared to other channels, followed by *'Display'.* However, when we consider each channel's overall contribution to revenues and traffic, the channel from '*Referral*' has the highest revenues and traffic volume. In addition, customers using '*Chrome*' and '*Windows*'contributed to greater revenues than other operating systems.

Model Training

To improve the performance of the initial model, except for adding new variables, we also changed the algorithm and hyperparameters of it. While the original algorithm XGBoost is commonly used, Lightgbm, the algorithm we decided to use instead of XGBoost has a better accuracy but more importantly saves more time than XGBoost. In Google's perspective, the variables that may be put into the models could keep on increasing, it is important to come up with an algorithm that is accurate but also efficient.

In order to get a better MAE, which is the main indicator for the model, we tried out different boosting type and decided to change the boosting type from "Gradient Boosting Tree" known as gbtree to "Dropouts meet Multiple Additive Regression Trees" known as dart. In addition, we tuned the original hyperparameters, including max_depth, n_estimators, reg_alpha, reg_lamba and added others such as

learning_rate, bagging_fraction, feauture_fraction, bagging_freq, bagging_seed and verbosity.

However, after we got a great MAE of about 15, we found out our Average Error is 8, which should be lower than 6. To lower our Average Error, we decreased our learning_rate from 0.005 to 0.00225 and ended up with an MAE of about 17 and an Average Error of about 5. Since the MAE of our initial model is already better than the "Naive estimate" and "Regression only" models, we did not make any changes to them but kept them as references to make sure we did not make any changes to the original dataset.

Main Takeaways and recommendations

From our analysis, we found that most of the customers are located in the United States, especially in areas like New York and California. Out of 903652 visitors, 40.36% are from the United States. 30% of said US customers are in California with $11.9 million future revenue and 14% are in New York with $28 million future revenue.

Only 10% of the customers are from mobile devices. However, we recommend more focus on AndroidOS than IOS as it has $15 million average future revenue in comparison with $1.1 million while holding 4% and 5% of customers respectively. We also noted that customers who are familiar with Google products, such as ChromeOS, are most likely to contribute to future revenues. Customers that use ChromeOS browser will bring $69 million to future revenue, even though they only count as 8% out of all customers.

Even though about half of the customers have purchased apparel. Customers who have purchased office supplies (20%), drinkware (17%), lifestyle products (8%), bags (8%) and electronics (6%) have a more positive impact on future revenue. Take electronic products as an example, customers who have purchased electronics will generate $98 million of future revenue.

Most of the customers are from referrals (47%), but it has a low impact on future revenue ($17 million) so we recommend investigating further as it wasn't translating well into revenue despite having a high customer rate. On the other hand, 16% of the customers with $60.7 million of future revenue are from direct channel grouping. Display channel is highly effective as it helps generate $40.3 million future revenue, while it only counts for 1% of channel grouping. Both affiliate channel grouping and social channel grouping attracts limited numbers of customers (<1%) that have low purchase intention. In conclusion, we recommend Google to allocate more budget to display channel and direct channel grouping and invest less budget in affiliates and social channel grouping, while further looking into referrals.

Appendix

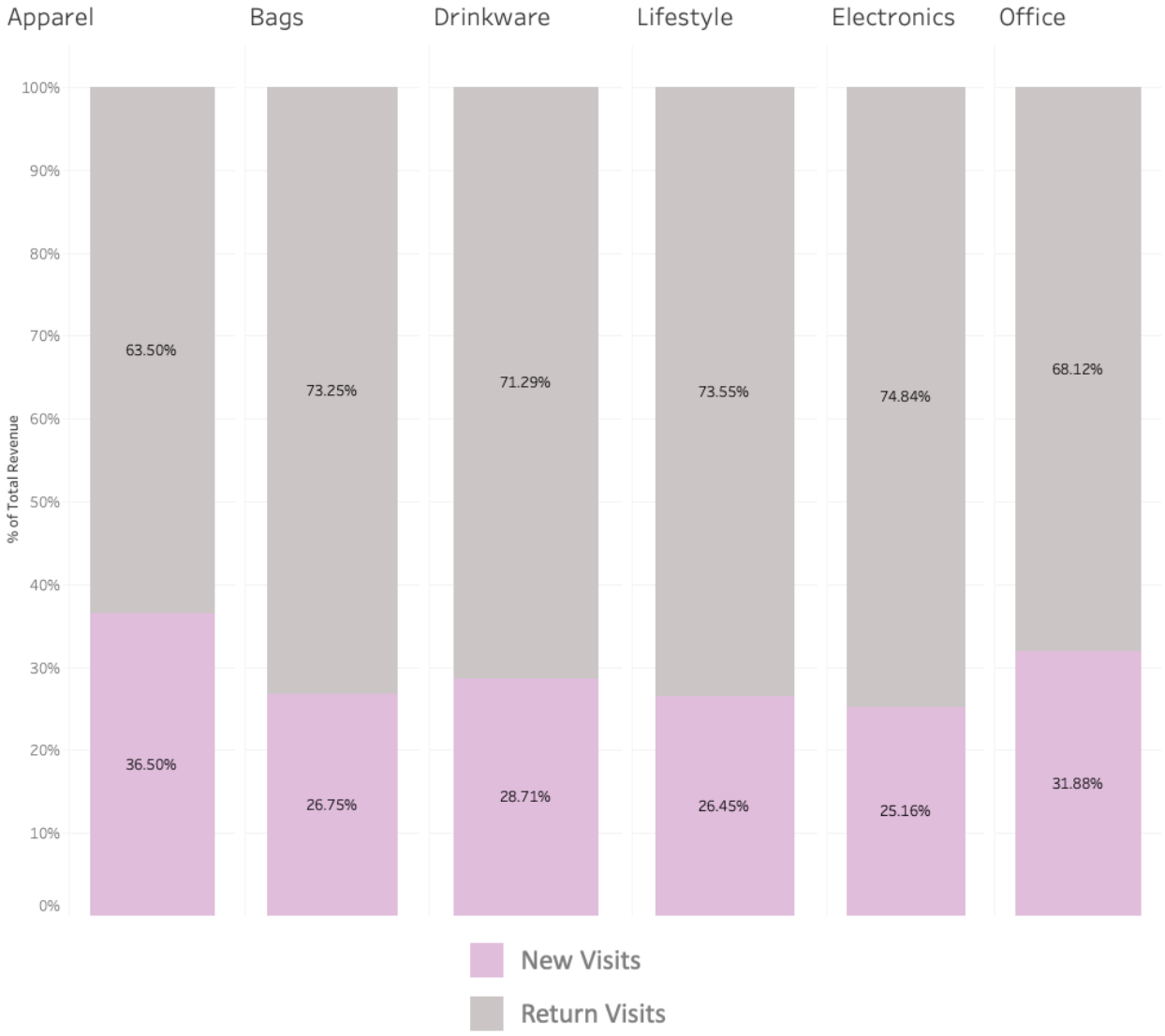
#Graph 1 - Revenue across Different Channels Point-of-Time

	Direct	Organic	Paid	Referral	Social
0	11,840	7,596	752	15,572	85.5
1	4,172	8,328	1,815	19,027	61.9
2	5,628	7,436	1,186	13,928	
3	7,573	4,761	971	6,223	255.7
4	4,500	6,074	1,363	8,443	155.6
5	3,287	4,100	223	7,219	84.5
6	1,709	2,307	157	5,194	45.0
7	2,026	1,319	260	3,306	102.9
8	1,276	1,118	287	1,582	
9	168	200	52	479	
10	737	712	133	300	
11	251	490		447	
12	1,530	912	917	5,601	
13	5,793	5,687	1,114	10,800	24.0
14	23,589	11,641	1,023	15,162	64.1
15	14,679	12,189	437	25,062	97.0
16	18,501	9,412	2,575	24,901	451.3
17	15,549	10,480	2,552	30,954	19.2
18	37,852	13,672	1,103	31,520	12.0
19	13,799	13,784	1,730	35,431	279.5
20	14,512	12,420	1,675	34,630	63.6
21	19,178	13,268	566	26,334	
22	13,388	8,863	1,057	25,881	134.4
23	8,498	8,390	823	21,575	590.1

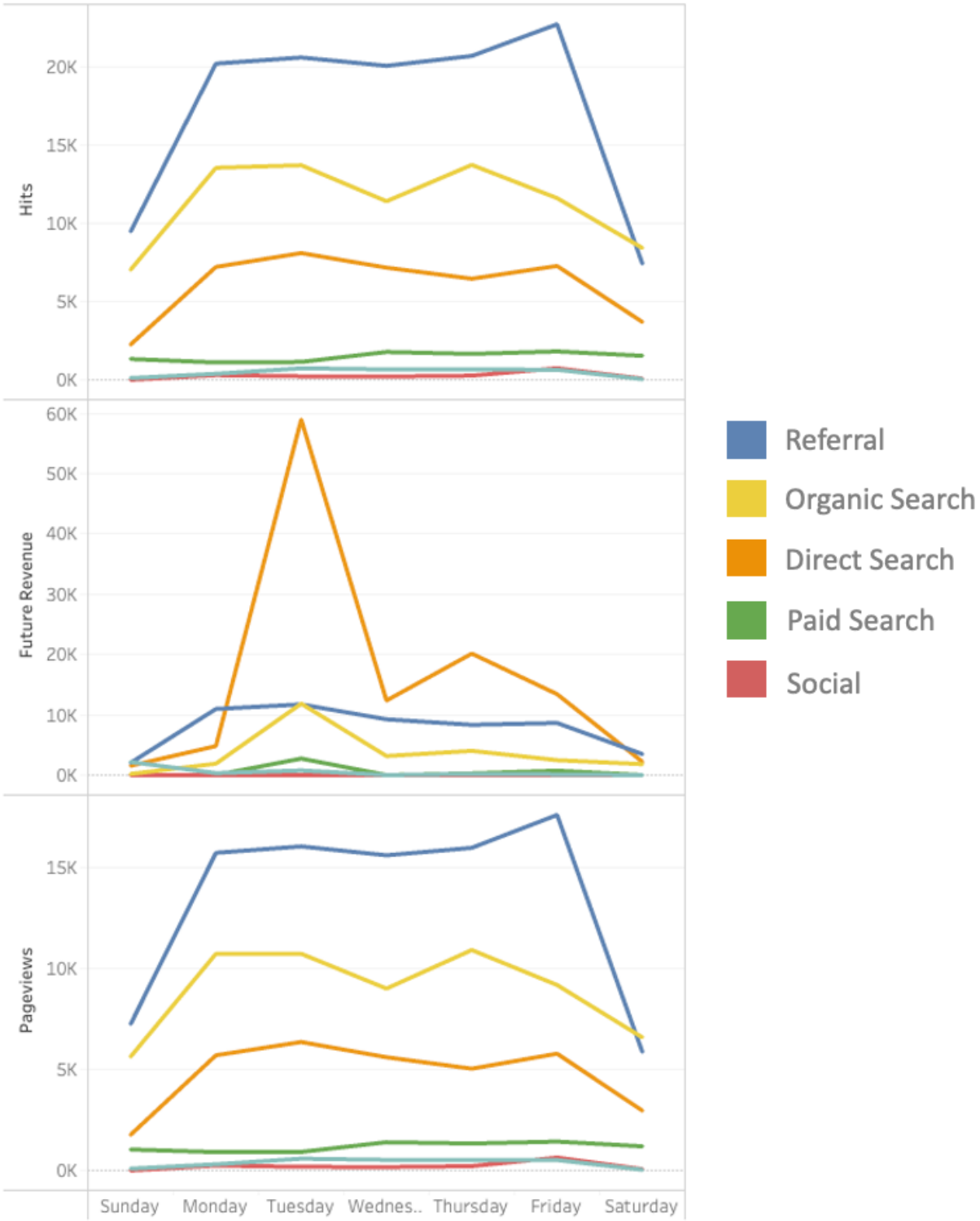
#Graph 2 - Traffic across Different Channels Point-of-Time

	Direct	Organic	Paid	Referral	Social
0	1,770	3,416	514	5,545	43
1	969	3,348	471	4,164	52
2	1,207	3,945	555	4,067	
3	1,495	2,532	437	2,686	153
4	1,227	3,216	425	3,979	95
5	972	2,172	239	2,173	44
6	334	1,163	88	1,386	39
7	371	698	133	617	28
8	665	489	226	423	
9	75	290	77	209	
10	212	282	79	188	
11	69	271		132	
12	328	699	206	933	
13	862	1,622	343	2,405	27
14	991	2,598	261	3,828	48
15	2,201	4,219	266	5,301	33
16	2,163	3,037	520	6,060	216
17	2,550	4,080	583	7,774	11
18	2,793	4,567	310	7,749	8
19	2,155	5,020	547	7,193	453
20	2,364	4,024	691	8,072	31
21	3,065	4,162	353	6,625	
22	2,564	3,617	456	6,792	55
23	1,781	3,213	467	5,595	215

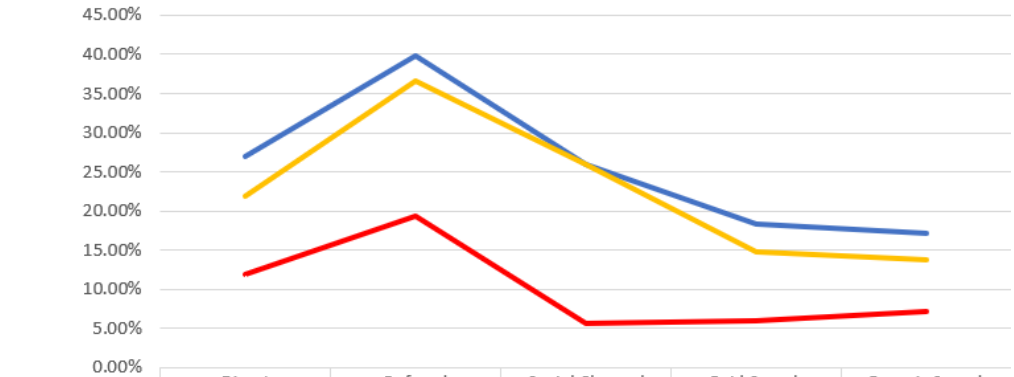
#Graph 3 - New Visits Conversion across Different Product Categories



#Graph 4 - Performance by Different Channels across Weekdays



#Graph 5- Channel Usage by Location



	Direct	Referral	Social Channel	Paid Search	Organic Search
California	27.03%	39.89%	25.93%	18.37%	17.11%
Bay Area	21.80%	36.56%	25.93%	14.84%	13.71%
New York	11.86%	19.32%	5.56%	6.01%	7.12%