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 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.

Figure Google Organization Chart

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. 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 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 construction zones or accident lanes to avoid. 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**

Google’s mission is to organize the world's information and make it universally accessible and useful (Google’s Approach, 2022). By connecting the geometrical content of its Google Maps databases to digital traces that it collects, Google can assign meaning to space, transforming it into place. While Google’s stated objective is “to organize the world’s information and make it universally accessible and useful,” its Google Maps endeavor allows it to organize your world’s information, making it personally accessible and useful (Anthony S., Andrew C., & Arie C., 2015).

The goal of this competition is to compute smartphone’s location down to the decimeter or even centimeter resolution which could enable services that require lane-level accuracy such as HOV lane ETA estimation (Kaggle, 2022). This brings up a few questions that need to be resolved in order to resolve this problem in a better way.

* What is the model of Google Maps? In the years ahead, how would Google Maps want to be recognized in society compared to other maps services?
* Would using the chipset to get accurate smartphone location reflect negative intensions to lose a respectable number of users?
* What type of data usage would be deemed appropriate to secure the privacy of consumers?
* How much would loosing certain information or part of data retrieved from the GNSS Chip reflect on different departments within Google Maps?
* To what extent would Google Maps go in accepting a solution to have their model be implemented in a manner that it would have least impact on how Google is seen and used in the communities?

The top requirement for Google Maps is to retain accurate information from the users around the globe, while keeping that data secure and making sure it does not get into the wrong hands. The criteria for a successful outcome to the project from Google Maps’ point of view would include a Machine Learning model based on raw location measurements from Android smartphones collected in open sky and light urban roads using datasets collected by the host. This will help produce more accurate positions, bridging the connection between the geospatial information of finer human behavior and mobile internet with improved granularity. As a result, new navigation methods could be built upon the more precise data. (Kaggle, 2022)

**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, domain experts for geographical data and usage of GNSS chipsets, 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 zip file with roughly 36 different sub-files representing data for different days, a train zip file with roughly 62 different sub-files, and a metadata collection that consists of 3 different files, which adds up to approximately 102 different files of data for the project. The collection consists of CSV, TXT, and other files to assist with 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.

**Requirements**

To have a successful model deployment, the organization funding this project has a few key requirements, assumptions, and constraints that must be addressed and complied by all devices to fulfill the objective of this project. The primary requirement needed for this task would be the full cooperation and a signed licensed agreement from all participants, whose devices were used to collect the GNSS Chipset data, and the top executives of Google, Google Maps, and other relate applications/services that are being utilized in the project, including CEO, Sundar Pichai. This agreement would emphasize that all parties will refrain from inflicting any type of harm (verbal or physical) towards one another, abstain from manipulating or attempting to seal even a single row of the current data presented for the project, abstain from sharing the gathered data and confidential information outside of the project’s framework, and fully cooperate towards the duration of the project.

For all participants in this project and their devices, they will be provided with an additional resource, which is a two-way authentication for a security measure to abstain from having any private, personal information not relevant to the project mistakenly leaked. Any participant or top executive who violates any of these terms will jeopardize the overall objective towards this project and will have to compensate the damage inflicted to the sponsor, Kaggle. Anyone who is reported as violating a single condition multiple times or multiple conditions once, will be ejected for the duration of the project. Which could also jeopardize the project, should key participants and executives tasked to evaluate the data are not present. For this project, the participants and Google will also hold one another accountable and compensate one another should either party attempt to inflict harm or manipulate a share of the given data towards the other.

This is a project that is looking to resolve a substantial matter in a universal manner for all participants and top executives of Google. Every top executive and board member representing Google will be involved throughout the project for a primary reason, which is to assess how the duration and the completion of this project could impact each department of Google, from a financial standpoint to marketing and technological aspects. It is the task for all employees working in the data mining department to maintain a consistent, daily pattern of evaluating and managing the gathered data and trends for this project before presenting the final model that can be comprehended through words and figures towards a diversity of individuals from different professional backgrounds.

Figures that include visualizations, will also be included to portray a physical, comprehendible analysis of the current data points of geolocation and the objective at stake. While the data mining experts, analysts, and statisticians manage most of the evaluation and analysis towards the data, the system administrator and head supervisor of the Android GPS team at Google will have the final word towards the approval of a model that will be directed for a full endorsement from CEO Sundar Pichai.

To fulfill the requirements needed towards this project, a schedule has been drafted and composed to complete each on a weekly basis and leading to the final stage of the final projected model.

|  |  |  |
| --- | --- | --- |
| Week | Stage | Objective |
| Week 1 | Business Understanding | Evaluate Google Map’s business segments, draft model approach, and gather necessary resources. |
| Week 2 | Data Understanding Pt. 1 | Access resources, consult with IT, and create dashboard. |
| Week 3 | Data Preparation | Build database through Colab Notebook. |
| Week 4 | Data Understanding Pt. 2 | Accumulate and assess data reports. |
| Week 5 | Data Modeling | Build a Machine Learning model and evaluate the model for any potential errors. |
| Week 6 | Data Evaluation | Evaluate the model for final approval from Android GPS team.  Present the first approved model to Google’s top executives and CEO for official deployment. |
| Week 7 | Data Deployment | Deploy the final approved model for Google to implement. |

The schedule outlined above has received approval from sponsor Kaggle and Google for the Android GPS team to acquire and analyze the data towards the objective of the project so long as the process remains confidential and that the data resources gathered and utilized do not exit the facility of where the data mining process will occur. While both Kaggle and Google have given the data mining project team a maximum of six months to fulfill the objective of this project, every member of the Android GPS team came to a united agreement with a goal to have it complete in half the given time (approximately three months) to not impose any interference amid Google’s annual business process. Any conflict related to technology or the data mining process from a certain stage, however, will result with the Android GPS team having to repeat the stage currently on or for added measure, assess the requirements from the previous step.

**Assumptions**

The assumptions towards this project include the following:

* In the current age of rising technological devices and features, where teens and young adults like to stay connected or use their devices for directions or check-ins, all devices gathered for this project are no older than 5 years.
* Google’s current algorithm is programmed to differentiate between certain express lanes along the route or toll roads to reflect faster/quicker routes.
* In terms of the quality of the data, Google provides precise records of all routes, public or private, pertaining to emergency events and any other event.
* Google also provides detailed statements pertaining to the cash flow and how every data point recorded for a single day affects the business’s income and economy.
* The data is easily accessible pending necessary compliances and usage agreements.
* In terms of external factors, Google is fed through alerted messages of any event or incident to boost their technological features against other geolocation platforms.
* Google’s source of power is generated through the non-stop constant flow of data records and connections being established virtually between each device through different cell phone towers, which fuels the company’s economy.
* The price to incorporate Machine Learning (ML) towards the final model is assumed to be no greater than $50. The price of a central processing unit (CPU) along with added computer data storage, Local Disk (GB), Random Access Memory (RAM), and a Google Drive combined are assumed to be no greater than $20,000.

The assumptions listed above pertain towards the whole duration of the data modeling processes and can be verified through each phase of the data modeling process. An example would be the first phase, where all the data records for this project are accumulated. From here, the first assumption that assumes that all the gathered data points were gathered through those devices that are less than 5 years of age could be assessed and verified to see if it is true. This could be a crucial step in identifying which specific areas of the data to focus more towards than others and see if the benefits and risks towards this project from a business perspective remain the same or would change.

**Constraints**

The constraints that are to be addressed throughout the project include the following:

* For legal constraints and purposes, Google holds all users accountable towards using their platform while complying towards the User Agreement. Google reserves the will to, “share personal information with companies, organizations or individuals outside of Google if we have a good faith belief that access, use, preservation or disclosure of the information is reasonably necessary to meet any applicable law, regulation, legal process or enforceable governmental request.” (Google Maps Privacy Policy, 2017).
* For timescales, the data modeling team voluntarily committed to develop a model to resolve Google’s issue in less than two months. With most data modeling processes taking about a minimum of one year to complete, every individual and supervisor of the team will now have to dedicate themselves to spend extra time each day of the week, managing through the given data and analyze what model to compose. All team members are to remain committed towards the requirement schedule listed above to have the project complete on time, starting with each stage of the week.
* For resources, the Random Access Memory (RAM) form cannot retain a data modeling progress made once a computer or technological system shuts down or restarts. This memory form is unstable, so this would require every team member working in the data mining field to retain an extra copy of a Local Disk drive, a Hard Drive, and download extra storage space through the computer systems. Unfortunately, most of the Hard Drives obtained for this project can carry up to 1 MB of data while the files for the project add up to approximately 1.43 MB of data. Even though the computer data storage and Local Disks contain enough storage for the csv files and the data mining process, the Hard Drives with added space are primary necessities for all members to retain for this project in the case that the project encounters technological issues and all progress towards a project stage ended up being deleted nor saved at a certain checkpoint.
* In terms of budget constraints, the data modeling team is running on a strict budget limit despite receiving a copy of all listed resources for free. Any purchase of additional resources, hardware, and the data files come with a hefty, six-figure price and the more errors are made towards the project, the more back-up resources will be needed and the most money it will cost for the team to complete the objective. The budget issue correlates to the issue of maintaining a steady progressive trend towards the project and having each stage completed within a tight deadline.
* For accessibility rights, both parties (the participants and Google’s top executives/board members) must sign an agreement to allow the data mining team to receive free and full access of the content of specific messages pertaining to the objective. Once the agreement is signed, Google will also support the project’s goal to obtain the data by providing unlimited access towards their databases, files, and contents of messages with no password needed.
* In continuation of accessibility rights through a technical perspective, a Database Browser is needed to retain or access all csv files in proper format without having the data files, columns, or rows manipulated and rearranged in a wrong order. There are no constraints towards operational or data management systems as these will be provided for the data mining team once given access to Google’s account and the messages for the project.
* Relevant knowledge towards this project is not accessible since the given data is limited to the hidden clues or resources provided through the csv files.

**RESOLVEDD Strategy**

Google is garnered with constraints to maintain its integrity and emphasis of promoting a high moral and healthy standard while in operations for their users. An ethical decision-making process, which will be utilized to assess Google’s past dilemmas while ensuring that the process of this project does not instigate any additional dilemmas towards the company, is the RESOLVEDD strategy. The RESOLVEDD strategy involves a step-by-step methodology with an intent for users to note, “ethical issues and tackling them through their own set of values or through an ethics theory of their choice” (Vakkuri & Kemell, 2019).

Diagram

Description automatically generated

Figure The Nine Steps of the RESOLVEDD Strategy

For the Review stage, what is already known about the objective towards this project is that Google has also become a breeding ground for misused data to be mixed with actual information. This is one of multiple issues that Google has had to endure in previous years, but the tech giant desires to amend their share of issues with advanced data-driven resources and a concise decision-making approach to distinguish and categorize emergency incidents from nonemergency events. This is a procedure Google is committed to follow in order to stand out as the preferred and most-trusted web mapping platform to use on a daily basis compared to others.

For the Estimate stage, the main conflict directly points to Google’s current algorithms that are unable to differentiate the distinct meanings of certain key words that sound identical but appear differently. The first solution from Google is that they are committed to rearranging and organizing all location points under certain categories instead of being jumbled up in one massive root. Another solution is that Google continues to operate with their current strategy without taking the time to address the potential issues the procedure could inflict towards others. For the first solution, Google would create a new algorithm that includes specific categories for certain incidents to access information only related to specific routes. For example, creating a HOV Lane Category on Google Maps would provide or allow families to know when they can be in the far left, fastest, lanes to help reach their destination quicker and avoid traffic jams at junctions or key exit points. This method would be beneficial in keeping most of the users satisfied by the application feature, but the probable consequence of this solution would involve both Google and the users should a certain category make its way towards a different category section or is incorrectly labeled due to a word that sounds familiar, which addresses Google’s issue in failing to distinguish different meanings towards sound-relatable words. The second solution would enable Google to continue to operate in their current state with no major effect implemented towards their algorithms or language processing sections. This draws a serious consequence of having misinformed lanes or incidents continuing to blend with informed lanes or incidents, resulting with agencies or residents to accept the content of information. The likely impact towards the first solution would make Google appear more organized when designating and categorizing within certain, but specific categories. The likely impact towards the second solution would make Google appear stricter and more limited towards the use of information and retained for the public to view.

The main value upheld by Google through the two solutions would be the platform’s stance to promote a healthy, well-informed environment for their users. Through a new data-driven approach, Google would be making their platform trustworthy and resourceful to use to accumulate and search for information through a faster process than the Internet. A common value violated by both solutions would be Google’s promotion towards a transparent environment for all users to represent any incident or route without limitations. Enabling an unlimited, transparent environment is a dangerous procedure in that this could enable some users to post false or harmful content towards others. These values would cover the Evaluation stage and the impact (positive or negative) each could make towards Google and the users.

For the Decision phase, the best solution would be the first solution that would introduce a new algorithm that includes content categories for certain users to access information only related to specific routes, a similar process made through the Internet or website. As mentioned previously, Google would create content categories designated for specific routes (e.g.: HOV Lanes, Express Lanes, Construction Lanes, Detours, Emergency/Incident, and others.). When users observe an emergency category, they would have the ability to both view and post emergency-related messages within the category. Each one of these categories can be made public for any Google Maps user to be a part of, but the newly advanced algorithms from Google would be able to recognize the content of messages that do not relate to the topic based on certain word choices.

Lastly for the Defend phase, a key objective or conflict towards this solution would be that it still fosters a transparent environment for users to post sensitive content that relates towards a certain category or is being directed towards certain users. While this solution would not entirely resolve some of Google’s internal issues related to restraints, limitations, or privacy matters, this solution would be effective in improving the platform’s structure through a data-driven strategy and designate messages pertaining specific routes within classified and separate categories.

**Risks and contingencies**

Location data has been central to the privacy debate for years now. Security risks towards privacy and data integrity pose as risk factor towards this project in the event of data being hacked or information being leaked unwillingly. First iOS and then Android have given us options to deny, restrict and approximate such data from the dozens of apps that would consume data if they were allowed to. Google plays down these privacy risks, stating that Google Maps is designed to protect users’ information. It provides controls to easily manage settings and use industry-leading technologies like differential privacy to keep users’ data safe. Google Maps continues to make the best and most accurate way to navigate and explore the world—providing rich local business information, best in class search and navigation, and helpful features like the COVID layer and live busyness information.

Google Maps and other such maps in Geographical Information System have a lot of significance in every one’s life for in modern world due to technological development as well as contemporary needs in travelling, business planning, agriculture, e-marketing supply chain management, census and planning and excessive use of mobile phones. Being a revolutionary technology, it attracts the users from its inception. It has been revolutionary in having an impact on one’s daily life by helping one explore geographical locations virtually anywhere on the whole planet. It has become a norm that people use Google Maps before or while commuting to a certain place as most of the people rely on it to provide the shortest or fastest route to a destination. Google Maps has had a profound impact not only one’s personal life, but has opened new avenues of marketing, business intelligence, urban planning, infrastructure strategy development, as well as traffic engineering. Hence, no one can deny the impact it has had on our society in a brief period. (Burney, A. , Asif, M. , Abbas, Z. and Burney, S., 2018).

A major risk factor towards this project Google Maps has security concerns associated with its use. This is because whenever a user is searching for a geographical location on Google Maps, there is no way to ascertain their intention. As result, the user requests whatever, it is provided without much security checks or personal specific logging history. Millions of users of Google Maps utilize the assistance provided by this outstanding technology in everyday life and get better solutions of their problems for identifying any location as per their requirements. But many potential threats in form of unethical users having bad intensions, achieve their goals without being noticed. Since its inception, such users are utilizing this technology for malicious acts.

Another risk factor towards this project would be that Google has the same problems with hacking and phishing that any other internet behemoth has. There’s not much that can be done other than taking a holistic security-first approach to data protection. In these respects, Google works hard to protect consumer data from hackers. One of the biggest headlines surrounding Google’s approach to privacy and security isn’t about an outside threat but an inside one. In 2018, reporting found that Google still stored user location data even when users paused data collection. Google then had to close Google+ posthaste after a report showed that Google left 500,000 users’ personal information out in the open for the world to see. Indeed, the problem with this approach (beyond the violations of basic security practices) is that Google’s place as an advertisement company means that its objective is to collect as much personalized and unique data from its users as possible. Over 80% of Google’s revenue depends on its ability to do that. (Frankie W., n.d.).

**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 an AI, its 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 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.
* Metadata: Metadata is "data that provides information about other data", but not the content of the data, such as the text of a message or the image itself.

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 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

**Data Mining Goals and Success Criteria**

The intended output of this project will be to develop a model based on raw location measurements from Android smartphones collected in open sky and light urban roads using datasets collected by the host. This will then be leveraged to meet and satisfy Google’s objective towards the intended project in a business perspective. Other goals include ensuring that the benchmarks for this project fit to standard needs, the model produces a result that is of statistical significance, and if the outcomes from the predictive model can be acted upon for future use.

Your work will help produce more accurate positions, bridging the connection between the geospatial information of finer human behavior and mobile internet with improved granularity. As a result, new navigation methods could be built upon the more precise data.

To generate an effective machine learning model, a data mining approach utilizing predictions as a starting point is highly recommended to evaluate how the current data generated for the project presents a visualization of the data accuracy. The prediction process is a procedure that can detect the information within the dataset and highlight any variables or factors that could be generating the basis or the issues of the objective. Another approach would involve segmentation of the locations that will help produce more accurate positions, bridging the connection between the geospatial information of finer human behavior and mobile internet with improved granularity. This step involves a description-type approach to extract specific knowledge from the data sets and utilize the extracted information to support or solve the basis of the project through multiple application domains.

The key predictions that will be needed for determination through Google Colab Notebook would involve coding and evaluating the location dataset to compare data between similar devices and see if it constructs a path with similar locations around same times.

**Project plan/ Order of tasks (Gantt Chart.)**

Chart

Description automatically generated

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

**Initial data collection report**

The project from Kaggle challenge provides data from a variety of instruments useful for determining a phone's position: signals from GPS satellites, accelerometer readings, gyroscope readings, and more. As the project’s design is focused on post-processing applications such as lane-level mapping, future data along a route will be available to generate positions as precisely as possible. To encourage the development of a general GNSS positioning algorithm, in-phone GPS chipset locations are not provided, as they are derived from a manufacturer proprietary algorithm that varies by phone model and other factors. The datasets provided along with the project included a set of train dataset csv file and txt files, a set of test dataset csv and txt files, a sample dataset csv file, and a set of metadata files. All four of these sets jointly contain an extensive list of files that contain different attributes and data values.

Looking at the sample\_submission.csv file (Kaggle, 2022), we can see the list of attributes in that file is exceedingly small, tripId, UnixTimeMillis, LatitudeDegrees, LongitudeDegrees. All these attributes are non-null, with tripId as Object, UnixTimeMillis as Int64 while Latitude and Longitude Degrees as Float. The UnixTimeMillis is an integer number of milliseconds since the GPS epoch (1970/1/1 midnight UTC). Converted from GnssClock and the Longitude and Latitude Degrees are the WGS84 latitude, longitude (in decimal degrees) estimated by the reference GNSS receiver (NovAtel SPAN). When extracting from the NMEA file, linear interpolation has been applied to align the location to the expected non-integer timestamps. (Kaggle, 2022).

Collectively, the train and test dataset files could contain more than 12k columns (as per the summary on Kaggle’s Competition Data Site (Kaggle, 2022)). But this list would be adjusted and defined more throughout the project evaluation process.

**Data description report**

The dataset acquired from Kaggle, accessed through Excel, and needed for the duration of the project contains test, train, and metadata folders along with a sample\_submission.csv file. Taking one of the Train datasets files for GooglePixel4XL, we have device\_gnss.csv – which is the data from GNSS Chipset, similar for test datasets as well. It contains 47 columns with about 90153 entries. Each row contains raw GNSS measurements, derived values, and a baseline estimated location. This baseline was computed using correctedPrM and the satellite positions, using a standard Weighted Least Squares (WLS) solver, with the phone's position (x, y, z), clock bias (t), and isrbM for each unique signal type as states for each epoch. Some of the raw measurement fields are not included in this file because they are deprecated or are not populated in the original gnss\_log.txt. The list of columns and how to use them is provided along with the project description as below:

* MessageType - "Raw", the prefix of sentence.
* utcTimeMillis - Milliseconds since UTC epoch (1970/1/1), converted from GnssClock.
* TimeNanos - The GNSS receiver internal hardware clock value in nanoseconds.
* LeapSecond - The leap second associated with the clock's time.
* FullBiasNanos - The difference between hardware clock (getTimeNanos()) inside GPS receiver and the true GPS time since 0000Z, January 6, 1980, in nanoseconds.
* BiasNanos - The clock's sub-nanosecond bias.
* BiasUncertaintyNanos - The clock's bias uncertainty (1-sigma) in nanoseconds.
* DriftNanosPerSecond - The clock's drift in nanoseconds per second.
* DriftUncertaintyNanosPerSecond - The clock's drift uncertainty (1-sigma) in nanoseconds per second.
* HardwareClockDiscontinuityCount - Count of hardware clock discontinuities.
* Svid - The satellite ID.
* TimeOffsetNanos - The time offset at which the measurement was taken in nanoseconds.
* State - Integer signifying sync state of the satellite. Each bit in the integer attributes to a particular state information of the measurement. See the metadata/raw\_state\_bit\_map.json file for the mapping between bits and states.
* ReceivedSvTimeNanos - The received GNSS satellite time, at the measurement time, in nanoseconds.
* ReceivedSvTimeUncertaintyNanos - The error estimate (1-sigma) for the received GNSS time, in nanoseconds.
* Cn0DbHz - The carrier-to-noise density in dB-Hz.
* PseudorangeRateMetersPerSecond - The pseudorange rate at the timestamp in m/s.
* PseudorangeRateUncertaintyMetersPerSecond - The pseudorange's rate uncertainty (1-sigma) in m/s.
* AccumulatedDeltaRangeState - This indicates the state of the 'Accumulated Delta Range' measurement. Each bit in the integer attributes to state of the measurement. See the metadata/accumulated\_delta\_range\_state\_bit\_map.json file for the mapping between bits and states.
* AccumulatedDeltaRangeMeters - The accumulated delta range since the last channel reset, in meters.
* AccumulatedDeltaRangeUncertaintyMeters - The accumulated delta range's uncertainty (1-sigma) in meters.
* CarrierFrequencyHz - The carrier frequency of the tracked signal.
* MultipathIndicator - A value indicating the 'multipath' state of the event.
* ConstellationType - GNSS constellation type. The mapping to human readable values is provided in the metadata/constellation\_type\_mapping.csv file.
* CodeType - The GNSS measurement's code type. Only available in recent logs.
* ChipsetElapsedRealtimeNanos - The elapsed real-time of this clock since system boot, in nanoseconds. Only available in recent logs.
* ArrivalTimeNanosSinceGpsEpoch - An integer number of nanoseconds since the GPS epoch (1980/1/6 midnight UTC). Its value equals round((Raw::TimeNanos - Raw::FullBiasNanos), for each unique epoch described in the Raw sentences.
* RawPseudorangeMeters - Raw pseudorange in meters. It is the product between the speed of light and the time difference from the signal transmission time (receivedSvTimeInGpsNanos) to the signal arrival time (Raw::TimeNanos - Raw::FullBiasNanos - Raw;;BiasNanos). Its uncertainty can be approximated by the product between the speed of light and the ReceivedSvTimeUncertaintyNanos.
* SignalType - The GNSS signal type is a combination of the constellation name and the frequency band. Common signal types measured by smartphones include GPS\_L1, GPS\_L5, GAL\_E1, GAL\_E5A, GLO\_G1, BDS\_B1I, BDS\_B1C, BDS\_B2A, QZS\_J1, and QZS\_J5.
* ReceivedSvTimeNanosSinceGpsEpoch - The signal transmission time received by the chipset, in the numbers of nanoseconds since the GPS epoch. Converted from ReceivedSvTimeNanos, this derived value is in a unified time scale for all constellations, while ReceivedSvTimeNanos refers to the time of day for GLONASS and the time of week for non-GLONASS constellations.
* SvPosition[X/Y/Z]EcefMeters - The satellite position (meters) in an ECEF coordinate frame at best estimate of “true signal transmission time” defined as ttx = receivedSvTimeInGpsNanos - satClkBiasNanos (defined below). They are computed with the satellite broadcast ephemeris, and have ~1-meter error with respect to the true satellite position.
* Sv[Elevation/Azimuth]Degrees - The elevation and azimuth in degrees of the satellite. They are computed using the WLS estimated user position.
* SvVelocity[X/Y/Z]EcefMetersPerSecond - The satellite velocity (meters per second) in an ECEF coordinate frame at best estimate of “true signal transmission time” ttx. They are computed with the satellite broadcast ephemeris, with this algorithm.
* SvClockBiasMeters - The satellite time correction combined with the satellite hardware delay in meters at the signal transmission time (receivedSvTimeInGpsNanos). Its time equivalent is termed as satClkBiasNanos. satClkBiasNanos equals the satelliteTimeCorrection minus the satelliteHardwareDelay. As defined in IS-GPS-200H Section 20.3.3.3.3.1, satelliteTimeCorrection is calculated from ∆tsv = af0 + af1(t - toc) + af2(t - toc)2 + ∆tr, while satelliteHardwareDelay is defined in Section 20.3.3.3.3.2. Parameters in the equations above are provided on the satellite broadcast ephemeris.
* SvClockDriftMetersPerSecond - The satellite clock drift in meters per second at the signal transmission time (receivedSvTimeInGpsNanos). It equals the difference of the satellite clock biases at t+0.5s and t-0.5s.
* IsrbMeters - The Inter-Signal Range Bias (ISRB) in meters from a non-GPS-L1 signal to GPS-L1 signals. For example, when the isrbM of GPS L5 is 1000m, it implies that a GPS L5 pseudorange is 1000m longer than the GPS L1 pseudorange transmitted by the same GPS satellite. It's zero for GPS-L1 signals. ISRB is introduced in the GPS chipset level and estimated as a state in the Weighted Least Squares engine.
* IonosphericDelayMeters - The ionospheric delay in meters, estimated with the Klobuchar model.
* TroposphericDelayMeters - The tropospheric delay in meters, estimated with the EGNOS model by Nigel Penna, Alan Dodson and W. Chen (2001).
* WlsPositionXEcefMeters - WlsPositionYEcefMeters,WlsPositionZEcefMeters: User positions in ECEF estimated by a Weighted-Least-Square (WLS) solver.

Another file that we have in the dataset for Train and Test is device\_imu.csv. These are the readings from phone's accelerometer, gyroscope, and magnetometer. It has 8 columns and 734856 records per device file. Some of the useful column information is provided below:

* MessageType - which of the three instruments the row's data is from.
* utcTimeMillis - The sum of elapsedRealtimeNanos below and the estimated device boot time at UTC, after a recent NTP (Network Time Protocol) sync.
* Measurement[X/Y/Z] - [x/y/z]\_uncalib without bias compensation.
* Bias[X/Y/Z]MicroT - Estimated [x/y/z]\_bias. Null in datasets collected in earlier dates.

Train data for each device also consists of ground\_truth.csv files. These are references for all the locations at expected timestamps. This contains 9 columns with 3362 records. Some of the column descriptions are:

* MessageType - "Fix", the prefix of sentence.
* Provider - "GT", short for ground truth.
* [Latitude/Longitude]Degrees - The WGS84 latitude, longitude (in decimal degrees) estimated by the reference. GNSS receiver (NovAtel SPAN). When extracting from the NMEA file, linear interpolation has been applied to align the location to the expected non-integer timestamps.
* AltitudeMeters - The height above the WGS84 ellipsoid (in meters) estimated by the reference GNSS receiver.
* SpeedMps\* - The speed over ground in meters per second.
* AccuracyMeters - The estimated horizontal accuracy radius in meters of this location at the 68th percentile confidence level. This means that there is a 68% chance that the true location of the device is within a distance of this uncertainty of the reported location.
* BearingDegrees - Bearing is measured in degrees clockwise from north. It ranges from 0 to 359.999 degrees.
* UnixTimeMillis - An integer number of milliseconds since the GPS epoch (1970/1/1 midnight UTC). Converted from GnssClock.

**Data exploration report**

Data exploration is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more. The first step in this section was to understand the type of data we have with this project and what it represents. After importing all the files related to one of the training dataset smartphones, Google Pixel 4XL, we can plot the values in a heatmap of San Francisco, CA, to see what the coordinates show us on a map. The Google Pixel 4XL dataset folder was selected for initial data exploration process since it consisted of the basic device IMU and GNSS chip dataset, along with the supplemental data and ground truth. The device\_imu data consisted of the main attributes, such as MessageType and Measurement Coordinates. Ground Truth consisted of Longitude and Latitude with Speed in miles per second (mps). Figure 2 shows us exactly how this training device and its GNSS device was used to collect data along the route and compare with the ground truth values to form a route.

Map

Description automatically generated

Figure Google Pixel 4XL Train Data Heatmap

As our goal of the project is to collect accurate driving data with the help of GNSS Chipset data from smartphones, we can use the supplemental data provided in the project to convert it into data frames and into specific points on the map and make a pattern of route out of the daily routine of a specific device user. Based on each device specific data from a certain date (as the data is divided into files/folders of specific days of the year), we can make a map from the coordinates provided. Figure 3 shows the data of Google Pixel 4XL in the form of a graph based on different longitude and latitude points in the dataset.

A picture containing chart

Description automatically generated

Figure Google Pixel 4XL Longitude and Latitude points.

To understand the data provided in the device\_imu files, we need to understand the data description or message within the MessageType attribute. The MessageType Attribute contains readings from the phone's accelerometer, gyroscope, and magnetometer. The accelerometer is an in-built comment of a smartphone to measure its acceleration. It tracks the different motion like shaking, tilting, swinging, and rotating and accordingly change the orientation of your app. Gyroscope in a smartphone provides a GUI that enables a user to select menus by tilting the phone. One can deflect the phone slightly to go up and down the contact list. It enables a smartphone to trigger preset commands basis different motions. Finally, the magnetometer sensor measures the magnetic field for all three physical axes (x, y, z) in μT (micro-Tesla). This specification defines two new interfaces: Magnetometer that reports calibrated magnetic field values, and. UncalibratedMagnetometer that reports uncalibrated magnetic field values. Figure 4 below shows the distribution of uncalculated messages from the above mentioned three sensors in the Training dataset of Google Pixel 4XL’s device\_imu.csv file.

Chart, bar chart

Description automatically generated

Figure Bar Chart of device\_imu MessageType Attribute for Google Pixel 4XL

The bar chart from Figure 4 reflects that the most data collected from the device is mostly through the Magnetometer Sensor. This can be useful to understand the orientation of the device to mark the direction points accurately when drawing a map using the longitude and latitude points.

**Data quality report**

The datasets provided for each device within the project consists of every variable and value needed to establish a machine learning model that can generate map routes for better detailed accuracy when generating maps for real world routes. The accuracy of the data can also assist in calculating lane-level accuracy, such as HOV lane ETA estimates. The dataset does not contain missing values that would become a hurdle in solving the priority concern. Some attributes within the ground\_truth.csv and device\_gnss.csv are null by default entirely, AltitudeMeters and LeapSecond, and can be ignored or dropped to avoid any issues in further analysis. There are a number of attributes that do not provide value to the project individually and can be consolidated or aggregated into one attribute of importance or be excluded entirely to avoid additional computations. These minor discrepancies can be avoided by performing few data preprocessing tasks such as dropping all nulls and unwanted attributes to clean the data before processing any model training. The featurettes are removed before being encoded and vectorized, so that when the modeling process occurs, the data will appear clear and concise. This procedure could also be beneficial for the machine learning algorithms to ensure their highest level of accuracy when analyzing and classifying the content of data without being manipulated by vague insertions of unwanted features.

Despite these issues, the project team have been allocated enough time, funding, and resources to undergo a series of data mining procedures to adjust each of the variable’s mass-collected data, address the ambiguous or missing information portrayed or excluded for this project, and develop as many ideas before establishing and selecting a precise model that satisfies Google’s or Google Maps’ business and their customers.

**References**

Google Maps (2022). Google Maps.

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

Elizabeth R. (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/>

Google’s Approach (2022). Our approach to Search: Maximize access to information. Retrieved from

<https://www.google.com/search/howsearchworks/our-approach/#:~:text=Google's%20mission%20is%20to%20organize,it%20universally%20accessible%20and%20useful>.

Anthony S., Andrew C., & Arie C. (2015). How – and why – Google is transforming the map. Retrieved from:

<https://theconversation.com/how-and-why-google-is-transforming-the-map-35238>

Burney, A. , Asif, M. , Abbas, Z. and Burney, S. (2018) Google Maps Security Concerns. Journal of Computer and Communications, 6, 275-283. doi: 10.4236/jcc.2018.61027

Frankie W. (n.d.) Data Privacy & Google: What Are the Security Risks? Retrieved from

<https://www.uscybersecurity.net/?p=10969>

Zak D. (2021). Why You Shouldn’t Use Google Maps On Your iPhone After Update. Retrieved from:

<https://www.forbes.com/sites/zakdoffman/2021/06/19/apple-iphone-warning-switch-google-maps-to-apple-maps-after-chrome-and-gmail-backlash/>

Google Maps Privacy Policy (2017). Privacy Policy. Retrieved from:

<https://maps.google.com/intl/en/policies/privacy/google_privacy_policy_en.pdf>

Vakkuri, V., & Kemell, K.K. (2019). Implementing AI Ethics in Practice: An Empirical Evaluation of the RESOLVEED Strategy. Lecture Notes in Business Information Processing.

<https://doi.org/10.1007/978-3-030-33742-1_21>.