TITLE: PREDICTION OF ARRESTS BASED ON TERRY TRAFFIC STOPS

Background information

A Terry Stop, also known as an "Investigative Detention" or "Stop and Frisk," allows law enforcement officers to temporarily detain a person for investigation when they reasonably suspect involvement in criminal activity. This stop is used when there is insufficient evidence for an arrest (i.e., no probable cause) but enough suspicion to justify a brief investigation. The primary purpose is to confirm or dispel the officer's suspicions. If probable cause for an arrest arises during the stop, the suspect is arrested. If no probable cause is found, the suspect is released.

Problem statement

Terry stops can disrupt the normal lives of individuals, especially when no arrests are made. During these stops, police officers may spend valuable time that could be used to address crime in other areas. Additionally, Terry stops initiated through 911 calls can lead to a waste of resources, particularly when they result in no arrests. These stops have also been plagued by concerns of racial profiling, with many individuals believing that the stops are disproportionately based on their race or gender. This has eroded the public's perception of law enforcement and diminished their confidence in the police.A model was proposed to be developed to predict whether a Terry stop will result in an arrest. Given that Terry stops can be initiated through various channels such as 911 calls, messages, alarms, and police interactions, this model can help schedule calls appropriately and allocate resources more efficiently.

Objective

- **To build and determine the best classifier model for predicting Terry stop arrests with an accuracy above 75% and an F1 score above 80%.**

Data Understanding

The data for this analysis was obtained from the Seattle police department. It represents Each row represents a unique stop. Each record contains the perceived demographics of the subject, as reported by the officer making the stop, and officer demographics as reported to the Seattle Police Department. The original dataset contained 62020 entries. the officer squad had missing values which were dropped resulting in 61459 entries with 23

columns. Due to the sensitivity and data ethics consideration, gender and race columns were not used in the analysis. the data can be obtained from:http://tiny.cc/l4hzzz and the columns description:http://tiny.cc/c5hzzz.

Data Preparation

The following were applied to prepare the data for analysis:

1. **Handling of missing values-** only the officer squad had missing values at 9% which were dropped.
2. **The reported date and time were converted to DateTime format**. The date was then extracted into components: day of the week, day, month, and year. The original "reported date" and "reported day" columns were dropped to avoid multicollinearity. Only the hour component was retained, and the "time reported" column was dropped.
3. **Column Preparation:** The "Weapon Type" column was grouped into three categories: "Weapon Found," "Weapon Unknown," and "No Weapon." This was done to simplify the analysis. The same approach was applied to other columns as appropriate.

N/B Not all columns were converted to integers for visualization purposes. The column values were later converted to integers for modeling.

4. **Irrelevant columns dropped**. Columns containing sensitive information, such as gender, race, and personal identifiers like Officer ID, were dropped. Gender and race were excluded from the analysis to ensure the model remains unbiased and does not inadvertently reinforce stereotypes or discriminatory practices. The goal was to build a fair and objective model that focuses solely on relevant factors, avoiding any potential influence of sensitive attributes.

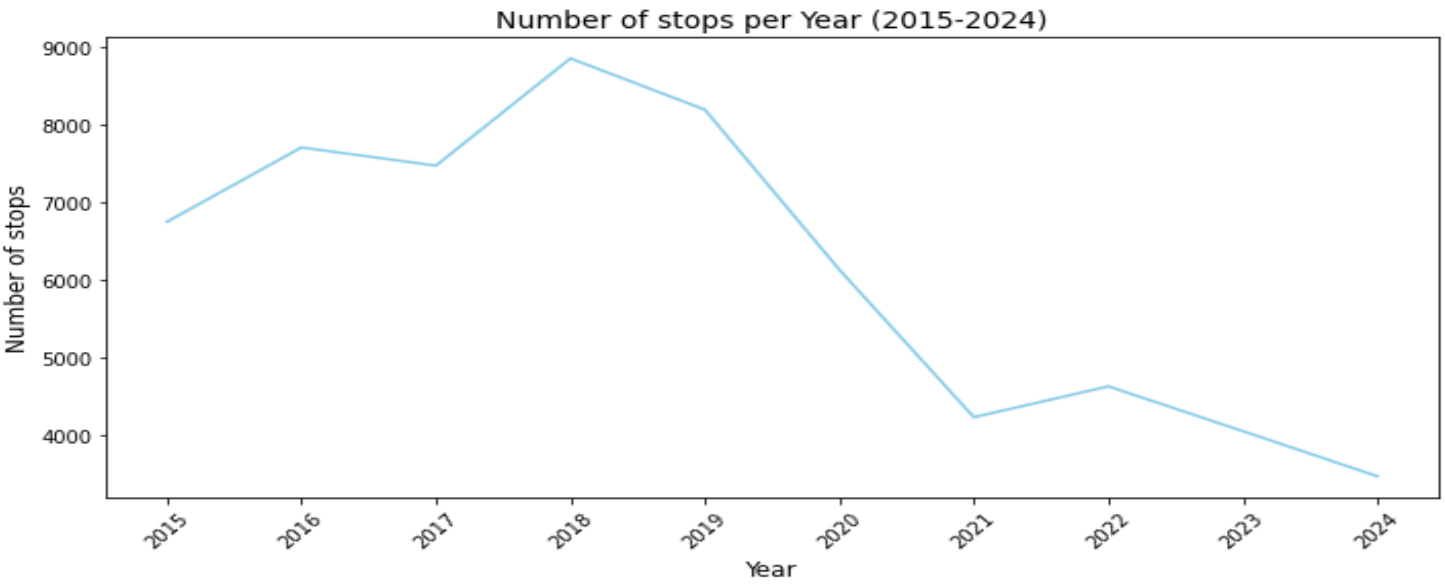
The resulting data frame contained 61459 rows and 16 columns. The data frame still contained object datatype and while some columns were transformed.

Data Analysis

The data analysis process focused on finding the relationship between the independent variables(all columns except the arrest flag column) and the target arrest flag column. The first step was to determine the trend in Terry stops over the years.

1. What is the distribution of Terry stops over the years

The purpose of this column was to inform us of the distribution of Terry arrests over the years.



Observation

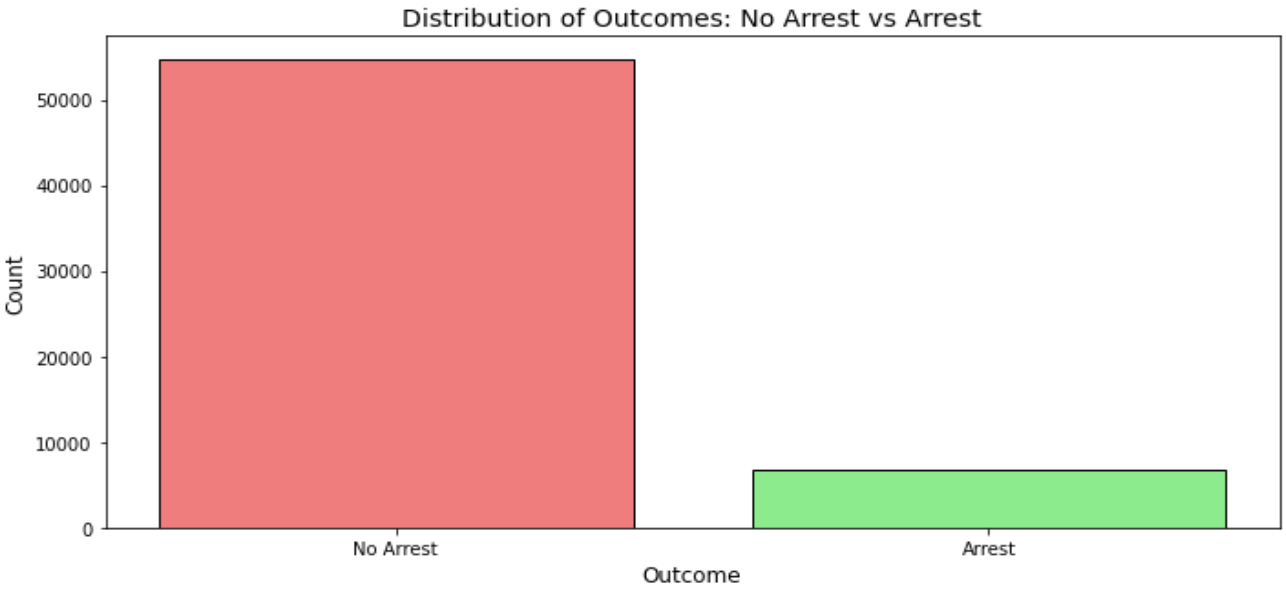
The data shows that the number of Terry stops steadily increased from 2015, reaching a peak in 2018. Thereafter, the number of stops declined, with the lowest recorded in 2024 at only 3,468 stops. This decrease can likely be attributed to the rapid adoption of stationary ALPR cameras starting in 2019. By April 2022, 1,500 cities across the United States had implemented Flock cameras. These cameras are used to scan license plates, and if a driver's license is found to be expired, a Terry stop is initiated. This technological shift may have contributed to the reduction in Terry stops.

2. How many Terry stops led to an arrest?

This analysis seeks to determine the number of arrests based on the Terry stops. Is it a significant number to the number of stops?

Observation

Between 2015 and 2024, a total of 6,760 arrests were recorded, accounting for 10% of all observations. For approximately 10 years, 61,459 Terry stops were conducted, but only 6,160 resulted in arrests.



3.What arrests were made based on the call type?

	Call Type	Arrest Count
0	911	4506
1	ONVIEW	1421
2	TELEPHONE OTHER, NOT 911	502
3	Call Type UNKNOWN	234
4	ALARM CALL (NOT POLICE ALARM)	88
5	OTHER CALLS	9

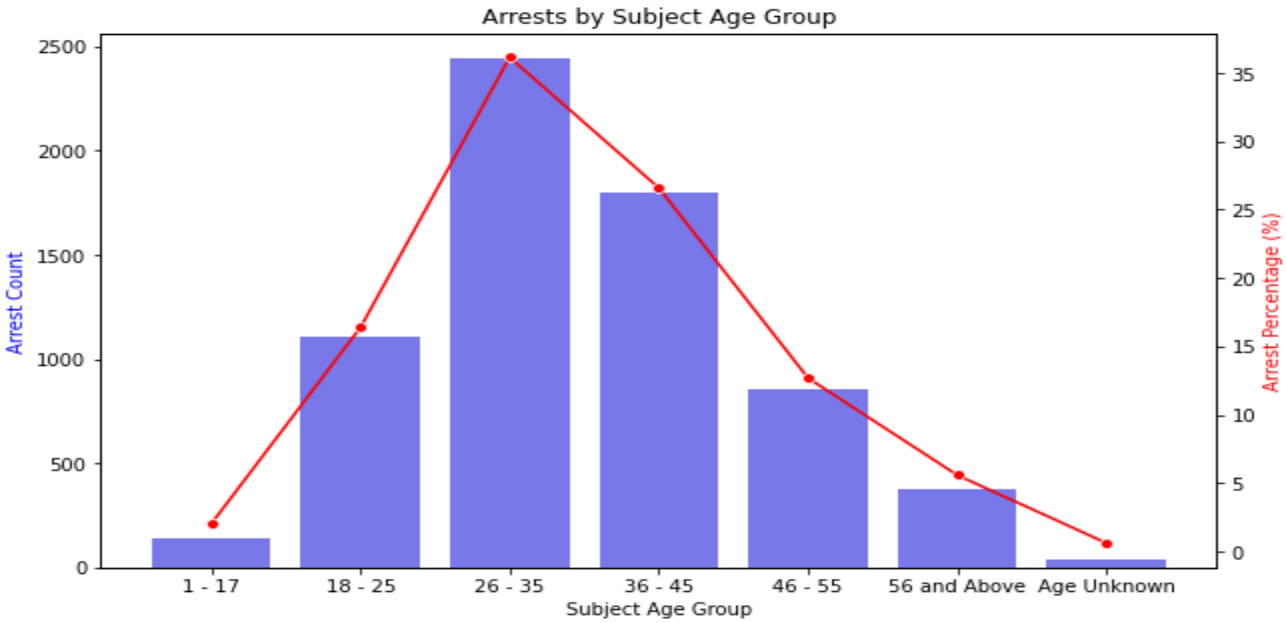
Observation

911 calls topped the list with 4,506 arrests followed by onview at 1,421, telephone other(not911) at 502, unknown at 234, and alarm call and other types of calls cumulated to 97.

4 . What is the distribution of arrests based on the subject's age?

Observation

Ages in the range of 26-35 contributed 36% of the arrests made followed by ages between 36 and 45 which contributed about 27%. Age group 1-17 accounted for the least arrests followed by ages 46-55 at 5.5%. The adult age in the USA is 18 years and the individuals in the subject groups were probably given a warning as they are still under parental protection and care. Individuals aged 56 and above may be perceived not to be a threat.

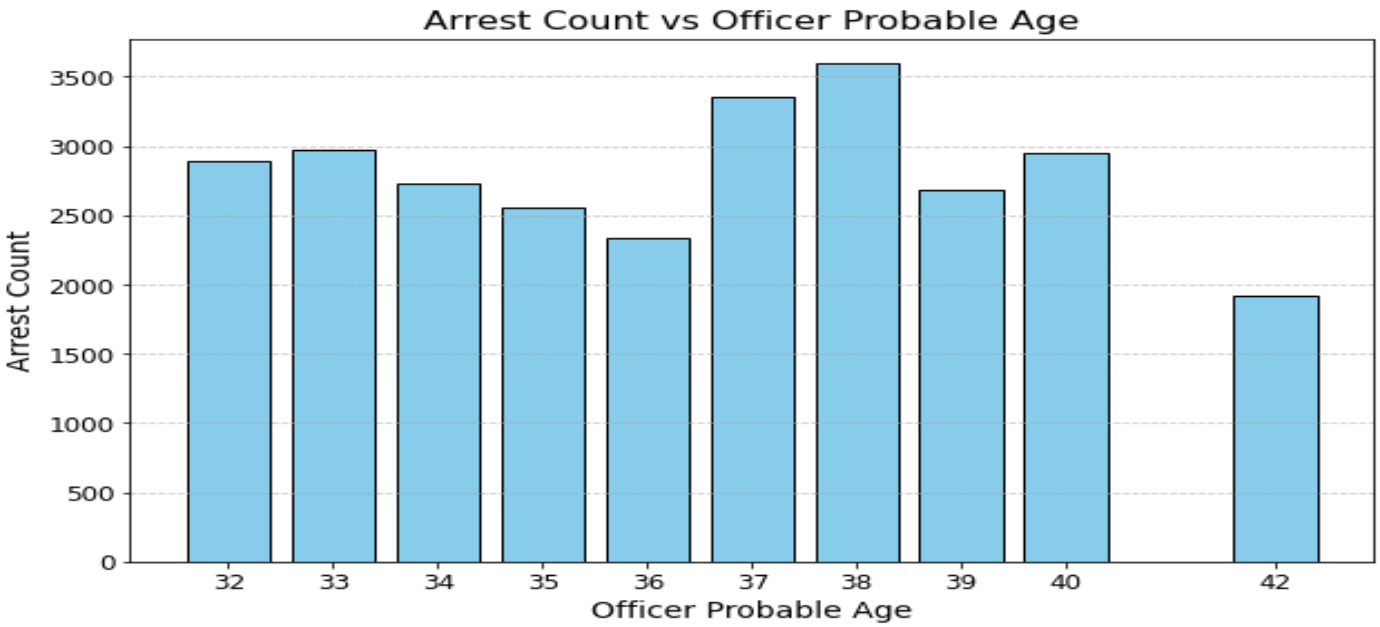


5. What is the relationship between officer experience and the arrest flag¶

For this analysis we assume that the retirement age for an officer in Seattle is 53 years, therefore limiting our analysis to only officers below the age of 53. Most arrests were made by officers between the ages of 29 and 38. Assuming an officer joins the force at a minimum age of 18 we can conclude that officers with more than 10 years of experience made correct judgments based on Terry traffic stops.

Officer YC	arrest_cou	Officer probable ag
1995	565	29
1989	479	35
1993	478	31
1992	455	32
1991	455	33
1987	451	37
1986	426	38
1990	416	34
1994	319	30
1988	277	36

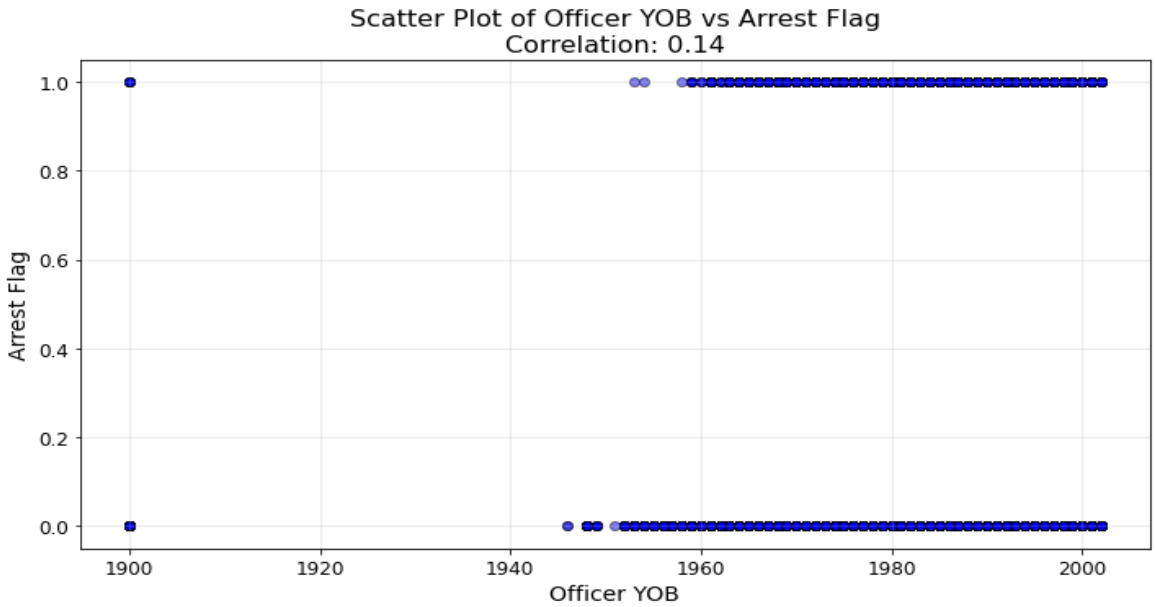
6. Officers who initiated a Terry stop but never made an arrest.



Observation:

Officers between the ages of 32 and 42 initiated the most terry stops but did not make an arrest. This leads to the question is there a relationship between officer age and terry stops initiated?

correlation between officer age and Terry stops initiated¶



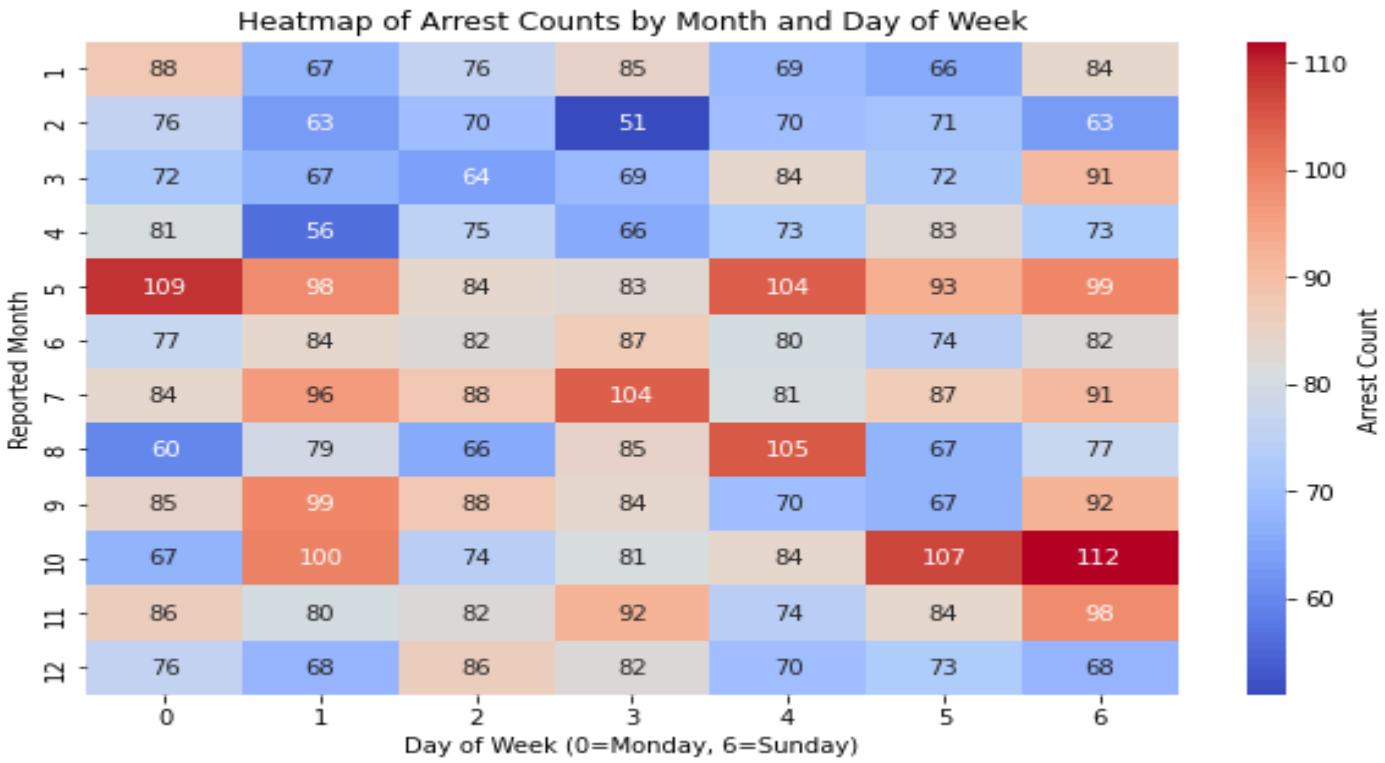
A correlation of 0.14 indicates a very weak relationship. This suggests there is no significant linear relationship between an officer's age and the number of Terry stops. Therefore, we can conclude that Terry stops were random and not influenced by the officer's age.

7 What is the relationship between date, day of the week, and arrest flag¶

	reported_Month	day_of_week	Arrests
0	5	3	670
1	7	6	631
2	10	4	625
3	11	3	596
4	9	5	585
5	6	6	566
6	8	1	539
7	1	6	535
8	12	0	523
9	3	4	519
10	4	0	507

Observation

The terry stops that led to arrests occurred on Fridays through Saturdays. May recorded the highest number of arrests with 670 arrests. August, September, October November also registered high numbers of arrests. The lowest number of arrests were made in December, January, February, March, and April. The month of December was expected to have the most arrests because of the festivities but the data showed otherwise.

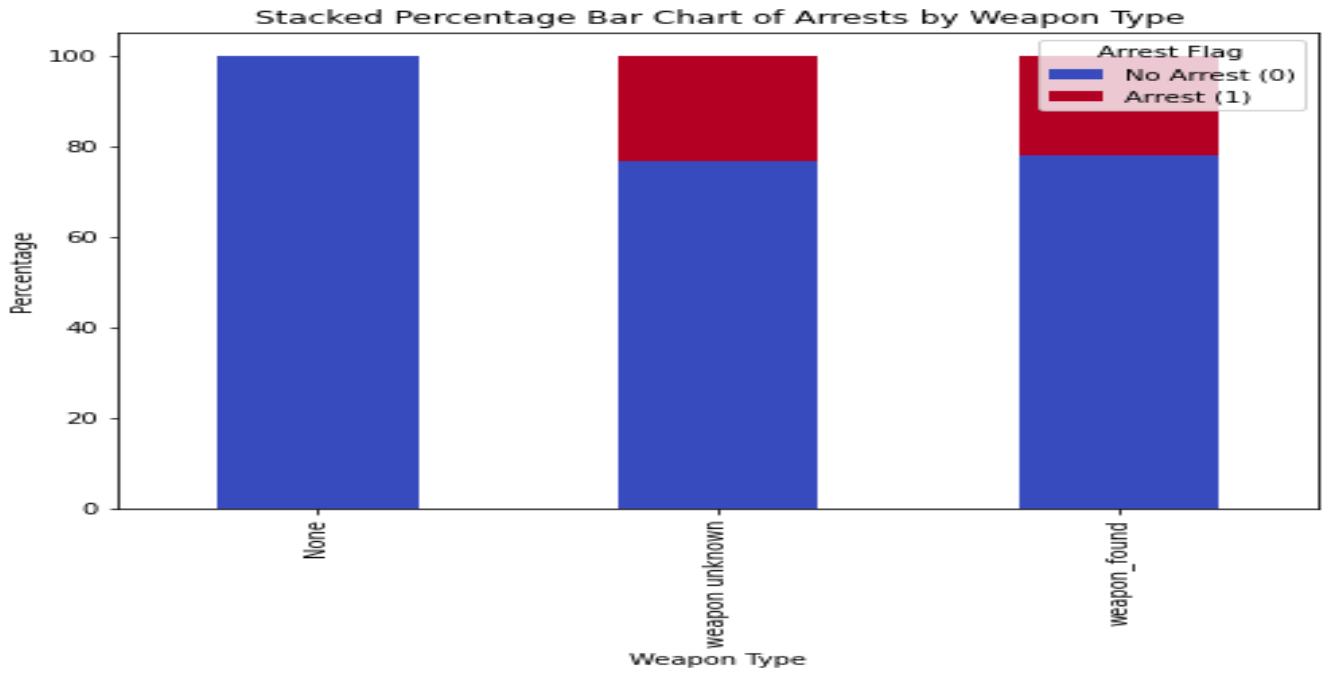


The heatmap indicates that Terry stops on weekends in October are more likely to result in arrests. Having identified the months and days of the week when arrests are most probable following a Terry stop, let's now explore whether being found with a weapon at specific times of the day further exacerbates the situation.

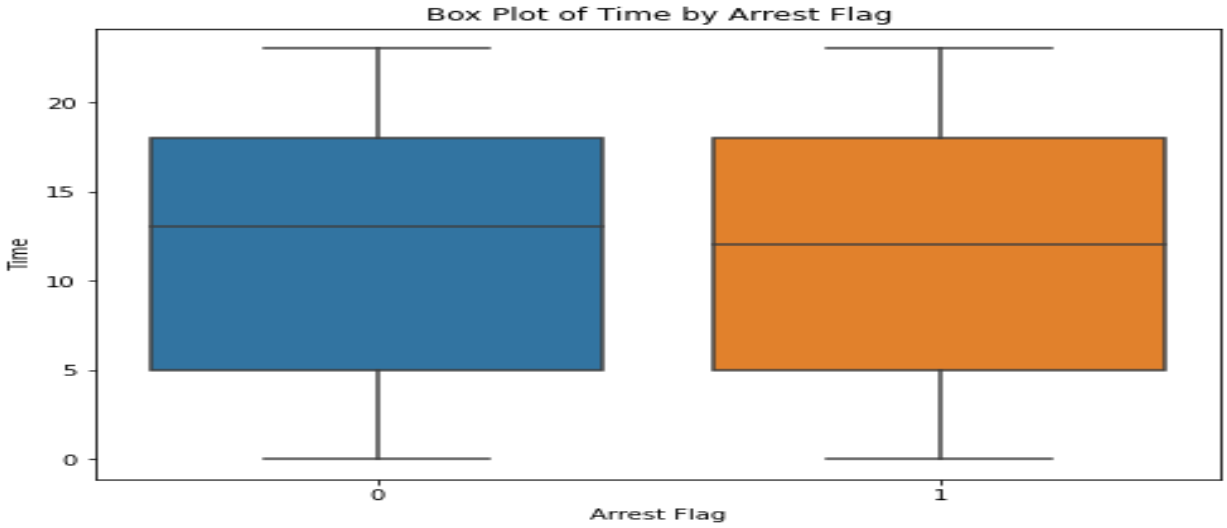
8 Does being found with a weapon or not likely to lead to an arrest?

Observation

When a suspect is not found with a weapon, they are released 99.97% of the time. Cases where the weapon type was unknown accounted for approximately 23% of arrests, while weapons being found contributed to 22% of arrests. Notably, only 0.022% of individuals not found with a weapon were arrested.



9 Distribution of arrests based on the time.



There seems to be no variance in the median therefore time does not influence the arrest flag.

10. Relationship between the location and arrests

	Precinct	COUNT("Precinct")	Beat
0	East	1021	G3
1	FK ERROR	4	99
2	North	1137	N2
3	OOJ	8	OOJ
4	South	967	R2
5	Southwest	658	F1
6	West	2606	M3
7	precinct Unknown	359	Beat UNKNOWN

The West M3 beat accounted for the highest number of arrests, with 2,606 arrests recorded, followed by the North precinct, which had 1,137 arrests. The location of arrests is a significant feature and is assumed to be highly relevant for modeling. However, including all three location-related attributes (beat, sector, and precinct) may lead to multicollinearity, as the model could infer a beat from the sector and precinct information. This redundancy might impact the model's performance and interpretation.

CONCLUSION

Over the years, the number of Terry stops has gradually decreased, with the lowest recorded in 2024. Factors such as the time of day and the officer's age did not appear to significantly influence the likelihood of arrests. However, key factors such as the subject's age, month, day of the week, call type, location (sector and precinct), and whether the subject had a weapon played a significant role in determining whether an arrest would be made. Among all factors, 911 calls accounted for the majority of arrests.

Recommendation

1. Build a model to predict Terry stops based on the observed features.

2. Focus should be put on 911 calls.

3 . Certain areas should be prioritized when handling Terry stops such as those in the West precinct as they saw most of the arrests.

Modelling

AIM: To create a model that accurately(85% and above accuracy) predicts the result of a Terry stop.

: Choose the best model among the evaluated models.

Modeling refers to the process of creating a simplified representation or abstraction of a real-world process, system, or phenomenon using mathematical, statistical, or computational methods. The goal of modeling is to understand, predict, or simulate the behavior of the system based on its key components and relationships.[1]

Classification model-Classification models are a type of machine learning model that divides data points into predefined groups called classes. Classifiers are a type of predictive modeling that learns class characteristics from input data and learns to assign possible classes to new data according to those learned characteristics[2].

What are the metrics for measuring a classification model's performance?

1. Accuracy: The proportion of correctly predicted instances out of all instances**. ((TP+TN)/(TP+TN+FP+FN)**

Disadvantage: Not always reliable with imbalanced datasets.

2. **Precision:** The proportion of true positive predictions out of all positive predictions made by the model.

   Precision=TP /(FP+TP) Answers the question of how many positives were predicted.

3. **Recall:**The proportion of true positive predictions out of all actual positives.

   Recall=TP/(TP + FN)

4. **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two.

5. Area Under the ROC Curve (AUC-ROC) Measures the model's ability to discriminate between classes across all thresholds. A higher AUC value indicates better model performance.

6. Confusion Matrix

- **Definition**: A table used to evaluate classification models, showing the counts of true positives, false positives, true negatives, and false negatives.
- **Usage**: Provides a clear view of model performance across different classes.

---

[1] https://www.databricks.com/glossary/machine-learning-models#:~:text=A%20machine%20learning%20model%20is,sentences%20or%20combinations%20of%20words.

[2] https://www.ibm.com/topics/classification-models

Checking for multicolinearity

Variance Inflation Factor was used to check for multicolinearity and columns with coefficients greater than 5 dropped this resulted in the columns as indicated below.

	Feature	VIF
0	const	0.000000
1	Subject Age Group	1.006357
2	GO / SC Num	1.103506
3	Stop Resolution	1.518859
4	Weapon Type	1.356148
5	Officer YOB	1.207795
6	Initial Call Type	8.417151
7	Final Call Type	8.015802
8	Call Type	1.277958
9	Officer Squad	1.051114
10	Arrest Flag	1.248224
11	Frisk Flag	1.178598
12	Precinct	1.410396
13	Sector	16.663518
14	Beat	17.388017
15	reported_Year	0.204092
16	reported_Month	1.001086
17	day_of_week	1.001510
18	Time	1.006695

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61459 entries, 0 to 61458
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Subject Age Group  61459 non-null  int64
 1   GO / SC Num        61459 non-null  int64
 2   Weapon Type        61459 non-null  int64
 3   Officer YOB        61459 non-null  int64
 4   Final Call Type    61459 non-null  int64
 5   Call Type          61459 non-null  int64
 6   Officer Squad      61459 non-null  int64
 7   Arrest Flag        61459 non-null  int64
 8   Precinct           61459 non-null  int64
 9   Sector             61459 non-null  int64
 10  reported_Year      61459 non-null  int64
 11  reported_Month     61459 non-null  int64
 12  day_of_week        61459 non-null  int64
 13  Time               61459 non-null  int64
```

The figure on the rightmost shows the resulting data frame after dropping the columns that are related.

class imbalance

The value counts was applied to the target and this resulted in an ambulance of about 90% for NO(majority class:0) arrest and 10% for (the minority class:1)

To address the class imbalance problem, SMOTE was applied .SMOTE oversamples the minority creating a balanced data frame. The result of applying smote was a balanced dataset of 41,062 for both classes. Caution was taken to ensure no data leakage by applying smote after the split and only to the training dataset.

```
#splitting the data into train and test data
X_train, X_test, y_train, y_test=train_test_split(X,y,test_size=0.25,random_state=42)
#using smote to handle the class imbalance.
from imblearn.over_sampling import SMOTE
# Initialize SMOTE
smote = SMOTE(random_state=42)
# Resample the dataset
X_resampled, y_resampled = smote.fit_resample(X_train, y_train)
from collections import Counter
print("Before SMOTE:", Counter(y_train))
print("After SMOTE:", Counter(y_resampled))
```

```
Before SMOTE: Counter({0: 41062, 1: 5032})
After SMOTE: Counter({0: 41062, 1: 41062})
```

Scaling

A logistic regression model is a distance vector model and scaling is applied to ensure that the prediction is not dominated by features of higher values.

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_resampled = scaler.fit_transform(X_resampled)

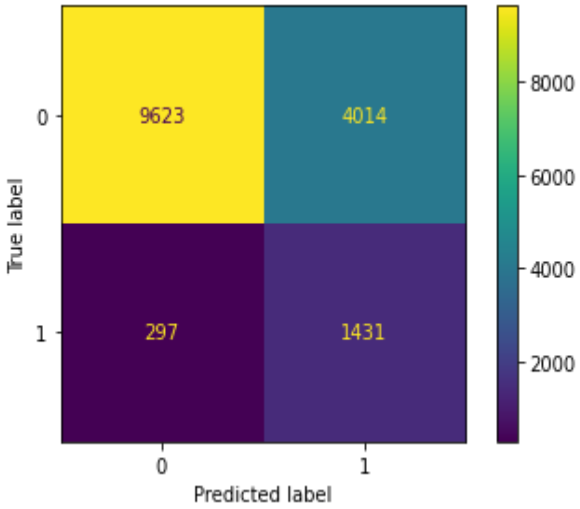
X_test = scaler.transform(X_test)

The scaler was fit on the training data but used to transform the test data. This also ensured no data leakage.

Building a Logistic regression model

Logistic regression is a classification model used to predict binary target. Logistic regression assumes linearity, and independence, and is widely affected by multicollinearity. Logistic regression works best for a linear dataset. A baseline model is a simple model without any tuning.

Results



	precision	recall	f1-score	support
0	0.97	0.71	0.82	13637
1	0.26	0.83	0.40	1728
accuracy			0.72	15365
macro avg	0.62	0.77	0.61	15365
weighted avg	0.89	0.72	0.77	15365

Evaluation:

The model achieves an accuracy of 72%. The model performs poorly in predicting the minority class and a macro average f1 score of 61%. A decision tree classifier model is recommended to improve the prediction.

Building a decision tree model(no hyperparameter tuning)

A Decision Tree Classifier is a type of machine learning algorithm used for classification and regression tasks. It models data by creating a tree-like structure where each internal node represents a decision (based on a feature), each branch represents an outcome of that decision, and each leaf node represents a class label or a predicted outcome.

Use cases

Decision trees can be used for other classifications such as Customer Churn Prediction, Credit Scoring (Loan Approval), Medical Diagnosis, Fraud detection,n and Sales and Marketing Campaign Effectiveness.

Advantages

1. Works well with frequency-encoded data such as the above dataset.

2. Easy visualization and interpretability as it mimics the human decision-making process.

Results

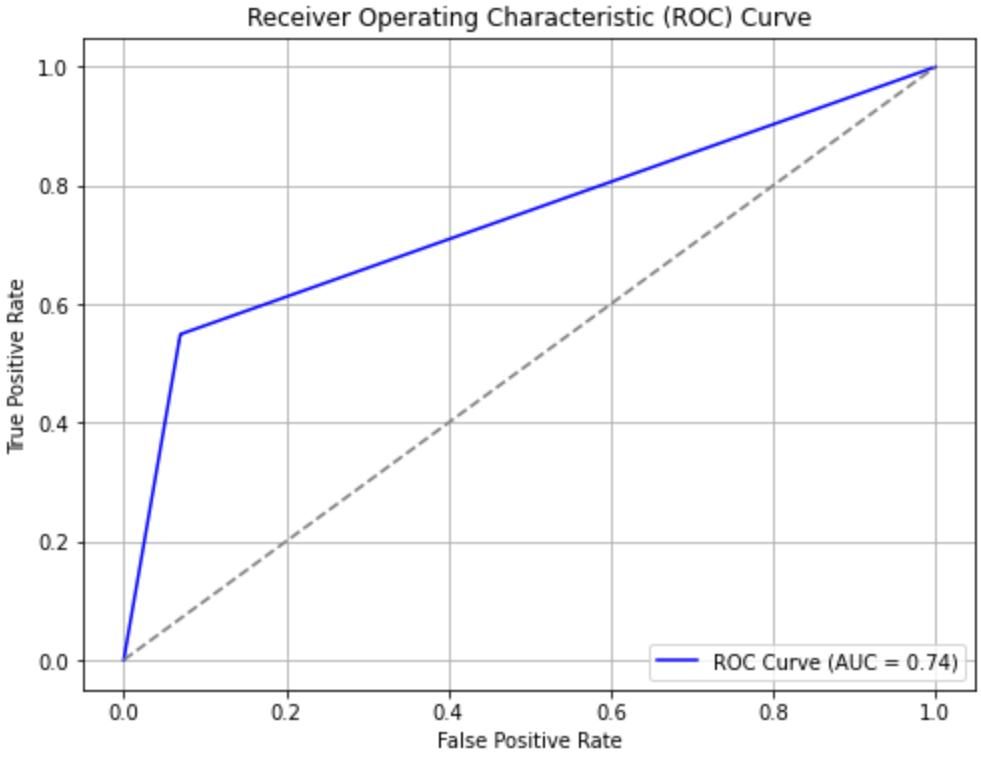
	precision	recall	f1-score	support
0	0.94	0.93	0.94	13637
1	0.50	0.54	0.52	1728
accuracy			0.89	15365
macro avg	0.72	0.73	0.73	15365
weighted avg	0.89	0.89	0.89	15365

Evaluation

The overall accuracy of the model has improved with the decision tree classifier to 89%. the ability of the model to predict the minority class has greatly improved with an f1 score of 73% confidence.Performance of the model on the training set(checking for overfit and underfit).This model performs better than the logistic regression model.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	41156
1	1.00	1.00	1.00	40968
accuracy			1.00	82124
macro avg	1.00	1.00	1.00	82124
weighted avg	1.00	1.00	1.00	82124

The model's performance on the training set with precision=recall=accuracy=f1score=100% is an indicator of overfitting. The model has memorized the data points rather than learning from them. The remedy is hypertuning the parameters to reduce overfitting.



Evaluating the models performance using ROC-AUC curve

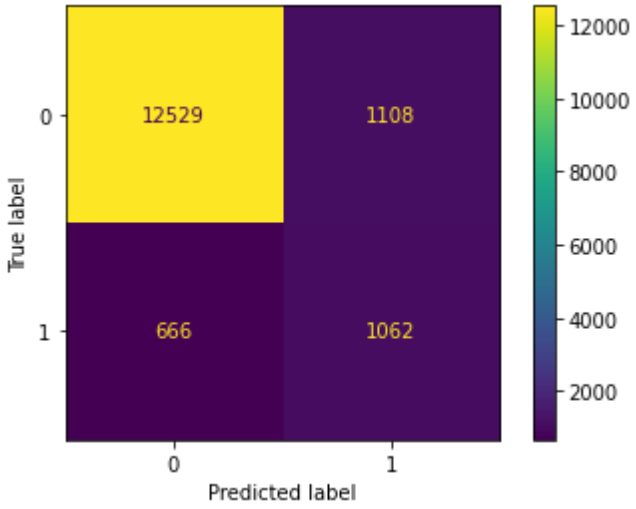
The Area Under the Curve (AUC) value is 0.74, indicating that the model has moderate performance

Hyperparameter-tuned Decision tree model

GridSearchCV was used to find the best parameters.it is a scikit-learn function that performs hyperparameter tuning by training and evaluating a machine-learning model using different combinations of hyperparameters. The best set of hyperparameters is then selected based on a specified performance metric.[3]The best parameters (criterion='entropy',max_depth =20,min_samples_leaf=1,min_samples_split=2)

[3] https://dev.to/anurag629/gridsearchcv-in-scikit-learn-a-comprehensive-guide-2a72#:~:text=GridSearchCV%20is%20a%20scikit%2Dlearn%20function%20that%20performs%20hyperparameter%20tuning,on%20a%20specified%20performance%20metric.

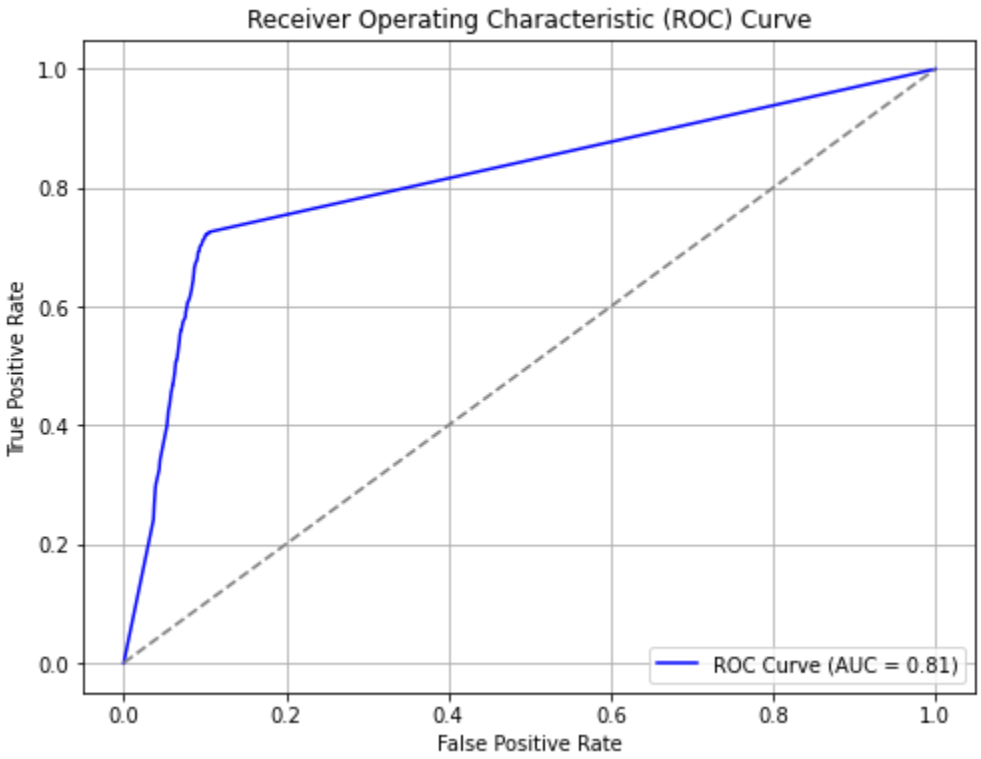
Results



	precision	recall	f1-score	support
0	0.95	0.92	0.93	13637
1	0.49	0.61	0.54	1728
accuracy			0.88	15365
macro avg	0.72	0.77	0.74	15365
weighted avg	0.90	0.88	0.89	15365

The tuned model's accuracy has decreased but the f1 score slightly improved. Since the f1 score is the most crucial metric for measuring the performance of the model then the tuned model performs better than the baseline decision tree model.

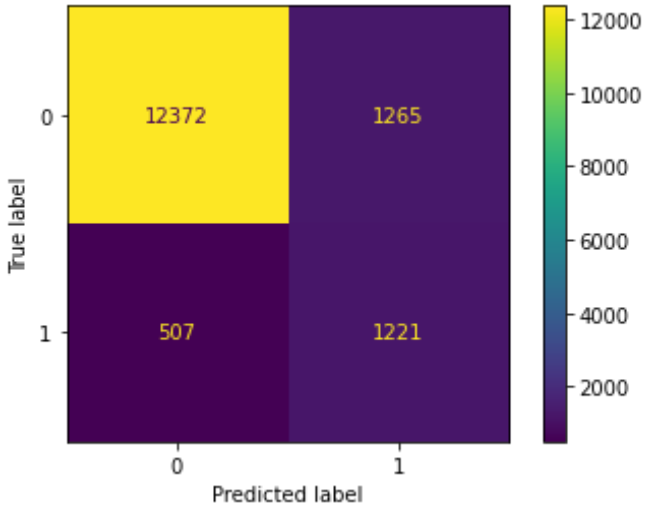
Evaluating the ROC-UAC curve



The model demonstrates moderate performance and can be improved further at an AUC score of 0.81 the model performs better than the one not tuned. A random forest classifier can improve the prediction.

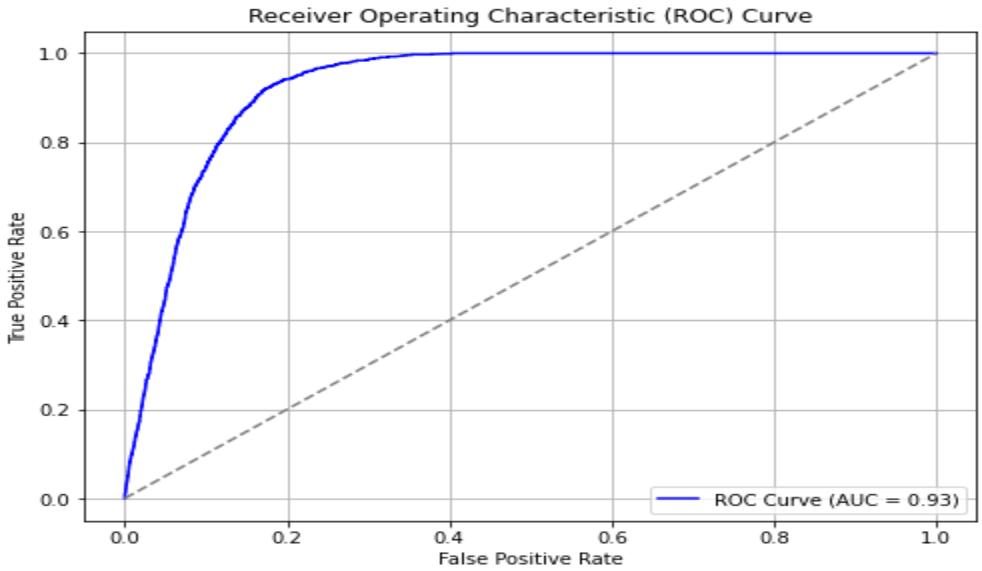
Building a random forest classifier

Results



	precision	recall	f1-score	support
0	0.96	0.91	0.93	13637
1	0.49	0.71	0.58	1728
accuracy			0.88	15365
macro avg	0.73	0.81	0.76	15365
weighted avg	0.91	0.88	0.89	15365

The model misclassified 507 instances as false negatives, meaning it predicted that a Terry stop would not lead to an arrest when it actually did. Additionally, it misclassified 1,265 instances as false positives, predicting an arrest when there was none. On the positive side, the model correctly identified 1,221 instances as true positives (arrests correctly predicted) and 12,372 instances as true negatives (correctly predicted no arrest).The F1 score of the model improved, while the accuracy remained constant. Since the F1 score is our chosen evaluation metric, the random forest classifier outperforms the tuned decision tree classifier, demonstrating better performance in balancing precision and recall.The curve below is close to the top-left corner, showing a high TPR and a low FPR, which signifies a good balance.AUC of 0.93 indicates a strong predictive power of the model.



CONCLUSION

The random forest model performs the best in terms of f1 score. It has an AUC score of 93% indicating a strong predictive power of the model. The objective of the analysis was achieved ie the random forest predicted the arrest flag at an accuracy of 89% and f1 score above 75%.

Recommendation

Despite achieving the desired accuracy, the random forest classifier, which was identified as the best model, performed poorly in terms of the F1 score. For this task of predicting whether a Terry stop would lead to an arrest, the F1 score is the most suitable evaluation metric, as it balances precision and recall, which are crucial for handling the imbalance between positive and negative classes.

To improve performance, an **XGBoost (Extreme Gradient Boosting)** model can be developed and its performance evaluated. XGBoost is known for its robustness and ability to handle imbalanced datasets effectively, which could potentially lead to better F1 scores compared to the random forest classifier. This approach might help in addressing the misclassifications and achieving better overall model performance.

References

1 Wikipedia contributors, *Automatic number-plate recognition*, Wikipedia, accessed December 6, 2024, https://en.wikipedia.org/wiki/Automatic_number-plate_recognition