Team – 09

Vidhi Shiyani   Komal Dodiya   Aniket Patel

Problem Statements

1. Regression Problem Statement: Predicting future total adult arrests in every county based on previous arrest patterns and category-level arrests statistics.

2. Classification Problem Statement: Identifying the hotspots in the five boroughs where the number of adult arrests (individuals 18 years or older) is significantly higher, so the government can implement initiatives to reduce the conversion rate of juvenile offenders.

Findings based on EDA:

• The data set consists of adult arrest rates by county and year for several categories of arrests.

• During data preparation, certain columns of arrest counts were initially read as text (due to the presence of commas) but could be successfully converted to numeric data types.

• Visualizations of annual arrests by type revealed considerable variation over offense types and time, shifts in crime trends and enforcement priorities.

Regression Problem Statement

Features considered:

• Features: Year, previous total arrests, and one-hot encoded county names.

• Target Variable: Total adult arrests.

Model evaluation and comparison:

1. Linear Regression

-> Mean Absolute Error (MAE): 678.42

-> Root Mean Squared Error (RMSE): 2166.42

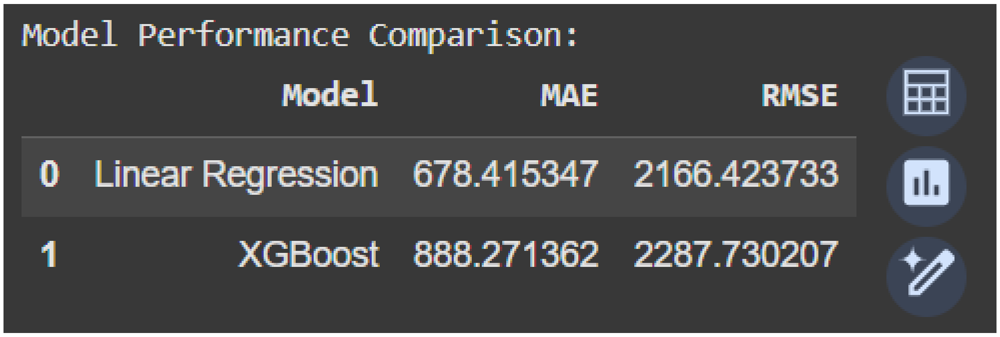
Therefore, we can say that on average total arrests in the test data 5109, the MAE of 13% is a great indicator of predictive accuracy.

2. XGBoost Regressor

-> Mean Absolute Error (MAE): 888.27

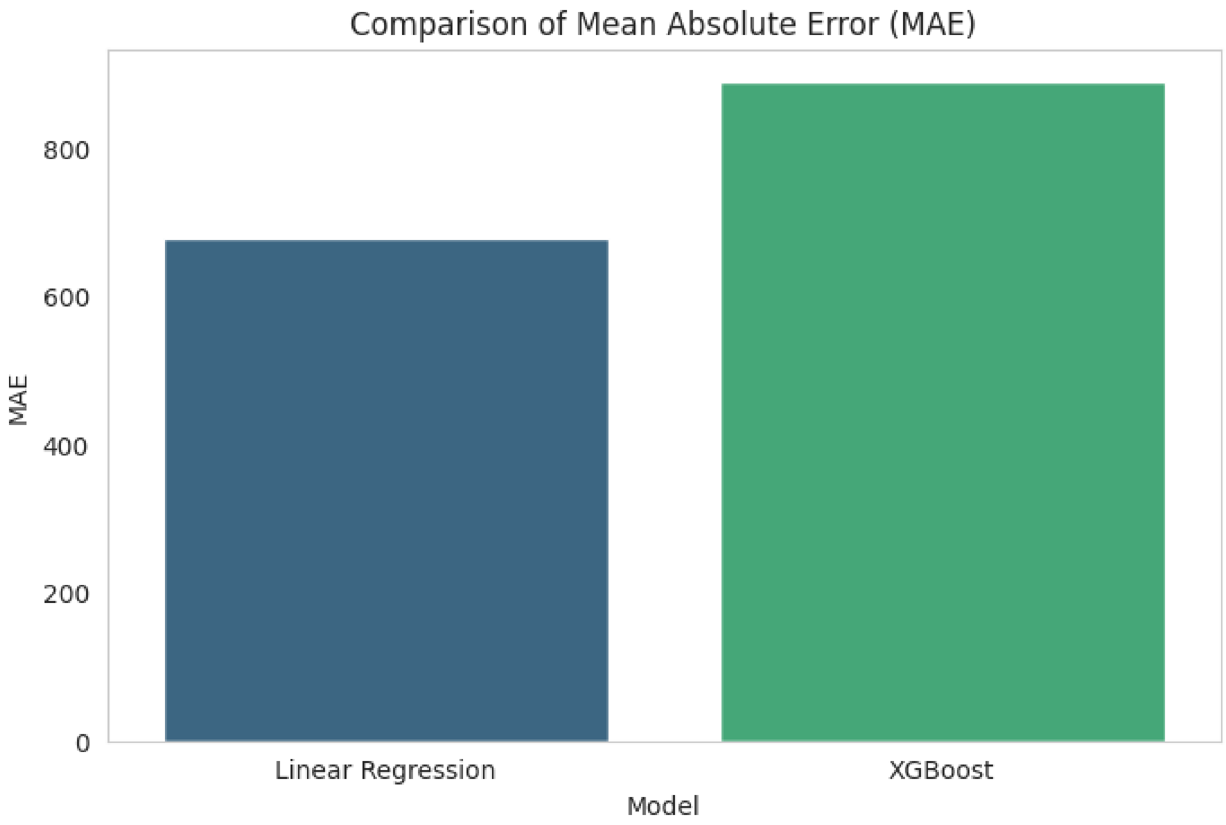
-> Root Mean Squared Error (RMSE): 2287.73

Hence, we can say that the linear regression model worked a better than the XGBoost model on both MAE and RMSE metrics.



```
Model Performance Comparison:
```

	Model	MAE	RMSE
0	Linear Regression	678.415347	2166.423733
1	XGBoost	888.271362	2287.730207



Classification Problem Statement

Features considered:

• Features: Year and one-hot encoded county names.

• Target Variable: Hotspot label (1 = Hotspot, 0 = Not Hotspot).

• Arrest category columns were removed to avoid data leakage.

Model evaluation and comparison:

1.Random Forest Classifier

-> Accuracy: 97%

-> F1-Score: 0.94

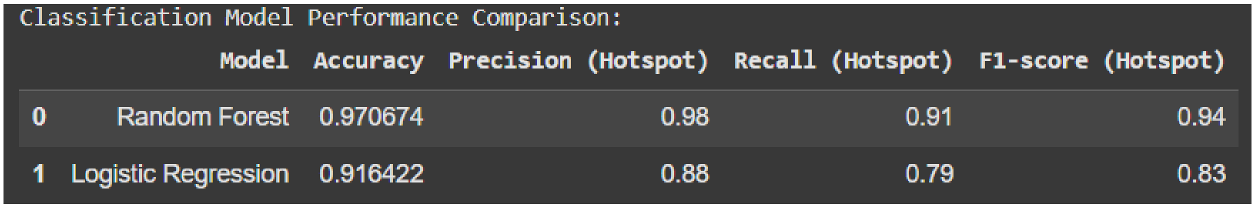
Here, random forest classifier resulted in high precision, recall, and F1-scores for both classes, indicating high performance.

2.Logistic Regression Classifier

-> Accuracy: 92%

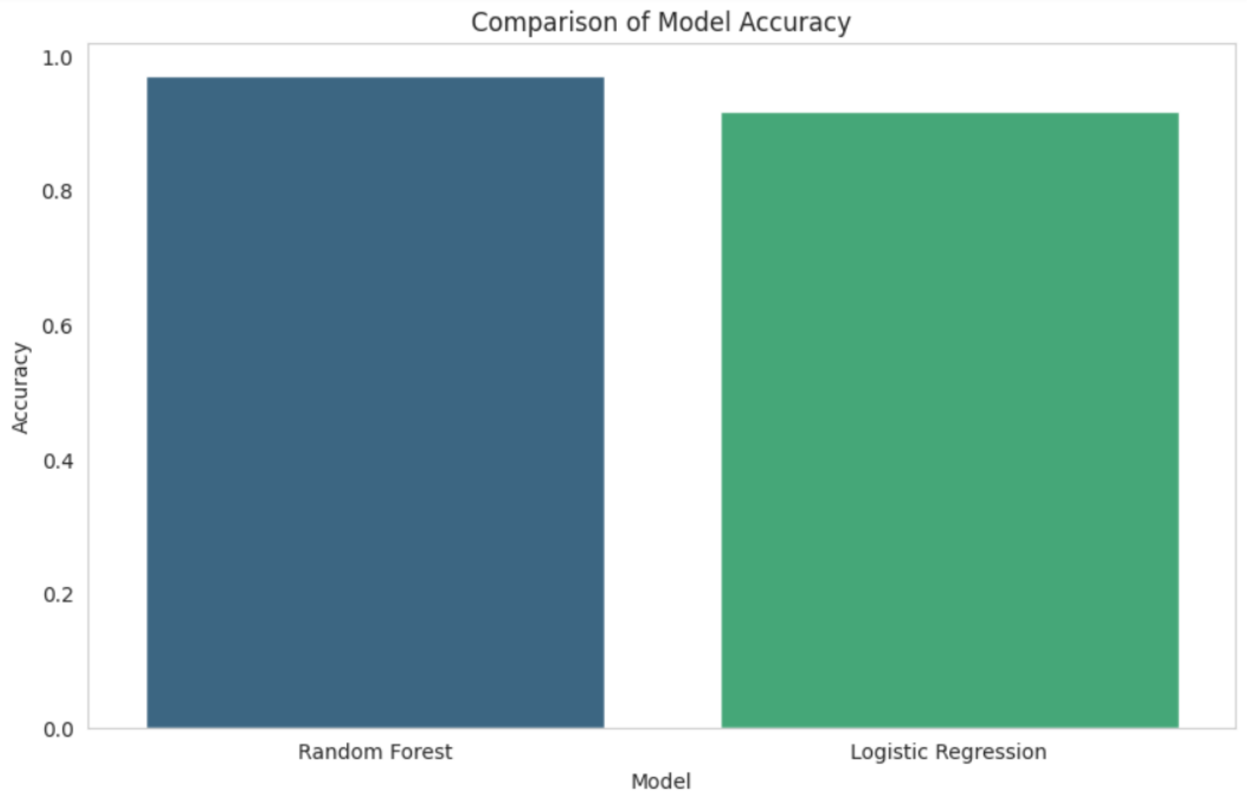
-> F1-Score: 0.83

We got a good overall performance but were not as good as the Random Forest model. Hence, the random forest classifier was more accurate and was able to identify hotspots more accurately.



Classification Model Performance Comparison:

	Model	Accuracy	Precision (Hotspot)	Recall (Hotspot)	F1-score (Hotspot)
0	Random Forest	0.970674	0.98	0.91	0.94
1	Logistic Regression	0.916422	0.88	0.79	0.83

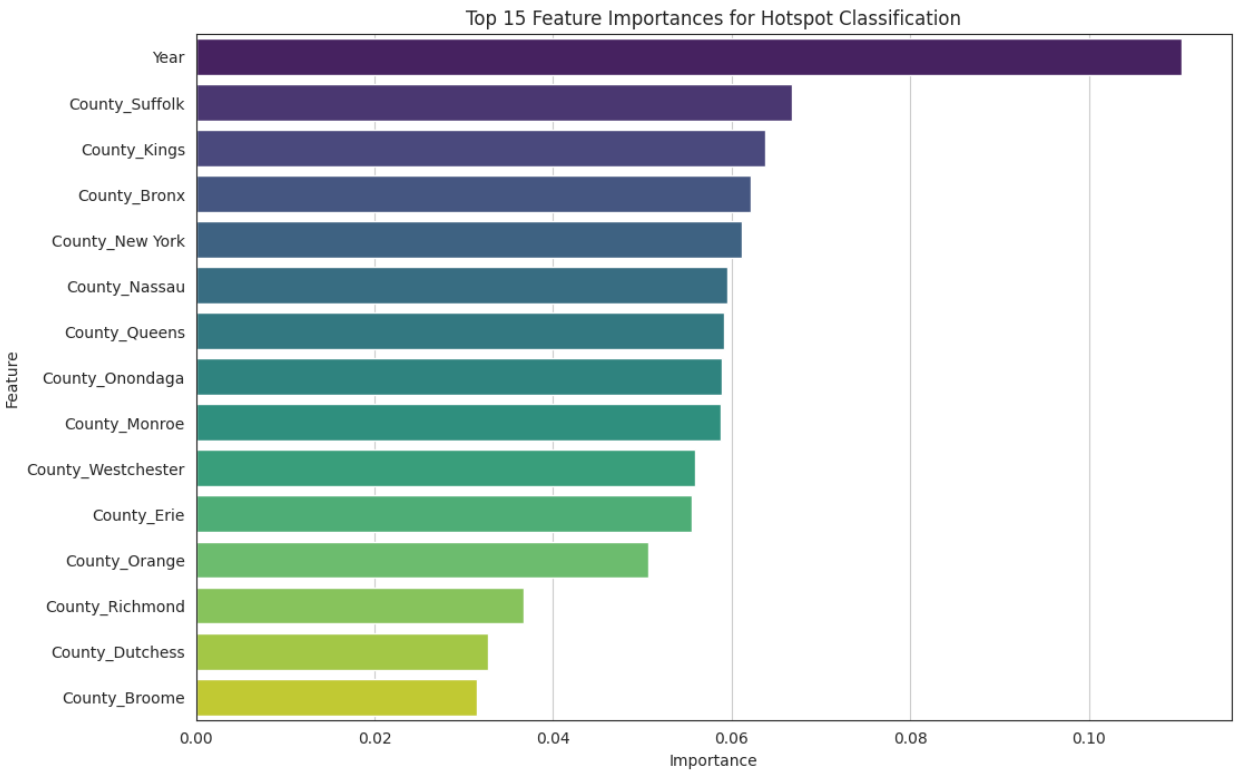


Feature Importance:

The feature importance analysis identified the following as the most significant predictors of hotspot classification:

• Year

• Counties: Suffolk, Kings, Bronx, New York, Nassau, Queens, Onondaga, Monroe, Westchester, Erie, Orange, Richmond, Dutchess, and Broome.



List of hotspots:

Ranking counties by the proportion of years as hotspots designated Bronx, Erie, Kings, Nassau, and Queens as persistent hotspots over the period under observation.

Overall Conclusion:

- The Linear Regression model provided a better prediction of total arrests by time.
- The Random Forest Classifier rightly identified some hotspot counties that had high predictability.
- Feature importance observations showed the significance of some counties and time variables (year) in defining hotspot status.