**Project Proposal**

**Team Members:**

1) Anurag Kunchi

2) Manisha Karnati

3) Poornima Kolasani

4) Snehal Choudhari

5) Suhana Shaik

**Title:**

311 City Service Requests in DC Completion Time Prediction

**Project Overview:**

"311" is a non-emergency telephone number and online service that residents can use to report non-urgent issues or request city services. These requests can cover a wide range of topics, including reporting issues such as potholes, graffiti, damaged street signs, or requesting services like trash collection or park maintenance. The goal of 311 service is to provide residents with a convenient way to connect with local government agencies to address and resolve issues in their community. The objective of this project is to predict the completion time for 311 City Service Requests in District of Columbia, D.C. The completion time refers to the duration it takes for the local government to address and resolve the reported issue or fulfill the service request.

**Research Questions:**

1. **What Factors Influence Service Request Completion Time?**

This is our primary task. We would need to identify the key factors that have a significant impact on the time it takes to complete service requests. This may include service type, location, priority, and some other relevant variables.

1. **Which Features Are Most Predictive for Completion Time?**

**We would need to analyse the dataset to identify which features** are the most informative for predicting completion time. Also, if there are some features that are more influential than others?

1. **How Does Request Priority Impact Completion Time?**
2. **Is There Seasonal Variability in Completion Time?**
3. **How Accurate Can Completion Time Predictions Be?**
4. **What are the Most Challenging Service Request Types to Predict Completion Time For?**

We would need to identify and analyze specific service request types that pose challenges in predicting completion time. We will check for the types that are consistently harder to estimate.

**Dataset:**

We are planning to use the following dataset for our project:

311 City Service Requests in 2021, D.C.

Link: <https://opendata.dc.gov/datasets/DCGIS::311-city-service-requests-in-2021/about>

Size: 132 MB

Samples: 360,817

**Results of Basic EDA:**

During the preliminary EDA, we found the following key results (PSB attached .ipynb file for the basic EDA):

The dataset contains 360,816 samples, where each sample represents entry of a call starting from the request came till the request is closed.

1. Our target variable would be “Timespan”, which will be a continuous variable.
2. We have identified continuous features in the dataset, they are:

OBJECTID, ADDDATE, RESOLUTIONDATE, SERVICEDUEDATE, SERVICEREQUESTID, LATITUDE, LONGITUDE, etc.

1. The dataset shows a portion of missing values, which we plan to handle through imputation. We have removed some columns after analysing that they won’t impact the task.
2. correlation among potential features:

A screenshot of a computer

Description automatically generated

Introduction:

Goal is to determine which Machine Learning Algorithm is best for our specific problem.

We cleaned the data, performed data exploration and visualization using Python and its libraries and used ML algorithms to make predictions.

We evaluated the performance of Logistic Regression, Gradient Boosting, Random Forest, Gaussian Distribution and XGBoost on a dataset with a target and set of features.

Accuracy/Performance depends on dataset, algorithms

Problem Statement:

Given a dataset containing information about **XXXXXX**, the goal is to determine best machine learning model that can predict the XXXXXXXXX based on various features. The features include details such as the XXXXXXX, geographical coordinates.

Summary of Data:

This dataset consists of 360816 rows with 36 columns.

A screenshot of a computer program

Description automatically generated A screenshot of a computer

Description automatically generated

The dataset is divided into training and test datasets which consists of 70% and 30% of dataset respectively. This is an unbalanced dataset.