Project Name: **Google Smartphone Decimeter Challenge 2022**

Author Name: **Ebad Akhter**

University of San Diego

Master of Science, Applied Data Science

Course: Foundations of Data Science

ADS 501, Section 1

Date: 9/12/2022

|  |
| --- |
| **Business Understanding** |

**Background**

Google, mostly used as a common search engine website, is the pioneer for Android Phones and Applications. “Googling something was all we once did with Google. Now we spend hours a day using its maps, videos, security cameras, email, smartphones and more.” (NYTimes, 2020) Almost 20 years ago, Google was used as one of the modern search engines which was simple and fast in fetching results. Fast forward 2 decades, and Google as a company is considerably different. There are multiple services, platforms, and devices that Google provides to its consumers and customers. Smartphones and various smart devices are becoming the norm of the modern era, with everything available at a touch/tap or simple “Hey, Google!” voice note, it is important for these devices to be as accurate and efficient as they can be.

One of these services from Google is Maps. (Google Maps, 2022) Maps was released in 2005 as a helpful way for people to navigate through their computers or smart phones, without the hassle of carrying paper maps. As the Maps application started to be used globally, new additions and extensions were added. It was open to the developers to add more features to the existing apps, as an open-source software. “About two years after launching Google Maps, we introduced real-time info on traffic conditions for more than 30 U.S. cities. And so, the phrase “there’s a lot of red...” was born.” (Elizabeth R., 2020)

With the traffic conditions getting worse as the population and usage of vehicles increase around the world, it is important to know the expanding roads and routes and to know the estimated time of arrival of specific lanes, like carpool or HOV. These and other useful features require precise smartphone positioning services. Machine learning models can improve the accuracy of Global Navigation Satellite System (GNSS) data. With more refined data, billions of Android phone users could have a more fine-tuned positioning experience. (Kaggle, 2022)

Figure below represents an organizational chart that identifies the names, divisions, and titles of the top ranked individuals who represent a respective department within Google. Each of the key individuals report directly to Google’s current CEO, Sundar Pichai.

An internal sponsor for this project is Kaggle Inc., which offers skills-based competitions to evaluate and enhance the field of data science. Kaggle provides the assigned dataset and materials for this project that was, “This year, co-sponsored by the Institute of Navigation, this competition continues to seek advanced research in smartphone GNSS positioning accuracy and help people better navigate the world around them. In order to build upon last year’s progress, the data also includes traces from the 2021 competition.” (Google Android Team, 2021.). Google’s Steering Committee consists of members that are employees of Wisconsin Skyward School Districts who volunteer their time and labor (Google Sites Steering Committee, 2022).

The business units most affected by this project include Google Maps Application development and management team, termed as “The Android GPS Team in Google” (Kaggle, 2022). Based on the objective of this project, this project would mainly be an additional feature to utilize the data from Global Navigation Satellite System (GNSS) and improve on the accuracy of smartphone location. This would highlight the features for the application and its strengths and weaknesses in managing and organizing the current usage of data. This could also involve the legal privacy protection team, as it would highlight their key roles in managing the appropriate usage of personal data and its content being shared within the Google Team and not being sold to a third-party data collector. For all the included Business Units, this project would serve as a proposal or a guide for the vice president of engineer and development to utilize the data from GNSS to improve estimated time of arrival and real-time traffic flow, while making sure the data being gathered from consumer smart devices is not being mishandled or misused that would damage the public trust for the company.

The problem areas include business management, operations, and customer care, which directly pinpoints towards the overall problem of the project, which is to access GNSS chipset’s raw measurements, which can be used to compute the smartphone’s position. The project has received clearance from Kaggle for data mining techniques and projects to be performed. For this project, it incorporates Machine Learning Algorithms that is designated to support and help machines learn from raw data. This could include usage of Deep Learning Algorithms as this data could include traces collected in harsher environments, such as deep urban areas with obstacles to satellite signals.

A solution being used to address the situation of the project is the use of previous 2021 Data Competition using Machine Learning and Deep Learning Algorithms, which is designed to support machines learn and analyze accurate positions, bridging the connection between the geospatial information of finer human behavior and mobile internet with much finer granularity. Building a machine learning model with a technique to process the dataset can highlight better road patterns or alternate routes that can also indicate possible construction zones or accident lanes to avoide. Mobile users could gain better lane-level coordinates, enhanced experience in location-based gaming, and greater specificity in the location of road safety issues. The solution would help consumers realize it's easier to get where they need to go.

**Business objectives and success criteria**

**Inventory of resources**

The personal resources that could be involved in this project include, but are not limited to, a project manager, data scientist, database administrator, IT staff member, market analysts, data mining experts, business experts, statisticians, and a legal and compliance team.

The hardware resources that will be provided and used for the duration of this project include an Excel csv dataset from Kaggle, Microsoft Excel or Google Sheets and Kaggle, a central processing unit (CPU), Dell PC installed with Windows program, Local Disk C with over 250 GB, Random Access Memory (RAM), Microsoft OneDrive or Google Drive, and Google Colab Notebook installed with Python 3. The data sources are the user sample data that were collected and included into Kaggle Inc.

Examples of data sources included for this project are a sample submission Excel csv file with 5.73 MB, a test csv file with roughly 36 different sub-files representing data for different days, and a train csv file with roughly 62 different sub-files, which adds up to approximately 98 different files of data for the project.

A relevant background knowledge involves information that is vital to analyze and comprehend a situation or problem. For the objective of this project, which is to calculate the precise location from the GNSS data, the use of relevant background knowledge allows readers to process and distinguish multiple meanings of words that sound familiar from others.

**Risks and contingencies**

**Terminology**

The following lists feature two glossaries of terminology relevant to the project and both glossaries were drafted with no prior availability of glossaries from Kaggle.

Glossary – Google terminology:

* 3D mesh: To make any 3D model in a computer, you need to give the computer spatial coordinates that it can understand.
* AI – which stands for artificial intelligence – is a blanket term given to any kind of intelligence that machines (usually computers) seem to demonstrate.
* ML (machine learning) is a kind of AI, it’s computers demonstrating the intelligent trait of learning from prior iterations and other inputs.
* Algorithm: An algorithm is a set of instructions for a computer or another kind of mathematical – that is, purely logical – brain to follow.
* API: An API – application programming interface – is a computer interface that gives people access to various computer programs or databases.
* Cartography: The science – and art – of studying and making maps.
* GNSS: Global Navigation Satellite System - chipsets that provide raw measurements, which can be used to compute the smartphone’s position.

These were gathered from Google Maps: Explained (Atlist, 2020).

Glossary – Data Mining terminology:

* Accuracy: The percentage of total correct predictions divided by the total number of instances.
* Algorithms: A technological process through calculations or problem-solving operations.
* Alternative hypothesis: A hypothesis that highlights a difference between two or more variables is anticipated by the researchers and that the observed pattern of a given data is not due to a chance occurrence.
* Analytics base table (ABT): A table designated for establishing analytical models and assesses the future behavior of a subject.
* Analytics solution: An approach made by an analytics practitioner involving machine learning.
* Average error: A contrast between the predicted and actual value.
* Bar chart: A chart that portrays categorical data through rectangular bars.
* Central Processing Unit (CPU): A circuitry that delivers computer programming instructions.
* Chi-squared test: A test to evaluate the occurrence of statistically significant difference between one or more variables or categories.
* Confusion Matrix: An analytical tool used to capture the processes in specific detail during an evaluation test. This serves as the basis for calculating additional performance measures.
* CRISP-DM (Cross Industry Standard Process for Data Mining): A model including six primary phases that highlights the lifecycle of a predictive data analytics project.
* Data: Facts or information that are collected through analysis and observation.
* Data Mining: The practice of analyzing large databases to predict the future based on the data.
* Data Pre-processing: The concept and procedure to transform raw data into a clean data set.
* Excel: A spreadsheet from Microsoft that equipped with computational or calculation capabilities along with graphing tools and creating tables.
* Gain: A measure that assesses how accurate predictions made by models were in comparison to random guessing.
* Histogram: A diagram that presents a visual representation of the distribution of numerical data.
* Colab Notebook: A web application for users to create and share computational documents” (Colab, 2022).
* Labeled Dataset: A dataset that labels the target feature with values.
* Linear Regression: A statistical analysis type that formulates a relationship between two variables: dependent and independent variable.
* Machine Learning: A data-pattern process.
* Machine Learning Algorithms: Algorithms that convert the process of learning a model and captures the relationship between the descriptive and the target feature in a dataset.
* Missing data: Blank or null values within a dataset that can generate a significant effect towards drawn conclusions from the data.
* Multinomial: Multiple target levels.
* Natural Language Processing (NLP): A text mining component that enables machines to read or analyze text.
* Null Hypothesis: The hypothesis that there is no statistical significance between a set of given observed variable, between two sets of data, or a measured phenomenon.
* Overfitting: A modeling error that indicates when a function is nearly identical or aligned to a limited sets of data points.
* Performance Measure: A measurement that can capture, numerically, how well the predictions made by the model match those that were expected.
* Precision: The number of true positives divided by the addition of true positives and false positives.
* Predictive Modeling: “A statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and exiting data” (Ali, 2020).
* Recall: The number of true positives divided by the addition of true positives and False Negatives.
* Resolved Strategy: A step-by-step, decision-making procedure used to evaluate ethical principles in practice.
* Segmentation: The process of dividing gathered data into separate categories or sections

**Project plan/ Order of tasks (Insert Gantt Chart here for Module 6 submission.)**

|  |
| --- |
| **Data Understanding** |

**Initial data collection report**

**Data description report**

**Data exploration report**

**Data quality report**

**References**

Brian X. Chen. (2020, October 20). It’s Google’s World. We Just Live in It.

<https://www.nytimes.com/2020/10/20/technology/doj-google.html>

Google Maps (2022). Google Maps.

Retrieved from <https://www.google.com/maps>

Elizabeth Reid (2020, February 06). A look back at 15 years of mapping the world

<https://blog.google/products/maps/look-back-15-years-mapping-world/>

Kaggle. (2022). Google Smartphone Decimeter Challenge 2022

<https://www.kaggle.com/competitions/smartphone-decimeter-2022>

Google Android Team (2021). Google Smartphone Decimeter Challenge 2021

<https://www.kaggle.com/competitions/google-smartphone-decimeter-challenge/overview>

Google Sites Steering Committee (2022). Steering Committee.

<https://sites.google.com/view/wiskywardusergroup/steering-committee>

The Org (2022). Google – Org Chart

<https://theorg.com/org/google/embed>

ATLIST (n.d.). Google Maps: Explained.

<https://www.atlistmaps.com/google-maps-explained/glossary>

Ali, R. (2020, September 23). Predictive Modeling: Types, Benefits, and Algorithms. Oracle Netsuite.

<https://www.netsuite.com/portal/resource/articles/financial-management/predictivemodeling.shtml#:~:text=Predictive%20modeling%20is%20a%20method%20of%20predicting%20future,is%20so%20commonly%20used.%20What%20Is%20Predictive%20Modeling%3F>

Colab (2022). Google Colab

<https://colab.research.google.com/>