



NYC Pothole Prediction

A Data-Driven Exploration of Potholes in NYC with
Predictive Modeling, Spatial Analysis, and Strategic
Recommendations

Sarmad Maqbool





Outline

- The Problem
- Solution Proposal
- Introduction to Data
- Exploratory Data Analysis
- Feature Engineering
- Initial Modelling



The Problem

In NYC's busy streets, where people spend 40 mins daily commuting, road quality is crucial.

Picture this: daily journeys disrupted by potholes and unpredictable hazards, endangering safety and delaying commutes.

This project is vital. No prior work in this area exists, and I aim to assist authorities in tackling this issue effectively.



The Solution

This project aims to predict zip codes of pothole locations across NYC using latitudes and longitudes leveraging machine learning classification problem solutions.

By analyzing historical data and road maintenance schedules, we'll develop a robust predictive model. Our objective is to provide real-time insights to municipal authorities, helping them proactively address pothole-prone areas. Ultimately, this initiative seeks to enhance NYC's urban transportation system and improve the quality of life for its residents and visitors.



The Impact

Potholes pose a common challenge for drivers and pedestrians, causing vehicle damage and traffic congestion. Efficient pothole management can save millions in repair costs and reduce environmental impact. This project aims to enhance NYC's transportation infrastructure, benefiting businesses, residents, and the environment.

Steps Involved in the Solution

- Acquiring Data and Intro
- Data Cleaning
- Feature Engineering
- Data Analysis
- Presenting Insights
- Statistical Analysis
- Modelling Data
- Model Evaluation



Acquiring Data

NYC Open Data 311 Service Requests Dataset: This dataset, available on NYC Open Data, contains information about service requests, including those related to potholes. It includes details such as the date of the request, location, and the status of the reported issue. • Street Pothole Work Orders - Closed (Dataset) | NYC Open Data,

<https://data.cityofnewyork.us/Transportation/Street-Pothole-Work-Orders-Closed-Dataset-/x9wy-ing4/data>



Introduction to Data

There are 17 columns in this dataset and 370,000 rows. Time range for this dataset is 2004 to 2023

- FID: Object, Unique Identifier of the Table.
- Shape Geometry, Polyline
- DefNum: Text, Defect Number.
- InitBy: Text, The unit that initiated the service action
- HouseNum: Text, House or building number on the street
- OnFaceName: Text, Pothole Location: Main Street
- OnPrimName: Text, Pothole Location: Main Street's Primary Name
- FrmPrimNam: Text, Pothole Location: From Street
- ToPrimName: Text, Pothole Location: To Street
- SpecLoc: Text, Defect Specific Location
- Boro: Text, Borough Code B – Brooklyn X – Bronx M – Manhattan
Q – Queens S – Staten Island
- Source: Text, Origin of the Report
- RepStatus: Text, Street Pothole Repair Status XCL = Closed
- RptDate: Date, Date the street pothole was reported
- RptClosed: Date, Date the street pothole report was closed
- Shape_Leng: Double, Length of Polyline in Feet



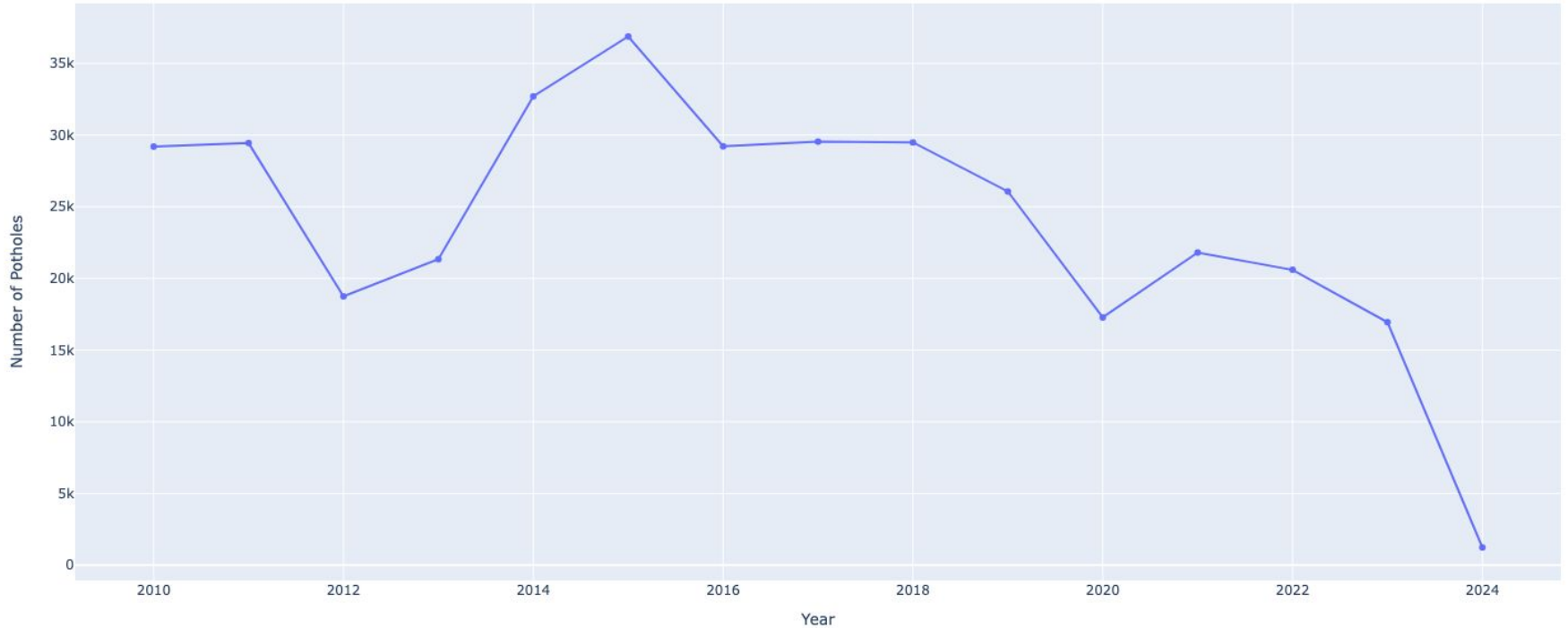
Data Cleaning

We have 124,790 nulls in the house_num column and 240,426 nulls in spec_loc column and few other nulls in other columns related to location.

The first approach would be to fill the missing data but we have large number of nulls we can't actually replace or impute the missing values from other columns.

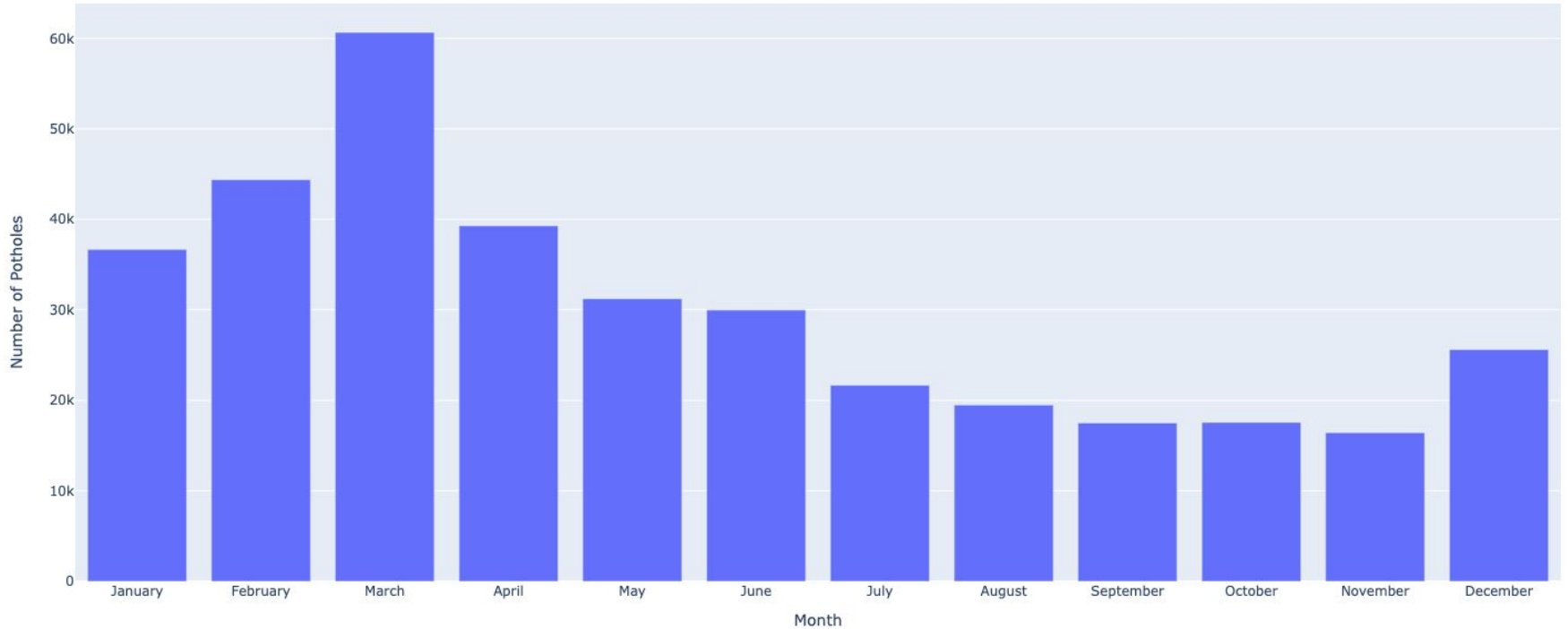
I'll create new columns from the existing data to handle nulls.

Number of Potholes on Yearly basis



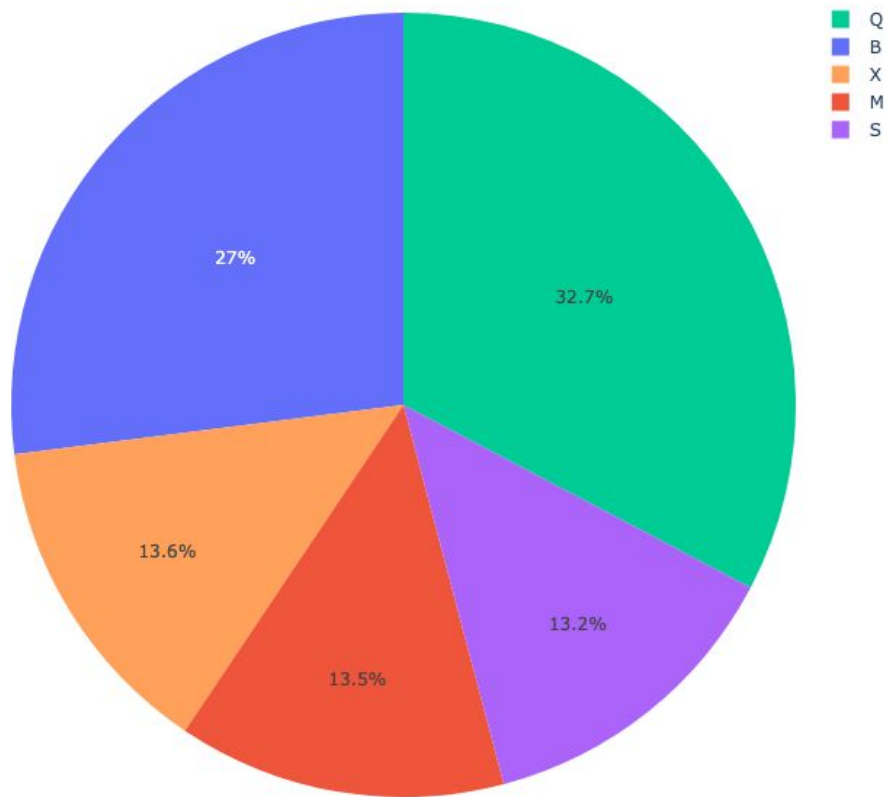
The highest number of potholes recorded in NYC in a calendar year was in 2015 with slightly over 36K potholes. Overall the yearly average is 24032 potholes per year. Whereas it took on average about 4 days to fix a pothole.

Number of Monthly Potholes Reports



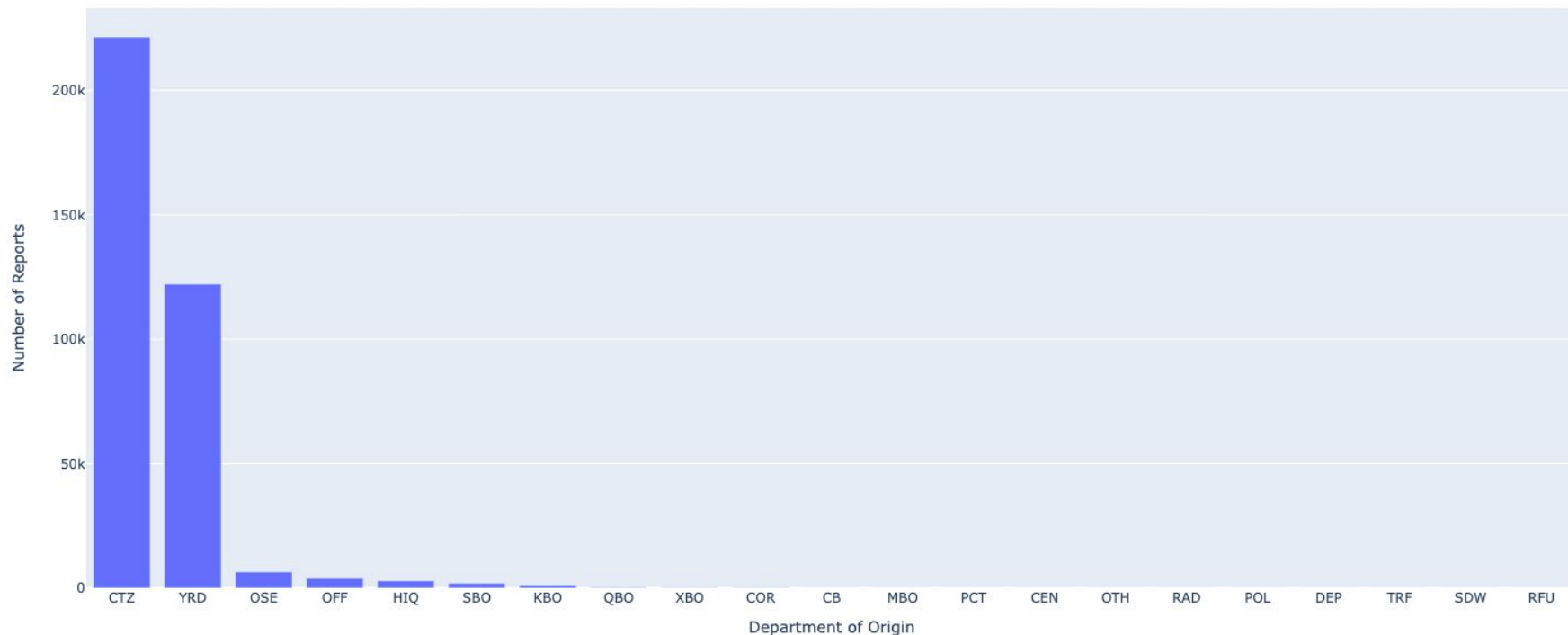
The trend to note here is that the pothole numbers are maximum in the spring season which helped me to unveil an interesting fact that If the water freezes and thaws over and over, the pavement will weaken and continue cracking.

Distribution of Potholes by Boroughs



We can see that the most number of potholes occurrences are in Queens and then Brooklyn. We can assume that number of potholes are dependent on the area of Boroughs and population.

Number of Pothole Service action initiated by Departments



The citizens department has handled the most number of complaints regarding to potholes and then Yard department contributed a lot with the pothole repair. Both departments contributed almost 90% to the reports generated for potholes in this dataset.

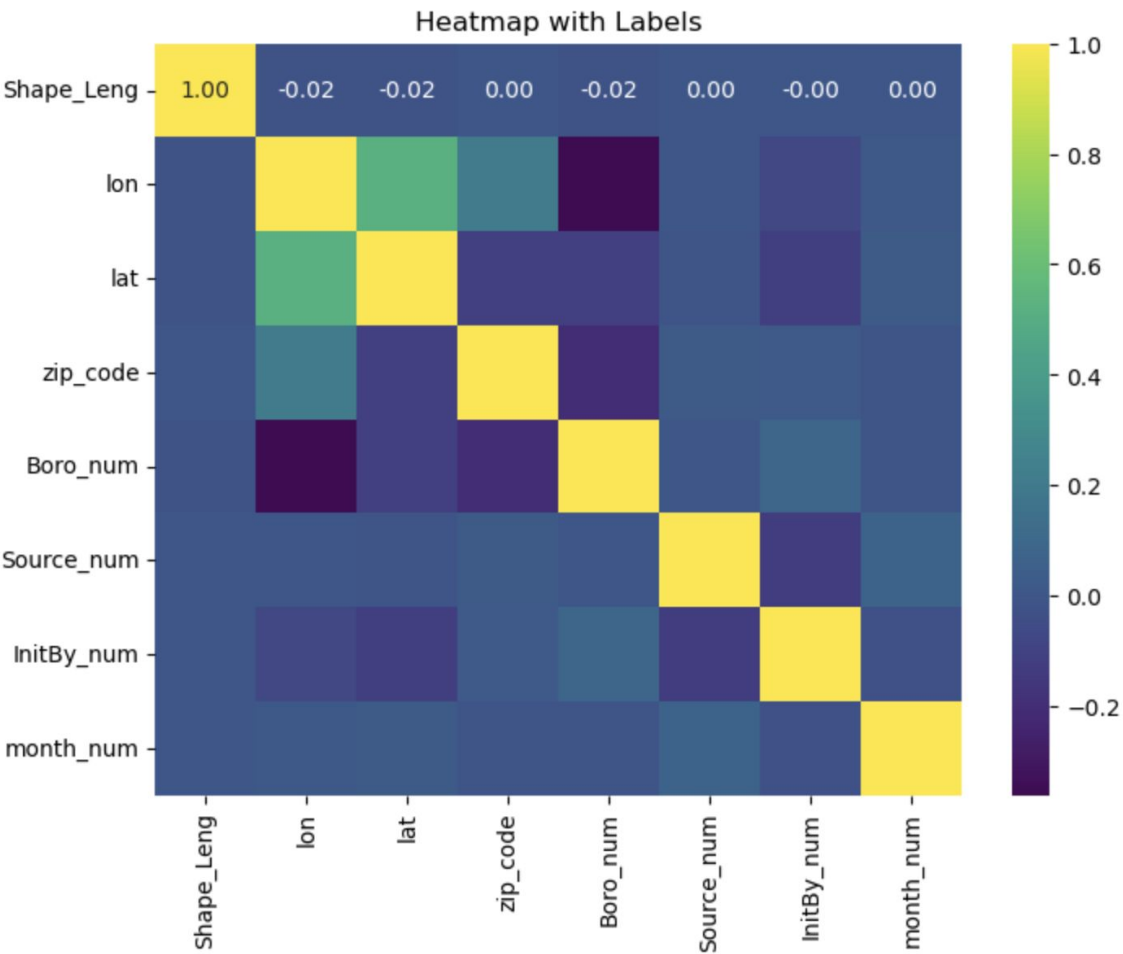


Feature Engineering

I have featured engineered zip_codes from “the_geom” column which is a string of longitudes and latitudes and it will be our target column. We have 228 unique zip codes (large number of classes to predict)

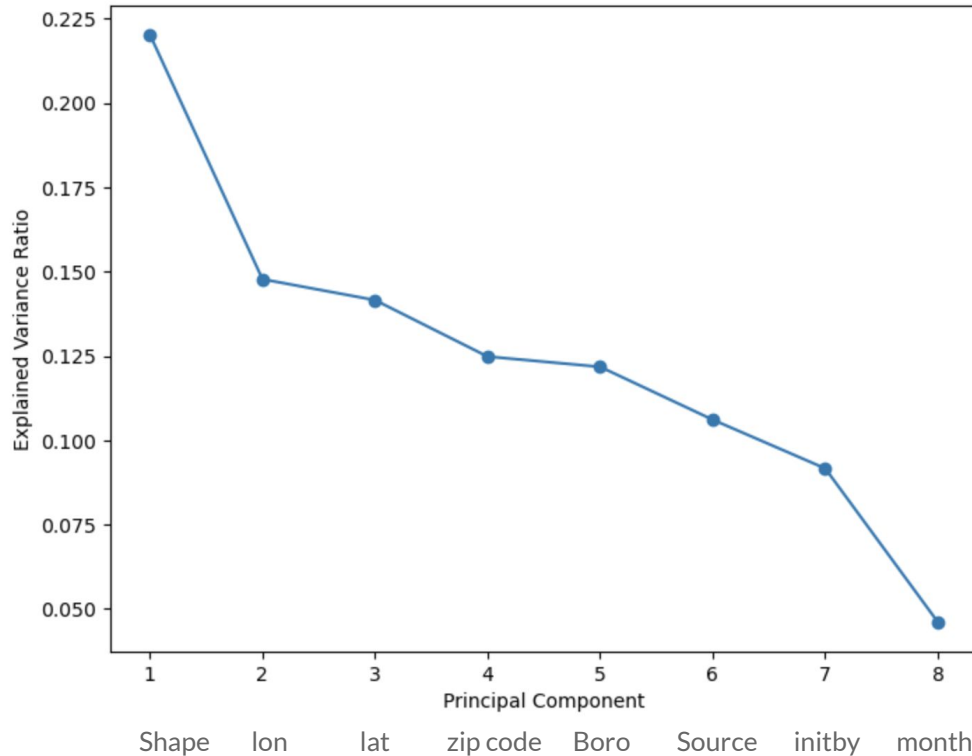
The process was reversing the latitudes and longitudes using Python’s GeoPy library.

I have also feature engineered the number of days took to fix a pothole.



Overall there is a trend of negative and zero co-relations between features which is not providing a useful insight to gather the importance of features for the occurrence of potholes. So far latitudes and longitudes have a good correlation with zip code. Feature name "Shape_Leng" is not actually helping at all. So I am going to drop this feature.

Contribution of each feature in terms of variance.



The feature with the highest variance is "Shape_Leng" as we have different sizes of potholes.


Other than that rest of the features are pretty close to each other in terms of variance. Which is in my opinion is not that good for modelling as we are predicting a large number of zip codes and the model is supposed to learn well.



Initial Modelling

For initial modeling I have tried the most common ML algorithms.

1. Logistic Regression
2. Random Forest
3. Support Vector Machines
4. Neural Networks
5. KNN



Evaluation Metrics for Baseline Model - Logistic Regression

Accuracy

- Out of all the instances in the test dataset, the model predicts the correct zip code for approximately 68% of them.
- The remaining approximately 32% of the instances are incorrectly predicted by the model.

Precision: which means that approximately 66.3% of the instances predicted as positive (belonging to a certain zip code) are indeed correct.

Recall: means that approximately 68.3% of the actual positive instances (belonging to a certain zip code) are correctly identified by the model.

F1 score: 64.9% means that the overall performance of the model is reasonably balanced between precision and recall.

Evaluation Metrics - Accuracy for other models

Random Forest: The model is slightly overfitting and it is learning too much from the data. I have to check for data leakage and optimization. Train and test 100% and 92%.

Support Vector Machines: This model is also not overfitting and performing quite good. Train and Test 76%.

Neural Networks: This algorithm has performed best among all with the train and test accuracy of 84%. The model is not overfitting and so far it looks more promising than others.

K-Nearest Neighbors: This model is definitely overfitting with the train accuracy of 83% and test accuracy of 74%.

Next Steps

→ Advance Modelling

Hypertuning of Parameters, Pipeline, PCA reduction and iterations

→ Discuss Model Performance and Errors

→ Handling Class Imbalanced Data

→ Enhance the Target by introducing population and area of zip codes



Questions?

