Google Analytics Customer Revenue Prediction

INTRODUCTION

- In this project, a **Python** based **Exploratory Data Analytics (EDA)** notebook is prepared to analyze the **Google Analytics Dataset** obtained from **Kaggle**. The **Data Science and Analytics** is one of the leading areas now-a-days. EDA is the foremost component of Data Science & Analytics, as it plays a vital role to general the initial insights from the data after performing descriptive statistics, inferential statistics, or both, and visualization techniques to generate actionable insights from data. Under this project, Python programming with Jupyter Notebooks IDE has been utilized, which are coupled with popular Python Libraries like Numpy, Scipy, Pandas, Matplotlib, Seaborn and many others. This project explores the data using data aggregation & grouping techniques, after the application of data cleaning and preparation phase.

PROBLEM STATEMENT

- AIM
  - Google Analytics track the client websites integrated on its platform by webmasters. The traffic on these websites in tracked in the real-time by Google Analytics, and advertisements are pushed to the designated containers. The advertisement revenue is predominantly generated by Google Adwords and Google Adsense.
- PROBLEM STATEMENT
  - In this project, the primary objective is to predict the revenue generated by analyzing the visitor environmental factors, such as GeoLocation, BrowserType, OSType, etc. The revenue generation prediction can possibly help the platform to push more attractive or lucrative ads on the requested page.

CONCLUSIONS

- Random Forest Regression is best compatible algorithm to predict Google Analytics Customer Revenue
- Most visitors uses Windows OS
- Most visitors originates from American continent
- Mac OS users generates the highest in-class revenue percentage
- North American continent generates the highest revenue in entire world

DATA PREPRATATION

```
In [75]: train_df[numerical_cols].isnull().sum()
Out[75]: visitNumber                      0
         totals.transactionRevenue        0
         _weekday                         0
         _day                             0
         _month                           0
         _visitHour                       0
         totals.bounces              109498
         totals.hits                      0
         totals.newVisits             57953
         totals.pageviews                46
         dtype: int64
```

percentage of missing value in catrgorical columns

```
In [82]: b
Out[82]: channelGrouping                 0.000000
         device.browser                  0.000000
         device.deviceCategory           0.000000
         device.isMobile                 0.000000
         device.operatingSystem          0.000000
         geoNetwork.city                 0.000000
         geoNetwork.continent            0.000000
         geoNetwork.country              0.000000
         geoNetwork.metro                0.000000
         geoNetwork.networkDomain        0.000000
         geoNetwork.region               0.000000
         geoNetwork.subContinent         0.000000
         trafficSource.adContent         0.000000
         trafficSource.campaign          0.000000
         trafficSource.isTrueDirect     63.001883
         trafficSource.keyword           9.904180
         trafficSource.medium            0.000000
         trafficSource.referralPath      0.000000
         trafficSource.source            0.000000
         dtype: float64
```

- Categorical column with 63% missing data must be deleted
- For all other missing categorical values, "other" category is generated
- Numeric variables "bounces" dropped due to high number of missing

FEATURE ENGINEERING

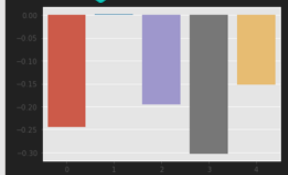
	date	_weekday	_day	_month	_year	_visitHour
0	2018-07-01	6	1	7	2018	
1	2018-07-22	6	22	7	2018	
2	2018-07-01	6	1	7	2018	
3	2018-09-27	3	27	9	2018	
4	2018-10-04	3	4	10	2018	

Feature engineered date column to generate more compatible and quanfied features like day, month, year and visitHour.

Index	mean_hits_per_networkDomain	mean_pageViews_per_networkDomain
0	6	5
1	4	3
2	6	5
3	6	5
4	1	1

networkDomain (Autonomous system) level segmentation is performed to engineer quantified features.

MODEL FITTING – FITTING THE WORST MODEL



A multiple linear regression model is fitted over the NON-Standardized data.

Cross validation score of worst model is actually worst.

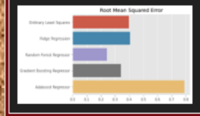
Wasn't expected this low score.

CONCLUSIONS

- HOWEVER ONLINE ECOSYSTEM IS RAPIDLY CHANING AREA, FOLLOWING ARE SOME GUIDELINES TO GENERATE HIGHER REVENUE
  - Your website must be compatible and well tested for Top 3 browsers (Chrome, Safari and Firefox), which has approx. 93% of internet traffic.
  - Focus more on Search Engine (44%) & Social Media (21%) marketing
  - An effective online marketing campaign should target more people on Workdays than Weekends
  - You have only 4 minutes to catch user's attention.

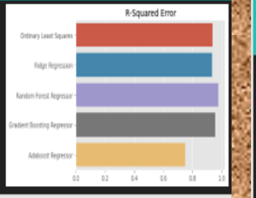
RESULTS & FINDINGS

- Lowest MSE & RMSE error is observed for Random Forest Regressor
- Second best regressor is Gradient Boosting Regressor



RESULTS & FINDINGS

- Highest R-squared error is observed for Random Forest Regressor (0.97).
- Random Forest Regressor accurately represents 97% variance in the data
- Gradient Boosting Regressor accurately represents nearly 95% of data
- Adaboost Regressor is worst performer, and accurately represents only 75% of data



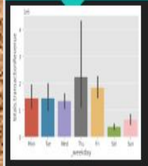
CONCLUSIONS

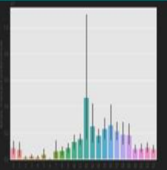
- Random Forest Regression is best compatible algorithm to predict Google Analytics Customer Revenue
- Most visitors uses Windows OS
- Most visitors originates from American continent
- Mac OS users generates the highest in-class revenue percentage
- North American continent generates the highest revenue in entire world

HYPOTHESIS QUESTIONS

- VISITOR-LEVEL HYPOTHESIS
  - **Most visitors use Windows OS**
    - H0: Windows OS generate highest visitor count
    - H1: Other OS generate highest visitor count
  - **Most visitors originate from North American continent**
    - H0: Highest visitor count originate from North America
    - H1: Highest visitor count originate from other regions
- REVENUE-LEVEL
  - **Windows OS visitors generate the highest revenue in OS category**
    - H0: Windows OS generates highest revenue
    - H1: Other OS generates highest revenue
  - **North American continent generates the highest revenue**
    - H0: Highest visitor count originate from North America
    - H1: Highest visitor count originate from other regions

EXPLORATORY DATA ANALYSIS

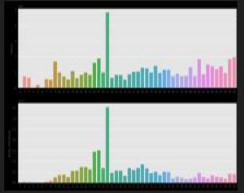




- Most of the revenue is generated during weekdays
- 7AM to 7PM is the most lucrative time segment in 24 hours a day
- Lowest revenue observed from 2AM to 6 AM

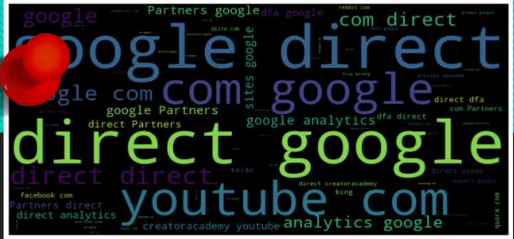
EXPLORATORY DATA ANALYSIS

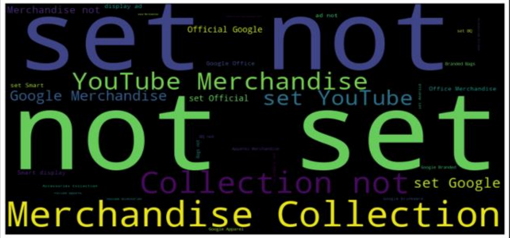
With increase in web traffic, the revenue increases as well



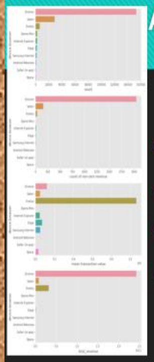
EDA: AD CONTENT & SOURCES

- Most ad content is not set and can be any random ad selection
- YouTube, Merchandise and Collection ads are ranked after random
- Most of traffic is either Direct traffic or generated by Google search engine
- YouTube also generates significant traffic

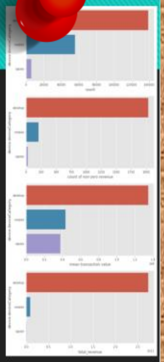




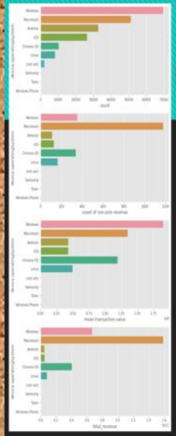
EXPLORATORY DATA ANALYSIS



- Though safari users are more , revenue generating customers are less
- Mozilla Firefox users generate a more mean value
- Only chrome users are major source of revenue generation
- Desktop users generate more revenue



EXPLORATORY DATA ANALYSIS



- Highest number of requests generated by Windows OS
- Higher percentage of Mac users generate revenue in comparison with windows
- Highest revenue contribution comes from MAC users followed by Windows
- Mean revenue of Chrome OS users is highest
- Concludingly, MAC, Windows, Chrome OS users are source of revenue generation

Google Analytics Customer Revenue Prediction

PRESENTED BY: **AJITPAL BRAR**

MAY 30, 2020

CONTENTS

INTRODUCTION

- In this project, a **Python** based **Exploratory Data Analytics (EDA)** notebook is prepared to analyze the **Google Analytics Dataset** obtained from **Kaggle**. The **Data Science and Analytics** is one of the leading areas now-a-days. EDA is the foremost component of Data Science & Analytics, as it plays a vital role to general the initial insights from the data after performing descriptive statistics, inferential statistics, or both, and visualization techniques to generate actionable insights from data. Under this project, Python programming with Jupyter Notebooks IDE has been utilized, which are coupled with popular Python Libraries like Numpy, Scipy, Pandas, Matplotlib, Seaborn and many others. This project explores the data using data aggregation & grouping techniques, after the application of data cleaning and preparation phase.

INTRODUCTION

- In this project, a Python based Exploratory Data Analytics (EDA) notebook is prepared to analyze the Google Analytics Dataset obtained from Kaggle. [1]
- This project is based upon the application of Exploratory Data Analysis & Model Fitting on the given dataset.
- Sampled dataset contains total of 34 columns and 200,794 instances. [1]
- The subset of 3.5 GB is randomly sampled from 24GB dataset. [1]
- Regression method would be used to predict the revenue, because Revenue (Dependent variable)

PROBLEM STATEMENT

- AIM
  - Google Analytics track the client websites integrated on its platform by webmasters. The traffic on these websites in tracked in the real-time by Google Analytics, and advertisements are pushed to the designated containers. The advertisement revenue is predominantly generated by Google Adwords and Google Adsense.
- PROBLEM STATEMENT
  - In this project, the primary objective is to predict the revenue generated by analyzing the web ecosystem factors, such as GeoLocation, BrowserType, OSType, etc. The revenue generation prediction can possibly help the platform to push more attractive or lucrative ads on the requested page.

HYPOTHESIS QUESTIONS

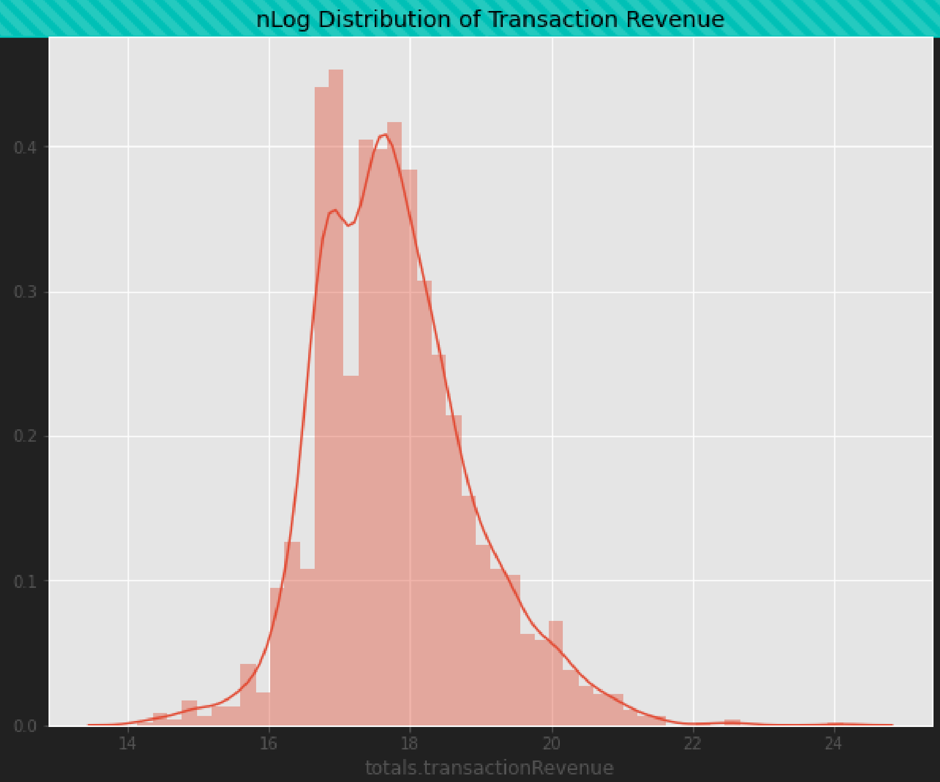
- VISITOR-LEVEL HYPOTHESIS
  - **Most visitors use Windows OS**
    - H0: Windows OS generate highest visitor count
    - H1: Other OS generate highest visitor count
  - **Most visitors originate from North American continent**
    - H0: Highest visitor count originate from North America
    - H1: Highest visitor count originate from other regions
- REVENUE-LEVEL
  - **Windows OS visitors generate the highest revenue in OS category**
    - H0: Windows OS generates highest revenue
    - H1: Other OS generates highest revenue
  - **North American continent generates the highest revenue**
    - H0: Highest visitor count originate from North America
    - H1: Highest visitor count originate from other regions

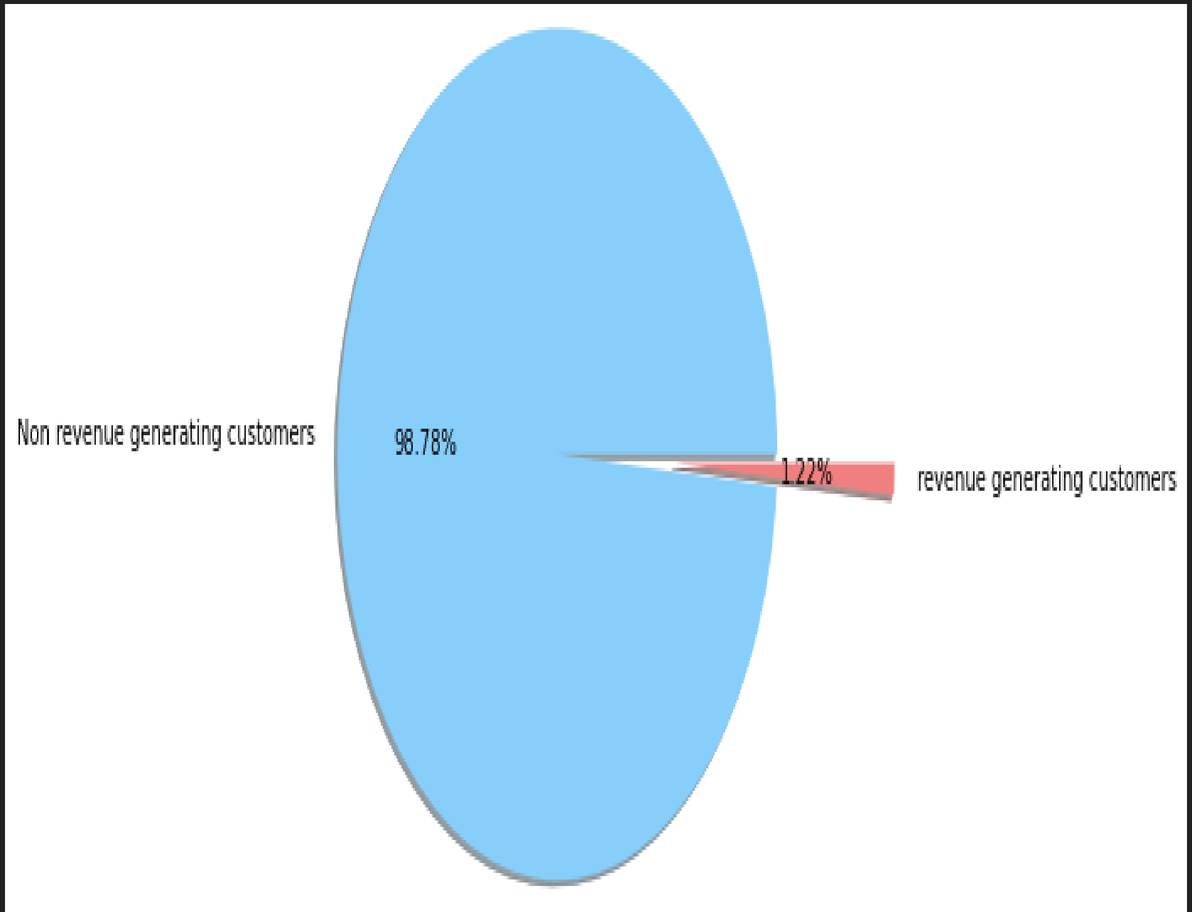
METHODOLOGY

- The storyboard method is used to plan the analytics timeline in this project.
- A primary storyboard was been prepared for the introductory EDA phase. The story board was basically divided into five stages, which describes requirement of univariate, bi-variate or multi-variate data analysis.
- The work has been performed on Google Analytics Customer Revenue Prediction Dataset, which is obtained from Kaggle competition. The prize money for this project was US$45000, which considered an impressive prize amount for Data Science competitions.

- Pandas with Numpy for the EDA.
- Sci-kit (Sklearn) package for Machine Learning (Regression)
- Python 3.7.3
- Pandas version 1.03
- Numpy version 1.18.2
- Sklearn 0.21.2

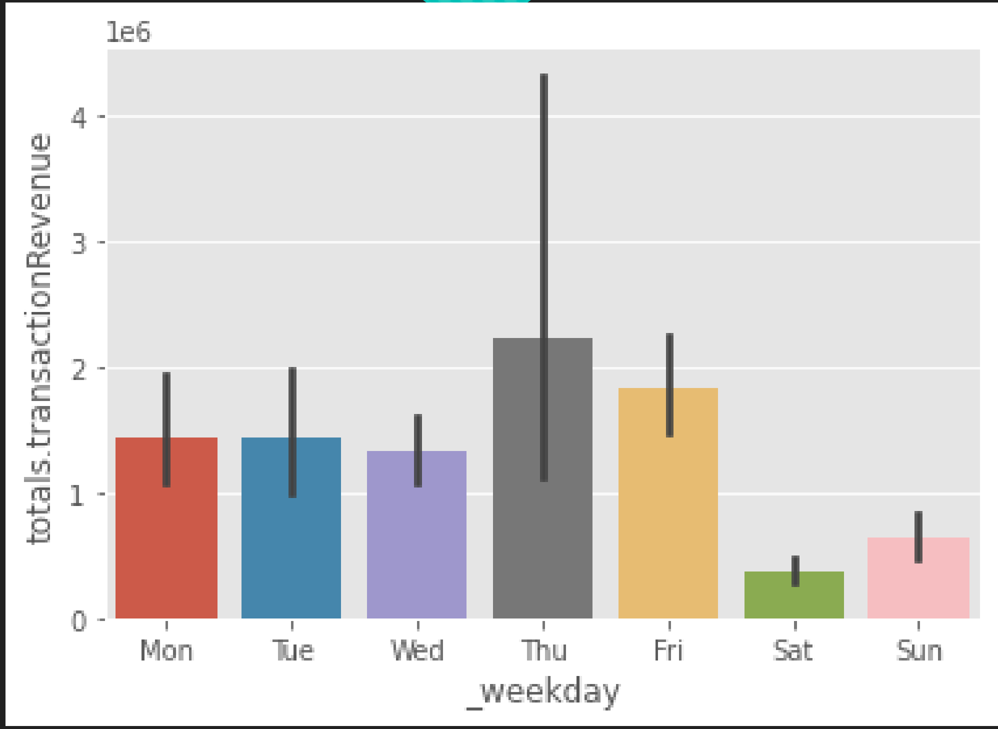
EXPLORATORY DATA ANALYSIS

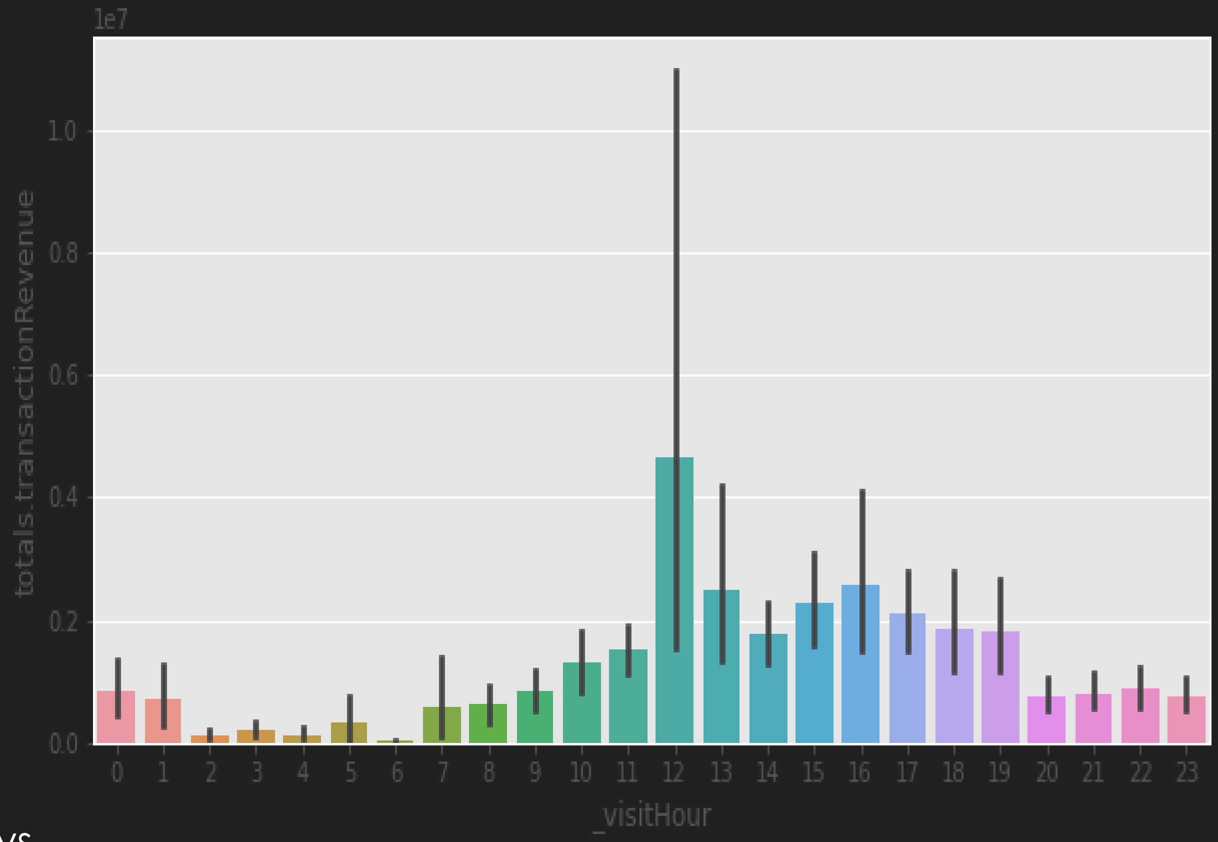




- 98.78% visitors haven't generated any revenue.
- Only 1.22% customers generated all the revenue.
- Revenue Distribution is slightly skewed to right.

EXPLORATORY DATA ANALYSIS

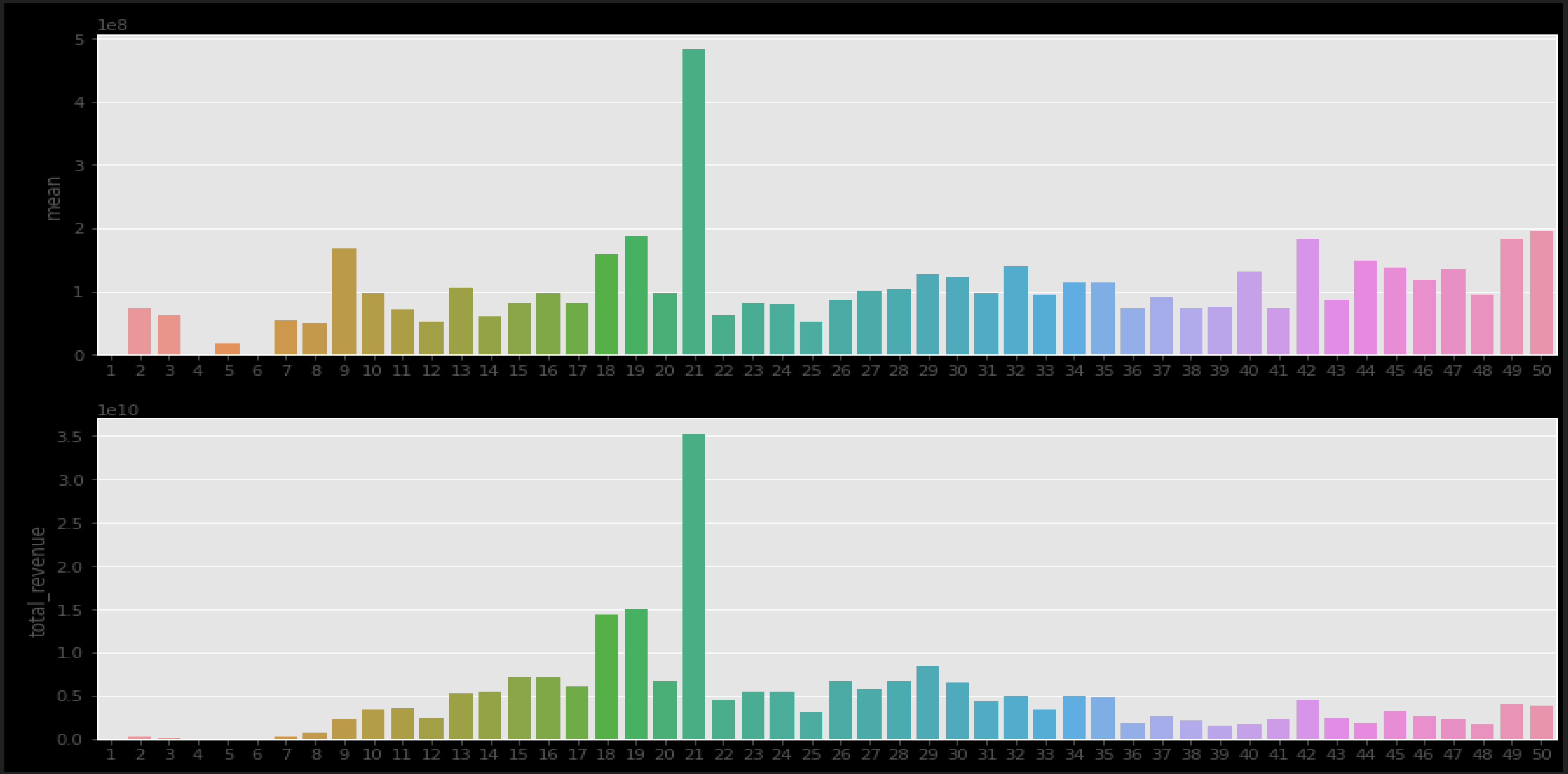




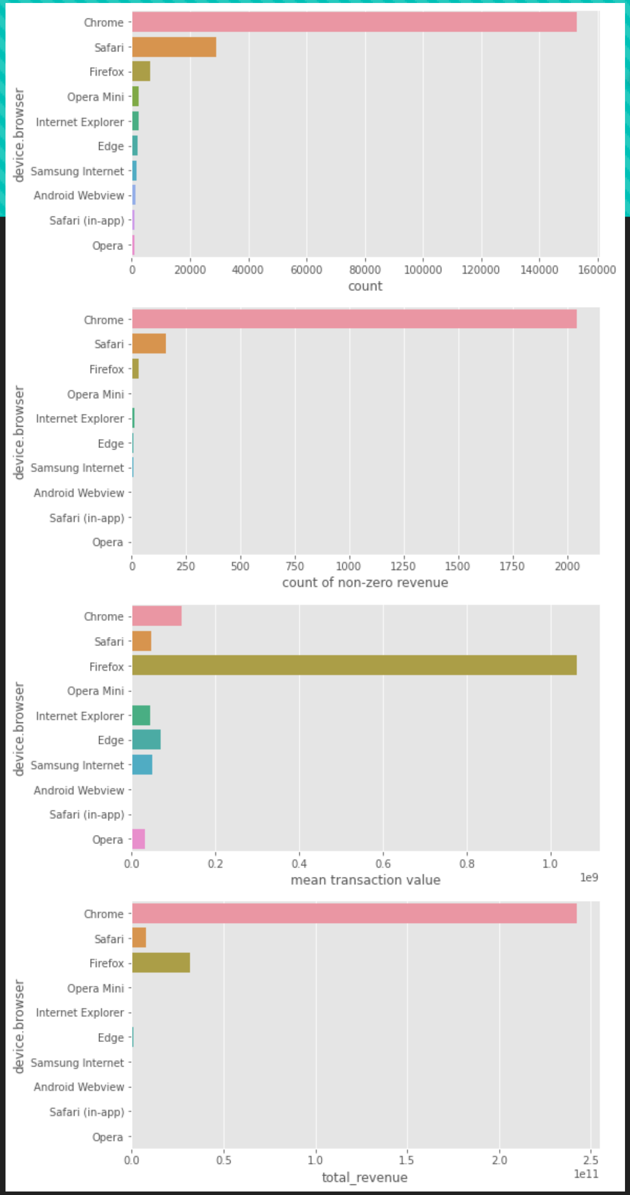
- Most of the revenue is generated during weekdays
- 7AM to 7PM is the most lucrative time segment in 24 Hours a day
- Lowest revenue observed from 2AM to 6 AM

EXPLORATORY DATA ANALYSIS

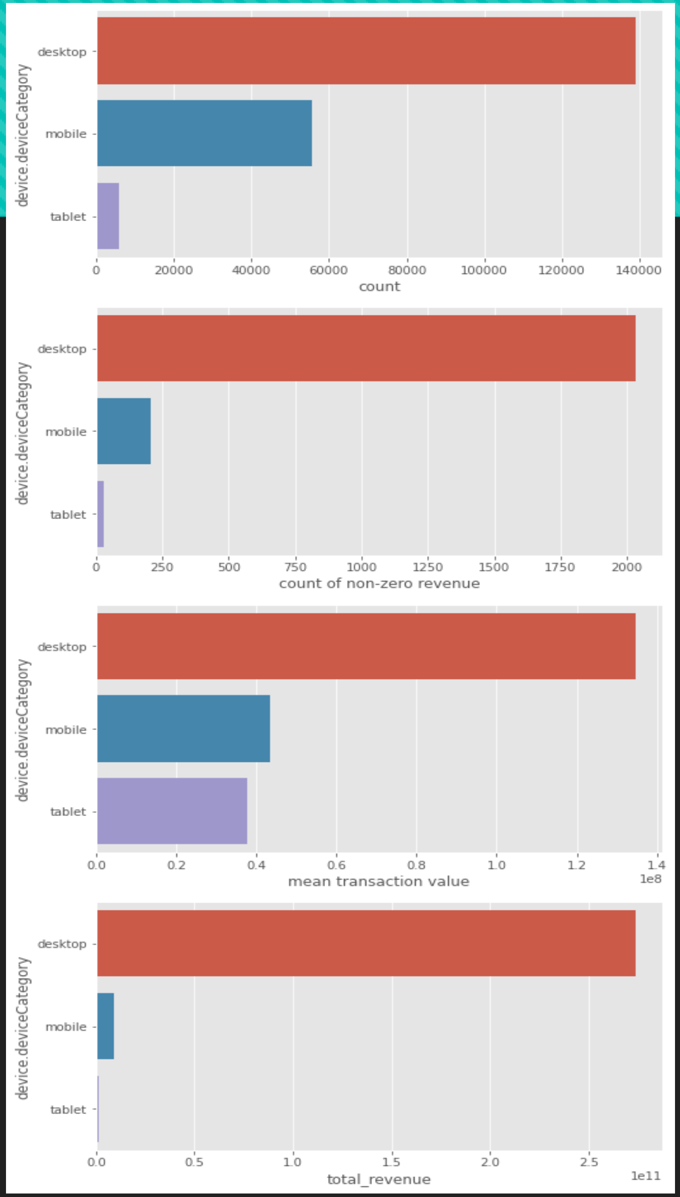
With increase in web traffic, the revenue increases as well

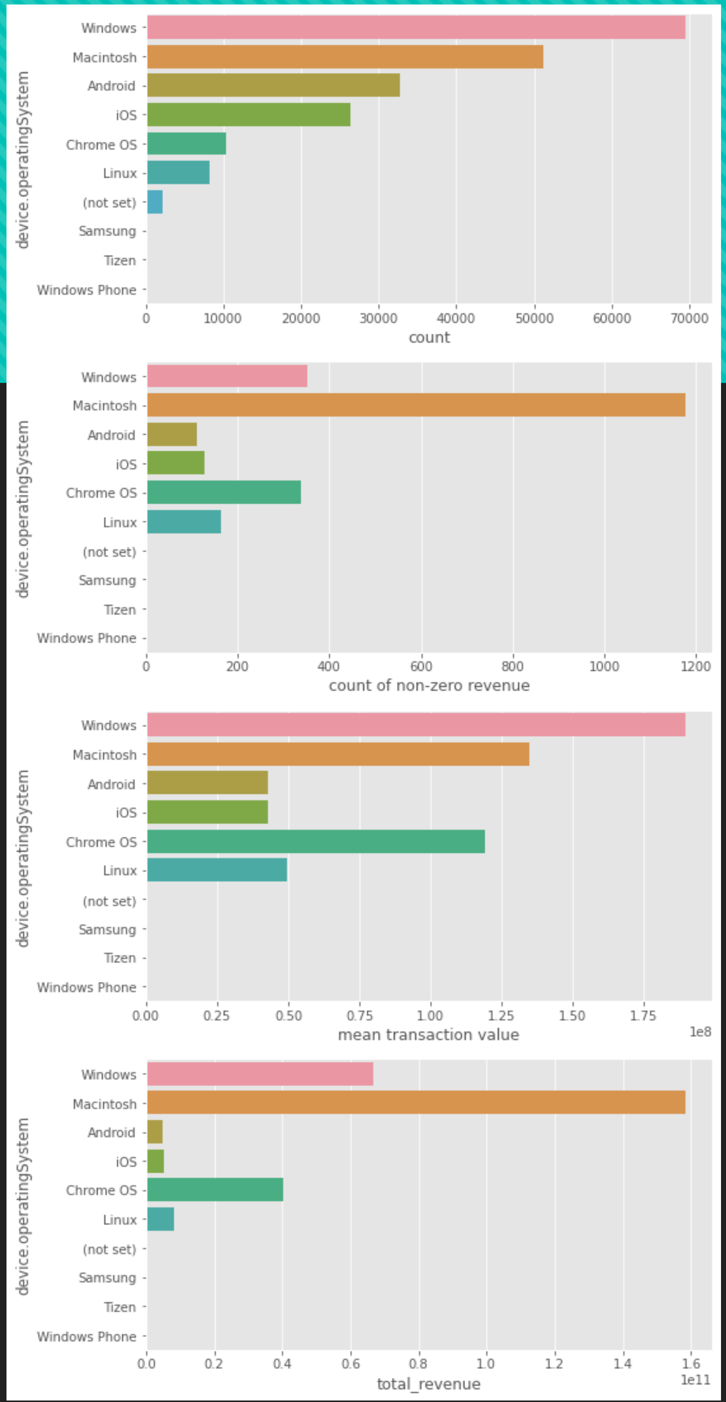


EXPLORATORY DATA ANALYSIS



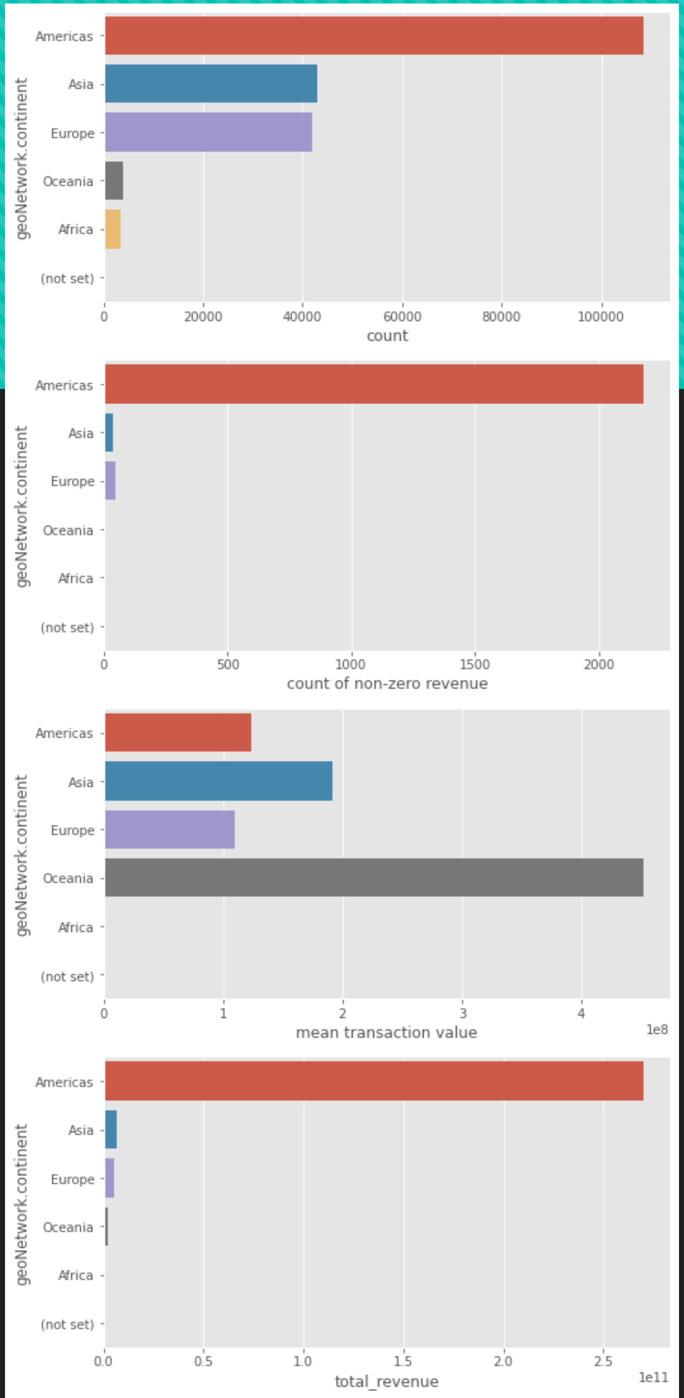
- *Chrome users are major source of revenue generation*
- *Desktop users generate more revenue than mobile & tablet*
- *Mozilla Firefox users generate a more mean value*
- *Less of Safari browser users generate any revenue*





EXPLORATORY DATA ANALYSIS

- *Highest number of visitors uses Windows OS*
- *Mac users generate comparatively higher revenue in comparison with other OS*
- *Highest revenue contribution also comes from MAC users followed by Windows*
- *Mean revenue of Chrome OS users is highest*
- *Concludingly, MAC, Windows, Chrome OS users are source of revenue generation*

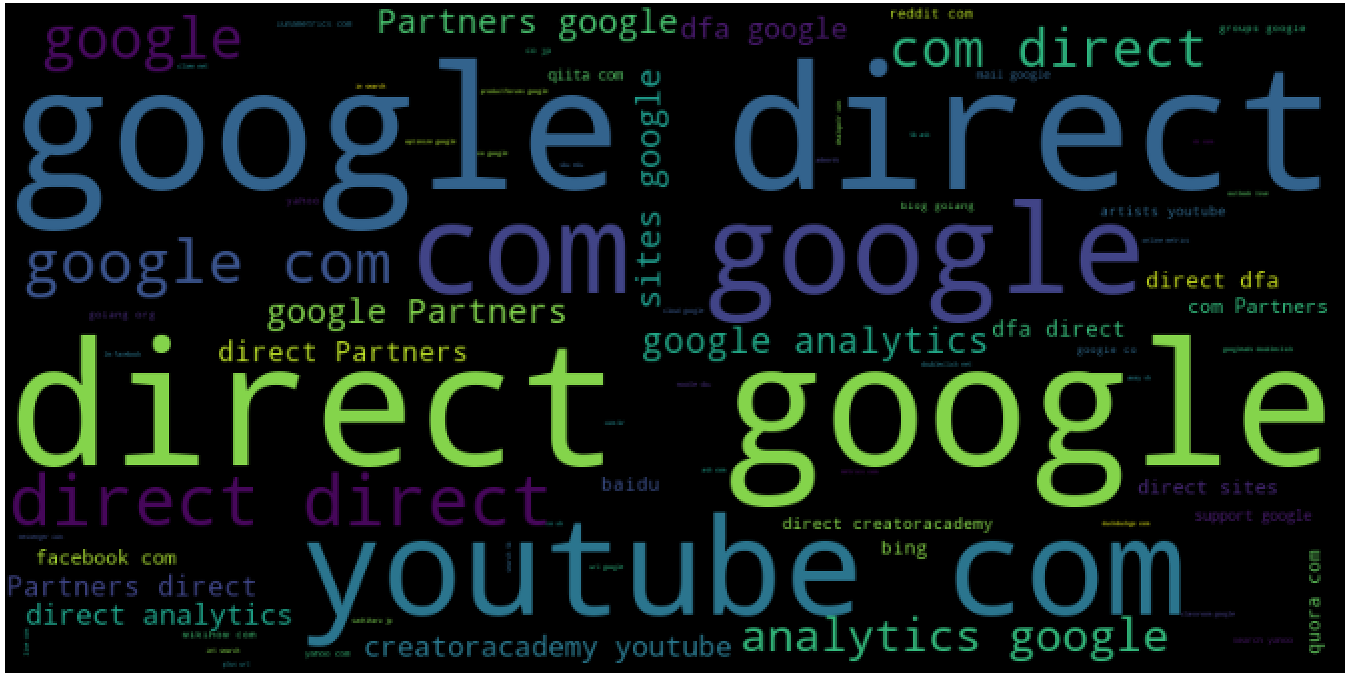


EXPLORATORY DATA ANALYSIS

- *Highest visitor count is received from American continent*
- *Highest Revenue is generated by American continent*
- *Highest Mean Transaction value is generated by Oceania*
- *Asia & Europe has high visitor count, and extremely low revenue*

EDA: AD CONTENT & SOURCES

- Most ad content is not set and can be any random ad selection
- YouTube, Merchandise and Collection ads are ranked after random
- Most of traffic is either Direct traffic or generated by Google search engine
- YouTube also generates significant traffic



DATA PREPRATATION

```
In [75]: train_df[numerical_cols].isnull().sum()

Out[75]: visitNumber                        0
         totals.transactionRevenue          0
         _weekday                           0
         _day                               0
         _month                             0
         _visitHour                         0
         totals.bounces                109498
         totals.hits                        0
         totals.newVisits               57953
         totals.pageviews                  46
         dtype: int64
```

- Categorical column with 63% missing data must be deleted
- For all other missing categorical values, "other" category is generated
- Numeric variables "bounces" dropped due to high number of missing

percentage of missing value in catrgorical columns

```
In [82]: b

Out[82]: channelGrouping                 0.000000
         device.browser                  0.000000
         device.deviceCategory           0.000000
         device.isMobile                 0.000000
         device.operatingSystem          0.000000
         geoNetwork.city                 0.000000
         geoNetwork.continent            0.000000
         geoNetwork.country              0.000000
         geoNetwork.metro                0.000000
         geoNetwork.networkDomain        0.000000
         geoNetwork.region               0.000000
         geoNetwork.subContinent         0.000000
         trafficSource.adContent         0.000000
         trafficSource.campaign          0.000000
         trafficSource.isTrueDirect     63.001883
         trafficSource.keyword           9.904180
         trafficSource.medium            0.000000
         trafficSource.referralPath      0.000000
         trafficSource.source            0.000000
         dtype: float64
```

FEATURE ENGINEERING

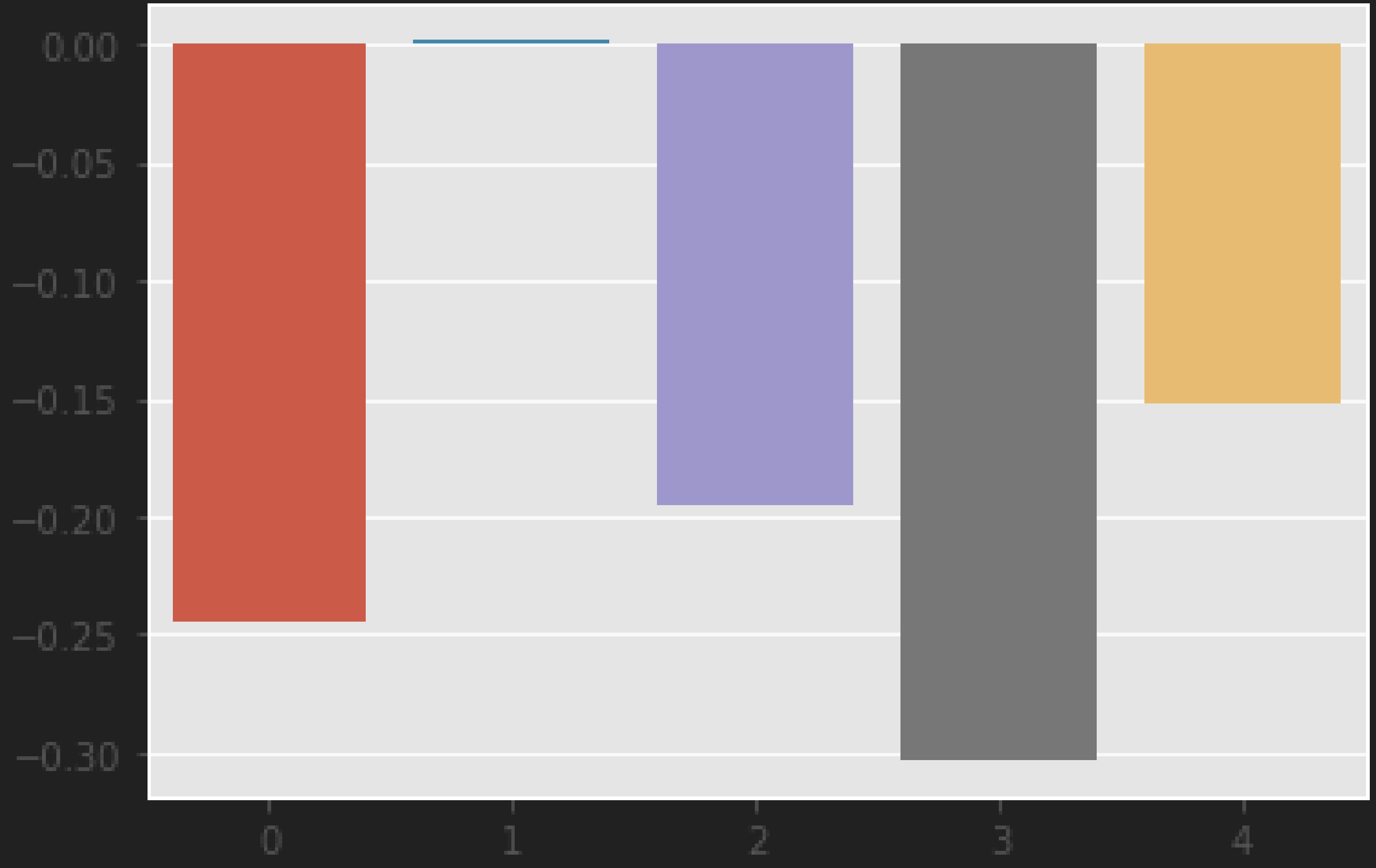
date	_weekday	_day	_month	_year	_visitHour
2018-07-01	6	1	7	2018	3
2018-07-22	6	22	7	2018	15
2018-07-01	6	1	7	2018	10
2018-09-27	3	27	9	2018	19
2018-10-04	3	4	10	2018	3

Feature engineered date column to generate more compatible and quanfied features like day, month, year and visitHour.

Index	mean_hits_per_networkDomain	mean_pageViews_per_networkDomain
0	6	5
1	4	3
2	6	5
3	6	5
4	1	1

networkDomain (Autonomous system) level segmentation is performed to engineer quantified features.

MODEL FITTING – FITTING THE WORST MODEL

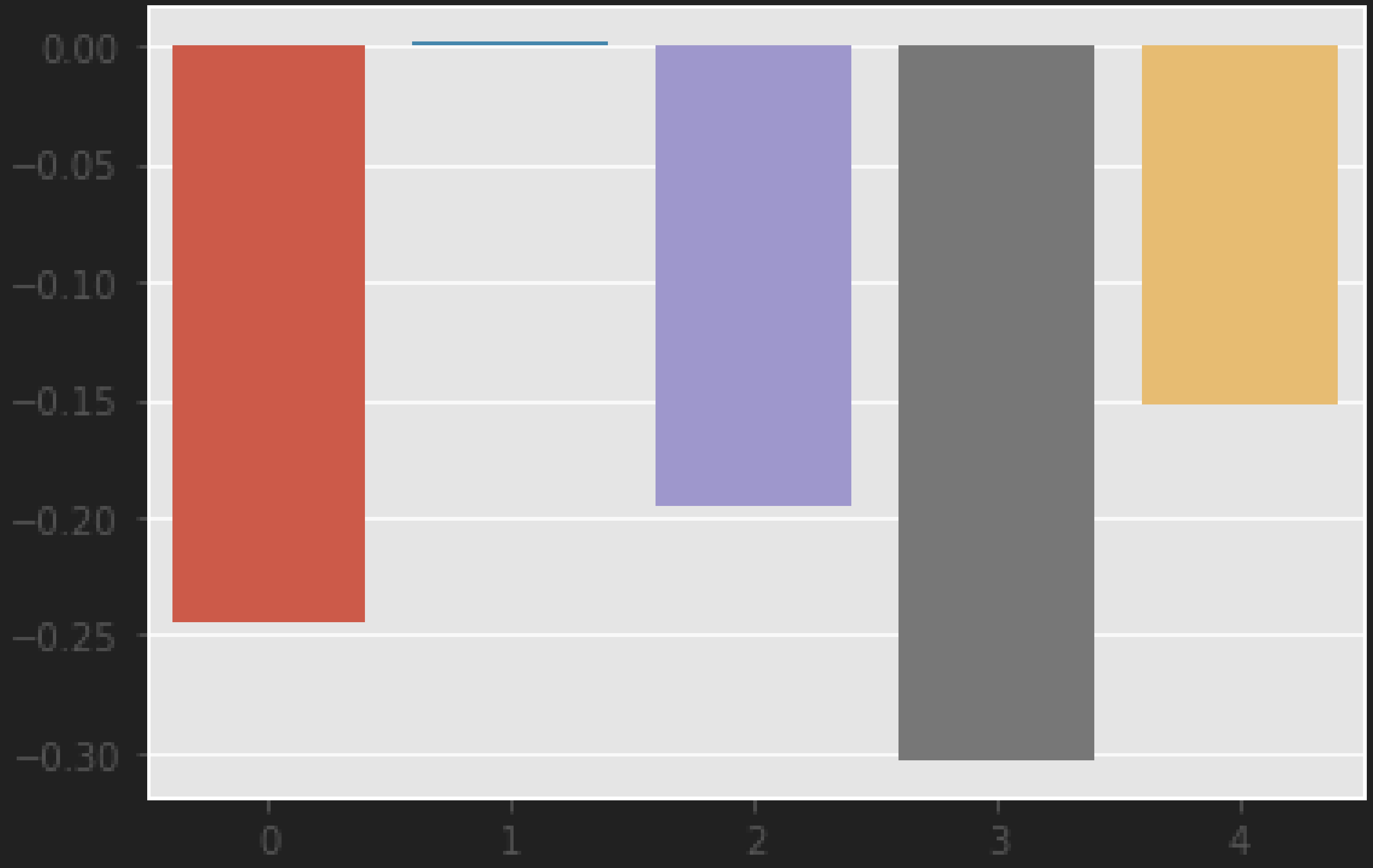


A multiple linear regression model is fitted over the NON-Standardized data.

Cross validation score of worst model is actually worst.

Wasn't expected this low score.

MODEL FITTING – FITTING THE WORST MODEL

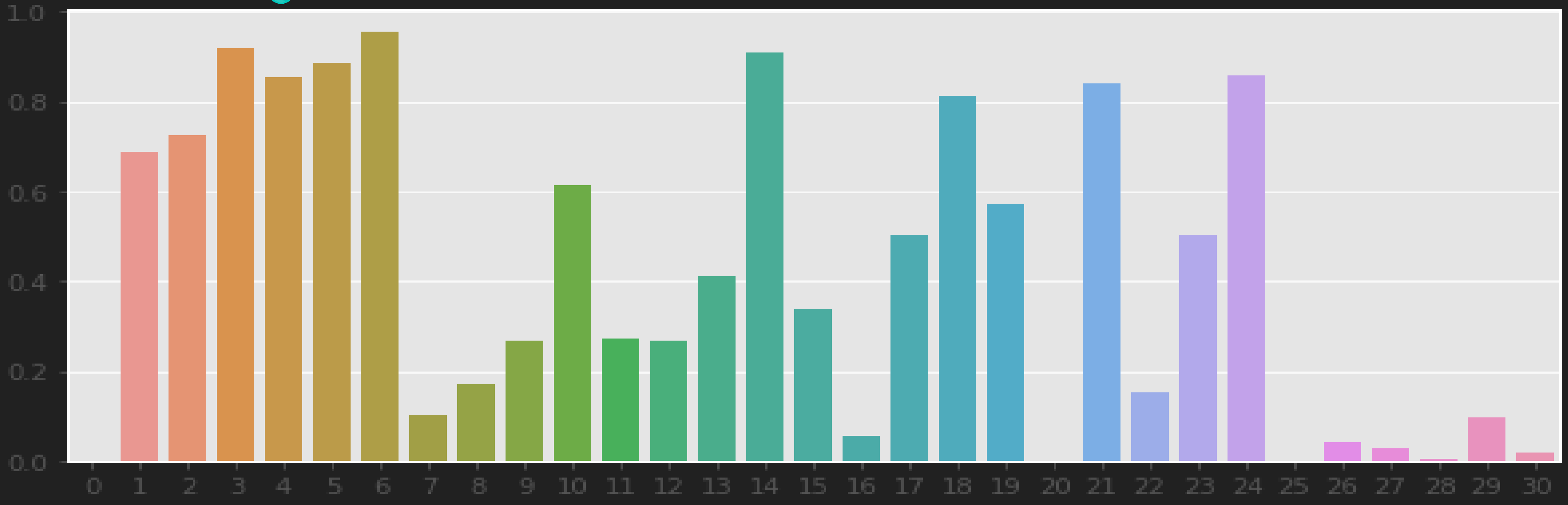


A multiple linear regression model is fitted over the Standardized data.

Cross validation score of worst model is still worst, and showed no improvement.

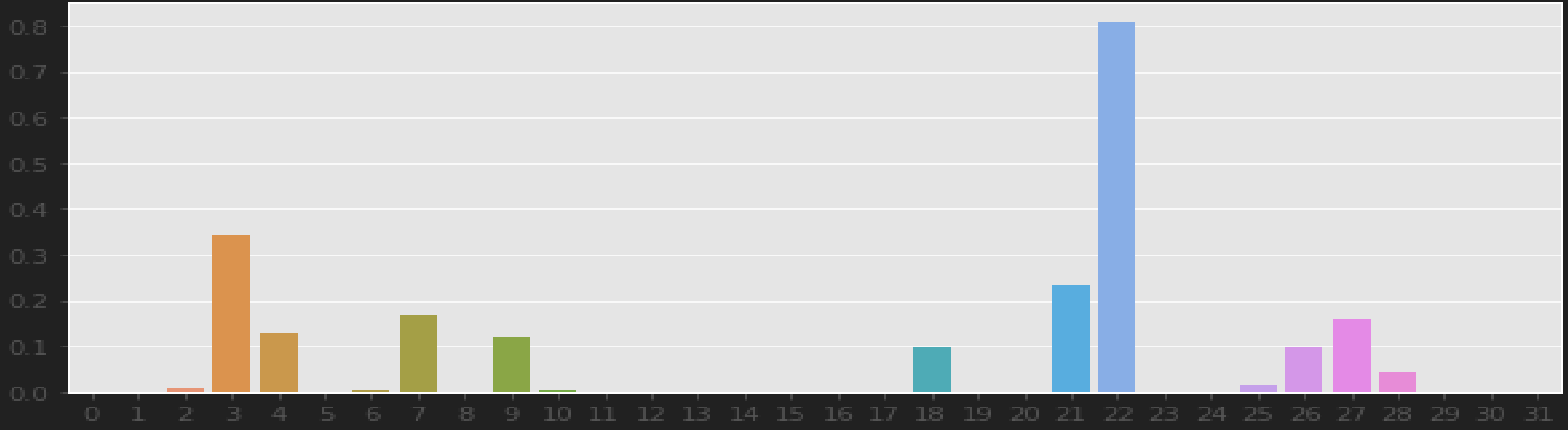
Let’s investigate this model

MODEL FITTING: INVESTIGATE THE WORST MODEL



Out of 30 X-variables, 28 are above alpha region.

MODEL FITTING: AFTER TREATMENT



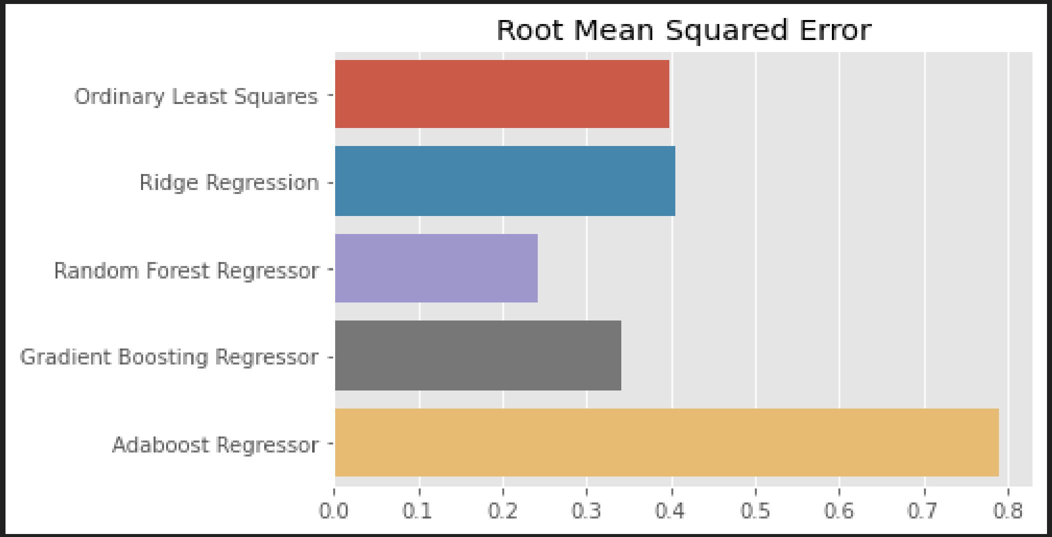
Removed date column: Already converted to weekday, dayOfMonth, month & year
Remvoed fullVisitorId: Indicates a particular visitor, don't signify any numeric relationship
Removed visitId: Indicate an individual visit, non-compatible candidate for regression
Removed visitStartTime: Not required because visit hour on 24-hour scale already in the stack

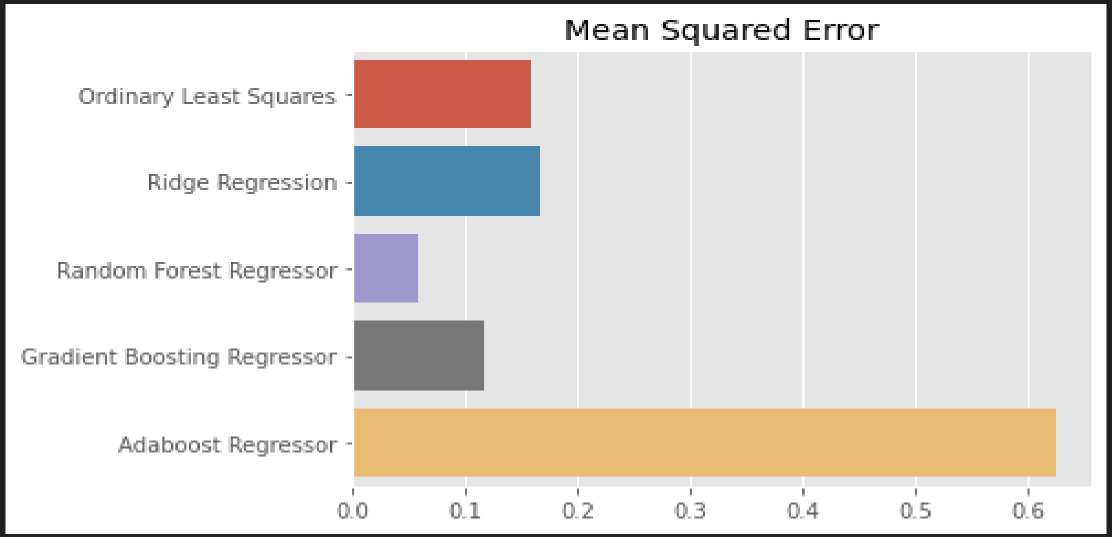
RESULTS & FINDINGS

- Finally five regression models are applied to data:
  - Ordinary Least Squares
  - Ridge Regression
  - Random Forest Regression
  - Gradient Boosting Regression
  - Adaboost Regression

RESULTS & FINDINGS

- Lowest MSE & RMSE error is observed for Random Forest Regressor
- Second best regressor is Gradient Boosting Regressor



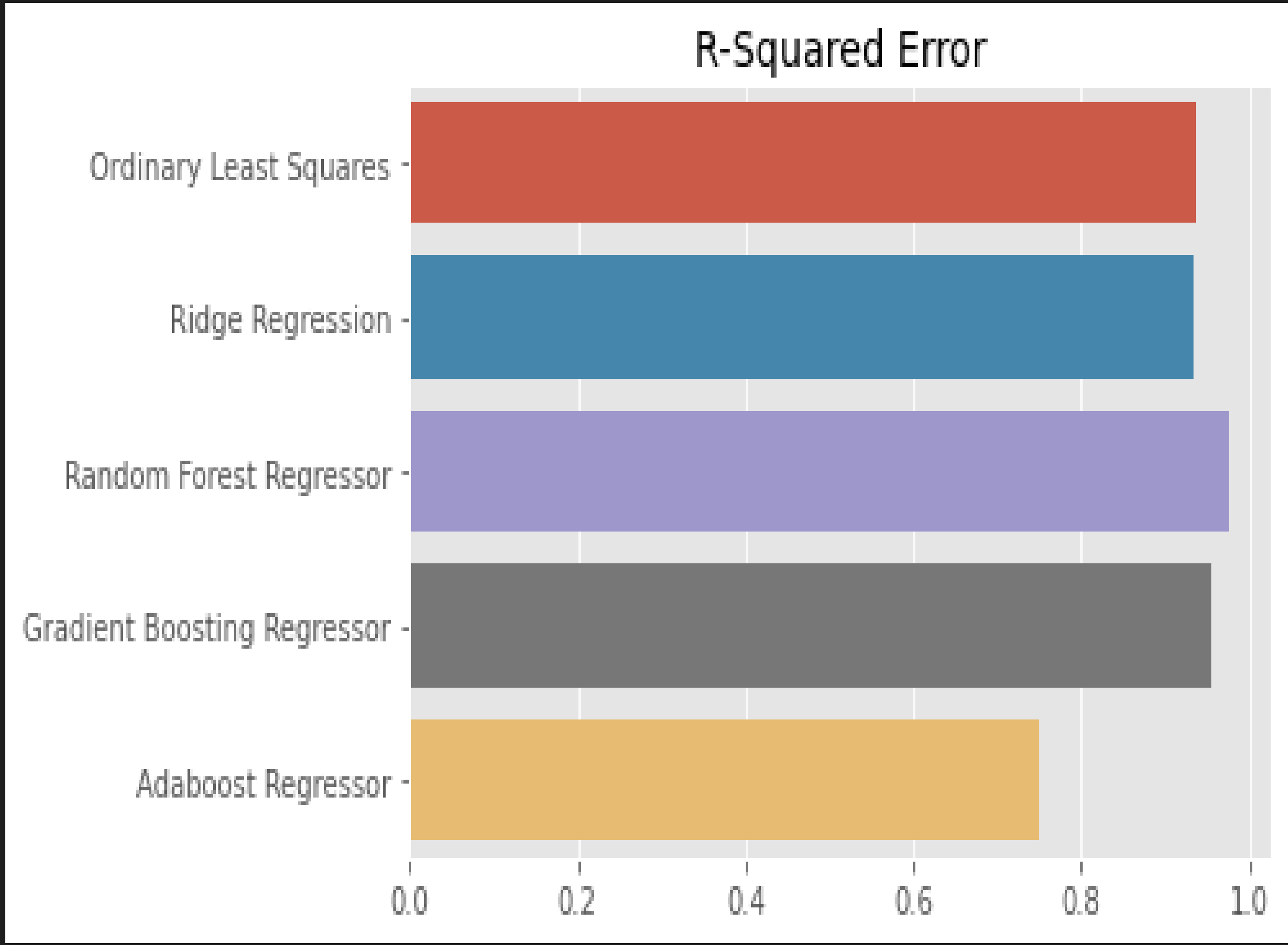


RESULTS & FINDINGS

- Highest R-squared error is observed for Random Forest Regressor (0.97).
- Random Forest Regressor accurately represents 97% variance in the data

- Gradient Boosting Regressor accurately represents nearly 95% of data

- Adaboost Regressor is worst performer, and accurately represents only 75% of data



RESULTS & FINDINGS: CAUSALITY ANALYSIS

- In this model, Adaboost algorithm seems not capable of identifying the shortcomings by high-weight data points, as dataset is highly random and high-weight data points doesn't correctly describe the variance. [1]
- On the other hand, gradient boosting depends upon controlling the step size in order to find the best Global Minima. Due to its global optimization, Gradient Boosting performed comparatively better in this case. [2]

CONCLUSIONS

- Random Forest Regression is best compatible algorithm to predict Google Analytics Customer Revenue
- Most visitors uses Windows OS
- Most visitors originates from American continent
- Mac OS users generates the highest in-class revenue percentage
- North American continent generates the highest revenue in entire world

CONCLUSIONS

- HOWEVER ONLINE ECOSYSTEM IS RAPIDLY CHANING AREA, FOLLOWING ARE SOME GUIDELINES TO GENERATE HIGHER REVENUE
  - Your website must be compatible and well tested for Top 3 browsers (Chrome, Safari and Firefox), which has approx. 93% of internet traffic.
  - Focus more on Search Engine (44%) & Social Media (21%) marketing
  - An effective online marketing campaign should target more people on Workdays than Weekends
  - You have only 4 minutes to catch user's attention.

REFERENCES

- [1] https://en.wikipedia.org/wiki/AdaBoost
- [2] https://en.wikipedia.org/wiki/Gradient_boosting
- [3] https://www.kaggle.com/c/ga-customer-revenue-prediction

THANK YOU

FOR YOUR ATTENTION