Customer Revenue Prediction using Google Analytics Data

Barathwaaj Parthasarathy[1],Pravin Sundar[1],Prashanth Thirukkurungudi Sekar[1]

Abstract
This Project is inspired from a live Kaggle contest on predicting the revenue of customers using Google Analytics data from a Google merchandise store. The modified project aim is three-fold, identifying revenue generating customers, predicting the revenue generated by the identified customers and finally computing the life time value of the said customers.
After processing and understanding the data through Exploratory Data Analysis, the task of identifying the revenue generating customers follows. This has been done by using Logistic Regression, Random Forest and Gradient Boosting Classifiers.
The second phase of the project involves using Linear, Random Forest and Gradient Boosting regressions to predict the revenue.
In the final phase, the Customer Lifetime Values are calculated using RFM(Recency, Frequency, Monetary) methods.
To evaluate the classification models AUC and Recall scores are employed. Similarly, for the regression models, Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) are used.

[1] *Data Science, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*

Contents

Introduction

In the current e-commerce world, not every customer produces significant amount of revenue to the business and a thumb rule of 80-20 holds good. This poses a significant challenge for marketing teams to understand the revenue generating base of the customers.

Deriving actionable insights from Google Analytics would help or enable them to optimize their re-targeting and re-marketing strategies and eventually improve their Return on Investment (ROI) from marketing. This can be scaled for those companies who choose to use data analysis on top of Google Analytics [1].

We will be deploying classification and regression algorithms (Supervised Learning) to identify revenue generating customers, forecast the revenue generated by the identified customer and finally formulate their lifetime value.

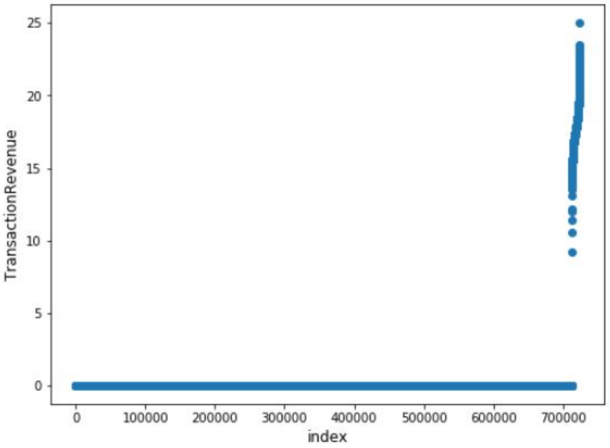
1. Background

Google Analytics is a free web analytics tool offered by Google to help analyze website traffic. For most companies, their website serves as a hub for all of the digital traffic. If any marketing activities such as search ads or social media ads are run, the users most likely visit the website somewhere along their user journey.

Given that the company's website is the central hub of their digital presence, the website is the best way to give them a holistic view of the effectiveness of all the campaigns they are running to promote their product/services online. Google Analytics is a free tool that can help them track your digital marketing effectiveness. Since this project is inspired by a live Kaggle contest [2], there is some work done on the revenue prediction phase of the project.

2. Exploratory Data Analysis

The extracted data is a transaction level data of Google Analytics belonging to the E-Commerce domain. The data contains 903653 transactions and 55 features. Out of these 55 features, 12 are numerical and 43 are categorical(including binary).
Upon first look, 12 features had more than 50 % missing values and these were removed from the feature set. Apart from this, 18 single level features were also dropped.
For the classification aspect of the project the data was grouped by visitor ID and for the target variable, a new flag was created for the customers who generated revenue.



As it can be seen from the above graph,there was imbalance in the data with 98.6 % of records belonging to a single class(0) [3].
Synthetic Minority Oversampling Technique would be performed during the feature engineering phase to overcome this issue of imbalance [4].

3. Identified Business Questions

After analyzing the data, we have chosen three linked business cases which would ultimately provide actionable insights to optimize marketing spends

1. Three binary classifier models were used to predict or identify customers who generate revenue
2. Using the results from the above business case, Regression models were used to forecast the revenue generated by the identified customers
3. The lifetime value of customers were predicted (or) in other words, the net profit attribute of the customers were evaluated to enable efficient re-targeting and remarketing strategies

4. Baseline Models

The results of the baseline models for the first phase classification with their performance are tabulated below

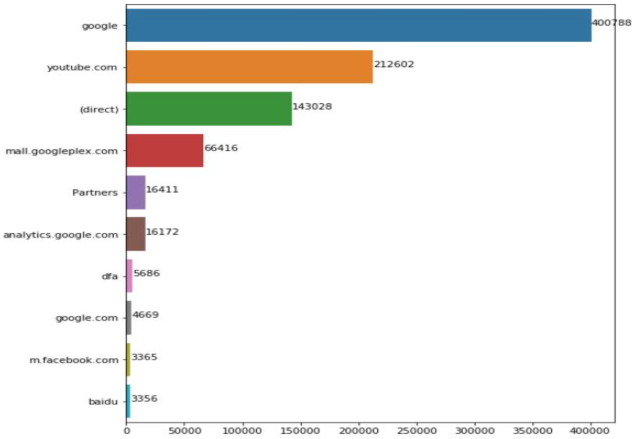
Model	AUC	Recall
Logistic Regression	0.964	0.126
Random Forest	0.911	0.361
Gradient Boosting	0.986	0.303

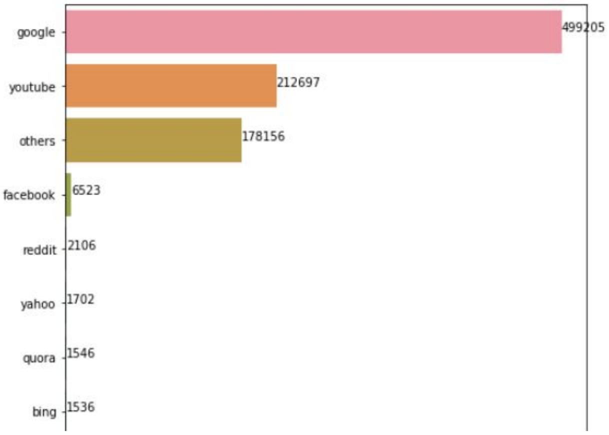
Table 1. Baseline Models

As seen above, high AUC scores were obtained from the baseline models.
But, the more important question is to find out who amongst all the customers are valuable, i.e. who generate revenue. The recall value of the predictions should considered since the proportion of revenue generating customers is very low and sometimes a revenue generating customer will be predicted as non profitable. The recall gives an indication of only the profitable customers which is the main objective.
Henceforth, the recall values would be the primary evaluation metric for the classification models.
As seen from Table 1, Random forest has the best recall value. But this is only the baseline model performance, feature engineering and hyper-parameter tuning would be performed to improve the performance of the models.

5. Feature Engineering

Scope for reducing the levels of few features was identified since these features contained high number of levels. Features like 'browsers', 'source', 'device os' and 'network domain' values were grouped together based on business understanding. For example, 'trafficSource' was modified by combining similar levels together. As it can be seen below, various levels that contained extensions of '../google/..' were combined as 'google'.





New features were created by joining two other features by means of business understanding. For example, ’Traffic Source’ and ’geonetwork country’; ’browser’ and ’OS’; ’browser’ and ’device category’ were combined to form new features that were used for the models.

Features which were of no use to the first phase models were removed from the feature set. Categorical features where factorized and added to the feature set. Synthetic Minority Oversampling Technique(SMOTE) was performed to get a better split. After this, a target variable split with 16.66 % was achieved.

6. Phase 1 - Revenue Generating Customers

Classification

After wrangling and cleaning was performed for the features, the model performance has improved with respect to recall scores:

Model	Recall
Logistic Regression	0.839
Random Forest	0.885
Gradient Boosting	0.834

Table 2. Tuned Models

As it can bee seen from performance of the models, Random forest has the highest recall value and a corresponding lowest false negative.

For the first business question, the revenue generating customers have been successfully identified.

Inference: The best model captured 88.5% of customers who are actually revenue generating. This resulted in a loss of approximately 7% of the total revenue.

Going forward in the project, the results of the Random Forest model were used to answer the other business questions.

7. Phase 2 - Customer Revenue Prediction

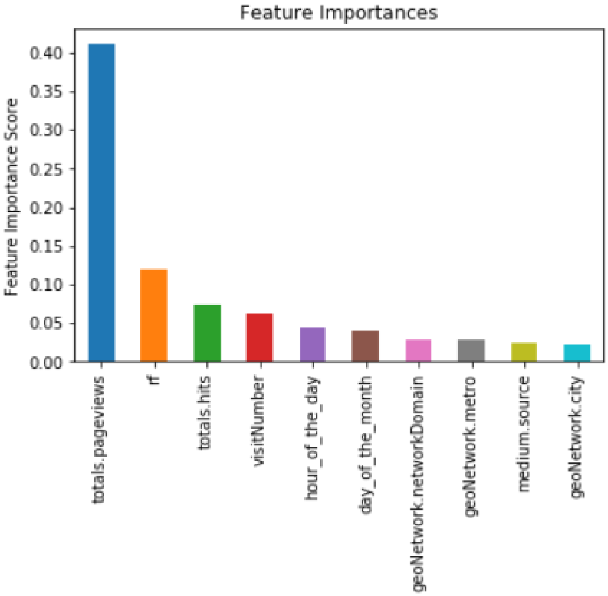
After classifying the customers as profitable or non profitable, the revenues generated by these customers were predicted. Here, the data was reverted back to the initial transaction level data to train the models. Linear Regression, Random Forest Regression and Gradient Boosting Regression models were used to make the predictions.

The results of the classification in the first business case was incorporated as an independent feature to the feature set. The time-stamp was also split into day of the week, hour of the day and day of the month to capture the trend and seasonality present in the data. Month was not used as a feature as the transactions were restricted to a year and adding this feature would be unwarranted.

After the models were trained, the revenues of the customers identified as profitable in the first phase of the project were predicted. Given the range of the predictor variable (Revenue generated), Root Mean Squared Error (RMSE) of the log of predicted revenue is a better estimate to evaluate the models when compared to Mean Absolute Percentage Error.

Model	RMSE
Linear Regression	1.804
Random Forest Regression	1.647
Gradient Boosting Regression	1.625

Table 3. Phase 2 - Regression Models



Inference: The newly created features were ranked amongst the most important features.

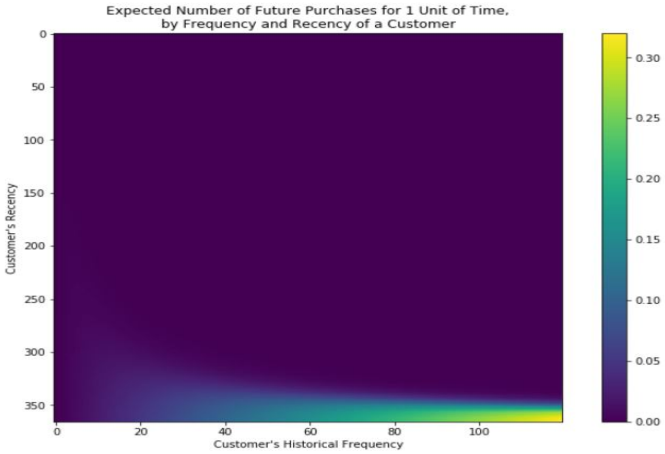
The Gradient Boosting Model predicted a total revenue of 2.7 billion dollars of revenue generated over the analysis period.

8. Phase 3 - Customer Lifetime Value Prediction

In E-Commerce business, customers go away silently without a necessity for them to inform the business. There arises a need to model (or) predict the customer's lifetime value by looking at the customer's last transaction, the revenue generated through the transaction and the frequency of the transactions [5].

Only 9% of the customers had made more than one purchase. Two models : Betageofitter (BG/NBD) model considering the frequency and recency of the customers followed by GammaFitter model which takes into account the frequency and monetary value of the customers [6].

First, a Betageofitter (BG/NBD) model was developed to identify the number of future purchases that would be made by the customer.

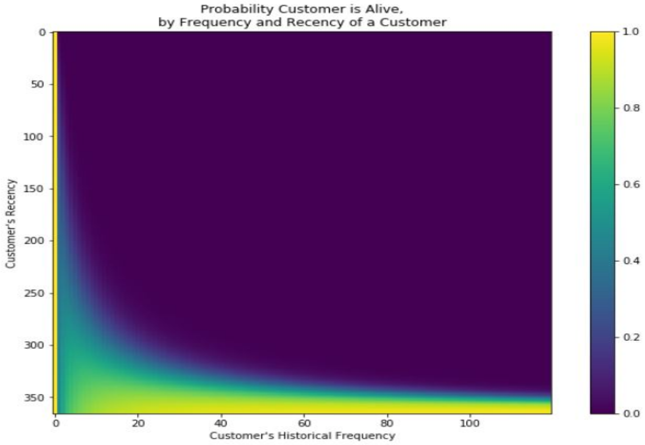


The above graph implies the following :

- A customer who has purchased a lot and most recently is the most important/profitable customer (Bottom right corner)
- A customer who has purchased a lot but not recently have probably stopped purchasing or churned (Top right corner)
- There are other customers who buy infrequently (with recency value between 40-300)

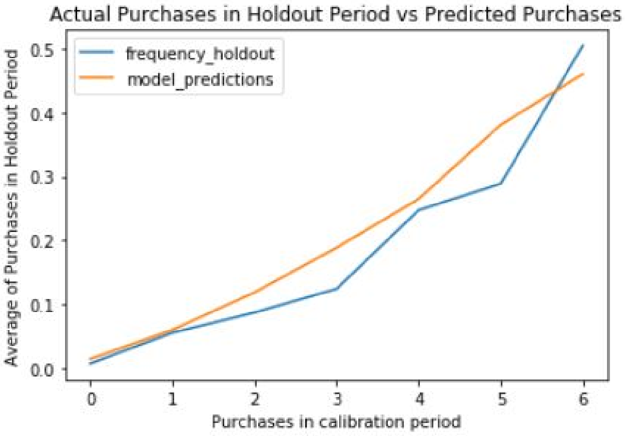
A customer who has a frequency (represents the number of repeat purchases the customer has made) of 120 and recency (represents the age of the customer when they made their most recent purchases. This is equal to the duration between a customer's first purchase and their latest purchase) value of 350 is the most profitable customer.

Predicting which customers are definitely alive :



A customer who has purchased recently are surely active customers. On the other hand, customers who purchased a lot but not in the recent past are likely to have become inactive.

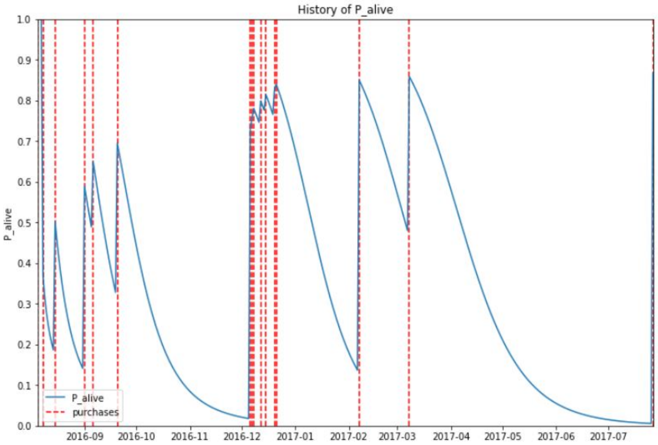
Before proceeding with the GammaFitter model, the Betageofitter model was evaluated using a holdout dataset with last three months of the customers transactions.



It can be observed that the model predicts pretty accurately for purchases ranging from 0 to 5 but over-estimates for 3 and 5 purchases.

Next, the GammaFitter model was developed by taking into account the frequency and monetary value of the customer and eventually utilizing the results obtained in the previous model to evaluate the value of the customer for a period of 12 months through the discounted cash flow method[7].

Purchase pattern and predicted value of the most valuable customer :



Even though this customer was not a frequent visitor, the customer brought in substantial value to the business and the predicted value of the customer for the next 12 months was 16 million.

9. Summary and Conclusions

9.1 Insights and Analysis

Through this project, by analyzing Google Analytics data, using various Machine Learning concepts, three specific business questions have been answered.

1. Who are the revenue generating Customers?
2. How much revenue are they generating?
3. What is their lifetime value to the business?

From this project the top prospective customers based on their revenue generating potential, their revenue sum and their total value with regard to the business were identified. This monetary based ranking analysis (Customer Lifetime Value) enables the business to optimize their re-targeting and remarketing strategies and eventually improve their Return on Investment (ROI).

9.2 Challenges and Improvement

Challenges:

1. The data acquired contained transactions only for a year which was restricting the analysis from capturing month level seasonality trends. This could be overcome by obtaining data externally to extend the period of analysis.
2. The underlying challenge of E-Commerce business with a very small proportion of the actual visiting customers generating revenue, creates an imbalance in the dataset. As seen in the analysis techniques like SMOTE were incorporated to overcome this challenge.

Improvements:

1. Given the size of the data, dynamic hyper-parameter tuning using GridSearchCV could not be performed due to lack of computational resources. With sufficient power and resource, effective hyper-parameter tuning may improve the performance of the models
2. In order to tackle the curse of imbalance in dataset other techniques like Cost Sensitive Learning and cluster based sampling could be incorporated to order to efficiently balance the data to obtain maximum performance of the models
3. Finally, more sophisticated models like Neutral Networks and Extreme Gradient Boosting could be implemented and further stacking of models could result in better predictions and classifications.

Acknowledgments

The success and outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have got this all along the completion of the project. All that we have done is only due to such supervision and assistance and we would not forget to thank them.

We respect and thank **Dr. Gregory J Rawlins** for taking us through the concepts of Applied Machine Learning and help us gain knowledge and understanding through his interactive lectures. We are also grateful for providing us with an opportunity to work on this project, **"Customer Revenue Prediction using Google Analytics Data"** which gave us a chance to explore the real-world applications of the concepts learnt.

We owe our deepest gratitude to our project guide **Christopher Torres-Lugo**, who took keen interest on our project work and guided us all along. He was always open for questions and clarifications during his office hours whenever we ran into trouble or had a question regarding our project. He consistently allowed this project to be our own work but steered us in the right the direction whenever he thought we needed it.

We would also like to extend our gratitude to **Manoj Joshi** and **Zeeshan Ali Sayyed** for their encouragement and more over for their timely support and guidance throughout the coursework.

References

[1] **Google Marketing Platform** – https://marketingplatform.google.com/about/analytics/

[2] **Kaggle Live Contest** – https://www.kaggle.com/c/ga-customer-revenue-prediction

[3] **Exploratory Data Analysis** – https://www.kaggle.com/sudalairajkumar/simple-exploration-baseline-ga-customer-revenue

[4] **Sampling Techniques** – https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

[5] **Customer Lifetime Value** – https://towardsdatascience.com/whats-a-customer-worth-8daf183f8a4f

[6] Peter S. Fader, Bruce G. S. Hardie, Ka Lok Lee. 2005. "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science* Vol. 24, No. 2

[7] **Discounted Cash Flow** – https://en.wikipedia.org/wiki/Discounted_cash_flow