# **Executive Summary**

Google Digital Marketing acquires new clients everyday, including Fortune 500 companies. These companies determine the financial value of new customers in order to help them to find balance in terms of short-term and long-term marketing goals and demonstrate a better understanding of financial return on your investments. Google needs to optimize a general training pipeline in order to use it for their clients. Finding the customer lifetime value (CLV) can be critical for companies because this can improve forecasting, customer retention, drive repeat sales, and many more benefits. CLV measures how much a business can plan to earn from the average customer over the course of the relationship. Using Big Query for the data, we were able to query the specific data needed to calculate the CLV. Colab was used to execute the Machine Learning (ML) models, more specifically Regression and Classification. The key results were that building a custom neural network and hyper parameter tuning the model had an ROC AUC score of 0.6, which was our best model. Our baseline of this classification method was 0.5 and the limitations of this model was the feature selection. While the regression model did not perform the best, the testing R-Square was 0.43, meaning that the model was only accurately predicting the future revenue about 43% of the time. The biggest limitation was the independent variables. Majority of the variables used were not highly correlated to the dependent variable, future revenue, but were the most correlated of all the fields in the dataset. The team believes that having more features that are highly correlated can potentially improve these models.

# **Background**

In this era, online marketing platforms have been on the rise in all shapes, forms, and sizes. From this scenario, customer lifetime value has become a necessary idea as it represents the total amount of money a customer is expected to spend in a business during his/her lifetime. There are several ways to predict customer lifetime value such as aggregate model and machine learning model. In this project, we chose a machine learning model because it provides a complete overview of the customers and other factors.

# **Problem Statement**

Similar to acquiring new customers, retaining existing customers is essential but costly. After companies acquire new customers, it is important to keep profitable relationships with customers. However, companies do not need to retain every single customer as some customers are loyal customers creating more value but some customers purchase only one product or service. At the same time they need to make sure that they will not overspend or underspend on marketing efforts. To focus on profitable customers and identify such groups of customers, companies need to predict the lifetime value of customers based on many factors such as purchase frequency, retention rate, churn, and profit margin. This can help businesses to sustain and enable companies to outperform their competitors. In addition, the companies will know where customers stand and the market appropriately, drive repeat sales and encourage higher-value sales, improve customer retention and foster brand loyalty, and ultimately increase profitability.

In this project, we scope the problems on Google Digital Marketing acquiring new customers everyday and finding how google smart programs can drive performance and achieve client business goals. We have 4 assumptions as follows;

1. Big query helps us identify insights through Exploratory Data Analysis
2. Based on big query, we can develop Machine Learning Models to predict customer lifetime value.
3. We can get the best model and apply them to Google Product Suite Applications
4. We can identify features used to predict customer lifetime value for the future projects or other clients

# **Methods**

**Overall approach:**

The main objective as described above is to Predict the customer lifetime value and whether the customer will purchase again within the next 90 day of his first purchase. Our approach was to use regression analysis for predicting the customer lifetime value and classification to predict the repeat purchase.

Our final prediction can be one of three options:

* If our regression model is good but our classification model is bad, the prediction will be [Regression Model Prediction] \* [Average Repeat Purchase Rate]
* If our regression model is bad but our classification model is good, the prediction will be [Average futureValue] \* [Classification Model Prediction]
* If our regression model is good and our classification model is good, the prediction will be [Regression Model Prediction] \* [Classification Model Prediction]

**Datasets:**

The sample dataset contains Google Analytics 360 data from the Google Merchandise Store, a real ecommerce store. This dataset contains transactions on the google merchandise store for the year 2017.

The Google Merchandise Store sells Google branded merchandise. The data is typical of what you would see for an ecommerce website. It includes the following kinds of information:

* Traffic source data: information about where website visitors originate. This includes data about organic traffic, paid search traffic, display traffic, etc.
* Content data: information about the behavior of users on the site. This includes the URLs of pages that visitors look at, how they interact with content, etc.
* Transactional data: information about the transactions that occur on the Google Merchandise Store website.

The dataset was stored on Google BigQuery data warehouse and was queried using the BigQuery client in Google Colab notebook.

we dropped outlier in product quantity purchased and total transactions revenue by dropping quantities greater than 1500 and revenue greater than 3000. The sample dataset looks as follows:

*A picture containing text

Description automatically generated*

*Fig. 1*

**Analytical Models:**

1. **Regression analysis:**

We used the Xgboost Regression model for predicting the customer lifetime value.

after dropping the outlier as shown in the following figure, we checked the correlation between target and independent variables and only kept highly correlated variables in the model :

Text

Description automatically generated

*Fig. 2*

The following figure shows the correlation heatmap :

Graphical user interface, application

Description automatically generated

*Fig. 3*

The following figure shows the Xgboost model fitted to the training data ;

Text

Description automatically generated

*Fig. 4*

The prediction R2 is as shown in the following figure:

Chart, scatter chart

Description automatically generated

*Fig*. 5

We tested multiple models and hyperparameter tuning each model with different parameters to optimize the model performance. and we were able to increase the r2 up to 40%. We think the absence of the better features that can explain the variance in the target variable is the main cause of such model performance.

1. **Classification:**

Next we tried classification using xgboost classifier for classifying the repeat purchase of the customer i.e. whether the customer will purchase again within 90 day of the first purchase. after creating a target variable with future revenue > 0 and then classifying it as repeat Purchaser. we dropped future revenue from the dataframe and kept most relevant features.

data preparation :

Text

Description automatically generated

*Fig. 6*

Addressing the Class imbalance issue with smote and random under sampling :

Graphical user interface, text, application, email

Description automatically generated

*Fig. 7*

Chart, pie chart

Description automatically generated

*Fig. 8*

Xgboost Classification Model :

Text

Description automatically generated

*Fig. 9*

Classification Report :

Table

Description automatically generated

*Fig. 10*

After multiple iteration with tuning the classification models we were able to increase the ROC AUC score by 20%. We think the absence of the better features that can explain the variance in the target variable is the main cause of such model performance.

# **Results and Conclusions**

We built up a model through XGBoost and achieved a ROC score of 0.6. The Models can be generalized on other datasets for predicting the customer lifetime value and whether the customer will purchase in the next 90 days. Google teams can take the hyperparameters from our best model and apply it to Google Product Suite Applications such as Search Ad 360, Tag manager 360 and Data Studio etc. We also found out a few topics that might be profitable to discuss, for example: the purchases usually happened on weekdays and man’s google products have much more views. However, this dataset that we used to build the model has a huge data imbalance issue which will cause bias or skew. Only 7% of customers are repurchasers. Therefore, we used two methods to deal with this issue, synthetic data points and undersampling, but it is better to implement our model to the dataset without this limitation.

# **References**

XGBoost paper:

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System.

[UA] BigQuery Export schema:

<https://support.google.com/analytics/answer/3437719?hl=en>

# Buy ‘Til You Die: Predict Customer Lifetime Value in Python

[Python - Buy Till You Die: Predict Customer Lifetime Value | by Heiko Onnen | Medium | Towards Data Science](https://towardsdatascience.com/buy-til-you-die-predict-customer-lifetime-value-in-python-9701bfd4ddc0)

## [Wide & Deep Learning: Better Together with TensorFlow](http://ai.googleblog.com/2016/06/wide-deep-learning-better-together-with.html)

[Google AI Blog: Wide & Deep Learning: Better Together with TensorFlow (googleblog.com)](https://ai.googleblog.com/2016/06/wide-deep-learning-better-together-with.html)

# Google Analytics Sample

[Google Analytics Sample | Kaggle](https://www.kaggle.com/datasets/bigquery/google-analytics-sample/code)

XGBoost Python Package:

<https://xgboost.readthedocs.io/en/stable/python/python_api.html>

# Neural Network Models for Combined Classification and Regression

[Neural Network Models for Combined Classification and Regression (machinelearningmastery.com)](https://machinelearningmastery.com/neural-network-models-for-combined-classification-and-regression/)

# **Appendix – Reproduction of Results**

Location of Data Source:

* Big Query
  + google\_analytics\_sample (08/01/2016 - 08/01/2017)

Tools/Platforms Used:

* Big Query
* Colab
* Pandas
* numpy
* matplotlib.pyplot
* pycaret.regression
* xgboost
* sklearn
  + LinearRegression
  + LogisticRegression
  + DecisionTreeRegressor
  + DecisionTreeClassifier
  + roc\_auc\_score
  + r2\_score
  + precision\_score
  + recall\_score
  + confusion\_matrix
  + classification\_report
  + accuracy\_score
  + f1\_score
  + mean\_squared\_error
  + train\_test\_split
  + prediction\_error
  + ResidualsPlot
  + GridSearchCV
  + Ridge
  + RidgeCV
  + LassoCV
  + binarize
  + GaussianNB
  + AdaBoostClassifier
* seaborn
* imblearn
  + RandomUnderSampler
* tensorflow
* tempfile

List of Scripts:

Option 1 - Google Analytics eComm Data] Lifetime Value Project.ipynb

EDA.ipynb

LifetimeValue.ipynb

How to run Scripts:

Load ipynb file into any python workspace. Run file.

This would require access to big query to run the queries that select the data that goes into the dataframes

If I have to point out my strength, I would use following Leadership principle to show my nature.

First would be Customer Obsession, Invent & Simply as I always believe in Smart work. Hire & Develop the Best , Think Big.

Weekness would be Frugal. Sometime this is bad.