Social Data Science – Exam Project

REPORTING ON MINORITIES

The implementation of machine learning models in predictive policing

Group 33:

Troels Christian Nielsen Boesen (277)

Frederik Niemann Kildedal (157)

Kristian Hamburger Holm (16)

Mathias Skaaning Bruun (184)

Institute for Economics / Centre for Social Data Science

University of Copenhagen

Character count: 35,861

Table of Contents

Allocation of Responsibility for Different Sections

The distribution of the overall responsibility for each section is as follows, but please note that all group members have contributed equally to both coding as well as the production of the paper. We all take full responsibility for the final product.

Troels (277): 2.2, 5.2, 6

Frederik (157): 4.1-44, 5.1, 6

Kristian (16): 2.1, 2.3, 6

Mathias (184): 1, 3, 6

1 Introduction

Nearly two decades ago, the movie *Minority Report* (2002) presented a dystopian view of the future of policing, in which technology has made it possible to predict crimes before they happen and thereby apprehend would-be offenders before the fact.

Previously, this has exclusively been the stuff of sci-fi novels and Hollywood movies, but recent technological advancements in computer science along with increased availability of diverse kinds of data is paving the way for real-life crime prediction. In February of this year it was revealed that 14 UK police forces either make use of predictive models already or plan to do so (Kelion, 2019). This begs the question: how great is the potential of these models, and how do we deal with the ethical issues raised by the prospect of predictive policing?

Several concerns are voiced within the literature, including the ambiguousness of temporality and causality introduced by the aforementioned notion of "pre-crime" as well the risk of implicit racial profiling and other civil rights violations (Asaro, 2019; Kaufmann, Egbert, & Leese, 2019; Perry, et al., 2013). For this project specifically, temporality and causality are issues of minor concern, whereas racial profiling plays a crucial role.

Using data gathered through the official API of the UK police, this paper attempts to answer the following research question:

Is it possible to predict the outcome of a police stop-and-search based on geographic and temporal data combined with personal characteristics of the suspect in question?

We train a logistic regression model to predict whether or not a given search will be "successful" in the sense that the suspect is found to be in possession of illegal goods or exhibiting other unlawful behaviour.

Following our initial analysis and evaluation of our model, we discuss the ethical implications of predictive policing, focusing particularly on the pitfalls of implementing models that make predictions about criminal behaviour based on sensitive personal information.

2 Data

We use open data about crime and policing in England and Wales, provided by the UK police on data.police.uk. The UK police delivers an API containing monthly data about every stop-and-search for each police force in England and Wales from July 2016 to April 2019. The content of each request contains data for all stop-and-searches for a given month in each police force.

2.1 Collection

In order to gather the data from every police force in each period into one DataFrame, we had to loop through every month from July 2016 to May 2019 (the inner loop), and then loop through every 44 police forces in England and Wales (the outer loop). In each inner loop we created a new column variable *force*, which indicates in which police force the stop and search took place. In total we made 1496 requests each containing 17 columns. The final concatenated dataset contains a total of 838,934 observations (rows).

The police API has a rate limit of 15 requests per second with bursts of 30, which means that we could reduce the waiting time to 0.1 seconds (https://data.police.uk/docs/api-call-limits/). After the first attempted scrape, it was noted that the connector had timed out while getting data from the Metropolitan Police Service. Since this force covers Greater London and thus captures a vast portion of the British population, this was judged to be the result of large datasets rather than substantive errors. Thus, the timeout was increased to 250, after which the scrape ran without error.

The reason for adapting our request process is illustrated by the two figures below, showing the call time and the response size over time, respectively:

Figure 2.1

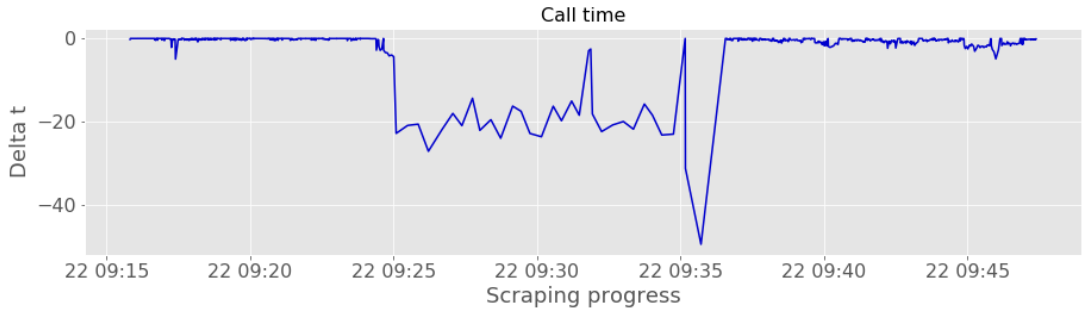
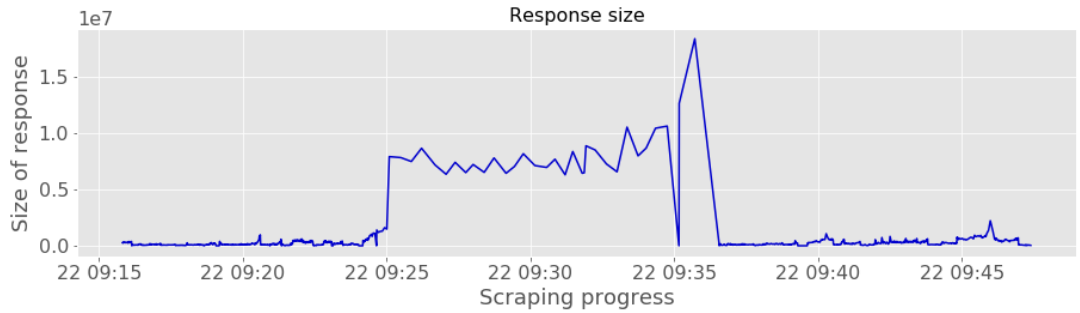


Figure 2.2



The first figure shows the time it took to make the call. From 09:25-09:35 the call time increased significantly when calling one specific force (the Metropolitan). This was due to an increasing response size which is illustrated in the bottom figure.

2.2 Research Ethics

In addition to the general significance of sound research ethics, the subject matter of the present project makes it even more important to consider ethical questions. Criminal records are broadly considered to be highly sensitive personal information. This paper relies entirely on data compiled by government-controlled bodies already required to fulfil ethical standards. The API documentation explains that both the Information Commissioner's Office and Data Protection specialists at the Home Office were consulted heavily in preparation for the launch of the API.[1] A short review of the data has thus been judged to suffice for the present purposes.

The data gathered does not contain any information that can be used directly to identify individuals. Neither is it possible from the available data to deduce the identity of individuals with any non-random success. This is partly due to the very broad categorisation of personal characteristics of the individuals searched (e.g. age category above 34), and partly due to anonymisation by the police of location data. All coordinates given are closest "anonymous location" defined as:

1. Appears over the centre point of a street, above a public place such as a Park or Airport, or above a commercial premise like a Shopping Centre or Nightclub.
2. Has a catchment area which contains at least eight postal addresses or no postal addresses at all.[2]

Further to the ethical concerns raised above, it is necessary to be aware of the potential legal issues. This is due to the fact that the project among other variables uses race to predict outcomes, and thus falls in the controversial category of racial profiling. The ethical concerns regarding racial profiling will be treated later but note initially that the topic is highly controversial even in legal terms. In the US, the courts have ruled that profiling based *partly* on race is *not* in violation of the constitution. (US vs. Weaver, 1990). In the UK, (Baker & Phillipson, 2011) argue that laws allow for stricter scrutiny and a higher threshold for the permissibility of racial profiling. This nevertheless implies *a fortiori* that it is allowed in some circumstances. As such, the present project is not legally speaking on shaky ground.

[1] https://data.police.uk/about/#anonymisation
[2] https://data.police.uk/about/#location-anonymisation

2.3 Cleaning and Transformation

Many of the observations for stop-and-searches had missing values. This is not the result of an error in the scraping process, but because the police have not filled in all information for all the stop-and-searches. It was thus not possible to gather the missing information in other ways. Furthermore, the nature of the relevant variables was such that it would be inappropriate to attempt to impute the missing values. Considering the volume of observations, the rows containing missing values for the variables included in our machine learning model were dropped. Under the assumption that it is not systematic which observations contain missing values, dropping them will not influence the estimates of the machine learning model. This assumption could reasonably be challenged, but it is outside the scope of this paper to do so.

In order to improve the prediction of our machine learning model, we wanted to create more features to include in the model. To use both the time of the day and the time of the year for the stop-and-searches as features, we had to subtract month and the hours from the *datetime* variable, creating two new variables *month* and *hours*. In order to plot the data geographically, we also needed to somehow extract the area of each stop from our locational variable. The *force* variable is geographic in nature, but it does not allow us to plot the data in a geospatial manner. In the variable *location* there is an estimated longitude and latitude of the anonymous point closest to the stop-and-search, which could be used to locate under which Local Authority (an administrative geographical unit) each stop-and-search took place. This was done by importing a CSV file from data.gov.uk[3] containing the longitude and latitude of a point in each Local Authority in England and Wales, provided by the ONS. By minimizing the distance between the location of the stop-and-search and the Local Authority we could approximate under which Local Authority the stop-and-search took place. This is only an approximation since it is not certain that the Local Authority located closest to the location of the stop-and-search is the actual authority in which the location belongs. It does, however, yield some smaller geographical areas which can be projected onto a map of Local Authorities. In order to use the longitude and latitude of the stop-and-searches, we first had to extract these from the dictionary variable *location*. This meant that we also got 4 new columns *latitude, longitude, street_id* and *street_name*.

Finally, we created a dummy variable *success,* indicating whether the stop was a success or not, meaning whether the police found something illegal on the searched individual. The variable is defined as a "failure" if the *outcome* equals "False" or "Nothing found – no further action" and otherwise a "success".

[3] https://data.gov.uk/dataset/6a4fcbf1-1562-4ab4-8623-6fd97b6deb5e/local-authority-districts-december-2017-super-generalised-clipped-boundaries-in-great-britain

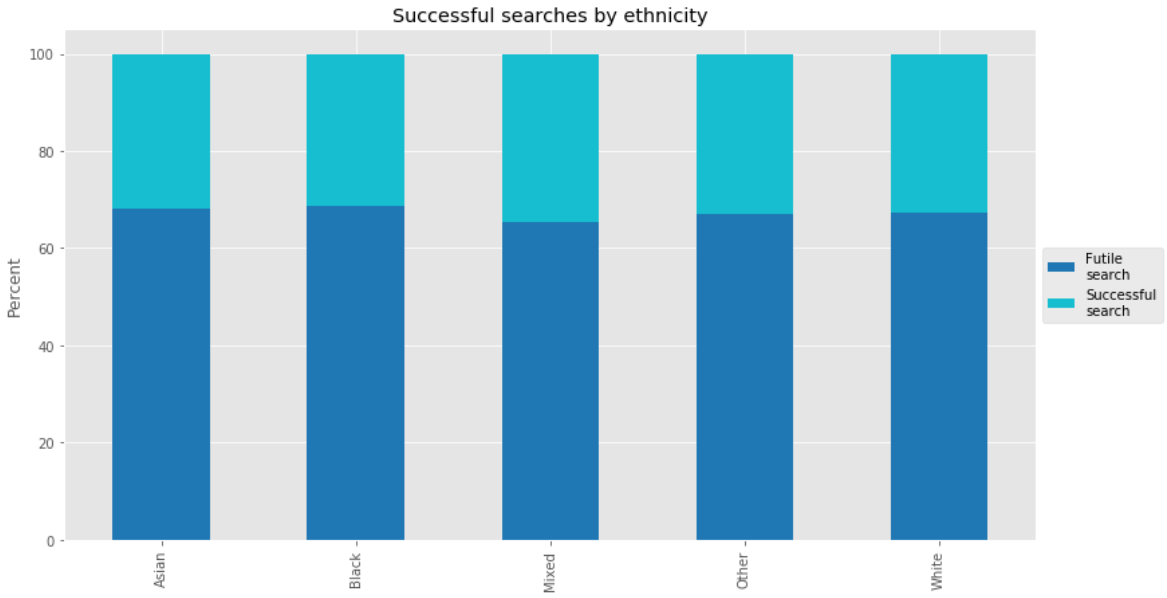
3 Descriptive Analysis of Our Dataset

In the following section, we present descriptive measures of our sample data, examining variables of interest: gender, ethnicity, and age group of the suspects along with success rates across different groups. Please refer to *Appendix A* for tables containing value counts and crosstabs of selected demographic features.

The largest group of suspects in our sample consists of white males, followed by black and Asian males. Overall, 91.77% of our sample is male, while only 8.17 of all suspects are female. 0.06% fall in the "Other" category. The most prevalent ethnicities are white (59.09%) and black (25.90%), and the age group 18-24 is the most well-represented at 37.43%.

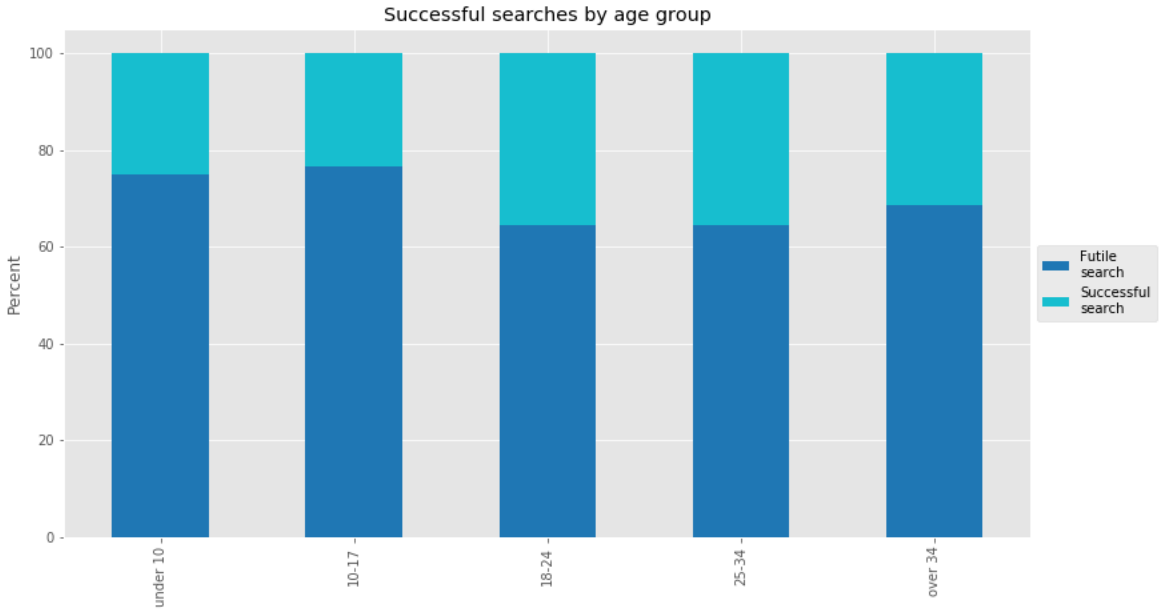
The overall mean success rate (i.e. proportion of cases in which the suspect exhibits some sort of unlawful behaviour) is 32.22%. Further exploratory plotting of the data shows the success rate across groups:

Figure 3.1



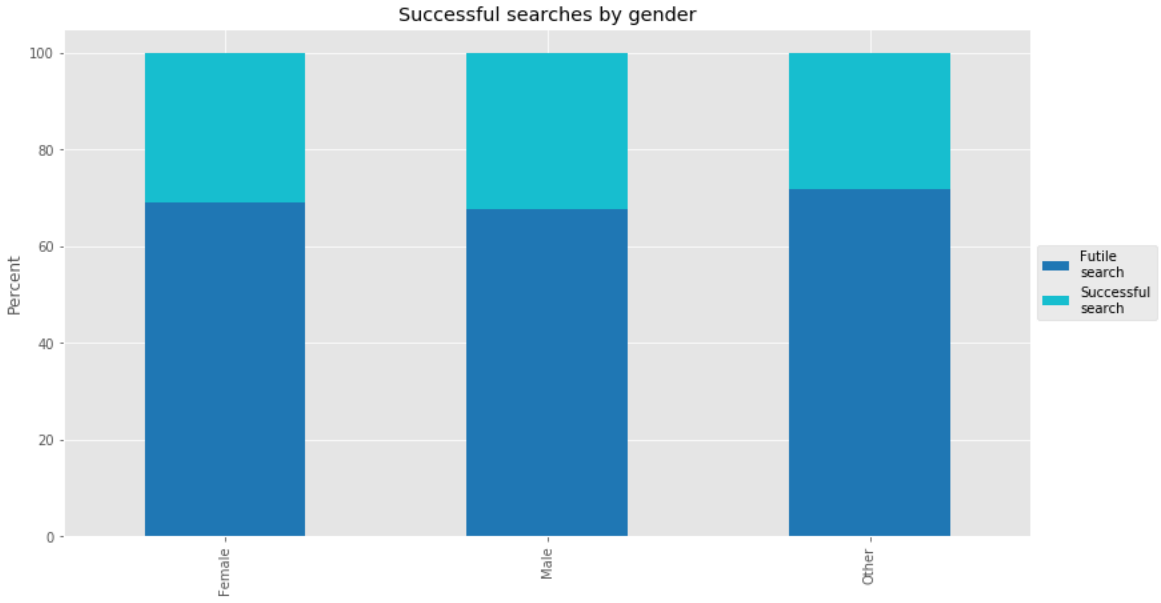
The success rate is almost identical across different ethnicities, ranging from 31.4% for black suspects to 34.5 % for mixed suspects. This shows that ethnicity is not very informative of whether the outcome is successful or not, hence it will most likely not have much explanatory power in a prediction model.

Figure 3.2



There is slight variation between age groups. The maximum success rate is about 35.4% for the ranges 18-24 and 25-34, which are also the most frequently stopped-and-searched age groups (see *Appendix A*). The success rate lies between 25.0% and 23.32% for children up to 17 years old, but please note that children under 10 years only make up 0.05% of the observations.

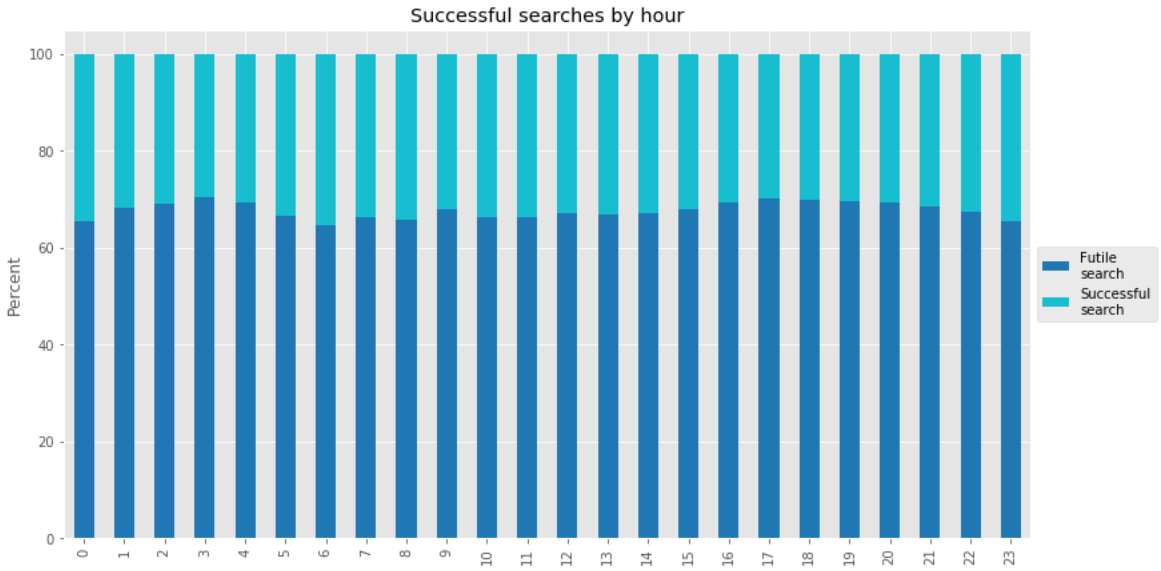
Figure 3.3



Once again, we do not observe much variation in the success rate, and thus gender is not very indicative of the outcome either. The success rate for females is 30.84% and 32.35% for males. However, the frequency of stops is much higher for men than for women, as 91,77% of individuals stopped-and-

searched by the police are men (see *Appendix A*). The "Other" category has the lowest success rate at 28.12%, but this only constitutes an extremely small part of the dataset.

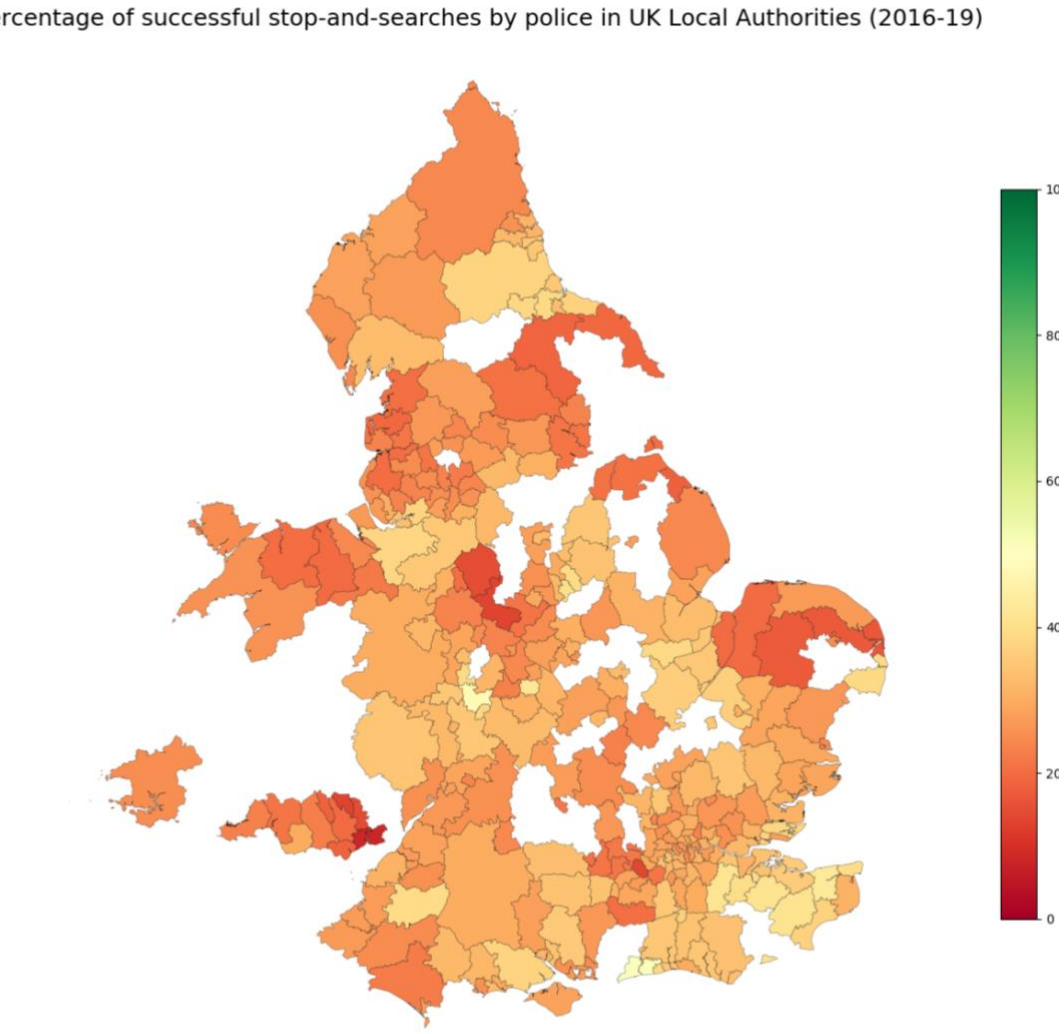
Figure 3.4



Furthermore, it is clear that the proportion of successful searches is quite stable across hours of the day, even though the absolute number of stops varies (the largest share of stop-and-searches (8.54%) is conducted between 11 PM and midnight).

It is evident from the cases in our sample that police stop-and-searches are typically futile – in the sense that the suspect is not actually found to be in possession of any illegal goods. To further illustrate this tendency, the choropleth below plots the proportion of successful stop-and-searches by UK police in each Local Authority. The colour varies from green to red depending on the success rate in the particular area. Red shading means a larger proportion of futile searches, while green shading would correspond to higher success rates:

Figure 3.4



The gaps in the plot are caused by removing areas with fewer than 100 observations, of which there are 45. Please note that low numbers of observations are not caused by less crime, but rather by missing *location* data. Consequentially, even though geospatial data is convenient for exploratory plotting such as this, a more aggregated "geographic" measure is used in the actual model by including only the *force* variable to capture any locational variance.

Regarding the choropleth, the proportion of successful stop-and-searches is remarkably low in most Local Authorities. Obviously, this discovery further motivates our research: if the UK police carry out such a large number of stop-and-searches in vain, it may be interesting to establish an effective way of predicting which searches will have a successful outcome.

It is important to note, however, that considerable difficulties with regards to prediction are to be expected due to the apparent lack of variation in success rates across the features available to us, and because the dataset has proven to be highly imbalanced with a majority of stop-and-searches being unsuccessful.

4 Model Training and Prediction

As mentioned, the outcomes of the searches are divided into two categories: successful searches, where stops lead to further action, and unsuccessful searches where no further action is taken with the stopped individuals. The successful stop-and-searches variable is converted to a dummy variable with a successful stop equal to 1 and an unsuccessful outcome equal to 0.

4.1 Logistic Regression

The prediction of the successful variable is a classification problem with the goal of correctly classifying the successful and unsuccessful outcomes with one and zeroes. To solve the binary classification problem the logistic regression model is used. The logistic regression evaluates the probabilities of positive/negative (1 or 0) outcomes by estimating a linear relationship between features (x values) and log-odds:

$$logit(p(y = 1|x)) = \boldsymbol{w}^T\boldsymbol{x}$$

From the log-odds calculation the probability of the outcome being positive/negative is estimated by applying the sigmoid function:

$$\phi(z) = \frac{1}{1 + e^{-z}}, \qquad where\ z = \boldsymbol{w}^T\boldsymbol{x}$$

This probability is then used to estimate the outcome based on a threshold (here 0.5):

$$\hat{y} = \begin{cases} 1 & if\ \phi(z) \geq 0.5 \\ 0 & if\ \phi(z) < 0.5 \end{cases}$$

When selecting the weights $\boldsymbol{w}$, one needs to choose $\boldsymbol{w}$ such that the likelihood function is maximized (or minimized if the negative likelihood function is used). This is done by choosing an algorithm that maximizes the log-likelihood function:

$$\log L(\boldsymbol{w}) = \Sigma_{i=1}^{n}\left[y^i \log\left(\phi\left(z^i\right)\right) + \left(1 - y^i\right)\log\left(1 - \phi\left(z^i\right)\right)\right]$$

In this paper, the limited-memory BFGS method is used and applied in the scikit learn logistic regression function for the optimization problem.

The target value is the outcome of the stop-and-search and is captured by the dummy variable *success*. Since the goal is to predict the outcome by using data on personal characteristics, time and geography, the following variables from the scraped data are applied in the machine learning model:

Variables in Model

	Label
1	age_range
2	gender
3	object_of_search
4	officer_defined_ethnicity
5	type
6	Force
7	hour
8	month

The personal characteristics are described by the variables *age_range*, *gender* and *ethnicity*. The variables *object_of_search* and *type* describe the characteristics of the search, while the geographical parameter is captured by the *Force* operating in the area of the stop and search. Finally, the *hour* and *month* variables seek to capture the temporal influence on the outcomes.

From the selected variables all missing values are dropped, as mentioned, so only observations with data for all of the categories above are included. This leaves 673,651 observations for the analysis where the mean of the success variable is equal to 0.3222. All the variables are categorical, and before running the logistic regression these are converted to dummy variables, so that each feature is a category within the variables shown above. A list of all features is provided in *Appendix B*.

4.2 Training and Hyperparameter Optimization

The data is split into a development set and a test set to avoid contamination. They are divided such that the development set contains 80% of the data, while the test set contains the remaining 20%. The development set is used to optimize the regularization parameter λ. The regularization is optimized to control the complexity of the model and thereby the variance and bias. The variance and bias are measures of the variability in the estimates across samples and systematic errors in the predictions. When modifying

the λ parameter by testing it across different values one can control the regularization and thus find an appropriate solution to the bias-variance trade off. In this estimation process the weight decay regularization method is used (in scikit learn "l2"). This adds the following term to the cost function (i.e. the negative log likelihood function from before):

$$\frac{\lambda}{2}\Sigma_{j=1}^{m} w_j^2$$

When this term is added, the weights will shrink. This ensures that the model can fit the data, without extreme weights and thus makes out of sample performance more robust. The regularization requires that feature scales are comparable (Raschka & Mirjalili, 2017). Therefore, all features have been adjusted using scikit learn's *StandardScaler* class. In scikit learn's logistic regression function, the λ is controlled by the C parameter, which is the inverse of λ. Therefore, decreasing the value of C will increase the regularization and vice versa. To select the optimal value of this regularization parameter, k-fold cross validation is applied. As mentioned, the hyperparameter tuning is done on the development set. This set is split into a training and validation set k times. This way the model can be fitted on the training sets and tested on the validation sets in each of the k iterations. In each iteration $\frac{k-1}{k}$ is used for training and $\frac{1}{k}$ is used for validation. In this paper the k in the k-fold cross-validation is set to 10, as empirical studies have shown that this gives a good bias and variance trade-off (Raschka & Mirjalili, 2017). This method is used to test which value of the hyperparameter C has the highest accuracy in its predictions and therefore should be used for fitting on the whole development set. The accuracy is defined as follows:

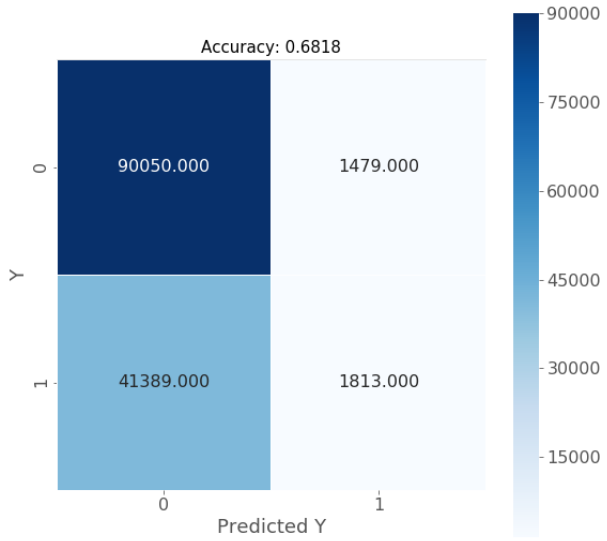
$$Accuracy = \frac{True\ Positive + True\ Negative}{False\ Positive + False\ Negative + True\ Positive + True\ Negative}$$

The accuracy measure shows the fraction of correct predictions out of the total predictions and thus tells how often one's prediction is true. The hyperparameter C, which gives the best accuracy across the k-fold cross validations are fitted on the development set and then used to predict the target variables based on the features in the test set.

4.3 Results

The results of the prediction are shown in the following confusion matrix, where the number of all correct and incorrect classifications of both positives and negatives can be seen.

Figure 4.1



The accuracy at first sight seems to be at an acceptable level of 0.6820. However, when comparing this value to what accuracy one would get by just classifying all stops and searches as unsuccessful (0.6778), the model does not do a significantly better job. To assess the quality of the model, it is necessary to take other performance measures into account. From the confusion matrix it is seen that the model mainly predicts the stop-and-searches to be unsuccessful. This is problematic as the bottom left part of the matrix shows that there is a huge number of false negatives. The right part of the matrix reveals that the model very rarely predicts that a stop-and-search will be successful, despite approximately 30% of the dataset being successful searches. To examine these predictions further, the metrics precision, recall, F1 and support are taken into account.

4.4 Performance Metrics

In the formulas below true positive is denoted as TP, false negative as FN and so on.

$$Precision = PRE = \frac{TP}{TP + FP}$$

$$Recall = REC = \frac{TP}{FN + TP}$$

$$F1 = 2\frac{PRE * REC}{PRE + REC}$$

These metrics are computed for the failures (negatives/0) and successes (positive/1) and are shown along with support in the table below. Support shows the actual number of failures and successes in the test set.

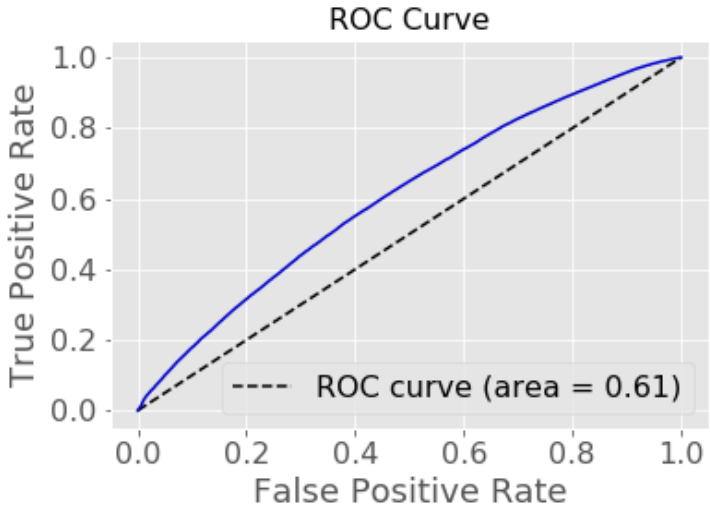
Model Performance Metrics

	Failure	Success
Precision	0.685109	0.550729
Recall	0.983841	0.041966
F1	0.807739	0.077989
Support	91529	43202

From these metrics it can be seen that the model performs very poorly on the success side. The model proves to have a precision of 0.55. This means that when it predicts a successful outcome, there is a higher chance of this being true than for a random stop-and-search. However, the model is terrible at capturing all of the successful outcomes, as it very rarely predicts that the outcome is successful. This leads to a very low recall on the success side of 0.0419, indicating that the majority of successful stop-and-searches are misclassified as failures in the model. This low recall also results in a very low score when the combination of precision and recall is measured in the F1 which is equal to 0.0779 for successes. For the failure side, the metric scores are much higher. However, this is mainly a result of the model generally classifying the outcomes as failures. This leads to an extremely high recall of 0.9838 and the model does capture most of the failed stop-and-searches.

The model mainly predicts unsuccessful outcomes and therefore does poorly on another important factor: capturing the successful outcomes. Of course, one can modify the hyperparameter to be measured by the scoring of other measures such as for example F1. However, this has been tested and does not yield significantly different results. Another improvement which could have altered the model, would be changing the threshold in the logistic regression and thereby trying to improve the recall, possibly at the expense of the accuracy score. However, this should have been done in the tuning of the hyperparameters and against another score than accuracy, for instance F1, so one does not just overfit on the estimated model using test set information post-learning. It is possible, though, to evaluate the model over different thresholds by plotting the ROC curve and seeing whether the model is better than random guessing. The ROC curve plots the true positive rate and false positive rate across different thresholds for the classification. For the estimated model this yields the following graph:

Figure 4.2



With

$$FPR = \frac{FP}{FP + TN}, TPR = REC$$

In the plot, the dashed line is the results of random guessing. Here it is worth noticing that the model does better than this, and the area under the curve is 0.61 compared to 0.5 for random predictions. Decreasing the threshold increases the true positive rate and thereby more of the successful outcomes are predicted, although this is at the expense of increasing the amount of false successful outcome predictions.

Despite the better-than-guessing performance shown in the graph, ROC curves can sometimes be misleading for imbalanced datasets such as this (Davis & Goadrich, 2006). Instead, a precision-recall curve can be used to examine the model's abilities. To illustrate the trade-off for varying thresholds, graphs with recall and precision for success are constructed below – here also with a graph showing the thresholds on the x axis. This shows how precision is highly affected when wanting to capture more of the successful outcomes.

Figure 4.3

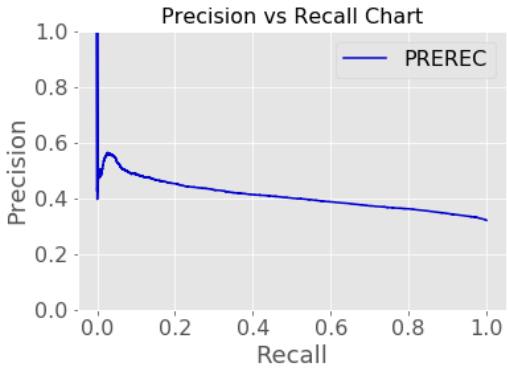
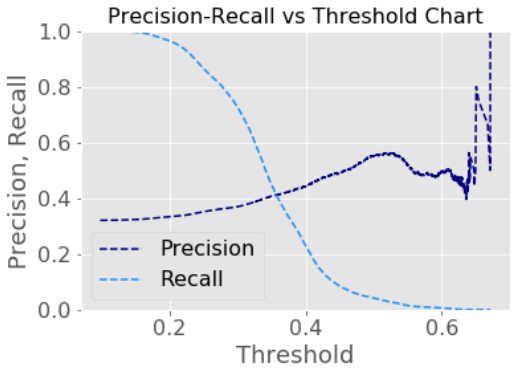


Figure 4.4



Here it is shown that recall is very low at high precision levels and decreases rapidly when increasing the threshold from 0.2 to 0.4, roughly speaking. Precision similarly increases rapidly when increasing the threshold from around 0.6 to 1. For the remaining part of the graph, variation in threshold only has minor impact on the precision. For the precision-recall curve, the AUC is only 0.4080, indicating that the model is neither particularly good at capturing the outcomes nor highly precise when predicting them.

Overall, the model does a bad job in predicting the outcome when keeping the imbalanced dataset in mind as the accuracy is evaluated. This is highly visible from the low recall and F1 scores, which suggest that the model captures only few of the successful stops and searches. Of course, this could have been altered by hypertuning the threshold, putting more focus on higher recall in order to catch all criminals. However, this was not apparent prior to the post-modelling descriptive analysis.

One parameter where the model does a slightly acceptable job is in its precision. When the model actually predicted that the stop-and-search would be successful, it proved right 55% of the time, whereas classifying all outcomes as successful would only have had a precision of approximately 30%.

5 Discussion

5.1 Data and Model Results

In the evaluation of the results in this paper, there are a few things that must be taken into account when assessing the validity of the results. It is not only the outcomes in the dataset that are imbalanced, the inequality is also highly present in the features. For example, the dataset consists of 91.77% males, while the gender distribution in the UK is 49.3% males and 50.6% females (Statista, 2018). This must be a result of a selection process where the male feature or unknown features correlated with it are used in the selection, leading to more imbalanced data on certain characteristics than for the general population. It follows that the data used in this paper is not an unbiased selection of stop-and-searches for random

people on the street. Factors such a suspicious appearance, shady behaviour etc. most likely affect the selection process of whom to stop-and-search. This means that the dataset is biased from the outset, and the selection criteria remain unknown. As a result, the model makes predictions within a dataset where the observations have been systematically collected in an unidentified manner. The implication of this is a non-random dataset which leads to the model trying to predict outcomes on different features, that might also have had an effect on the data being generated in the first place. In other words, police officers may have already selected the suspects in our sample based, in part, on the same features as those included in our model. As a result, the sample is non-random, and the interpretation of a model estimated within this sample will be biased.

Another factor to consider is what one actually wants to capture in the data. It is possible to argue that there is a trade-off between catching the majority of the criminals and minimizing resources spend on unsuccessful stop-and-searches. This discussion is important for the model criteria and could have been established in a more elaborate manner before the analysis. The scoring parameter, accuracy, which is used in this paper, mainly considers whether the predictions are correct on a general level. If, instead, one wanted to be sure of not letting a considerable number of criminals pass by, which is probably preferable for the police, a larger focus could be put on increasing the number of success predictions. Altering the threshold and other factors in the estimation process could have led to different results and a larger recall. This would have been preferable if another specific goal was set for the prediction process instead of just optimizing accuracy.

Alternatively, both the accuracy and other performance measures could all have been sought improved by trying out different models and tuning more hyperparameters. The optimization method used only allows for weight decay regularization, while applying different optimization methods – along with changing the regularization mechanisms – might have led to improvements. Other classification approaches such as naïve Bayes, *k*-nearest neighbours, random forest etc. might have improved the model's predictive ability across all measures.

The greatest potential of predictive policing lies, perhaps, not in street-level operations such as stop-and-searches which can be spontaneous in nature, rely heavily on a few external characteristics of the suspects and are infeasible to conduct completely at random. Rather, more promising results might be achieved through the application of machine learning models on realms of crime fighting wherein, for instance, registry data can be used.

5.2 The Ethics of Crime Prediction Software

The model this project attempts to construct draws on a number of personal characteristics for suspected offenders. It is generally settled in the literature on applied ethics that predictions on the basis of e.g. gender or age are ethically sound. However, there is vast disagreement on the question of *racial* profiling. In line with the literature on the subject it is assumed that the generalisations are factually sound, and that no other, better selection tool is available. (Applbaum, 2014; Hellman, 2014; Altman, 2019) These assumptions can reasonably be challenged, but under non-conforming circumstances, the main issue for racial profiling would not essentially be ethical.

Relative to the cases considered in the established literature in applied ethics, this project distinguishes itself primarily in the use of machine learning. It is thus necessary to ask both (1) whether racial profiling *simpliciter* is ethically defensible and (2) whether there are any salient features that distinguishes machine learning based racial profiling (MLRP) from human racial profiling (HRP)?

For the present purposes, the most applicable ethical theories are consequentialism, deontology, and contractualism.[4] Generally speaking, the Kantian focus on the dignity of the individual makes deontology the easiest moral philosophy from which to condemn racial profiling. Any argument in favour of racial profiling based on deterrent effect would further be a classic example of treating humans as means rather than ends, which is categorically forbidden in Kantian deontology (Johnson & Cureton, 2019). Contractualism represents an intermediate step. While (Applbaum, 2014) and (Hellman, 2014) disagree on whether racial profiling can be allowed under a contractualist framework, their disagreement hinges on *empirical* facts about the current state of affairs in American racial tensions. Thus, if racial profiling can be defeated in a consequentialist analysis, it holds *a fortiori* that it generally cannot be defended.

In a consequentialist analysis, racial profiling would simply be allowed[5] if it produced more positive outcomes than negative ones.[6] It is a well-known problem in consequentialism that strong pain for a few individuals is outweighed by mild pleasure for many people. Building on the argument of Kennedy (1999), it is, however, doubtful whether this would in fact be the case for racial profiling. Kennedy rightly argues that those subjected to racial profiling are *repeat players* who on several occasions will have to endure the inconvenience of "false-positive searches". Furthermore, Hellman's objection (2014) can be adopted for a consequentialist analysis. She holds that false-positive searches based on racial profiling do not just

[4] These (and virtue ethics) are the main branches of contemporary normative ethical theory. Since virtue ethics almost tautologically focuses on individual morality and character development, this is judged to be irrelevant for the present purposes (Baron, Pettit, & Slote, 1997).

[5] Potentially required. The discussion of supererogatory acts in consequentialism is beyond this discussion.

[6] Negative consequences, displeasure, disutility, pain etc. (and the parallel for pleasure) are used interchangeably below. It is immaterial to the present discussion which welfare theory is used.

cause inconvenience but demean and insult the relevant individuals. In consequentialist terms, she contends that historic factors have the effect of deepening the intensity of disutility felt by the innocent targets.

While the argument given above shows the depth of displeasure felt, it can also be argued that the positive consequences from racial profiling are limited. Note here that this project *only* deals with street-level stop-and-searches, and that the ethical discussion is limited to cases where no other approach is feasible. It holds almost *ex hypothesis* that the criminals who can *only* be stopped with this type of selection mechanism are minor offenders. Consider terrorism, major drug operations, or organised violent crime. For all of these types of crime, it is almost *a priori* true that their very nature makes them enforceable through other means. As such, the benefits to the average citizen are arguably fantastically small.

The analysis above naturally forms the basis for the conclusion that racial profiling *simpliciter* is not morally defensible. This is undisputedly a controversial conclusion that would both find support and resistance in the existing literature. However, the argument greatly illustrates the salient features that determine in any individual case whether racial profiling can be allowed.

First, Kennedy's argument: All other things equal, circumstances where false-positive stops are *more* likely to be endured by the same individuals make racial profiling *less* ethically defensible. Similarly, for Hellman: In countries and cultures with *more* negative stereotypes about racial minorities, racial profiling is *less* morally defensible. Finally, under circumstances where *more* high-risk offenders are likely to escape apprehension without the use of racial profiling, racial profiling is *more* ethically defensible. This would e.g. mean that racial profiling (all other things equal) is *more* morally defensible in the UK than the US due to a different history of segregation and *more* defensible in risk-scenarios with greater likelihood of lone wolf terrorists than otherwise.

The difference between MLRP and HRP does not significantly influence neither the first nor the third point above. It is less clear how it impacts Hellman's dignity objection. It could be argued that the use of machine learning removes the risk of generalisations carrying any historical or social meaning. However, this would miss the fact that these historical and social factors have shaped the data on which the machine is trained. Thus, using machine learning risks masking historical injustices as mathematical truths. In addition to this, MLRP based on biased data risks creating a vicious circle which in fact would *deepen* the racial injustice. If the police continuously retrain the model and the training data is created by a racially biased model, it would necessarily self-reinforce its bias. As such, the presence of historical or instutionalised injustices are especially important when considering MLRP as opposed to HRP.

6 Conclusion

In this paper we have examined the relationship between the outcome of a police stop-and-search and features of the specific search. We have sought to limit the original selection made by the police further by trying to predict whether a given stop-and-search will be successful or not.

The initial exploratory analysis of our data revealed only miniscule variation in the success rate across most personal characteristics of the suspects, which seems to be an indication of the fact that the people in our dataset are exactly that – *suspects*. That is, they have already been singled out for searching by the UK police based on suspicion.

Based on this non-random subsample of the population selected for stop-and-searches by the police, we are unable to meaningfully predict the outcome of a given search, conditional on geographic and temporal data combined with personal characteristics of the suspect. The accuracy of our model does not vary significantly from simply predicting all the stop-and searches as failures. Neither does the model capture many of the successful outcomes and the recall is extremely low.

Although this paper is a contribution towards testing the limits of prediction modelling in practice, we must emphasize that our results do not indicate that all attempts at predictive policing, in a broader sense, will be futile. Specifically, prediction models may produce better results when trained on features beyond those that are physically observable, as well as under conditions where population data – or unbiased samples thereof – can be applied. In any case, the implementation of predictive policing must be preceded by careful consideration of the ethical implications involved.

7 References

Altman, A. (2019, 08 29). *Civil Rights*. Retrieved from Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/civil-rights/

Applbaum, A. I. (2014). Bayesian Inference and Contractualist Justification on Interstate 95. In C. a. (eds.), *Contemporary Debates in Applied Ethics* (pp. 219-231). Wiley-Blackwell.

Asaro, P. (2019). AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care (vol. 38, no. 2). *IEEE Technology and Society Magazine*, pp. 40-53.

Baker, A., & Phillipson, G. (2011). Policing, profiling and discrimination law: US and European approaches compared. *Journal of Global Ethics*, 105-124.

Baron, M. W., Pettit, P., & Slote, M. (1997). *Three Methods of Ethics.* Oxford: Blackwell.

Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd international conference on Machine learning*, 233-240.

Hellman, D. (2014). Racial Profiling and the Meaning of Racial Categories. In C. a. (eds.), *Contemporary Debates in Applied Ethics* (pp. 232-243). Wiley-Blackwell.

Johnson, R., & Cureton, A. (2019). *Kant's Moral Philiosophy*. Retrieved from Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/kant-moral/

Kaufmann, M., Egbert, S., & Leese, M. (2019). Predictive Policing and the Politics of Patterns. *The British Journal of Criminology, Volume 59, Issue 3*, Pages 674–692.

Kelion, L. (2019, February 4). *Crime prediction software 'adopted by 14 UK police forces'*. Retrieved from bbc.com: https://www.bbc.com/news/technology-47118229

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., Hollywood, J. S., & Perry, W. L. (2013). *Predictive Policing : The Role of Crime Forecasting in Law Enforcement Operations.* The RAND Corporation.

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning.* Packt Publishing.

Statista. (2018). *Mid-year population estimate of the United Kingdom (UK) in 2018, by gender and country (in million people)*. Retrieved from Statista: https://www.statista.com/statistics/281240/population-of-the-united-kingdom-uk-by-gender/

US vs. Weaver, No. 89-2887 (8th US Circuit Court of Appeals 1990).

8 Appendix A

Distribution of age groups in dataset

Age group	Count	Percentage
18-24	252150	37.43
25-34	158611	23.54
10-17	136474	20.26
over 34	126048	18.71
under 10	368	0.05
Total	673651	100

Distribution of genders in dataset

Gender	Count	Percentage
Male	618198	91.77
Female	55044	8.17
Other	409	0.06
Total	673651	100

Distribution of ethnicities in dataset

Ethnicity	Count	Percentage
White	398090	59.09
Black	174490	25.90
Asian	82014	12.17
Other	18060	2.68
Mixed	997	0.15
Total	673651	100

Stop-and-searches by hour of the day

Hour	Count	Percentage
23	57547	8.54
0	49022	7.28
16	44944	6.67
15	44622	6.62
17	40260	5.98
14	39156	5.81
20	38832	5.76
22	37587	5.58
19	36815	5.46
21	36441	5.41
18	36260	5.38
13	33385	4.96
1	29180	4.33
12	28709	4.26
11	24563	3.65
2	20414	3.03
10	19139	2.84
3	14173	2.10
9	12626	1.87
4	8891	1.32
8	7757	1.15
7	5126	0.76
5	4724	0.70
6	3478	0.52
Total	673651	100

Crosstab of age group percentages per ethnicity

	Age group					
Ethnicity	**Under 10**	**10-17**	**18-24**	**25-34**	**Over 34**	**Total**
Asian	0.02	14.41	48.85	25.11	11.61	100
Black	0.04	24.03	41.79	20.53	13.61	100
Mixed	0.20	25.38	37.91	25.08	11.43	100
Other	0.06	20.54	44.67	22.00	12.74	100
White	0.07	19.79	32.84	24.61	22.70	100

Crosstab of gender percentages per ethnicity

	Gender			
Ethnicity	**Female**	**Male**	**Other**	**Total**
Asian	3.00	96.92	0.08	100
Black	4.70	95.26	0.04	100
Mixed	8.63	91.37	0.00	100
Other	4.48	95.40	0.12	100
White	10.92	89.01	0.06	100

Crosstab of gender percentages per age group

	Gender			
Age group	**Female**	**Male**	**Other**	**Total**
10-17	7.70	92.24	0.06	100
18-24	6.67	93.27	0.06	100
25-34	8.75	91.18	0.06	100
Over 34	10.95	88.99	0.06	100
Under 10	9.78	89.67	0.05	100

9 Appendix B

Feature List 1

age_range_18-24
age_range_25-34
age_range_over 34
age_range_under 10
gender_Male
gender_Other
object_of_search_Article for use in theft
object_of_search_Articles for use in criminal damage
object_of_search_Controlled drugs
object_of_search_Crossbows
object_of_search_Detailed object of search unavailable
object_of_search_Evidence of offences under the Act
object_of_search_Evidence of wildlife offences
object_of_search_Firearms
object_of_search_Fireworks
object_of_search_Game or poaching equipment
object_of_search_Goods on which duty has not been paid etc.
object_of_search_Offensive weapons
object_of_search_Psychoactive substances
object_of_search_Seals or hunting equipment
object_of_search_Stolen goods
officer_defined_ethnicity_Black
officer_defined_ethnicity_Mixed
officer_defined_ethnicity_Other
officer_defined_ethnicity_White
type_Person search
type_Vehicle search
Force_bedfordshire
Force_cambridgeshire
Force_cheshire

Feature List 2
Force_city-of-london
Force_cleveland
Force_cumbria
Force_derbyshire
Force_devon-and-cornwall
Force_dorset
Force_durham
Force_dyfed-powys
Force_essex
Force_gloucestershire
Force_greater-manchester
Force_gwent
Force_hampshire
Force_hertfordshire
Force_humberside
Force_kent
Force_lancashire
Force_leicestershire
Force_lincolnshire
Force_merseyside
Force_metropolitan
Force_norfolk
Force_north-wales
Force_north-yorkshire
Force_northamptonshire
Force_northumbria
Force_nottinghamshire
Force_south-wales
Force_south-yorkshire
Force_staffordshire

Feature List 3
Force_suffolk
Force_surrey
Force_sussex
Force_thames-valley
Force_warwickshire
Force_west-mercia
Force_west-yorkshire
Force_wiltshire
hour_1
hour_10
hour_11
hour_12
hour_13
hour_14
hour_15
hour_16
hour_17
hour_18
hour_19
hour_2
hour_20
hour_21
hour_22
hour_23
hour_3
hour_4
hour_5
hour_6
hour_7
hour_8

Feature List 4
hour_9
month_10
month_11
month_12
month_2
month_3
month_4
month_5
month_6
month_7
month_8
month_9