CUSTOMER LIFE TIMEVALUE

GOOGLE MERCHANDISE

ACTION LEARNING

MAR 11 2022

BUMK 776
Action Learning Project

Chu-Hsuan Tsao
I-Ju Lin
Naila Sharmin
Yiling Kang
Zhaoying Ren

Introduction & Background

Predicted Customer lifetime value(pCLV) is an essential index for a company to forecast profitability, set customer acquisition budgets, and determine goals for growth and improvement. It represents the future revenue generated by newly acquired customers throughout their relationship with the business. This project aims to select the best model for Google Merchandise Store to predict the pCLV of their newly acquired customers within 90 days after being acquired. The model can enable a value-based bidding strategy for Google's clients who struggle with a low repurchase rate or want customers to take direct action on their site. Also, through this analysis, we identify important factors that are linked to higher future revenue, which could help google merchandise and other google clients target the right customers and develop a smart bidding strategy on paid search promotions.

Variable Selection

In order to build a model for customer lifetime prediction, we sum the customers' total purchase revenue over the next 8-90 days after the first purchase week as the dependent variable. We did feature engineering in the independent variables selection process by transforming the Google Analytics raw data into variables for predicting customers' lifetime value.

For variable selection, we stand in customers' viewpoint and try to determine what features might affect them to make their purchase on Google Merchandise Store, considering factors such as geography, time, action, psychology, customer browsing habits, customer action, etc. And we created these new variables based on speculation to confirm the effects in the model.

- **Geographical variables:** We are interested in identifying whether the area where most customers are located relates to future revenue or repeat purchase tendency, so we also add some area-related variables. (e.g., New York/California/Bay Area)
- **Time Variables:** Considering time may be one of the determinant elements for predicting future revenue or identifying customers characteristics, we create new time variables (e.g., week, Quarter(Q1-Q4), month, Weekday/Weekend, Daytime/Nighttime).
- **Psychology**: As psychological factors or pricing strategy might somehow affect customers' purchase or even purchase more on the website, we also add some variables related to these aspects. (e.g. Max product price/social referral/Promotion)
- **Customer Browsing Habits:** As we want to understand our customers' characteristics, we aimed to learn more about their website browsing habits. These variables could also help us understand how we can reach out to those customers. (e.g., channel grouping/operation system/browsers)
- **Customer Action-related Variables:** To better learn customers' behaviors on the website, several related variables are made. (e.g. bounce rate/product categories/frequency/ add & remove to cart).

We used Google BigQuery to look into details for each feature and check the number of records in the raw data. Dummy-coded the variables by transforming the value into 0 and 1 and dealt with missing data by making the Nan value to 0. We also match strings with their categories and turn them into dummies for product categories. After selecting variables, we examine customers' behaviors based on their first purchase and a week after the first purchase. Accordingly, put the variables into five following groups: First purchase, First purchase sessions, First purchase hits, First week sessions, First-week hits, and Future purchases (**Table1**).

Pandas profiling is employed to assist exploratory data analysis. After cleaning raw data and creating new variables, Pandas data frame 'df' has been created and imported to Google Collab. There are 53 variables in our final dataset. To build up models, we divided the final dataset into development and holdout datasets by April first in 2017.

Modeling

1. Initial model

Based on the exploratory data analysis, we found a few extreme values for variables such as futureRevenue, revenue, visitNumber, and productQuantityPurchased. For better modeling, we try to drop the records containing these outliers first (Table 2). After cleaning the extreme data, we put all variables except 'fullVisitorId,' 'firstPurchaseSessionTime' in the XGBoost regression model to generate our predicted result. Unfortunately, the R-squared is extremely low(0.1), indicating that most predictions are close to 0. As a result (Figure 1), we decided to split this model into two parts and combined them to make the final prediction more precise. The first regression model is trained on only users whose future revenue is not equal to zero, and the second part will be a classification model trained on all users predicting the probability that they will make at least one future purchase.

2. Regression model

We improve our regression model to predict future revenue more accurately by dropping more outliers and normalizing data according to the initial model result. First, we choose only users who purchase at least once in the 90-day, and we also changed the measure of the outlier to remove more extreme values without dropping too much data (Table 3). Moreover, we undergo data normalization for the numeric variables to get rid of various irregularities that can make it more challenging to interpret the data and get a properly constructed and well-functioning database.

Then, we calculate the correlations between different variables and the dependent variable future revenue, looking for the factors that may influence the future revenue. After various attempts, we decided to regard 0.16 as a relatively high correlation level and chose variables with higher than 0.16 correlation in our regression model to predict future revenue (Table 4). We tried to use various models for the regression model choice, such as LinearRegression, DecisionTree, XGBoost, RandomForest, and Gradient Boost. Based on the model prediction accuracy (RMSE & R-squared), we finally choose XGBoost as our regression model. To further improve our regression model, we also tried to change the hyperparameters like learning rate (Table 5) and finally got the better and reasonable Training and Testing R-squared (Training R-squared:0.58 Testing R-squared:0.36) (Figure 2).

3. Final Predict model

We also applied different types of machine learning methods for the classification model. Finally, we chose the Gaussian Naive Bayes one with the relatively high ROC(0.587) (Figure 3) to use in the final prediction. However, using the classification model prediction multiplying with average future revenue or regression model prediction, we found that both result in substantial prediction errors and RMSE (Figure 3). Thus, we use the average repeat purchase rate as an alternative to present the probability that people will purchase in the future. And we use [Regression Model Prediction] * [Average Repeat Purchase Rate] as our final model to make future predictions. By running the final model, the result shows that the regression only model trained on repeat purchasers performed much better than the naive prediction, with a significant decrease in MAE from 31.59 to 18.76 (Figure 3). Meanwhile, the final model also has a lower RMSE(69.25) and absolute average error(3.7) compared to the naive estimate, which means less prediction bias (Figure 3). To conclude, using the regression only predict model, we successfully reduce the MAE by more than 40%, allowing the model to predict the 90-day revenue of each new acquisition more accurately. We also identified the determinant factors for future revenue, assisting Google Merchandise Store to boost revenue in the future. More insights are generated for companies leveraging Google ads to optimize the digital advertisements toward driving acquisitions with higher future revenue.

Conclusion & Insights

Acquiring customers with higher pCLV is a challenge for not only Google Merchandise Store but also clients that utilize Google Ads & Google Analytics for online promotion. We developed a few insights through the analysis process that will benefit those clients to boost revenue. In our final model, we identified that the below 14 variables have a mentionable impact on predicting future revenue; therefore, we came up with the following recommendations.

'directChannelGrouping', 'productQuantityPurchased', 'officePurchased', 'chromeOS', 'hits', "Q4", "firstWeekRevenue", "timeOnSite", 'electronicsPurchased', 'Maxproductprice', "revenue", 'pageviews', 'lifestylePurchased', 'referralChannelGrouping'

- **Focus on providing special offers to specific product categories - office goods, lifestyle items, and electronic products**

According to our analysis, specific categories such as Office goods, lifestyle items, and electronic products from Google Merchandise Store have more probability of generating higher future revenue, and investing more in these specific categories buyers can pay off in terms of increased revenue. Google can provide offer-bundles for office, lifestyle, and electronic products and feature deals of the month for high-selling items from these categories in Q4.

- **Targeting the “Chrome” user base**

Visitors using Chromebook devices or having ChromOS as their default OS is highly related to generating higher future revenue. This is a self-explanatory fact as people owning Chromebooks or using the ChromOS operation system might be already loyal users of Google-branded products and more likely to purchase from Google merchandise stores. Even though this might be a niche segment for Google to focus on, prioritizing this specific user segment for targeted promotion and advertising can increase future revenue.

- **Efficient allocation of ads budget by focusing more on direct visitors**

One fascinating insight that we found is that visitors coming to the site by clicking through links (referralChannelGrouping) displayed on the other sites (not social media rather blog/article or content websites) do not contribute to future revenue. Even though these visitors clicked on the link that shows some sort of interest on their behalf, such visitors do not purchase or contribute to the revenue in the long run. So Google Merchandise can focus more on visitors who search directly for the store or have pre-determination on visiting the site (directChannelGrouping) rather than randomly clicking on links that show up on other websites to increase future revenue.

- **Provide pop-up promotions/coupon when users have interacted with the website to a certain threshold level**

Our analysis indicates that higher user interaction with the website leads to increased future revenue. People who spend more time on in-general site exploration, clicking on products pictures/descriptions, and different category tabs are more probable to purchase. Providing pop-up coupons/promotional discounts/offers to encourage them to click on more pages and stay longer on the website.

- **Increased allocation of Advertising budget on the 4th Quarter**

Our analysis indicates that customers who are acquired in the 4th quarter are more likely to generate higher revenue in the future than the other three quarters. The fourth quarter consists of holiday seasons, e.g., Christmas, Halloween, Thanksgiving, Black Friday, Cyber Monday, etc. So, for eCommerce sites, intensifying promotions at this specific time of the year would attract potential customers to the site.

- **Targeting high spenders and buyers of expensive products**

Our analysis indicated that those who have bought relatively expensive products are more likely to contribute to the site's future revenue among all the purchasing customers. So, businesses can segment their customer base based on their share of wallets. Those spending a significant amount on expensive products are definitely more valuable to the company for maximizing future revenue. Promoting products that have a similar expensive price range to this customer base or offering them personalized deals for such products on-site can be a recommended way to retain this profitable segment.

An automated Smart bidding strategy is most advisable for eCommerce sites like Google Merchandise store, where clients can set up contextual signals like device operating system, time of the day/week, etc. Such smart bidding considers whether a user has browsed a product during a previous site visit, is on a loyalty program list the client has uploaded or has a profile similar to existing customers.

Limitation:

Although we developed a well-performed prediction model for this analysis project, there are limitations to this model. Given that not many people are motivated to buy branded merchandise repeatedly, Google Merchandise Store only obtained 7.2% of newly acquired customers who have repurchase behavior in the first week after their first purchase. It is hard for us to make an accurate prediction based on only 7% of the data. Since they did not revisit the website, about 93% of customer data cannot be put in our pCLV prediction model. This fact might influence the accuracy of our prediction.

Appendix

- Table 1:

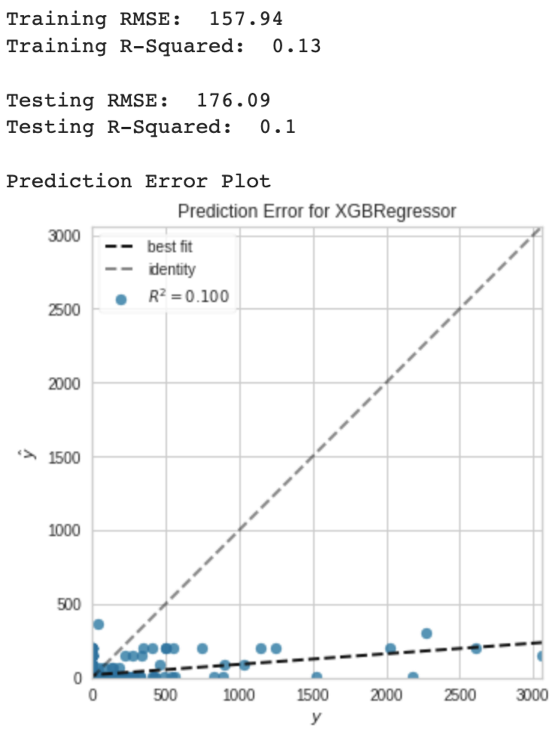
Groups	**Variables**
First Purchase	● fullVisitorId ● firstPurchaseSessionTime
First purchase sessions	● fullVisitorId ● revenue ● visitNumber ● hits ● pageviews ● timeOnSite ● referralChannelGrouping ● organicSearchChannelGrouping ● directChannelGrouping ● paidSearchChannelGrouping ● chromeBrowser ● safariBrowser ● macintoshOS ● windowsOS ● chromeOS ● LinuxOS ● california ● newYork ● bayArea ● Weekend ● TrafficSourceTrueDirect ● Q1-Q4 ● Month

	● Week
First purchase hits	● fullVisitorId ● productQuantityPurchased ● apparelPurchased ● officePurchased ● drinkwarePurchased ● lifestylePurchased ● bagsPurchased ● electronicsPurchased ● BrandAndroidPurchased ● BrandGooglePurchased ● BrandYoutubePurchased ● addedItemtoCart ● HasSocialSourceReferra
First week sessions	● fullVisitorId ● firstWeekVisits ● firstWeekTransactions ● firstWeekRevenue ● bounceRate ● Frequency
First week hits	● fullVisitorId ● removedItemFromCart_week ● addedItemtoCart_week ● HasSocialSourceReferral_week ● Promotion
Future purchase	● fullVisitorId ● future revenue

- Table 2

Variable name	**Drop value**
futureRevenue	>=10000
revenue	>=10000
visitNumber	>=150
productQuantityPurchased	>=1500

- Figure 1



```
Training RMSE:  157.94
Training R-Squared:  0.13

Testing RMSE:  176.09
Testing R-Squared:  0.1

Prediction Error Plot
```

- Table 3

Variable name	**Drop value**
futureRevenue	>=10000
revenue	>=10000
visitNumber	>=500
productQuantityPurchased	>=1100
hits	>=380
pageviews	>=400
timeOnSite	>=15000
futureeRevenue	==0

- Table 4

Varaiables	**Correlation with future revenue**
revenue	0.362663
hits	0.187032

pageviews	0.200819
firstWeekRevenue	0.194428
Maxproductprice	0.277352
directChannelGrouping	0.274020
chromeOS	0.207824
officePurchased	0.235239
lifestylePurchased	0.187917
electronicsPurchased	0.260611
productQuantityPurchased	0.541093
Q4	0.168452
timeOnSite	0.168338
referralChannelGrouping	-0.196190

- Table 5

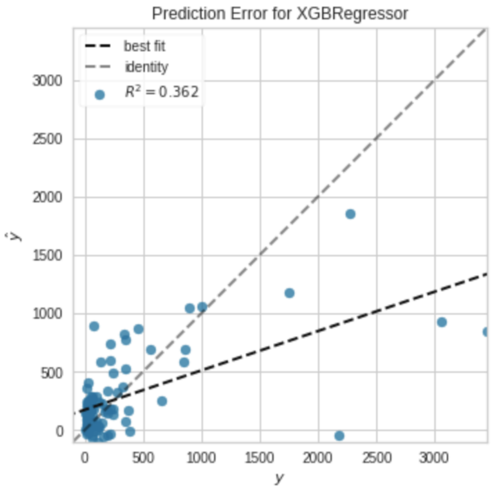
hyperparameter	**value**
max_depth	1
n_estimators	155
reg_alpha	700
reg_lambda	29
objectvie	'reg:squarederror'
booster	'gbtree'
random_state	123
learning_rate	0.45

- Figure 2

```
Training RMSE:  360.71
Training R-Squared:  0.58

Testing RMSE:  481.13
Testing R-Squared:  0.36

Prediction Error Plot
```



- Figure 3

```
clasification report:
              precision    recall  f1-score   support

           0       0.93      0.93      0.93      1125
           1       0.24      0.25      0.24       102

    accuracy                           0.87      1227
   macro avg       0.58      0.59      0.59      1227
weighted avg       0.87      0.87      0.87      1227


confussion matrix:
[[1045   80]
 [  77   25]]

ROC AUC:
0.5869934640522876
```

```
Naive MAE:  31.59
Naive RMSE:  70.2
Naive Avg Error:  10.83

Initial Model MAE:  26.55
Initial Model RMSE:  72.86
Initial Model Avg Error:  -5.54

Regression Only Model MAE:  18.76
Regression Only Model RMSE:  69.25
Regression Only Model Avg Error:  -3.7

Classfication Only Model MAE:  30.98
Classfication Only Model RMSE:  102.72
Classfication Only Model Avg Error:  10.21

Predict_model_MAE MAE:  37.37
Predict_model_MAE RMSE:  141.6
Predict_model_MAE Avg Error:  16.5
```