StacyMogeni / **phase3_project**

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

phase3_project Public

1 Branch 0 Tags Go to file t Go to file Add file Code

StacyMogeni Add presentation slides bd1be06 · 1 minute ago

.gitignore	Initial commit	last week
README.md	Update README.md	8 minutes ago
Terry_Stops_20250...	Add the initial dataset	last week
Untitled.ipynb	Add Conclusions and Rec...	14 minutes ago
cleaned_Terry_Sto...	Add Data Preparation	2 days ago
notebook.pdf	Add jupyter notebook pdf	13 minutes ago
presentation.pdf	Add presentation slides	1 minute ago

README

Terry Stops Arrest Prediction

About

No description, website, or topics provided.

Readme
Activity
0 stars
0 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

📌 Project Overview

This project analyzes Terry Stop data to predict whether an arrest was made after a stop, based on details of the stop, subject, and officer. The analysis is sensitive because insights may inform policy and operational decisions regarding police interactions.

🎯 Problem Statement

Police departments and policymakers need insights into the factors influencing arrests during Terry Stops. By leveraging machine learning, we aim to build a predictive model that:

- Identifies key factors influencing arrests.
- Ensures fairness by avoiding the use of sensitive attributes like race and gender.
- Provides actionable insights to support ethical and data-driven decision-making.

📂 Dataset

The dataset includes information on:

- **Stop details** (e.g., reported year, presence of weapons, frisk flag).
- **Subject details** (e.g., age group).
- **Outcome** (arrest made or not).

Target variable: `arrest_flag` (1 = Arrest made, 0 = No arrest).

🔄 Data Preprocessing

Steps applied:

- **Encoding:** Label Encoding (`subject_age_group`), One-Hot Encoding (other categorical variables).
- **Scaling:** StandardScaler applied to numerical features.
- **Class imbalance:** Addressed using **SMOTE** to balance arrests vs non-arrests.
- **Sensitive attributes:** Race and gender were excluded to avoid bias in modeling.

🤖 Modeling

We compared four models:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost

Key Results

Model	Accuracy	Recall (Arrest=1)	ROC-AUC
Logistic Reg	0.931	0.58	0.914
Decision Tree	0.934	0.52	0.726
Random Forest	0.948	0.68	0.931
XGBoost	0.941	0.69	0.933

Final Model

- **Chosen Model:** XGBoost
- **Reasoning:** Balanced performance across metrics with strong recall (important for correctly identifying arrests).
- **Hyperparameter Tuning:** Improved accuracy to 0.944 and ROC-AUC to 0.934, though recall slightly decreased to 0.68.

🔑 Key Insights

- **Weapon presence** was one of the strongest predictors of arrest.
- **Frisk flag** and **subject age group** also played important roles in predicting outcomes.
- Sensitive features (race, gender) were tested but excluded from the final model to maintain fairness.

📊 Conclusions

- Arrest outcomes are strongly influenced by risk-related factors such as weapon presence.
- Machine learning models, especially **XGBoost**, can effectively predict arrests with high accuracy.
- Fair modeling practices are essential — excluding sensitive variables avoids biased outcomes.

✅ Recommendations

- **Policy:** Focus resources on higher-risk stop situations (e.g., weapon-related) where arrests are more likely.

- **Ethics:** Continue avoiding sensitive demographic features in predictive modeling.
- **Future Work:** Explore methods to further improve recall (catching more true arrests) such as adjusting thresholds or applying cost-sensitive learning.