IN3062: Introduction to Artificial Intelligence Coursework

Kamal Kirosingh

Mohammed Alqumairi

Contents

Introduction

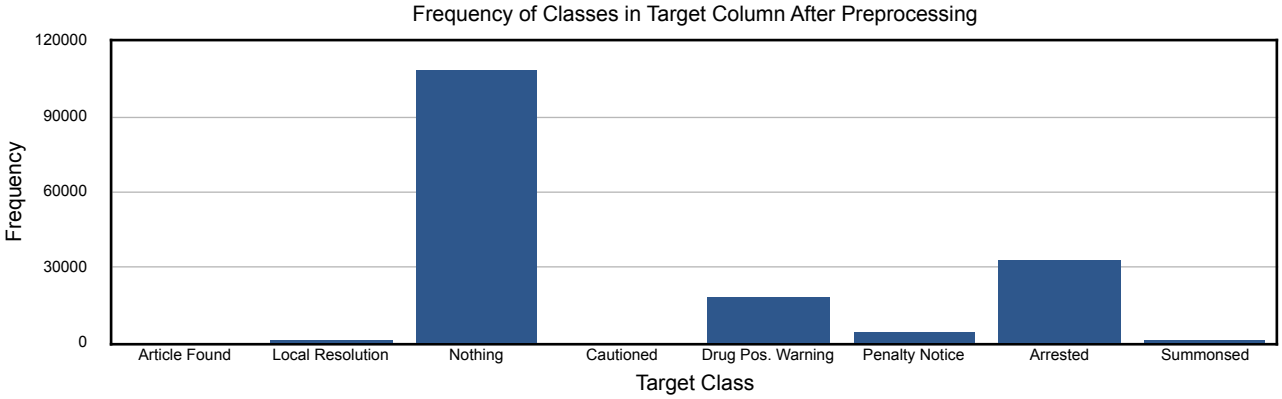
In this report, machine learning techniques will be used to predict the outcomes of police stop and searches, in London, spanning from late 2014, to 2017. The problem domain is law enforcement, and is therefore highly relevant today given the increased scrutiny on policing practices in 2020. The dataset was retrieved from Kaggle, originally taken from the British Home Office (Dane, 2017).

The objective of research is to predict what police officers do to a suspect following from a stop and search. It is as such a classification problem.

Data Exploration

The raw dataset is comprised of seventeen columns, and 302,623 rows. Out of the seventeen columns, one specifies the date and time of the stop and search, two are numerical (specifying the latitude and longitude of the area wherein the stop and search occurred), and the rest are categorical.

The target column is the “Outcome” column (the outcome of the stop and search), containing eight unique values, including “Suspect Arrested”, “Offender given penalty notice”, and “Nothing found - no further action”. The latter target class became a source of great difficulty, as most samples (approximately 65% of all samples after preprocessing) belonged to this "Nothing found - no further action” class. This imbalance was the chief difficulty faced in the completion of this report.



Preprocessing

Cleaning the dataset from columns irrelevant to the research objective was the first order of business. The only column judged to be irrelevant in this regard, was the suspect’s self-defined ethnicity, as this is not something that the police officer would have been aware of before the stop and search. For data on the suspect’s ethnicity, another column, “Officer Defined Ethnicity”, defines the suspect’s ethnicity from the officer’s perspective.

After removing irrelevant features, we ensure no empty cells exist in the dataset. Out of the sixteen remaining columns, only two were numerical (“Latitude” and “Longitude”), and so empty cells in each were replaced with a median. As for the remaining categorical columns, any row with an empty cell under any of them, would need to be removed, as it made no sense to find a median to categorical data. To avoid needing to discard most samples in the dataset, six columns were deleted because they were mostly comprised of empty cells. The final result, is a dataset with nine features, a target, and 165,651 samples.

After cleaning the dataset, all columns- apart from the two numerical ones- needed to be encoded before passing the data into machine learning models for training. Scikit Learn’s “LabelEncoder” was used for this.

To deal with the severe dominance of a single class in the dataset, several preprocessing techniques were used before training each model, tuned iteratively to improve model performance. Those include over sampling using Synthetic Minority Oversampling Technique (SMOTE), and under sampling. This will be elaborated on when discussing the development of each model.

References

Dane, S., 2017. London Police Records. [online] Kaggle.com. Available at: <https://www.kaggle.com/sohier/london-police-records?select=london-stop-and-search.csv> [Accessed 19 November 2020].