Report 4

PREDICTIVE MODELLING NEW YORK CITY'S STOP-QUESTION-FRISK (SQF) DATA

By

PRAJWAL NAGARAJ

Data Preparation

In the first phase of the CRISP-DM framework, data preparation involves defining and preparing class variables, removing unnecessary variables, and describing the final dataset used for classification.

Class Variables:
For the classification tasks, we have chosen to predict whether a person is armed and whether an arrest will be made. The target variable for the first task is 'armed', which we derived by combining information from various weapon-related columns. We assigned a value of 1 if any weapon-related column had a 'Y' entry and 0 otherwise.

Data Cleaning and Preparation:
Initially, we loaded the dataset and examined weapon-related columns for missing values. We created the 'armed' variable as described above, and afterward selected relevant features such as 'trhsloc', 'perobs', 'frisked', 'searched', 'contrabn', 'inout', 'sex', 'race', 'height', and 'build' for further analysis.

We performed data transformation and preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features. Categorical variables were one-hot encoded, and numerical features were standardized to ensure uniformity in scale.

Modelling

In this phase, we built and evaluated multiple classification models for each task.

Classification Models: For predicting whether a person is armed, we created three different classification models: Random Forest Classifier, Logistic Regression, and Decision Tree Classifier.

For predicting whether an arrest will be made, we built a Logistic Regression model.

Advantages of Each Model:

- **Random Forest Classifier:** This ensemble learning technique provides robust performance, handles high dimensional data well, and is less prone to overfitting. It's effective for handling imbalanced datasets and capturing complex relationships between features and the target variable.
- **Logistic Regression:** Known for its simplicity and interpretability, Logistic Regression is efficient for binary classification tasks. It provides probabilities for class membership and can handle linearly separable data.
- **Decision Tree Classifier:** Decision trees are intuitive and easy to understand, making them useful for exploring feature importance. They can handle non-linear relationships and are robust to outliers.

Evaluation: We assessed the performance of each model using various metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. Cross-validation techniques were employed to ensure generalizability of the models. The results indicated the effectiveness of each model in capturing patterns in the data and making predictions.

Model Evaluation

For the 'armed' prediction task, the Random Forest Classifier achieved an accuracy of approximately 92%, with precision, recall, and F1-score of around 8%, 53%, and 14%, respectively. The Logistic Regression model for predicting arrests achieved an accuracy of approximately 96%, with precision, recall, and F1-score for the positive class (arrest made) at 80%, 40%, and 53%, respectively. Finally, the Decision Tree Classifier achieved an accuracy of around 90%.

Evaluation

The models demonstrate promising performance in their respective tasks. For policing purposes, these models can aid law enforcement agencies in making informed decisions, such as allocating resources effectively and prioritizing interventions. The value of the models can be measured by their ability to enhance decision-making processes, reduce response times, and potentially prevent crime.

Model Implementation

To implement the models effectively, continuous data collection is crucial to keep the models updated and reflective of changing trends and patterns. Other relevant data that could be collected include socio-economic factors, historical crime data, and demographic information. The frequency of model updates depends on the rate of change in the underlying data and the criticality of timely decision-making.

Conclusion

In conclusion, predictive modeling offers valuable insights for law enforcement agencies to improve policing strategies and enhance public safety. By leveraging machine learning techniques and robust evaluation methods, these models can assist in predicting armed encounters and the likelihood of arrests, enabling proactive and targeted interventions.