# Gstore Revenue Prediction

Group1:  Andy Li; Jiaming Guo; Minyang Wang; Shaojie Zhang

## Problem

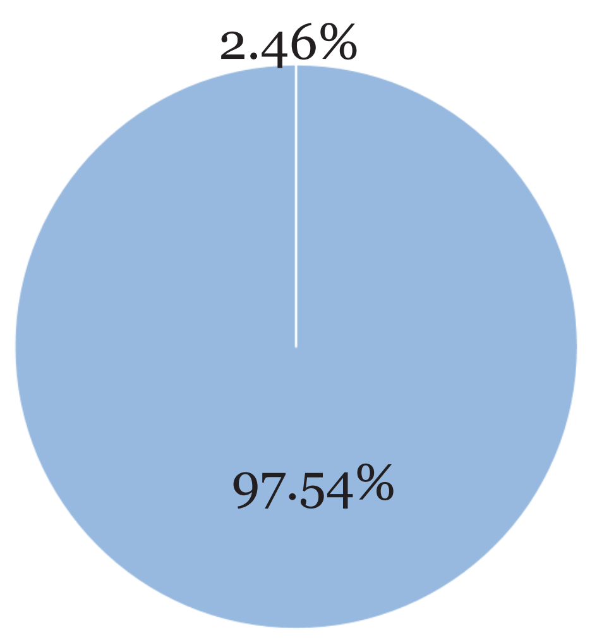
Learning the customers' behaviors and their abilities to purchase is always one of the most important methods for the merchant to increase their profit. Our project stands at the point of Google Merchandise Store (Gstore) trying to maximize the profit by predicting the revenue that each individual user may contribute to Gstore. As our main goal is to efficiently invest budget to target or potential customers, it is necessary to perform such analysis to predict whether this person is a potential customer.

Based on the result, we might provide google some ideas in how to adjust their promotional strategies. This model can help not only google but also other retail companies to adjust their promotional strategies. Making their budget into largest profits.

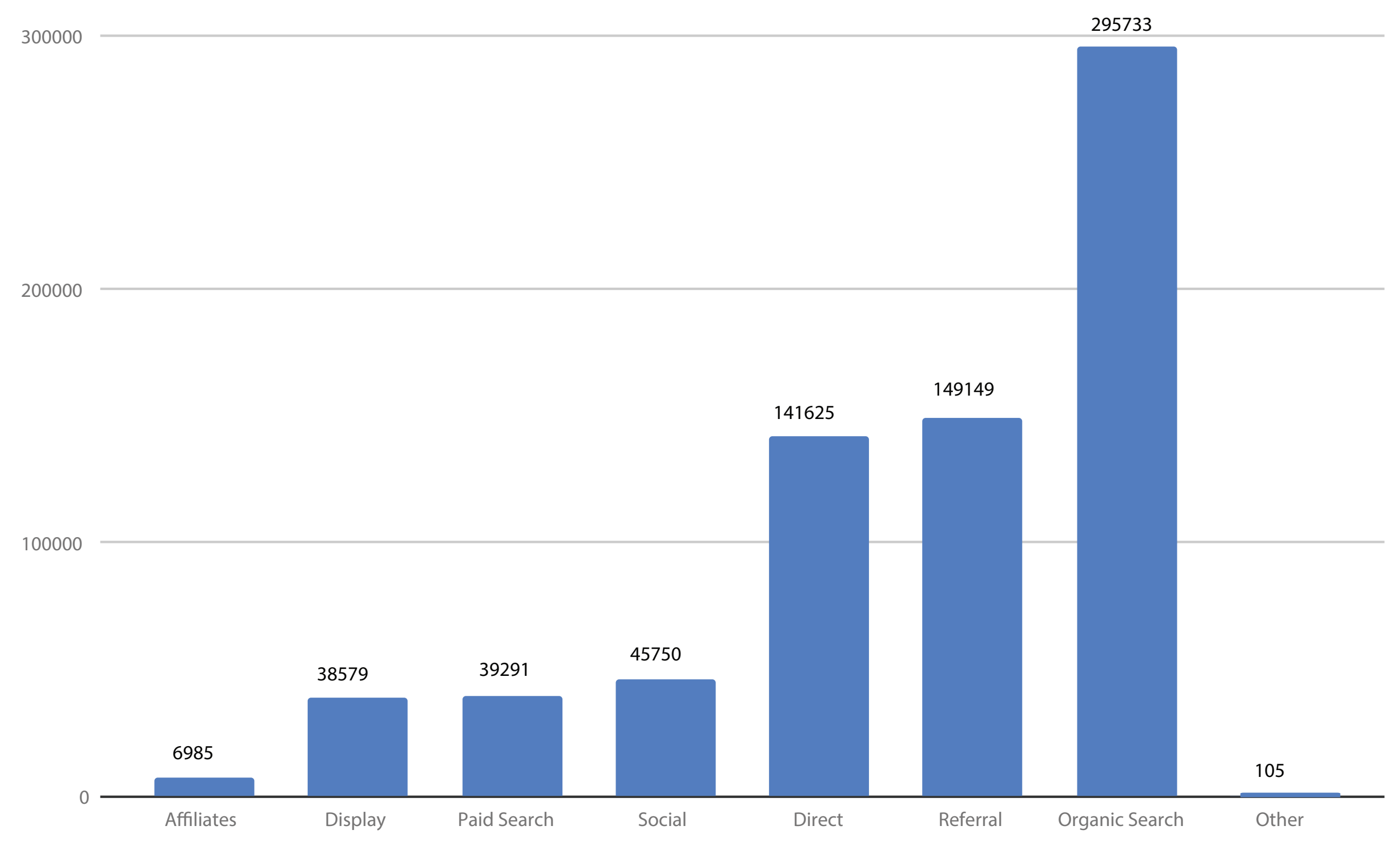
## Data Description

- 1 708 337 rows
- 33+ features
- label: totalTransactionRevenue
- Test, Validation, Training Dataset
- Focus on US area data

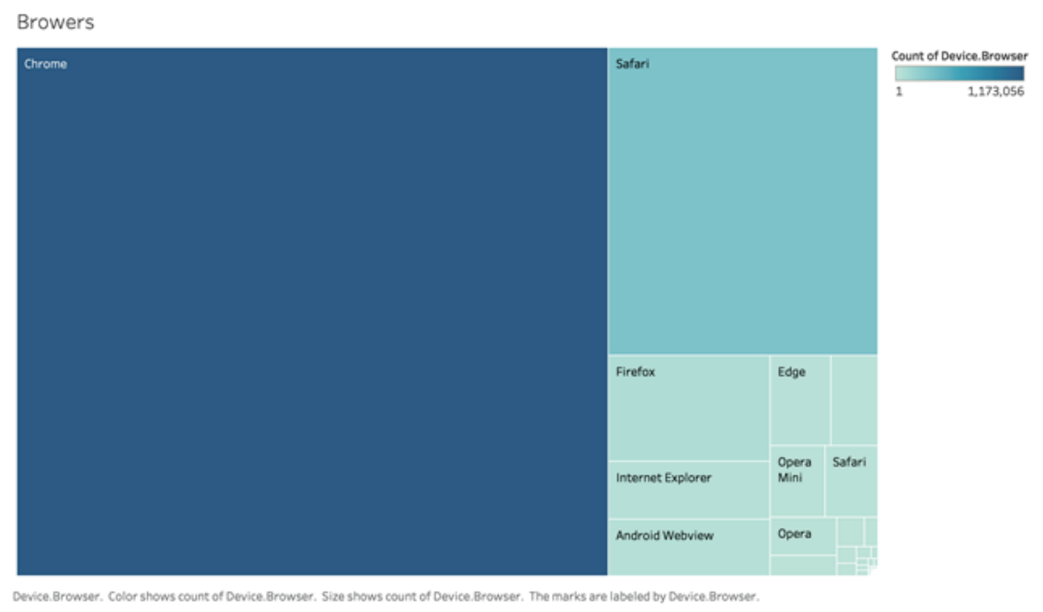
Destribution of label



Histogram of channle grouping

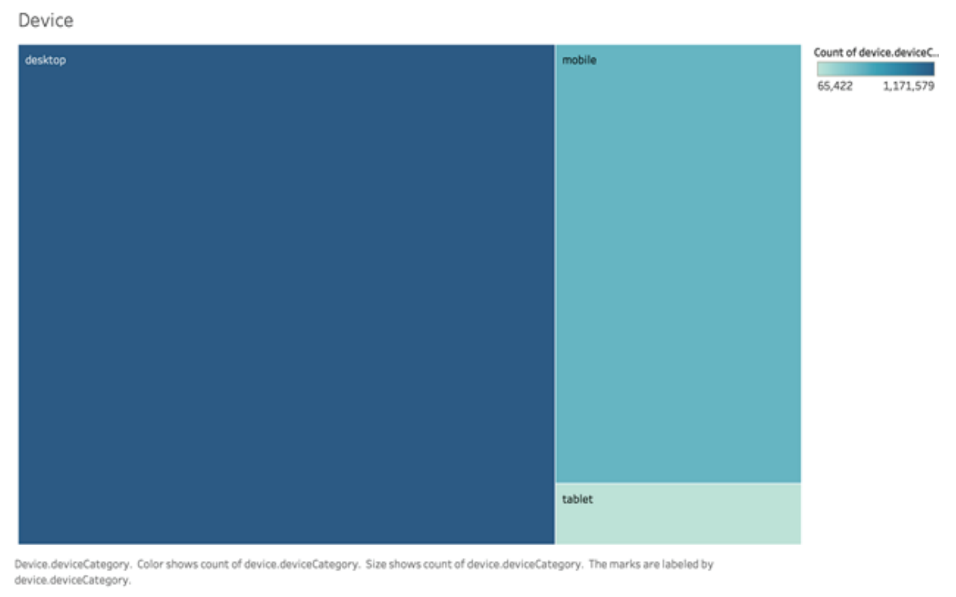


Aggregated Browser



Device.Browser. Color shows count of Device.Browser. Size shows count of Device.Browser. The marks are labeled by Device.Browser.

Aggregated Device



Device.deviceCategory. Color shows count of device.deviceCategory. Size shows count of device.deviceCategory. The marks are labeled by device.deviceCategory.

## Models

Oversampling training dataset in order to deal with imbalance data, and we trained following models to make prediction

i Logistic Model
- with PCA component from 1 to 31(step 3)
- without PCA

ii Random Forest
- Max Depth from 1 to 15
- Number of Trees from 1 to 101(step 5)

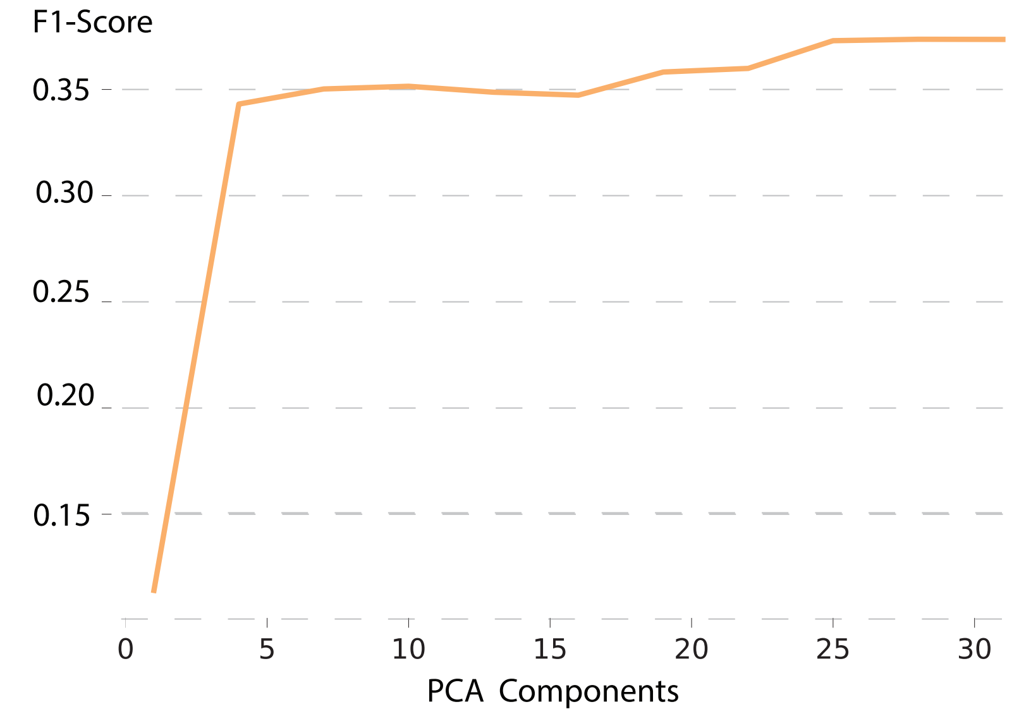
iii GBM Classifier
- Max Iterator from 1 to 21(step 2)
- Max Depth from 1 to 5

## Model Comparison (Validation)

Since this is a imbalance data, F1-Score is used to validation

Logistic Regression

***With PCA***

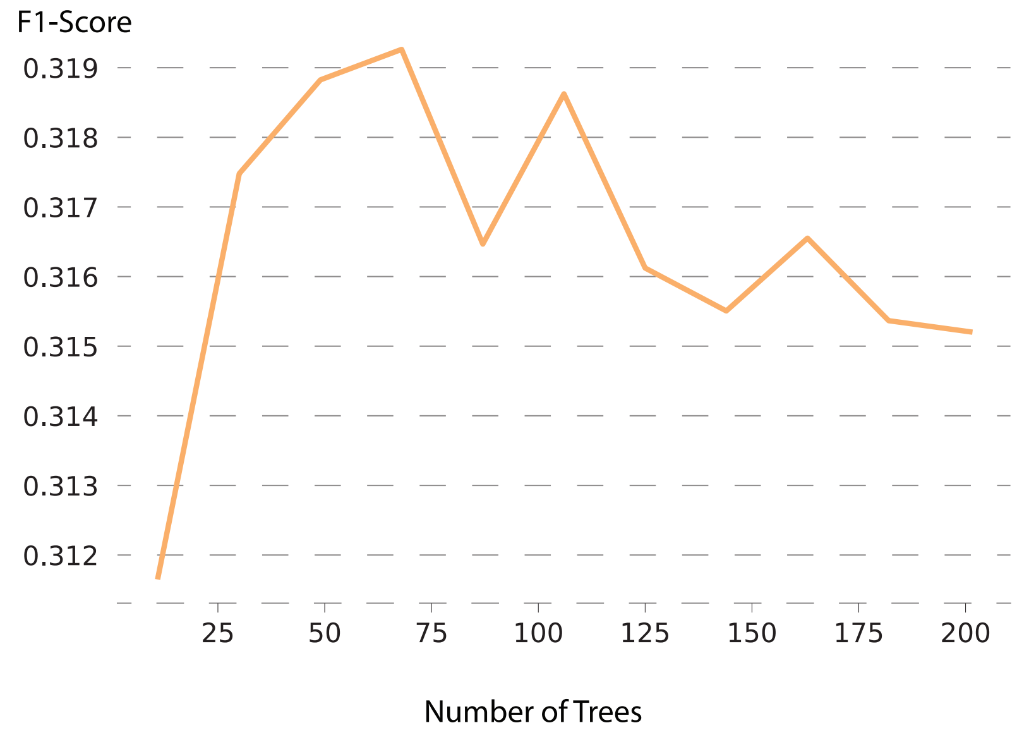


***Without PCA***

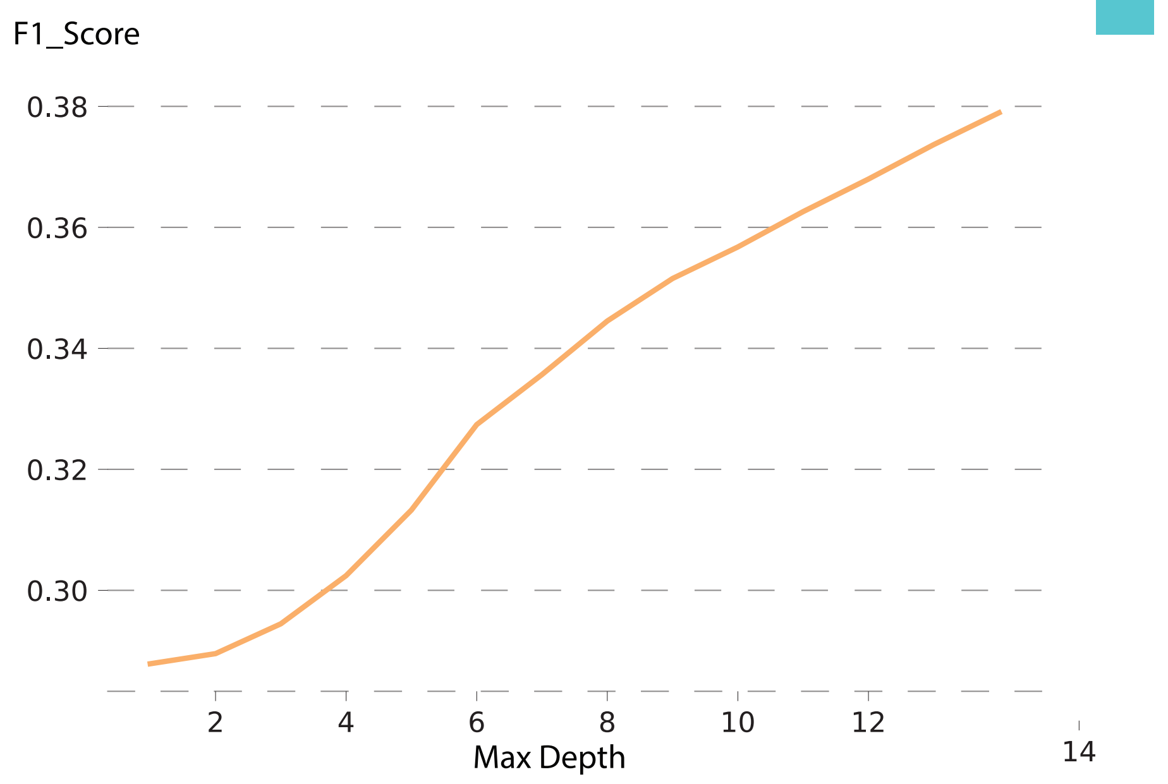
F1
0.37741014683

Random Forest

***With Number of Trees***



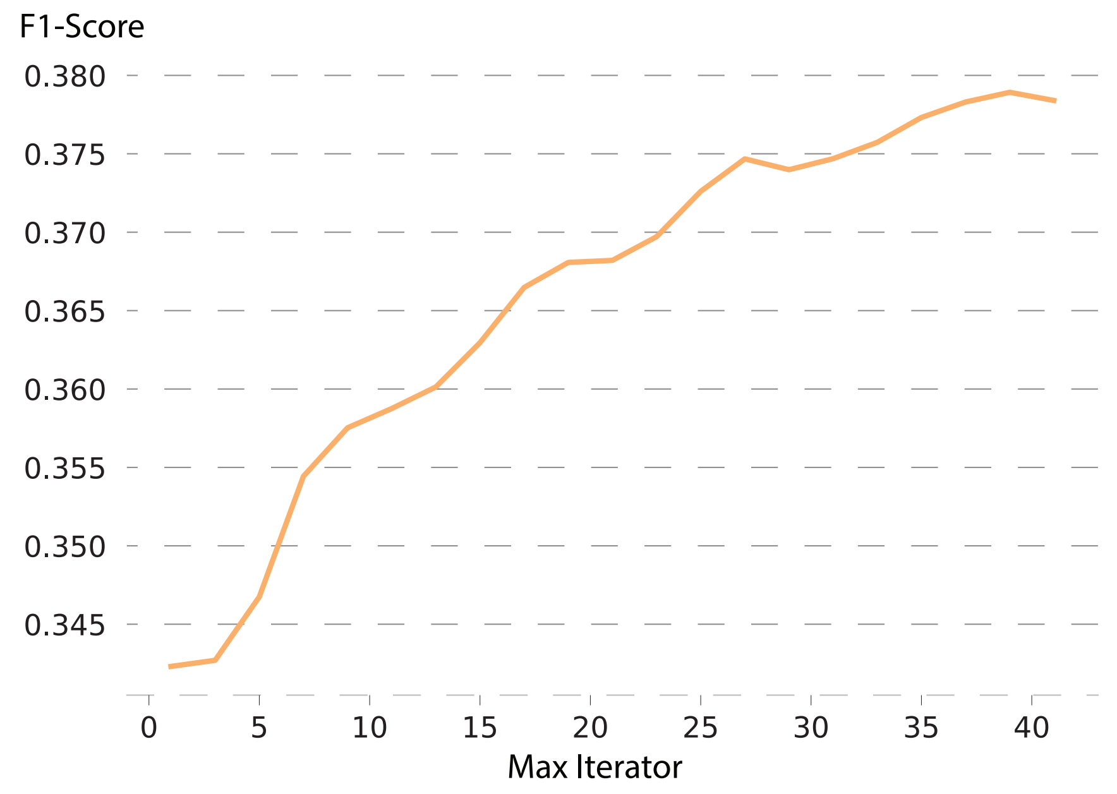
***With Max Depth***



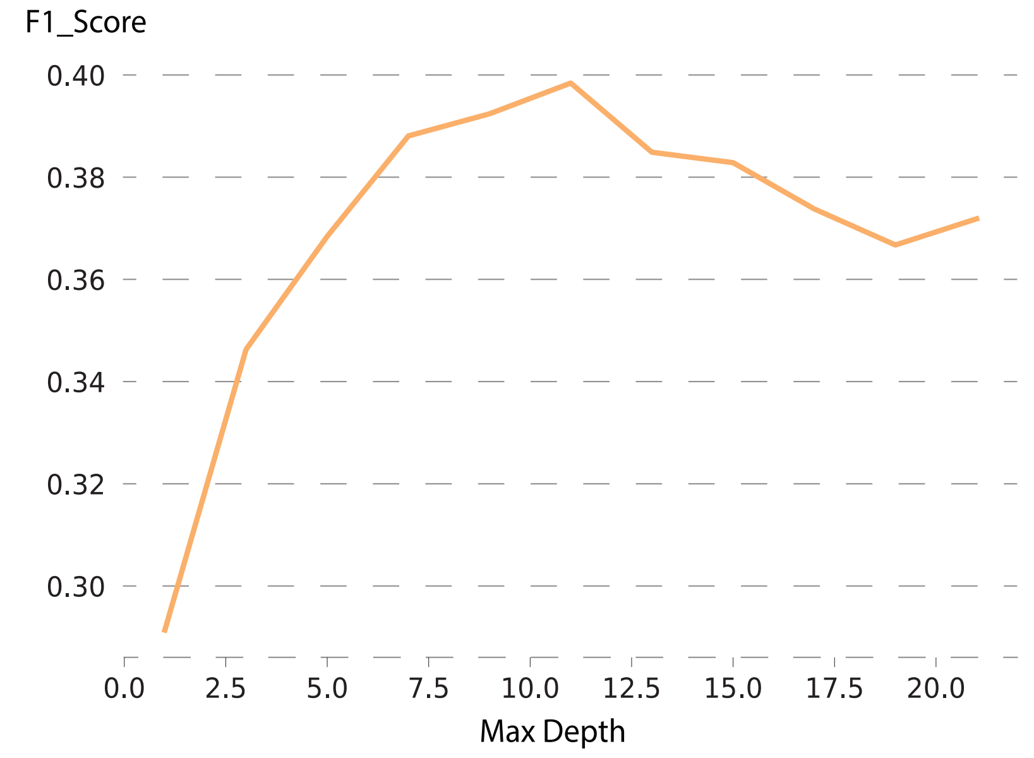
***Best Model:***
***Number of Tree = 60***
***Max Depth = 15***
***F1-Score = 0.373***

GBT Classifier

***With Max Iterator***



***With Depth***



***Best Model:***
***Max Iterator= 38***
***Max Depth = 11***
***F1-Score = 0.4054***

## Testing

GBT Classifier

***Confusion Matrx***

|  | 0 | 1 |
|---|---|---|
| 0 | 66349 | 3658 |
| 1 | 347 | 1368 |

F1=0.40587

| Column | Weight |
|---|---|
| Page views | 0.55179 |
| Time On Site | 0.085920 |
| Hits | 0.070044 |
| Session Quality Dim | 0.060760 |
| Operating system | 0.054816 |

## Conclusion

Based on the confusion matrix of GBT Classifier(40%) , GBT model does perfect in predicting customers. Page views and time on site are two most important deciding features. The number of page views have the highest weight(55%) . However, this model will mis-predicted a lot of non-customers, which will cause a lot of extra spend on campaign.

Future work: We are still working on reduce the number of False Positive, so we can target the right customer. Also, we can use cloud computing to gain more computing power to build more complicated model.

## References

Google Swag Sotre: https://shop.googlemerchandisestore.com/; Link to Kaggle Dataset: https://www.kaggle.com/c/ga-customer-revenue-prediction/overview; Detailed explaination of the variables: https://docs.google.com/document/d/13_9QdF3SxdWD3_XAG8MimiJjPxzKoBJ2H5P0bnw27XM/edit?usp=sharing; Detailed explaination of the feature engineering processes: https://docs.google.com/document/d/10ZC_bP1aIbMcPkw_uDw8VtivAeTy71lY7g0CSwpq4dc/edit?usp=sharing; Dedicated to our beloved professor - Daniel Acuna