CSE558 : Data Science

Project Report Submission

Team Name : Big Dawgs
Ishwar Babu, Mohit Bhardwaj, Priyash Shah

Introduction

The dataset was taken from the link below, and the dataset contains over a million records detailing individual arrests made in New York city from 2006 to 2020. The dataset has 19 features some of the important ones are stated below

- **Geospatial Data :** Latitude, Longitude, X_COORD_CD, and Y_COORD_CD maps for visual analysis.
- **PERP_RACE :** Offers insights into racial data and, when combined with other features, can reveal patterns in serious crimes
- **OFNS_DESC :** General offense description, broader than the detailed PD_DESC.
- **LAW_CODE :** Provides formal offense charges.
- **ARREST_BORO :** Identifies boroughs for potential neighbourhood safety analysis in NYC.

Dataset Link : NYPD_Arrest

Our main aim behind this project was to analyze arrest trends, demographic disparities, and geospatial patterns in NYC, aiming to identify crime hotspots, predict arrest trends, and explore demographic impacts on arrest rates. This analysis will help in law enforcement in resource allocation, policy-making and target interventions. Along with this would support safer neighborhood planning, fair policing and informed policy decisions.

Problems Faced

Due to the raw nature of the dataset there were a few major issues in the dataset that needed to be addressed with the help of preprocessing methods, few of which are mentioned below :

a) Presence of Missing Values : There were many columns in various features that had many missing values (NULL/NaN) which is important to handle as it leads to biased analyses or errors in the model training. In order to tackle this we identified such columns and replaced the NULL/Nan values with appropriate replacements such as average, predictions or categorical values like “Unknown”.
b) String Date format : The initial data had the “Arrest Date” column in string format (eg. “06/18/2008”) which is not feasible in order to provide a detailed analysis such as seasonal or yearly trends. We approached this problem by breaking down the date string into 3 different columns for date, month and year. This facilitates

time-series analysis, enabling insights into temporal patterns and arrest trends over time.

c) Data Type Filtering : The initial data frames had inconsistent data types within a single column which may occur due to incorrect data entries or improper data import. We filtered out the rows with data types different from the primary data type of a specific column and reset the index in order to maintain continuity thus removing mixed data types leading to minimal errors.

Hypothesis tests and their Conclusion:

Below is a list of Null Hypothesis that we proposed followed by experiments to test them along with corresponding results and analysis.

i) There is **no significant difference** in the counts of males and females in arrest data : In order to test this hypothesis we did the z-test for proportions and obtained a p-value of 0.0, which is less than alpha(0.05). Thus **rejecting** the null hypothesis and indicating significant difference in arrest count between genders.

ii) Arrest distribution is **uniform** across racial groups : A chi-squared goodness of fit test was done which gave a p-value of 0.0, thus **rejecting** the null hypothesis and indicates variation across racial groups. x

iii) Crime rates are **uniformly distributed** across all months : Doing chi-squared goodness of fit test gives us a p-value of 0.000977, less than alpha (0.05) thus **rejecting** the null hypothesis, indicating monthly variations in crime rate.

iv) Drug-related offenses are **uniformly distributed** over the years : We got a p-value of 0.0 on doing the chi-squared goodness of fit test indicating that the trend in drug offenses over the years has varied, thus **rejecting** the null hypothesis.

v) Weapons and felony offenses are **uniformly distributed** over the years : Applying the chi squared test, we get a p-value of 0.335 for weapons and 0.642 for felony. Since both the p-values are more than alpha (0.05) thus we **fail to reject** the null hypothesis, indicating no significant trend in weapon and felony offenses.

Model Selection :

We tried various models on our dataset like Logistic Regression, Support Vector Classifier(SVC), Random Forest, Linear Regression and Gradient Boosting and made the following observation based on its accuracy

- Random Forest consistently performs better across all target variables, showcasing its strength in both classification and regression tasks for PERP_RACE and PERP_SEX as targets.
- Logistic Regression and SVC are competitive for simple classification tasks like PERP_SEX but fall short for complex targets like PERP_RACE.
- For regression tasks, linear regression struggles as compared to gradient boosting due to its linear assumptions.

Thus, we select Random Forest classifier as our trained model as it effectively handles high-dimensional datasets and its ensemble nature ensures better accuracy and generalization by reducing overfitting.

Performance on Non-scaled dataset :

Our chosen Random Forest classifier model performance on our non-scaled dataset is shown in the below figure (Fig1).

Target Variable	Accuracy	F1- Score	Training Time (s)	Prediction Time (s)
PERP_RACE	0.58	0.57	437.91	29.94
PERP_SEX	0.82	0.81	461.98	25.22
AGE_GROUP	MSE: 459.69	R^2: 0.0155	6.62	0.015

Fig1 : Performance of RFC model on non-scaled dataset

Scaling methods :

We used two scaling methods, Principal Component Analysis (PCA) and Gaussian Random Projection (GRP). In terms of efficiency, PCA achieved a slightly better accuracy and F1-scores for classification tasks as compared to GRP while for regression tasks the mean squared error and R^2 scores were similar for both GRP and PCA. Also, GRP slightly increases training and prediction time for classification tasks due to its randomized projection step although the difference is negligible for regression tasks.

Performance on Scaled Dataset :

Our chosen Random Forest classifier model performance on our scaled dataset is shown in the below figure (Fig2).

Target Variable	Method	Accuracy	F1- Score (Weighted)	Training Time (s)	Prediction Time (s)
PERP_RACE	PCA	0.57	0.56	964.84	27.50
PERP_SEX	PCA	0.82	0.80	1035.32	24.13
AGE_GROUP	PCA	MSE: 460.92	R^2: 0.0128	2.52	0.0184
PERP_RACE	GRP	0.55	0.53	1037.077	28.49
PERP_SEX	GRP	0.82	0.80	1127.57	24.99
AGE_GROUP	GRP	MSE: 461.19	R^2: 0.0123	2.615	0.02

Fig2 : Performance of RFC model on scaled dataset

Conclusion and Future Work :

A few conclusion that we drew from our analysis were that:

- Random Forest was the best-performing final model, demonstrating robustness across all target variables.
- It effectively handles high-dimensional data and provides a balance between accuracy and computational efficiency.
- Dimensionality reduction techniques i.e PCA and GRP scaled the model for large datasets but introduced slight trade-offs in accuracy.

The future goals of our work are as follows :

- Incorporate more advanced feature engineering, such as geospatial and temporal variables, to improve predictions.
- Investigate alternative dimensionality reduction methods to further optimize model efficiency while maintaining accuracy.
- Explore Random Forest enhancements like hyperparameter tuning or transitioning to Gradient Boosting frameworks (e.g., XGBoost) for potential performance gains.
- Perform deeper analysis of demographic factors to explore actionable insights for addressing disparities.

Bibliography:

Dataset citation:

https://www.kaggle.com/datasets/okettaeneye/nypd-arrests-data-historic-2006-2020

https://www.kaggle.com/datasets/okettaeneye/nypd-arrests-data-historic-2006-2020

CSE558 Data Science Lectures