**High Level Design**

Google Analytics Customer Revenue Prediction

|  |  |
| --- | --- |
| Written By | Teena Jain |
| Document Version | 1.0 |
| Last Revised Date | 25-March-2022 |

**Document Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| Date Issued | Version | Description | Author |
| 25-March-2022 | 1 | Initial HLD -V 1.0 | Teena Jain |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

[Abstract: 4](#_Toc99083091)

[1. Introduction 4](#_Toc99083092)

[1.1 Why this High-Level Design Document? 4](#_Toc99083093)

[1.2 Scope 4](#_Toc99083094)

[1.3 Definitions 4](#_Toc99083095)

[2. General Description 5](#_Toc99083096)

[2.1 Problem Statement 5](#_Toc99083097)

[2.2 Proposed Solution 5](#_Toc99083098)

[2.3 Data Requirements 5](#_Toc99083099)

[2.4 Tools Used 5](#_Toc99083100)

[2.5 Constraints 6](#_Toc99083101)

[3. Design Details 6](#_Toc99083102)

[3.1 Process Flow 6](#_Toc99083103)

[3.1.1 Model Training and Evaluation 6](#_Toc99083104)

[3.1.2 Deployment Process 7](#_Toc99083105)

[3.2 Event Log 7](#_Toc99083106)

[3.3 Error Handling 7](#_Toc99083107)

[4. Performance 7](#_Toc99083108)

[4.1 Reusability 7](#_Toc99083109)

[4.2 Application Compatibility 7](#_Toc99083110)

[4.3 Resource Utilization 7](#_Toc99083111)

[5. Key Performance Indicator 8](#_Toc99083112)

[6. Conclusion 8](#_Toc99083113)

[7. Reference 8](#_Toc99083114)

# Abstract:

Create an automated system for predicting potential future business, finding potential customers based on the various parameters as decided by the machine learning algorithm. The purpose of the document is to explain the High architecture that would be used for developing the Google Store revenue prediction system. we are given with the user's past data and transactions (when they logged into G-store). so, by using this data we need to predict the future revenue will be created by those customers. We will build a predictive model using G-store data set to predict the total revenue per customer that helps in better use of marketing budget, and we will also interpret the most impacting element on the total revenue prediction using different models.

# Introduction

## Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary details to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

* Present all the design aspects and define them in detail.
* Describe the user interface being implemented.
* Describe the needed Python libraries for the coding.
* Describe the performance requirements.
* Include design features and the architecture of the project.
* List and describe the non-functional attributes like:
  + Security
  + Reliability
  + Maintainability
  + Portability
  + Reusability
  + Application Compatibility
  + Resource Utilization
  + Serviceability

## Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture(layers), application flow (Navigation), and technology architecture, The HLD uses non-technical and mildly technical terms which should be understandable to the administrators of the system.

## Definitions

|  |  |
| --- | --- |
| TERM | Description |
| DB | Database, the cloud platform where the data will be stored. Can be considered  cloud storage. |
| ML | Machine Learning |
| API | Application Programming Interface can be considered a website link from there we can extract information. |
| ide | Integrated Development Environment |

# General Description

## Problem Statement

We’re challenged to analyse a Google Merchandise Store customer dataset to predict revenue per customer. In every business it was proven about 80–20 rule., this rule tells us 80% of our revenue will be generated by only 20% of our potential customers. So, our goal is to predict the revenue that is going to be generated by those potential customers in the future.

## Proposed Solution

This project works with a two-year data set with user transactions in the Google Merchandise Store. We perform a thorough data pre-processing and aggregation, engineering features such as recency, frequency, and page visit statistics. We build an ensemble of LightGBM models to predict future revenues generated by the existing Google customers.

## Data Requirements

* + we need to extract data from Cassandra database train set and test set.
  + we will be predicting the target for all users in the posted test set: test set for their transactions in the future time period of December 1st, 2018, through January 31st, 2019.
  + Each row in the dataset is one visit to the store. Because we are predicting the log of the total revenue per user, not all rows in test.csv will correspond to a row in the submission, but all unique **fullVisitorIds** will correspond to a row in the submission.
  + some of the features are in JSON format so we need to parse those JSON columns., regarding this we will see in brief at the time of data reading.

## Tools Used

   Logo, company name

Description automatically generated 

**Circle

Description automatically generated A picture containing text, porcelain

Description automatically generated Logo

Description automatically generated**

. PyCharm is used as IDE

. For Visualisation of plot Matplotlib, Seaborn is used.

. Heroku is used for deployment

. Cassandra Database is used to retrieve, insert, and delete the database.

. Front end development is done using HTML/CSS.

. Python Flask is used for backend development.

. GitHub is used as version control system.

## Constraints

The System is user-friendly, the user will get all proper messages while using the web app. He/she also will get a proper error message if he/she has done something wrong on the web-app page. All the errors and results will be delivered in the easiest possible way and all the buttons are going to insert on the webpage are be labelled properly, so the user will not get confused to use the system.

# Design Details

## Process Flow

Below is the processed flow diagram is as shown.

Proposed Methodology

## **Model Training and Evaluation**

New Data Set

Prediction

Result of Evaluation

Test Set

Train Set

Model

Data Set

Prediction

## **Deployment Process**

Diagram

Description automatically generated

## Event Log

The system is logged every event so that the user will know what process is running internally. Logging is just because we can easily debug the issue.

## Error Handling

An error is defined as anything that falls outside the normal and intended usage.

# Performance

The result of google analytics Customer Revenue prediction project is more than 88% accurate. Its performance is better using tree-based algorithms.

## Reusability

The code and the modules created at the time of building the project is required to maintain all coding guidelines and full project code is written in a modular fashion. Our system incorporates the flexibility to work properly from any location. And it should handle any improper input value from the user by providing a meaningful error message so the user can correct his/her mistake and enter valid input to get the desired results.

## Application Compatibility

The different libraries in Python programming language, CSS, and HTML have been used to build the system. Flask has been used for making the web APIs and HTML/CSS has been leveraged to make the web application. All the components of the application are supposed to work properly, and it is required to produce a result without any major impediments in place.

## Resource Utilization

Our application utilizes the given resources efficiently and effectively. It uses an optimal amount of internet to work and call the APIs on the web page. Our system has been designed to use less computational to make the application faster. Our application will be deployed on cloud platform, and it is utilizing the resource given on the cloud and work properly.

# Key Performance Indicator

* Number of transactions of user.
* Session Quality Dimension.
* First session from the period start.
* Number of page views.
* Operating System.
* Country.

# Conclusion

The Google Analytics Customer Revenue Prediction is for assisting business owners and predict future revenue of potential customer and help them grow the business profits. The idea is to analyse the past data and use a machine learning induced application to predict the target value.

# Reference

* Google image for collection the logos and images.
* https://en.wikipedia.org/wiki/High-level\_design