DEPARTMENT OF STATISTICS 2019

EXPLORING ML FAIRNESS IN IDENTIFYING VICTIMS OF CRIME IN BRITAIN

Candidate Numbers

47516

35661

39657

Submitted for the Master of Science,

London School of Economics, University of London

SEPTEMBER 5, 2020

Acknowledgments

This report would not have been possible without the profound contributions of our collaborators, namely Microsoft and the Thames Valley Violence Reduction Unit - we would like to particularly thank Mr. Matt Bishop and Detective Lewis Prescott-Mayling for their time and guidance throughout our period of conducting this research.

Additionally, this report would not have been possible without the consistent guidance from our LSE Supervisor, i.e. Dr. Kostas Kalogeropoulos who has been supporting us from day one - thank you Dr. Kalogeropoulos. We would also like to extend our sincere and heartfelt thanks to Professor Milan Vojnovic and Ms. Sarah Mcmanus, for their commitment and effort in making sure we have the necessary resources to complete this project. Finally, we would like to thank our families who have supported us throughout the duration of this project.

Contents

List of Figures

List of Tables

Executive summary

The topic of this research is "Exploring ML Fairness in Identifying Victims of Crime in Britain". Our project is focused on coming up with a holistic framework to develop a fair screening test to identify individuals vulnerable to experiencing four different types of crime, namely household crime, personal crime, threats, and domestic & acquaintance violence, in Britain. We structured our problem statement based on the EpiCrim concept, by viewing crime as a social disease and thus, approached the crime victimisation problem by coupling public health methodologies with fair statistical tools. In line with the proactive preventative policing plan announced under The Policing Vision 2025 in the UK, we hope that our findings would be able to complement this vision by providing our targeted end user, i.e. the Thames Valley Violence Reduction Unit (TVVRU) with the tools to identify susceptible individuals from the society before providing them with the necessary support.

Crime victimisation is an important but often neglected research area, as most studies on crime aim to reduce the general crime rate by concentrating on the traits of crime perpetrators. In addition, the authorities themselves approach preventative policing by trying to understand the social and demographic factors that motivate crime perpetration; The current joint research between Microsoft and the TVVRU is also centered on understanding crime perpetrators. While these efforts are undoubtedly crucial to reduce recidivism and crime in general, we feel that alternatively understanding crime from the lens of the victims would complement these current efforts, as crimes (especially the ones we are analysing) typically involve two parties, i.e. victims and perpetrators. Additionally, historical evidence also suggests substantial proneness to crime by the same type of crime (Reiss, 1981). Thus, preventative policing should also include identifying factors that correlate highly with an individual's susceptibility to that type of crime, so that early support can be provided to them, therefore effectively reducing victimisation and revictimisation particularly by the same type of crime.

Given the stated problem, any statistician would naturally engage Machine Learning (ML) methods and tools to uncover the aforementioned factors and build classifiers to segregate potential victims of crime from those who are not. However, as pointed out by Barocas, Hardt and Narayanan (2019), when using ML to model human behaviour and characteristics especially, historical examples of the relevant outcomes will almost always reflect historical prejudices against certain social groups and prevailing cultural stereotypes, thus finding patterns in these data may likely result in the replication of the very same dynamics. Furthermore, in light of the current strong push for fair policing from various movements around the globe (including the UK) such as the 'Black Lives Matter' movement, it is imperative to include fairness into the equation of our

problem.

However, injecting fairness into traditional ML methods is not something trivial - numerous choices and judgments have to be made, starting from choosing between using either observational criteria of fairness or causal reasoning, or a combination of both. In this report, we focus on the former, leading us to another set of choices, i.e. selecting between the Independence/ Separation/ Sufficiency observational criteria of fairness. Upon deciding on the final criterion of fairness, choosing the best technique (between pre-processing, post-processing, and optimisation during training) to achieve the selected fairness criterion is another task requiring thorough deliberation. In this report, we explain in detail each set of decisions made, addressing the objectives of this study step-by-step.

We began developing our framework through a robust variable selection process, reducing the number of independent covariates from as high as 2,962 to only 21, 13, 15, and 10 respectively across our chosen Logistic Regression classifiers for household crime, personal crime, threats, and domestic & acquaintance violence respectively. This way, our end users can utilise our models without needing to conduct an extensive data acquisition process, thus reducing the monetary and time cost of gathering inputs to feed into our models.

Interestingly, we found that there is a regional effect across all four types of crime, suggesting that susceptibility to crime varies by location. We also found higher risk of experiencing the four types of crime, the greater the severity of antisocial behaviour witnessed by an individual. Frequent visits to night clubs/ pubs are associated with greater exposure to experiencing personal crime/ threats, but not necessarily household/ domestic crimes. Additionally, having more cars seems to increase the propensity to experience household crime, but not necessarily other types of crime. Another interesting finding - where age is significant (all crimes except threats), it has a negative association with crime victimisation, suggesting that older people are less at risk. Unsurprisingly, having worse health conditions/ a mental or physical disability/ poor financial condition are all associated with higher risks of experiencing all four types of crime.

Since the main theme of this report is to ensure our final models are fair, we then examined our selected models for unfairness towards any of the subgroups belonging to our two chosen "protected classes"/ "sensitive attributes", i.e. race and gender; We ended up with four subgroups, namely White males, Non-white males, White females, and Non-white females. Our main focus is to check whether or not the False Negative Rates (FNRs) are vastly different across the four subgroups, thus disadvantaging any of the subgroups when it comes to receiving the support they need as individuals who are susceptible to crime. We found that our classifiers for all four types of crime treat the different subgroups discriminatorily, with all of them being most unfair to the

Non-white males, establishing a solid case to use ML Fairness methods in order to eradicate this unfairness within our baseline Logistic Regression models.

Our chosen measure of fairness is the Separation observational criterion, since by design, it ensures all subgroups achieve equal False Positive Rates (FPRs) and FNRs, in line with the purpose of our study. Next, we implemented two techniques to achieve Separation, i.e. post-processing and optimisation during training. Under the former technique, we considered two methods, namely Equalised Odds and Equal Opportunity. Correspondingly, we trained two models for each crime under the latter technique, one with both FPR and FNR constraints, and another one with only FNR constraint, to compare with Equalised Odds and Equal Opportunity respectively. To measure unfairness between the subgroups, we introduced two measures called D_{FNR} and D_{FPR}, which correspond to the largest absolute difference in FNR and FPR between any pair of the four subgroups.

Our results suggest a possible trade-off between accuracy and fairness; Although the unconstrained baseline models registered higher accuracy values compared to the fair models, the latter recorded much lower values of FNRs, $D_{FNR}s$ and $D_{FPR}s$. Our highly imbalanced dataset makes accuracy an unsuitable performance metric, and since our priority is to classify most victims correctly and fairly across the subgroups, the latter three are our chosen performance metrics - thus, the fair models outperformed the baseline models in our case. Although focusing on classifying most victims correctly will be traded off with a higher misclassification rate for the non-victims, in line with our EpiCrim approach to this issue, it is more morally acceptable than to let individuals susceptible to crime be left without the support that they deserve.

Next, we found that between the two fairness methods, although the optimisation during training method gave slightly higher accuracies compared to the post-processing method, the latter outperformed the former in terms of our three chosen performance metrics. Furthermore, the latter technique can be applied after any classifiers without having to retrain them, making the fairness implementation process less invasive, easier to implement, and cost-savvy. Thus, our final chosen models across the four crimes are the fair Logistic Regression models achieved via post-processing techniques. Between the two post-processing techniques, Equalised Odds recorded smaller $D_{FPR}s$ but lower overall accuracies when compared to Equal Opportunity, since the former includes an additional constraint on FPR, but the latter does not. Thus, selecting between one of these two methods would be up to our end users' goals - whether they prefer to sacrifice some accuracy to gain fairness in terms of FPR, or vice versa.

We hope that our results would benefit our targeted end users, i.e. the TVVRU (also the UK police in general) in terms of fairly screening individuals within the UK demographics for susceptibility

to experiencing these crimes, so that they are given the necessary support, effectively reducing victimisation and revictimisation. Upon acquiring necessary data from the targeted societies and inputting them into our models to obtain the probabilities of experiencing the different types of crime, the end users of our post-processed fair models can just follow our guidelines in Section 3.2 of this report to classify those individuals into potential victims of crime or not, based on the probability thresholds obtained from our results (Table 6). Should the end users be interested in the alternative method, i.e. optimisation during training method, they can again refer to Section 3.2 of this report, as well as our file named 'Opt_during_training.ipynb' accessible on our GitHub page.

1 Introduction

1.1 Objectives and Research Questions

Research Topic and Context

Our chosen research topic is "Exploring ML Fairness in Identifying Victims of Crime in Britain". ML Fairness (Machine Learning Fairness) is a hot and relatively new research area in the Machine Learning field that is gaining considerable traction following its high emphasis on ensuring decision-making algorithms lead to fair and reliable decisions without prejudice against any social groups belonging to certain protected classes such as race and gender. Crime victimisation is often a neglected research area as most studies on crime are focused on analysing perpetrators of crime to reduce recidivism and crime rates in general; Analysing this alternative perspective of crime (victimisation) is crucial to detect susceptible individuals from the society for the rightful authority to provide early support to, thus reducing the risk of victimisation and revictimisation.

In this report, we draw upon the theory of Epidemiological Criminology (EpiCrim), i.e. the melding of Public Health with Criminology, to define our problem statement by viewing crime as a social disease. We address this social problem using statistical tools by exploring multiple ML models to classify individuals into those who are susceptible and not susceptible to various types of crime, while ensuring that all social groups belonging to our chosen protected classes are not unfairly disadvantaged by our classifiers. With this framework, we hope to complement our partners, i.e. Microsoft and the Thames Valley Violence Reduction Unit (TVVRU) in their joint research that is focused on crime perpetration, by providing this alternative lens to viewing crime.

Main Objectives

One of the main social objectives of this report is to provide our targeted end user, i.e. the TVVRU with a fair ML algorithm that doubles as a screening test to detect risk factors for susceptibility to crime, and to segregate individuals vulnerable to crime from those who are not. We hope that the findings from this report will enable the police department to utilise their resources more effectively in terms of providing support to the community, especially those who are vulnerable to crime. This is in line with our adoption of the EpiCrim theory, in which one of its main emphasis is the concept of "prevention is better than cure"; By providing support to potential victims of crime at an early stage, the rates of victimisation and revictimisation can be reduced.

Besides, we also aim to build models which are easy and less costly to use in terms of the data that the police department needs to acquire to feed into our algorithms, so that they can efficiently focus their resources on providing support to those who need them. Thus, one of our technical

objectives is to implement the best variable selection technique(s) to efficiently remove variables that are irrelevant, non-significant, and not highly correlated with crime victimisation across the different types of crime being explored.

To address one of our social objectives i.e. not systematically disadvantaging any social groups belonging to our chosen protected classes, our next technical objective is to explore the various ML Fairness techniques on our dataset to build fair classifiers. We aim to achieve one of the three fundamental observational criteria of fairness, namely Independence, Separation, and Sufficiency. Consequently, we aim to explore the three methods associated with achieving any one of the observational criterion, i.e. pre-processing, post-processing, and optimisation during training methods, and then evaluate their performances and suitability in the context of our study.

We also aim to explore in depth the trade-off between fairness and our chosen performance metrics, to compare the performances of models with and without fairness on a real world dataset. Our final technical objective is to be able to advance and extend the implementation of our fair classification techniques to more than one protected class - current fair classification techniques, namely classification without disparate mistreatment (Zafar et al, 2017a) and classification with Equalised Odds and Equal Opportunity (Hardt et al, 2016) only implemented their methods on one protected class.

Research Questions

Below are our research questions, aimed at addressing our main objectives:

1. How does the EpiCrim framework fit the context of our problem?
2. What are the attributes that correlate significantly with an individual's experience of different types of crime?
3. How do we deal with an imbalanced dataset (highly likely to be the case for crime data)?
4. Are our classifiers treating all social groups belonging to multiple protected classes fairly? Are there any specific groups being systematically disadvantaged by our classifiers?
5. If not, what are the necessary and suitable adjustments that can be added to our existing classification methods? What are the strengths and weaknesses of each technique in the context of our problem?
6. How does fairness trade-off with our chosen performance metric(s) across the different fairness methods? Considering all performance measures, what are the final selected models for each type of crime that fit our social and technical objectives the most?

1.2 Summary of Main Results

Based on the data available to us, we built victim classifiers for 4 main types of crime, namely household crime, personal crime, threats, and domestic & acquaintance violence. We began with a thorough variable selection process, reducing the number of attributes for the 4 types of crime from nearly 3,000 to only 21, 13, 15, and 10 respectively. We then observed obvious unfairness in our models across our 4 protected subgroups, namely the White males, Non-white males, White females, and Non-white females, with all of our models (for the 4 crimes) disadvantaging the Non-white males the most. Therefore, we introduced several fairness constraints on our chosen baseline Logistic Regression models via 2 main methods, namely post-processing and optimisation during training. The post-processing method includes two techniques namely Equalised Odds and Equal Opportunity. Results from these two techniques were compared to their corresponding optimised during training models.

We found that all fair models effectively removed unfairness, often at a small cost in terms of accuracy, as expected. However, in our victim identification task, we place a bigger emphasis on achieving the lowest possible false negative rate (FNR) (classifying most victims correctly) than achieving a high overall accuracy (classifying most observations correctly). Thus, after comparing all performance measures and the corresponding trade-offs, we decided that indeed, the fair models outperformed the unfair baseline models in terms of our research objectives.

We then compared the fair Logistic Regression models achieved via post-processing to those achieved via optimisation during training. The latter recorded slightly higher accuracies and was able to eliminate both disparate mistreatment and disparate treatment. However, the post-processing method registered lower values of overall FNR, D_{FNR} and D_{FPR} (D_{FNR} and D_{FPR} refer to the biggest absolute difference in terms of false positive rate (FPR) and FNR between any pair of the 4 protected subgroups), which are our main performance criteria. Moreover, post-processing can be applied after any classifier without modifying it. Therefore, we decided that the fair Logistic Regression models achieved via post-processing is more suitable for our victim identification task.

With respect to the two post-processing techniques, Equalised Odds predictors recorded smaller $D_{FPR}s$ but lower overall accuracies when compared to Equal Opportunity predictors. Therefore, choosing among these two methods would depend on our end users' goals - whether they prefer more fairness with respect to FPR or higher overall accuracy.

1.3 Outline

Our report starts with a literature review in Section 2.1, whereby we explored four main themes, namely 'Epidemiological Criminology', 'Victim Proneness', 'ML Fairness - Motivation', and 'ML Fairness - Implementation'.

Section 2.2 contains information on our chosen dataset, obtained from the Crime Survey for England and Wales (CSEW). After summarising the guidelines and background of CSEW, we explained the process of combining data from different years, selecting relevant variables from more than 2,000 of them, and eventually pooling the levels of some variables. We then included the description of variables for our final dataset.

The next section (2.3) discusses our research problems and chosen methodology in detail, and in the same order as our data analysis process. We began with explanations on the methods used for exploratory data analysis (EDA), the various classification algorithms and variable selection techniques considered, as well as the resampling and threshold-moving techniques used to deal with our imbalanced dataset. After justifying our choice of classifier for each type of crime i.e. Logistic Regression, we introduced the general ML Fairness methodology, followed by the two techniques to include fairness, i.e. post-processing and optimisation during training.

Section 2.4 contains details on our analysis process and results. We first presented results and visualisations from our EDA before discussing the various ML methods that we implemented on our dataset. Then, after deciding on the best classification method, we reported our findings from the variable selection process and from our experiments with resampling and threshold-moving methods to deal with our imbalanced dataset. Next, after evaluating the fairness of our chosen models and discovering that they do treat different sensitive attribute groups discriminatorily, we implemented the two chosen fairness methodologies, namely post-processing and optimisation during training. We then compared the performance of our unconstrained (unfair) baseline models for each type of crime to their corresponding fair models generated via both methods. Finally, we compared and contrasted the pros and cons of both fairness techniques implemented.

The final part (Section 3) of our report is our conclusion section. It starts with a summary of our main findings with respect to the objectives of this report. We then included guidelines to access and utilise the framework implemented in the previous section. Finally, we proposed several potential avenues for future research.

2 Main Report

2.1 Literature Review

In this section, we review 23 research articles by segregating them into 4 main themes, beginning with the main social approach to the problem (EpiCrim), followed by our more specific topic, i.e. crime victimisation. Next, we dive deeper into the ethical challenges of implementing traditional ML methods, i.e. unfairness to different sensitive attribute groups, and end our review with the various methods to implement ML Fairness.

2.1.1 Epidemiological Criminology

Both Public Health and Criminal Justice disciplines deal with marginalised populations, i.e. people exposed to risks of drug use, incarceration, health problems, and other difficulties. The distinctions between these two fields are somehow blurred, as they increasingly overlap. However, Lanier and Akers (2009) claimed that explicit theoretical and methodological linkages between the two disciplines are still uncommon. Thus, they proposed a framework called "Epidemiological Criminology" or in short, EpiCrim that merges Epidemiology (methodology and factors that affect disease spread) with Criminal Justice theory, methods and practices. Via this concept, Lanier (2010) suggested that apart from its legal definition, crime can be alternatively viewed as a disease, and we could reduce criminal behaviour to a set of measurable variables.

EpiCrim argues that the Criminal Justice system should design policies that would reduce rather than exacerbate social harm; EpiCrim aims to be preventative and protective, by identifying, applying, and evaluating prioritised interventions, resulting in better outcomes for the community (Mcmanus, 2018). Criminal Justice practitioners should leverage on epidemiological methods to identify factors leading to crime exposure before identifying individuals susceptible to this disease called 'crime'. According to Lanier and Akers (2009), these susceptible individuals can be both perpetrators and victims of crime, and the rightful authorities should step in by providing support to them. This is deemed to be a more effective way to break the chain of 'infection', control the spread of crime within society, and reduce recidivism as well as revictimisation rates.

Christmas and Srivastava (2019) pointed out a public health approach to policing that is already in use in the UK, called the SARA problem-solving model, whereby 'S' stands for scanning potential crime and problem, 'A' stands for analysing underlying causes of the problem, 'R' stands for response to address the causes of the problem, and 'A' stands for assessment of the success of the response. This model uses a social approach to understand the drivers of crime within populations.

The UK currently has a ten-year plan for policing, known as The Policing Vision 2025 (Christmas and Srivastava, 2019), that ensures policing is increasingly focused on proactive preventative activity. In line with the EpiCrim approach, the police are using an improved understanding of vulnerability, both in physical and virtual locations, to improve and differentiate service and protection. Besides, the UK police are also in the process of improving data sharing and integration to establish joint technological solutions to enable the transfer of learning between agencies, thus utilising digital transformation to better curb crime perpetration and victimisation. Linking back to the EpiCrim theory, this would enable them to track the spread of specific types of crime similar to tracking the spread of diseases, thus curbing the spread of this societal disease called crime.

2.1.2 Victim Proneness

The core of utilising the EpiCrim approach to reduce crime victimisation is to first be able to identify and segregate those who are susceptible to crime from those who are not. There is a vast literature on victim proneness such as the ones written by Bellis et al (2015) and Marmot (2010) suggesting that the circumstances of a person's life have a cumulative impact on his/ her life chances and life expectancy; Events such as adverse childhood experiences, emotional, sexual and substance abuse, domestic violence, and mental illness are likely to play a larger role in a person's susceptibility to crime. Reiss (1981) suggested that there is substantial proneness to crime by the same type of crime, and that the probability of consecutive victimisation by the same type of crime varies directly with the probability of occurrence of that type of crime among all victims. Thus, to prevent revictimisation particularly by the same type of crime, it is imperative to identify factors that correlate well with an individual's susceptibility to that type of crime.

This is where Machine Learning steps in - given the right set of data, ML can provide the tools to uncover factors relevant to crime victimisation among individuals. Barocas, Hardt and Narayanan (2019) argued that ML promises to bring greater discipline to decision-making because rather than starting with some intuition about the relationship between certain factors and an outcome of interest, ML lets us defer the question of relevance to the data themselves: which factors - among all that we have observed - bear a statistical relationship to the outcome. Thus, ML serves as a technical platform that provides statistical and objective solutions to social problems, such as crime victimisation.

2.1.3 ML Fairness - Motivation

Previous section highlights the importance and benefits of ML methods to uncover factors relevant to decision-making that humans might overlook; However, Barocas, Hardt and Narayanan

(2019) pointed out that although ML is "evidence-based", it is by no means an assurance that it will lead to accurate, reliable, or most importantly, fair decisions. When using ML to model human behaviour and characteristics especially, historical examples of the relevant outcomes will almost always reflect historical prejudices against certain social groups, existing demographic inequalities, and prevailing cultural stereotypes - finding patterns in these data may likely result in the replication of the very same dynamics (Barocas, Hardt and Narayanan, 2019; Binns, 2018). As ML begins to play a monumental role in consequential decision-making across various spheres such as predicting the risk of recidivism (Criminal Justice), filtering resumes of job applicants (Commercial), and even in granting loan approvals (Banking), there is a growing need to come up with more reliable and fair ML algorithms.

Barocas, Hardt and Narayanan (2019) argued that ML methods would generally work less well (higher error rates) for members of minority groups since there would typically be fewer data points for them (sample size disparity). A utility-maximising (utility here usually corresponds to accuracy) decision maker that does not consider the unequal distribution of errors between demographic groups can be deemed as an unfair classifier, despite achieving high utility overall; When decision-making systems are discriminatory, they create "allocative harms", i.e. withholding certain groups an opportunity/ resource, and when they perpetuate stereotypes, they create "representational harms", i.e. reinforcing the subordination of some groups along the lines of identity - race, gender, etc.

Barocas, Hardt and Narayanan (2019) also emphasised the possibility of introducing bias in predictive systems via feedback loops. This suggestion is well supported by Lum and Isaac (2016) who analysed PredPol, a predictive policing algorithm. They found that in Oakland, Black people would be targeted for predictive policing of drug crimes at roughly twice the rate of Whites, although both groups have roughly equal rates of drug use. The simulation performed by them highlighted that the initial bias would be amplified by a feedback loop, with policing increasingly concentrated on targeted areas, even though the PredPol algorithm does not explicitly take demographics into account.

The concept of ML Fairness is largely centered on making decisions that are fair to different groups based on certain "protected classes" or sometimes denoted as "sensitive attributes". Two of the most common "protected classes" (also legally recognised) are race and gender. In the Criminal Justice system, one of the most popular research on racial discrimination is the one carried out by the ProPublica team on the COMPAS algorithm that is largely used in the US to predict potential recidivism risk. Angwin, Larson, Mattu and Kirchner (2016) proved that the algorithm was likely to falsely flag Black defendants as future criminals, wrongly labeling them this way at al-

most twice the rate as White defendants. White defendants on the other hand, were mislabeled as low risk more often than Black defendants. These findings corroborate the existence of unintended racial discrimination in the Criminal Justice system, similar to the findings by Lum and Isaac (2016).

When it comes to gender discrimination in crime victimisation particularly, there are multiple viewpoints of the issue. Fox, Nobles and Piquero (2009) suggested a peculiar paradox; Although males are more likely than females to be victims of crime, females are substantially more fearful of crime than males. This paradox has been mentioned across several earlier studies done on the same topic (Fisher, 1995; Gibson et al, 2002; Jennings et al, 2007). However, Mallicoat (2018) suggested that although this fact is true, women are more likely to experience certain kinds of violent crimes, such as sexual assaults and stalking. Hedderman and Hough (1994) pointed out that in the UK, while men and women are treated differently by the Criminal Justice system, these differences largely favor women. Thus, depending on the type of crime, location, and the Criminal Justice system in place, there are various potential gender biases, either towards males or females in different ways. These findings emphasise the need for a fair ML algorithm with respect to both race and gender, especially when identifying potential victims of crime so that no individual is disadvantaged on the basis of their race or gender.

2.1.4 ML Fairness - Implementation

In response to the growing need for fair algorithms, there is a branch under ML that focuses on this, with 3 main methods introduced to implement fairness, i.e. via pre-processing the training data, post-processing the outcome, or optimisation during training process (Barocas and Hardt, 2017; Barocas, Hardt and Narayanan, 2019). These methods are aimed at achieving certain well-defined measurements of fairness; From the perspective of Anti-discrimination laws, fairness of a decision-making system is measured using the concepts of disparate treatment and disparate impact (Zafar et al, 2015; Binns, 2018). Disparate treatment is a form of intentional discrimination, whereby the decision-making process is (partly) based on the individuals' sensitive attributes, whereas disparate impact is often referred to as unintentional discrimination, whereby the outcomes of a decision-making process disproportionately disadvantage (or, benefit) a group of people with certain sensitive attribute value (e.g. females).

By excluding sensitive attribute information in the decision-making process, it is possible to avoid disparate treatment, but not necessarily disparate impact, as historical biases towards groups belonging to certain protected classes may likely be picked up in future predictions. Zafar et al (2015) designed convex margin-based classifiers that control for both disparate treatment and disparate impact simultaneously, while fulfilling the 'business necessity' clause of disparate impact that re-

quires meeting performance constraints by allowing the least possible degree of disparate impact (Barocas and Selbst, 2016). Although these classifiers by Zafar et al (2015) are able to achieve fairness in more than one sensitive attribute (which is desirable), this measure is only suitable to be used with datasets for which the "ground-truth" decisions are not available, i.e. one cannot tell whether a historical decision was right or wrong by referring to the labels for each observation.

In some decision-making processes, fairness notions such as equality in treatment or impact might be too stringent, precluding more accurate decisions, which may also be desired by every sensitive attribute group. To tackle this problem, Zafar et al (2017b) proposed preference-based notions of fairness in classification - instead of focusing on achieving parity across all groups, it allows any group of individuals to collectively select their own preferred treatment or outcome. This method was shown to always outperform parity-based notions of fairness in terms of accuracy. However, unless we have access to the preference of each sensitive attribute group, this notion of fairness is not practical.

Since we have access to information on whether or not a person is a victim of crime during our training process, disproportionality in outcomes can be justified and explained by the means of the ground-truth, thus the notions of fairness suggested by Zafar et al (2015) are not suitable. Additionally, we do not have information on the treatment preference for the different sensitive attribute groups, making the preference-based notion of fairness suggested by Zafar et al (2017b) to also be non-viable. An alternative notion of (un)fairness that is not limited by these issues is the one proposed by Zafar et al (2017a), called disparate mistreatment.

Disparate mistreatment occurs when the misclassification rates across different sensitive attribute groups are not equal; Zafar et al (2017a) proposed a method that can effectively eliminate both disparate treatment and disparate mistreatment simultaneously from decision boundary-based classifiers, by converting either one of the constraints on overall misclassification rate (OMR), FPR, or FNR (depending on the needs of the user) into a Disciplined Convex-Concave Program (DCCP), which can then be solved using well-known heuristics. However, unlike convex optimisation, this method does not provide any guarantees on the global optimality of the solution, and because the analytical covariance used in this method is approximated through Monte Carlo covariance on the training set, it might be inaccurate for smaller datasets, but work well on larger datasets.

Disparate mistreatment is an equivalent notion of Separation (requires all groups to experience the same FPR and FNR); Hardt et al (2016) introduced two other equivalent and relaxed notions of Separation i.e. Equalised Odds and Equal Opportunity respectively, in which the former requires having equal true positive rate (TPR) and FPR across all groups, while the latter only requires hav-

ing equal TPR across all groups. These two notions of fairness are achieved by post-processing the outcomes of existing unfair classifiers without requiring retraining them, unlike the method used to achieve disparate mistreatment (Zafar et al, 2017a). Zafar et al (2017a) suggested that results from these two methods are similar, but the optimisation during training method (Zafar et al, 2017a) is likely to outperform the post-processing method marginally (Hardt et al, 2016) for larger datasets, and vice versa. The former method also uses much more resources as it requires retraining classifiers, whereas the latter does not.

Literature Review - Conclusion

We first explored the topic of EpiCrim, establishing the importance of public health approach to crime, and how the UK police are already adopting proactive preventative policing. Next, we reviewed factors leading to victimisation and the importance of implementing ML methods to identify these factors and segregate individuals based on their proneness to crime. We then analysed studies that emphasised the main ethical challenge of traditional ML classifiers, i.e. embedded bias towards certain sensitive attribute groups, and reviewed multiple methods to implement fairness; The comparisons suggest that the Separation notion of fairness suits our dataset the best - this notion is achievable via one of two methods, namely post-processing and optimisation during training. Thus, in this report, we aim to implement these methods on our crime victimisation dataset, by extending these methods to deal with more than one sensitive attribute (sex and gender).

2.2 Data

The data used in this research is from the Crime Survey for England and Wales (CSEW), particularly data collected from adults aged between 16 to 74. The CSEW is a face-to-face victimisation survey in which people resident in households in England and Wales were asked about their experiences of a range of crimes in the 12 months prior to the interview, as well as perceptions of crime and anti-social behaviour. Most importantly, this survey collects information on the demographic characteristics of the respondents and households. Towards the end, adult respondents were asked to self-complete a series of sensitive questions such as their experience of abuse during childhood and illicit drug use. Although this module contains data that might be beneficial for our study, we were unable to obtain this part of data, as only researchers who have undergone training with UK Data Service can be granted access to this part of the dataset.

The CSEW has been successful in maintaining high response rates (70% - 75% over the past 10 years), and is designed to reflect the general population profile. With the exception of the City of London Police Force Area (PFA), the sample is designed to yield a minimum of 650 interviews with adults in each one of the 42 territorial PFAs. This questionnaire has a core set of modules asked of the whole sample, a set of modules asked only of different sub-samples, and self-completion modules asked of all respondents (as mentioned above). We decided to not include modules that were asked only to sub-samples, as this would discard $^{7}/_{8}$ of the sample and the modules are also very niched and irrelevant to our topic.

Interviews are done on a rolling basis over the course of a year, and the data captured in the 2018/2019 survey covers the period from April 2018 to March 2019, likewise for the earlier years. From 2017 onwards, a new approach was used to calculate the number of incidents for each major crime type and so, any datasets after 2017 are not comparable with previous years. Therefore, only datasets from 2017/2018 and 2018/2019 are included in our analysis and no time series analysis with incident variables was engaged. Within the 2017/2018 and 2018/2019 CSEW datasets, there are 34,715 and 34,163 observations, as well as 2,962 and 2,215 attributes respectively. Considering the number of variables in our dataset, having all features will most likely overfit and incur an expensive computational cost. Thus, duplicated questions and questions unrelated to our study such as those concerning the experience of other members within the households are removed, narrowing the number of attributes to 73.

We decided to structure our problem as 4 separate binary classification problems - classifying potential victims and non-victims for 4 different groups of crime, namely household crime, personal crime, threats, and domestic & acquaintance violence. This is because the factors that correlate with an individual being a victim of one type of crime may vary from another type of

crime. Consequently, the target variables are “totalhh_nocap”, “totperls_nocap”, “threats_nocap”, and “dom_acq_nocap” (Table 1) which correspond to whether or not the respondent has experienced household crime, personal crime, threats, or domestic & acquaintance violence respectively. Household crimes include all vehicle and property-related crimes, and the other crimes are self-explanatory.

After deciding on the target variables, we inspected other independent shortlisted attributes. We noticed several binary variables that could be merged into single multilevel categorical variables. For example, under the antisocial module, the interviewer recorded Yes/No answers to the respondent’s experience of several types of antisocial behaviour such as drinking, harassment, etc. separately and assigned a binary variable to each type of the antisocial behaviour. Thus, we merged these binary variables into a single multilevel categorical variable called “antisocial” (Table 1), where we pooled mild antisocial behaviours such as loud music and litter into one level, and more moderate and severe antisocial behaviours into two other levels. This way, we managed to reduce the number of levels for this variable from 18 to 4. We used similar practice to come up with the variables “mobproc”, “homesecurity” and “pubclub”.

Variable Name	Interpretation	Levels/ Values
sex	Sex of respondent	**1 - Male**; 2 - Female
age	Age of respondent	16 - 80 (80 corresponds to >=80)
nchil	No. of children under 16 in household	0 - 7
yrsarea	How long lived in this area	1 - < 12 months; 2 - >=12 months - < 2 years; 3 - >= 2 years - < 3 years; 4 - >= 3 years - < 5 years; 5 - >=5 years - < 10 years; 6 - >= 10 years - < 20 years; **7 - >= 20 years**
genhealt	General health	1 - Very good; 2 - Good; 3 - Fair; 4 - Bad; 5 - Very bad
onsdisab	Physical or mental health conditions	1 - Yes; **2 - No**; 3 - Refused/ Don’t know
relig3	Religious group	**1 - No religion**; 2 - Christian; 3 - Buddhist;4 - Hindu; 5 - Jewish; 6 - Muslim; 7 - Sikh; 8 - Any other religion; 9 - Refused/ Don’t know
educat2	Highest qualification	1 - Higher degree; 2 - First degree; 3 - Diplomas; 4 - A/AS levels; 5 - Trade Apprenticeships; 6 - O Level/ GCSE grades A-C; 7 - O Level/ GCSE grades D-G; 8 - Other qualifications; **9 - No qualification** 10 - Refused/ Don’t know
managhh2	Ability to find £100 to meet an unexpected expense	1 - Impossible to find; 2 - A bit of a problem; **3 - No problem**; 4 - Refused/ Don’t know
rural3	Rural or Urban Area	1 - Rural; **2 - Urban**

gor	Region	1 - North East; 2 - North West; 3 - Yorkshire and The Humber; 4 - East Midlands; 5 - West Midlands; 6 - East of England; **7 - London**; 8 - South East; 9 - South West; 10 - Wales
eincdc15	Index of Deprivation: Income	1 - 10 with 1 corresponding to 10% most deprived and 10 corresponding to 10% least deprived
ecridc15	Index of Deprivation: Crime	
eenvdc15	Index of Deprivation: Environment	
acctyp	Accommodation type	**1 - Detached house**; 2 - Semi-detached house; 3 - Mid-terrace; 4 - End of terrace; 5 - Maisonette; 6 - Flat - purpose built; 7 - Flat - converted; 8 - Rooms, bedsitter; 9 - Caravan/mobile home/houseboat
rnssec5	Socio-economic classification	**1 - Higher managerial, administrative & professional**; 2 - Intermediate occupations; 3 - Small employers and own account workers; 4 - Lower supervisory and technical; 5 - Semi-routine and routine; 7 - Not labelled 6 - Never worked and long-term unemployed;
nsethgrp	Ethnic group	**1 - White**; 2 - Mixed/multiple ethnic groups; 3 - Asian/Asian British; 4 - Black/African/Caribbean/Black British; 5 - Other ethnic group; 6 - Not labelled
livharm1a	Marital status	**1 - Married/civil partnered**; 2 - Cohabiting; 3 - Single; 4 - Separated; 6 - Widowed; 7 - Not labelled; 5 - Divorced/Legally dissolved partnership
tenharm	Tenure type	**1 - Owners**; 2 - Social rented sector; 3 - Private rented sector
rlstweek	Economic status in last week	**1 - Paid work**; 2 - Government training scheme; 3 - Away from job/waiting to take job up; 4 - Unpaid work; 5 - Looking for work; 6 - Student; 7 - Looking after family/home; 8 - Temporarily sick; 9 - Long-term sick/ill; 10 - Retired; 11 - Other
bikowner	Bike owner	**0 - Not a bike owner**; 1 - Bike owner
numcar3	Number of cars owned	**0 - None**; 1 - One; 2 - Two; 3 - Three or more
antisocial	Antisocial behaviour experienced or witnessed	**1 - None observed**; 2 - Observed mild antisocial behaviour; 3 - Observed moderate antisocial behaviour; 4 - Observed severe antisocial behaviour
mobproc	Mobile phone protection	**1 - Has mobile phone protection**; 2 - Has mobile phone but no protection; 3 - No mobile phone in the first place
nationality	Nationality	**0 - UK national**; 1 - Europe excluding UK; 2 - Africa; 3 - The Americas and the Caribbean; 4 - Middle East and Asia; 5 - Antartica and Oceania; 6 - Other categories

homesecurity	Visible home security	**1 - None observed**; 2 - Basic security observed; 3 - Moderate security observed; 4 - Other form of security observed
pubclub	Frequency of visit to pub / nightclub in the last month	**1 - No visit**; 2 - Less than once a week; 3 - Once to twice a week; 4 -About 3 times a week; 5 - Almost every day
totalhh_nocap	Experienced any household offence	0 - Not victim; 1 - Victim
totperls_nocap	Experienced any personal offence	
threats_nocap	Experienced any threats	
dom_acq_nocap	Experienced any domestic & acquaintance violence	

Table 1: Description of significant variables; emboldened levels are baseline levels, attributes colored in orange are our target variables, and variables without any emboldened levels are continuous variables

Variable Name	**Interpretation**
nadults	Number of adults in household
dftdrive	Person has driven a motor vehicle within a year or not
weekday	Hours spent away from home during the day
unoccl	Period home is left unoccupied on an average weekday
mobwh	Whether owns a mobile phone
eempdc15	Index of Deprivation: Employment
eedudc15	Index of Deprivation: Education
eheadc15	Index of Deprivation: Health and Disability
ehoudc15	Index of Deprivation: Housing and Services
rubbcomm	How common is litter or rubbish in immediate area
vandcomm	How common is vandalism or graffiti in immediate area
poorhou	How common are homes in poor condition/run down
vehowner	Vehicle ownership
reverw	Whether respondent ever had a paid job

Table 2: Description of Non-significant Variables

2.3 Research Problem and Methodology

2.3.1 Exploratory Data Analysis (EDA)

Before beginning the process of building models, we first performed EDA. EDA is required to develop an understanding of the data, identify potential correlations between the variables (to avoid potential confounding problems), as well as explore the distributions of the variables (especially the target variables). Performing EDA also enables us to spot patterns and trends in data through various plots.

Dealing with missing values

One of the major drawbacks of using survey data is that there will be instances of non-response bias. Thus, we first looked into missing values, as well as values that are coded as "refused to answer" or "don't know". There are two methods we considered to deal with the issue of non-response - imputation and deletion. Our decision on whether or not to delete rows with missing values was based on whether the values were missing at random or not, and the percentage of observations that was missing - we set our threshold to 5%, i.e. if the total number of observations with missing values are less than 5% and if it was safe to assume that they were missing at random, we would prefer deletion over imputation.

We conducted our check by each column (variable) and analysed the reason behind the missing values before deciding on our course of action. To illustrate, for most of the variables, there were very few missing entries (at most about 0.6% and we could not establish a systematic reason or pattern behind the missing values, suggesting that they were missing at random). Thus, for these observations, we decided to remove them via deletion. However, for some of the sensitive variables such as religion/ ethnicity/ mental state to name a few, we performed imputation by creating another category for the missing values and used them as a different level. Eventually, we were left with 67,178 observations from the original 68,878 observations - that totals to only about 2.5% of observations that were dropped, which is way below our threshold of 5%.

Visualising distributions and correlation analysis

Next, we explored the distributions of the variables. We were particularly interested to look at the spread of observations across our response variables as well as across our two sensitive attributes (for Fairness implementation part in Sections 2.3.5 - 2.3.7), namely race and gender. This is important as for crime datasets, we would typically have datasets that are skewed - there would normally be more observations for non-victims than victims of crime, resulting in an imbalanced dataset.

Next, we performed correlation analysis for an initial look at variables that may be correlated with

the response variable, as well as to identify potential correlations between the independent variables to avoid the confounding problem later at the modelling stage. For correlations between continuous covariates, we used a correlation plot, whereas for correlations between continuous and categorical covariates, we used boxplots. To investigate potential correlations between categorical variables, we used the Goodman Kruskal association measure via a package in R.

2.3.2 Classification Problem - ML Methods Explored and Chosen

We began our binary classification task by exploring several ML classification techniques, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), and Boosting. From our initial exploration, we found that Logistic Regression and Random Forest always outperformed the other two methods with respect to our dataset. Therefore, we will focus on these two methods for the rest of the report.

To understand Logistic Regression, we first look at the equation of Linear Regression with p covariates:

$$Y = \beta_0 + \sum_{j=1}^{p} (\beta_j x_j) + \varepsilon$$

Then our predictor will be:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} (\hat{\beta}_j x_j)$$

We estimate coefficients $\beta_0, \beta_1, ..., \beta_p$ by minimising the RSS, where:

$$RSS = \sum_{i=1}^{n} (Y - \hat{Y}_i), \quad \text{giving} \quad \hat{\beta}_j = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}.$$

Logistic Regression is an extension of this; It can be used to predict the probabilities of binary dependent variables $y_i \in \{0, 1\}$, given a set of independent variables $\{X_j\}_{j=1}^{p}$. In Logistic Regression, log-odds are assumed to be a linear combination of the covariates, i.e.:

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

The estimates of the coefficients are chosen to maximise the likelihood function:

$$L = \prod_{i=1}^{N} Pr(y_i = 1|x_i)^{y_i} \left(1 - Pr(y_i = 1|x_i)\right)^{1-y_i}$$

where $Pr(y_i = 1|x_i)$ computes a probability that each sample x_i belongs to class 1.

Next, we introduce the theory behind Random Forest. To understand how it works, we need to understand how Bagging works, as Random Forest is a special case of Bagging; Bagging is an ensem-

ble of decision trees, usually combining a large number of trees to improve prediction accuracy, however at the cost of interpretation. Bagging is formulated as follows:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_{*b}(x)$$

Where $\hat{f}_{*b}(x)$ is the predictor for an individual tree.

The base idea behind Bagging is that we generate several bootstrapped training datasets before training decision trees using them. Then, taking the average over all of the tree predictors, we obtain our bagging predictor with reduced variance. Random Forest works using the same concept, except that it builds trees using subsets of predictors (usually $\sqrt{predictors}$) rather than all of them. Therefore, the trees are more uncorrelated with each other, resulting in better predictive performance due to better variance-bias trade-offs. The Random Forest model usually performs well in terms of accuracy, while achieving lower variance. Another advantage of Random Forest is its robustness to outliers, and so, is comparatively less impacted by noise. However, it is computationally expensive to perform this method.

Comparing both methods, Logistic Regression seems to have more to offer with respect to the context of our study. Firstly, it outputs probabilities, giving the likeliness of a person being a victim of crime rather than merely classifying that person as a victim or non-victim like in Random Forest. It also offers better interpretability, since it outputs coefficients that can be directly interpreted with respect to the outcome variable. Finally, Logistic Regression is also less prone to overfitting; Nonetheless, in high dimensional datasets, it can still overfit, suggesting the need for a thorough variable selection process.

2.3.3 Variable Selection

One of our main objectives is to build a parsimonious model with good interpretability, so that our target user of the models, i.e. the TVVRU can utilise the models while minimising the cost incurred. Removing non-significant and irrelevant variables is crucial for this purpose; To use our models, our user would need to acquire information on all variables used to build them - thus, larger models would incur larger cost since more data would need to be collected. Furthermore, it helps in building more parsimonious and interpretable models while reducing the risk of overfitting.

In Section 2.3.2, we mentioned that our chosen model is Logistic Regression, since it outperformed all other models explored. One of the main advantages of Logistic Regression is that we can perform variable selection simultaneously using a range of techniques. We considered 5 techniques altogether, i.e. Best Subset Selection, Stepwise Selection (Backward and Forward), LASSO, and

Elastic Net, and eventually decided that LASSO worked best for our dataset. We could not utilise the Best Subset Selection method as our total number of variables is too large (> 40) and so, it is not computationally efficient to use this method that would need to fit 2^p models, where p corresponds to the total number of variables. For the same reason, both Backward and Forward Stepwise Selection methods are also inefficient.

Finally, we considered LASSO and Elastic Net. LASSO (Tibshirani, 1996) estimates coefficients by minimising $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$. Elastic Net (Zou and Hastie, 2005) is an extension of LASSO that includes the Ridge Regression penalty into the equation above, estimating coefficients by minimising $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2$. We tuned the parameters λ (for LASSO), λ_1 and λ_2 (for Elastic Net) and applied these parameter values to run our penalised Logistic Regression models. Comparing both methods, we ended up with more parsimonious models with lower AIC (Akaike Information Criteria) and BIC (Bayesian Information Criterion) values using the LASSO method, thus in line with our aim mentioned earlier, we chose the LASSO method for our variable selection problem.

2.3.4 Dealing with Imbalanced Data - Selecting Suitable Performance Metrics, Resampling Techniques and Threshold-Moving

After obtaining the Logistic Regression models post variable selection process, we carried out predictions on the test data. Since our data is highly imbalanced (Figure 1) with many more non-victims (negative class, y=0) compared to victims (positive class, y=1), having the probability threshold set to 0.5 resulted in all 4 classifiers classifying almost all of the negative (positive) observations correctly (wrongly). Thus, although we recorded accuracies in the range of over 90%, it is not a good measure for an imbalanced dataset.

Instead, we chose to look at the confusion matrix for each of the classifiers (with format as displayed in Table 3) and based our performance assessment on achieving high TPR (sensitivity) while minimising the FPR (1 - specificity) simultaneously. This is aligned with the aim of our study, i.e. creating a screening test to identify individuals susceptible to being victimised rather than a diagnostic test. Since there is an inherent trade-off between sensitivity and specificity, we would expect to achieve this at some expense of specificity, i.e. we would also be classifying some non-victims wrongly as victims, given the imbalanced data distribution.

To select the best trade-off between these two, we carried out a method called "threshold-moving", which is highly effective at dealing with imbalanced data, without needing much statistical intervention. Typically, probability thresholds of 0.5 do not work well with imbalanced data. Thus, there is a need to move the probability thresholds to achieve the best trade-offs between sensitivity and specificity for each of the four models. This can be done manually or automatically in R,

depending on the cost functions selected.

We unfortunately do not have access to historical or well-defined cost of wrongly classifying victims as non-victims and vice versa, as there are no centralised guidelines on the ideal TPR / TNR for detecting potential crime victims. Since we are using the EpiCrim framework whereby we view crime as a social disease, we referred to the guidelines used in clinical practice to decide on the acceptable FNR for our models. Although the acceptable FNRs vary by disease, several journal articles (Kim et al., 1977; Badrick and Turner, 2015) suggest that in clinical practice, FNRs are still generally between 20% - 30% in disease detection, and so we decided that as long as our TPRs (sensitivities) are above 70%, we should be doing relatively well.

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = 0$	
True Label	$y = 1$	True Positive (TP)	False Negative (FN)	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate (FNR)
	$y = 0$	False Positive (FP)	True Negative (TN)	$P(\hat{y} \neq y \mid y = 0)$ False Positive Rate (FNR)
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = 0)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Table 3: Confusion Matrix with Different Types of Classification Errors

Besides threshold-moving, we also considered two resampling techniques, namely Upsampling and Downsampling to deal with the data imbalance. Downsampling works by reducing the number of observations from the majority class randomly to match the number of observations in the minority class, to balance the dataset. This method is best used when the dataset is very large, to improve run time and storage issues. However, since our dataset is not that large (especially the minority class), this method would reduce our observations significantly, causing a serious loss in data. Thus, we only explored the Upsampling technique on our data, which works by replicating random observations from the minority class to match the number of observations in the majority class, again to balance the data.

2.3.5 ML Fairness - Key Concepts and Overview

After selecting the final classifiers for all four types of crime, we would be reaching the end of a typical ML workflow for designing models. However, with social data especially, there is the problem of unjustified differentiation between certain social groups with different sensitive attributes

such as race and gender - ML models are either intentionally or unintentionally discriminating against certain social groups. One would question, isn't discrimination the very point of Machine Learning?

ML Fairness practitioners are increasingly bringing up the issue of practical irrelevance, i.e. discrimination should not be based on the personal characteristics of a person, such as race and gender. There is also the issue of moral irrelevance - even though we might be able to show that the sensitive attributes have some statistical relevance to the prediction problem, it is not sound to consider discriminating individuals on the basis of their sensitive attributes as a matter of moral judgment.

Most of the existing notions of fairness are inspired by anti-discrimination laws. Those laws prohibit unfair treatment towards people from protected social groups (e.g. females, black people). A large number of recent studies have proposed mechanisms to achieve fairness in decision-making processes. Disparate treatment and disparate impact are the two most popular (un)fairness notions being used (Barocas and Hardt, 2017; Zafar et al, 2015), and they serve as the basis for more recent notions of (un)fairness, such as disparate mistreatment (Zafar et al, 2017a).

Disparate treatment is intentional discrimination that occurs when sensitive attributes are involved in the training process. In our case, disparate treatment arises when a female is classified as a person who is vulnerable to experiencing crime, whereas a male with the same values of non-sensitive attributes is not. To ensure fairness in treatment, two people with otherwise similar values for non-protected characteristics should not be treated differently solely because they have different values in socially salient groups such as race and gender (Barocas and Selbst, 2016). Thus, disparate treatment is relatively easy to notice and avoid by making sure none of the sensitive attribute information is used during decision-making.

Disparate impact on the other hand, may still exist even though sensitive attributes are not included in the training system. This is because there might be other features that are proxies of them - for example region may correlate with race, and this unintentional bias is likely to be picked up in the training process. Disparate impact happens when the outcomes disproportionately benefit (hurt) people with certain sensitive attribute values (Barocas and Selbst, 2016). In our victim detection task, the decision-making system is unfair if the proportions of males and females being classified as victims of crime are different.

Thus, achieving fairness is not an easy feat - it is not as simple as excluding sensitive attributes from the training process. Therefore, ML Fairness practitioners have come up with three fundamental criteria to establish fairness, namely Independence, Separation, and Sufficiency (Barocas

and Hardt, 2017; Barocas, Hardt, and Narayanan, 2019). Consider the setup below:

- X features of an individual (region, number of cars owned, etc.)
- A sensitive attribute (gender/ race)
- $C = c(X, A)$ predictor (susceptible to crime or not)
- Y target variable (victim of crime or not)

Independence is when C is independent of A, i.e. for all sensitive groups a, b and all values c, $P_a\{C = c\} = P_b\{C = c\}$. However, Independence ignores the possible correlation between Y and A - in particular, it rules out the perfect predictor $C = Y$, and it also permits laziness, i.e. it is possible to classify the actual victims from one group and random people in the other group, just to satisfy the same classification rate. Thus, researchers came up with a second criterion, i.e. Separation whereby C is independent of A conditional on Y; For all sensitive groups a, b and all values c and y, $P_a\{C = c|Y = y\} = P_b\{C = c|Y = y\}$. Under this criterion, the predictor C is allowed to be equal to the target variable Y, which means that Separation allows correlations between A and Y and between R and A, as long as it is intrinsic in the target variable. Thus, this formally resolves the optimality compatibility issue present under the Independence criterion. Besides, it penalises laziness by ensuring that all groups have the same TPR and FPR, creating an incentive to reduce errors uniformly across all groups.

The third and final criterion is Sufficiency, whereby Y is independent of A conditional on C, i.e. C is sufficient for predicting Y - we do not need to also look at A, because the sensitive attribute A is irrelevant if we have C, and C is sufficient. This is desirable only in certain cases, especially for legal reasons - for example when we are given a credit score and it is sufficient for race, then we do not need to actively look at race to make decisions since it is already subsumed by the score. However, there is a caveat - any two of the three criteria are mutually exclusive except in degenerate cases, which means that trade-offs are necessary and we can only satisfy one criterion at a time. After careful deliberation, we decided that the most suitable fairness criterion with respect to the aim of our study and the dataset that we have is the Separation criterion.

To achieve Separation, there are three main methods (Barocas and Hardt, 2017; Barocas, Hardt, and Narayanan, 2019) namely pre-processing, post-processing, and by adding optimisation constraints during training. In this report, we consider the latter two, as the pre-processing method requires collecting more data to improve fairness, and we currently do not have the capacity to acquire more data, as explained in Section 2.2. The two other methods will be explained in detail in Sections 2.3.6 and 2.3.7 respectively.

2.3.6 ML Fairness via Post-processing – Equalised Odds, Equal Opportunity

Hardt et al (2016) proposed two techniques namely Equalised Odds and Equal Opportunity, with the former being an equivalent notion of Separation and the latter being a relaxed notion of Separation. Both techniques are achievable by post-processing the outcomes of existing classifiers by moving the probability thresholds on the ROC curves belonging to different sensitive attribute groups to satisfy the constraints set by both techniques. The idea is to find a suitable threshold using the predictive score function $R = f(X, A)$ for each sensitive attribute group, whereby higher values of R correspond to greater likelihood of predicting $Y = 1$.

We say a predictor $\hat{Y}$ satisfies Equalised Odds with respect to a protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y, i.e. $P(\hat{Y} = 1|A = 1, Y = y) = P(\hat{Y} = 1|A = 0, Y = y)$, for $y \in \{0, 1\}$. On the other hand, we say that a predictor $\hat{Y}$ satisfies Equal Opportunity with respect to A and Y if $P(\hat{Y} = 1|A = 1, Y = 1) = P(\hat{Y} = 1|A = 0, Y = 1)$. In other words, Equalised Odds takes into account both FNR and FPR constraints, whereas Equal Opportunity only takes into account FNR constraint.

Using our trained Logistic Regression models for the different types of crime, we can input a person's characteristics into the models and obtain a real-valued predictive score R for each model. Then, we can create a binary classifier for each crime based on: $\widetilde{Y} = \mathbb{I}\{R > t\}$, where t is some threshold. If $\widetilde{Y}$ is only dependent on (R, A), it is known as a derived predictor.

We can define $C_a(t)$ as the ROC curve for group $A = a$. Let us define D_a as:

$$D_a = convhull\{C_a(t) : t \in [0, 1]\}$$

Any point in the convex hull D_a represents the false/true positive rates, conditioned on $A = a$, of a randomised derived predictor based on R. A predictor $\widetilde{Y}$ can always be taken to be a mixture of two threshold predictors. Therefore, the feasible set of false/true positive rates of possible Equalised Odds predictors is the intersection of areas under the A-conditional ROC curves, but above the main diagonal. Since the optimal predictor will always be on the top left boundary of this feasible set, we can effectively construct a new ROC curve that is the pointwise minimum of all A-conditional ROC curves.

To construct an Equalised Odds predictor, we choose a point on the boundary of this feasible set. If this point is on the ROC curve of a group, we choose a fixed threshold for that group t_a. Otherwise, we need to select two thresholds $\overline{t}_a > \underline{t}_a$, and use a mixture to obtain the predictor $\widetilde{Y} = \mathbb{I}\{R > T_a\}$, where T_a is a randomised threshold assuming the value $\overline{t}_a$ with probability $\overline{p}_a$ and the value $\underline{t}_a$ with probability $\underline{p}_a$. This way, each group has its own threshold. Hardt et al (2016) mentioned that we can use randomisation to achieve Equalised Odds but did not outline a process

for finding the thresholds and probabilities at which we randomise between the thresholds. Thus, we have created a process and derived an equation for the process of constructing an Equalised Odds predictor, as below:

1. Pick a point c on the worst-performing A-conditional ROC curve that we want to achieve, with corresponding TPR and FPR values: TPR_c and FPR_c.

2. For each group a we need to find two thresholds $\underline{t}_a$ and $\overline{t}_a$. Let the points $X_{\underline{t}_a}$ and $X_{\overline{t}_a}$ correspond to points on the A-conditional ROC curve for that group.

3. To find these points we then look at the intersection between the A-conditional ROC curve and the line $TPR = FPR + TPR_c - FPR_c$.

4. This gives us $X_{\underline{t}_a}$ and $X_{\overline{t}_a}$, enabling us to obtain $\underline{t}_a$ and $\overline{t}_a$ too.

5. To find the probability at which we randomise between the two thresholds, we need to solve the following equation for p: $p * X_{\underline{t}_a} + (1-p) * X_{\overline{t}_a} = c$.

Next, to construct an Equal Opportunity predictor, we choose a point on the boundary of the feasible set. Across all groups, it is sufficient to find points in D_a that are on the same horizontal line on the TPR vs FPR curve. Each group will have its own singular fixed threshold of t_a. Below is an outline of the process:

1. Pick a point c on the worst performing A-conditional ROC curve that we want to achieve, with corresponding TPR and FPR values: TPR_c and FPR_c.

2. For each group a we need to find a threshold t_a. Let the point X_{t_a} correspond to a point on the A-conditional ROC curve for that group.

3. To find this point we then look at the intersection between the A-conditional ROC curve and the horizontal line $TPR = TPR_c$.

4. This gives us X_{t_a}, enabling us to obtain the corresponding threshold t_a.

Among the benefits of the post-processing method is that it has been shown to record comparatively good performance without having to modify the existing classifier nor collect extra training data. Besides, it can be applied to a variety of classifiers, as long as we have access to the classifiers' ROC curves. However, it cannot be applied to certain fairness measures such as counterfactual fairness, since we need access to the protected attributes during the testing stage.

2.3.7 ML Fairness via Optimisation during Training

Another way to introduce fairness is via optimisation during training. Such an approach attempts to add fairness constraints or regularisation terms to the optimisation objective of the algorithm. In our case, we focus on achieving fairness in decision boundary-based classifiers, specifically Logistic Regression.

Introduction to disparate mistreatment

A classifier suffers from disparate mistreatment if different sensitive attribute groups do not achieve identical classification accuracy/ misclassification rate (Zafar et al, 2017a). For instance, if the victim misclassification rates are different between males and females, disparate mistreatment arises.

There are multiple measures of disparate mistreatment in binary classification, namely OMR, FPR, FNR, false omission rate (FOR), and false discovery rate (FDR) as highlighted in Table 3. Selecting a measure for disparate mistreatment depends on our real-world application scenario. For instance, in the criminal justice system, the cost of false positive is higher, since wrongly charging the innocent is deemed worse than overlooking a guilty person. In our case however, the cost of false negative is more, because it is more acceptable for the police force to wrongly provide support to those who are not susceptible to crime than let an individual vulnerable to crime be left without support.

To avoid disparate mistreatment in binary classification, we consider 3 main measures of misclassification, formalised as below, where we let z correspond to the sensitive attribute values:
overall misclassification rate (OMR):

$$P\left(\hat{y} \neq y \mid z = 0\right) = P\left(\hat{y} \neq y \mid z = 1\right) \tag{1}$$

false positive rate (FPR):

$$P\left(\hat{y} \neq y \mid z = 0, y = -1\right) = P\left(\hat{y} \neq y \mid z = 1, y = -1\right) \tag{2}$$

false negative rate (FNR):

$$P\left(\hat{y} \neq y \mid z = 0, y = 1\right) = P\left(\hat{y} \neq y \mid z = 1, y = 1\right) \tag{3}$$

Overcoming disparate mistreatment with respect to FDR and FOR has not been explored by Zafar et al (2017a), mostly because they would result in significant computational complexities - in this report, we would also not consider these two measures, but instead focus on eliminating disparate mistreatment quantified as measurements in Eqs.(1-3). This method also enables us to remove

disparate treatment simultaneously, and we would be pursuing fairness across multiple sensitive attributes.

Classification without disparate mistreatment

To design a fair Logistic Regression classifier that satisfies disparate mistreatment, the corresponding convex loss function $L\{\boldsymbol{\theta}\}$ is minimised under the chosen fairness constraint(s) from Eqs.(1-3). Taking OMR as an example, we would:

$$\begin{aligned} &\text{minimise} && L\{\boldsymbol{\theta}\} \\ &\text{subject to} && P\left(\hat{y} \neq y \mid z=0\right) - P\left(\hat{y} \neq y \mid z=1\right) \leq \varepsilon \\ & && P\left(\hat{y} \neq y \mid z=0\right) - P\left(\hat{y} \neq y \mid z=1\right) \geq -\varepsilon \end{aligned} \tag{4}$$

where the binary sensitive attributes $\{z_j\}_{j=1}^{m} \in \{0,1\}$ and $\boldsymbol{x}$ are disjoint. $\varepsilon \in \mathbb{R}^{+}$ indicates the difference in mistreatment between two sensitive attribute groups. Thus, the smaller the ε, the fairer the classifier would be. Disparate mistreatment criteria in equations Eqs.(1-3) are non-convex functions of the classifier parameters $\boldsymbol{\theta}$, therefore resulting in non-convex formulations like in Eq.(4). Unlike convex optimisation, non-convex optimisation may have multiple locally optimal points, hence it is difficult to solve it efficiently.

To resolve this problem, we use a tractable proxy $\text{Cov}(z, g_\theta(y,\boldsymbol{x}))$ to evaluate disparate mistreatment. $\text{Cov}(z, g_\theta(y,\boldsymbol{x}))$ measures the covariance between $\{z_j\}_{j=1}^{m} \in \{0,1\}$ (the respondents' sensitive attributes) and $g_\theta(y,\boldsymbol{x})$ (the signed distance between the decision boundary and the feature vectors of misclassified respondents). It can be formulated as (let $\mathscr{D}$ refer to the training dataset):

$$\begin{aligned} \text{Cov}(z, g_\theta(y,\boldsymbol{x})) &= \mathbb{E}\left[(z-\bar{z})(g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) - \bar{g}_{\boldsymbol{\theta}}(y,\boldsymbol{x}))\right] \\ &\approx \frac{1}{N} \sum_{(\boldsymbol{x},y,z)\in\mathscr{D}} (z-\bar{z}) g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \end{aligned} \tag{5}$$

where $g_\theta(y,\boldsymbol{x})$ can be defined as:

$$g_\theta(y,\boldsymbol{x}) = \min(0, y d_{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{6}$$

$$g_\theta(y,\boldsymbol{x}) = \min\left(0, \frac{1-y}{2} y d_{\boldsymbol{\theta}}(\boldsymbol{x})\right), \tag{7}$$

$$g_\theta(y,\boldsymbol{x}) = \min\left(0, \frac{1+y}{2} y d_{\boldsymbol{\theta}}(\boldsymbol{x})\right) \tag{8}$$

for OMR, FPR, and FNR respectively. Note that the covariance defined in Eq. (5) will be approximately equal to zero if a classifier satisfies any of the misclassification measures defined in Eqs.(6-

8). Using the proxy, Eq.(4) can be rewritten as:

$$\begin{aligned} \text{minimise} \quad & L\{\boldsymbol{\theta}\} \\ \text{subject to} \quad & \frac{1}{N}\sum_{(\boldsymbol{x},y,z)\in\mathscr{D}}(z-\bar{z})g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \leq c \\ & \frac{1}{N}\sum_{(\boldsymbol{x},y,z)\in\mathscr{D}}(z-\bar{z})g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \geq -c \end{aligned}$$

where the threshold $c \in \mathbb{R}^+$ controls the trade-off between accuracy and fairness.

However, the above covariance constraints for disparate mistreatment are still not convex. Next, we convert these constraints into a Disciplined Convex-Concave Program (DCCP). Convex-concave programming is an organised heuristic for solving non-convex problems where functions in objective and the constraints are a combination of a convex and a concave term (Shen et al, 2016).

Since we have two binary sensitive attributes $\{z_i\}_{j=1}^m \in \{0,1\}$ in our dataset, for each sensitive attribute, the training dataset $\mathscr{D}$ can be splitted into subsets $\mathscr{D}_0$ and $\mathscr{D}_1$, which represent individuals from different sensitive attribute groups $z=0$ and $z=1$ respectively. Let $\sim$ indicate $\leq$ or $\geq$:

$$\frac{1}{N}\sum_{(\boldsymbol{x},y)\in\mathscr{D}_0}(0-\bar{z})g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{1}{N}\sum_{(\boldsymbol{x},y)\in\mathscr{D}_1}(1-\bar{z})g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \sim c,$$

Let N_0 and N_1 be the number of people in subsets $\mathscr{D}_0$ and $\mathscr{D}_1$ respectively. Then $\bar{z} = \frac{(0\times N_0)(1\times N_1)}{N} = \frac{N_1}{N}$ and the constraint now becomes a convex-concave function:

$$\frac{1}{N}\left(\frac{-N_1}{N}\right)\sum_{(\boldsymbol{x},y)\in\mathscr{D}_0} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{1}{N}\frac{N_0}{N}\sum_{(\boldsymbol{x},y)\in\mathscr{D}_1} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \sim c,$$

In Logistic Regression, the convex loss $L(\theta)$ equals to $-\sum_{i=1}^{N}\log p\left(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)$. Therefore, we find the decision boundary parameters $\boldsymbol{\theta}$ for Logistic Regression without disparate mistreatment as a Disciplined Convex-Concave Program (DCCP):

$$\begin{aligned} \text{minimise} \quad & -\sum_{(\boldsymbol{x},y)\in\mathscr{D}}\log p\left(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}\right) \\ \text{subject to} \quad & \frac{1}{N}\left(\frac{-N_1}{N}\right)\sum_{(\boldsymbol{x},y)\in\mathscr{D}_0} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{1}{N}\frac{N_0}{N}\sum_{(\boldsymbol{x},y)\in\mathscr{D}_1} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \leq c \\ & \frac{1}{N}\left(\frac{-N_1}{N}\right)\sum_{(\boldsymbol{x},y)\in\mathscr{D}_0} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) + \frac{1}{N}\frac{N_0}{N}\sum_{(\boldsymbol{x},y)\in\mathscr{D}_1} g_{\boldsymbol{\theta}}(y,\boldsymbol{x}) \geq -c \end{aligned} \tag{9}$$

One of the main benefits of this fairness method is that it would also remove disparate treatment, since sensitive attributes $\{z_i\}_{j=1}^m$ are not used during the training process. Additionally, this type of classifier can also remove disparate mistreatment with respect to both FPR and FNR by applying separate constraints with $g_{\boldsymbol{\theta}}$ for both constraint types in Eqs.(6-8).

2.4 Analysis and Results

2.4.1 Exploratory Data Analysis

Figure 1 highlights the distribution of observations across the 4 types of crime being considered. We can clearly observe the data imbalance, with many more respondents being non-victims than victims, as expected. The data imbalance is worst for domestic & acquaintance violence, as there are relatively very few reported cases of victimisation for this crime, followed by threats and personal crime. The data imbalance is not that stark for household crime, with about 8.4 times more non-victims than victims recorded in the sample. In Section 2.4.4, we will explore techniques to deal with this data imbalance.

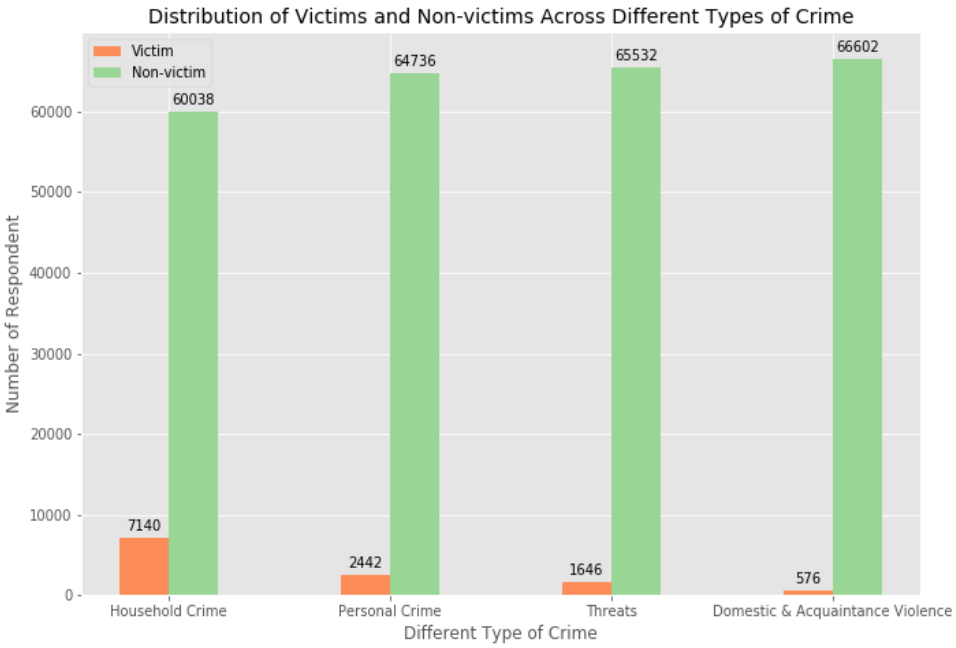

Figure 1: Distribution of Victims and Non-victims Across Different Types of Crime

We also looked at the distribution of respondents by our chosen sensitive attributes, i.e. gender and race. From Figure 2, we can observe that despite the imbalance proportion of victims and non-victims, there are no notable differences in the spread of respondents across the two groups by both gender and race, for household crime. Additionally, we could see that our sample has a very good balance in terms of gender, with almost equal numbers of males and females being sampled. However, there is a large imbalance by race, with only slightly more than 10% of the sample consisting of Non-whites. This may unintentionally result in our final classifier being unfair to the minority group, so it is a point worth noting. We do have the breakdown of the Non-whites by their major ethnicities, but because the sample size is already so small for these minorities, we decided to perform our analysis by comparing the Whites to the Non-whites as a whole rather than their specific ethnicities. As a similar trend is observed for the other 3 types of crime, their visualisations

are not included in this report.

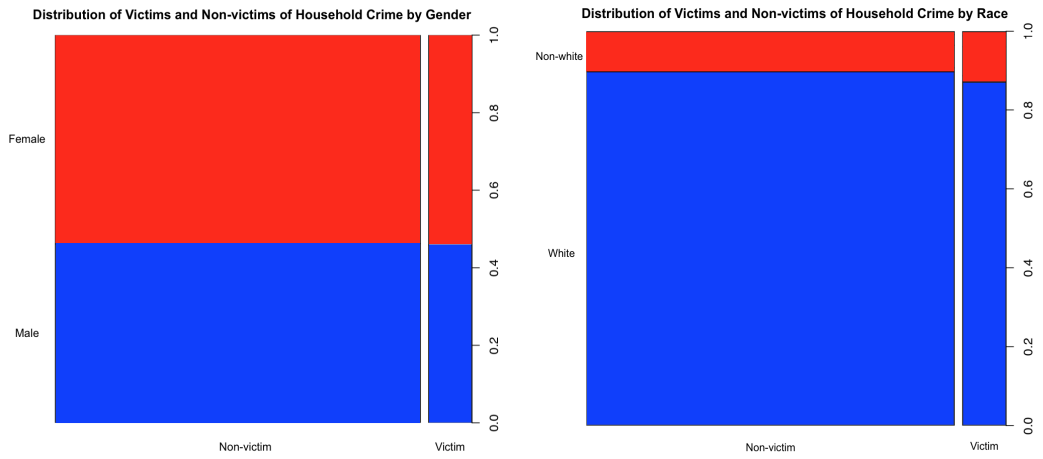

Figure 2: Distribution of Victims and Non-victims of Household Crime by Sensitive Attributes (Gender and Race)

Figure 3 highlights the correlations between all continuous variables in the dataset. As observed, there are high correlations between some of the Indices of Deprivation ("eincdc15", "eempdc15", "eedudc15", "eheadc15") as well as between "rubbcomm", "vandcomm", and "poorhou". After a thorough check (we also ran several Logistic Regression models to analyse the impact of removing some of the highly correlated variables on the significance and magnitude of the coefficients of other variables), we decided to drop "eempdc15", "eedudc15", and "eheadc15" and keep "eincdc15" because for the former 3 variables, there are other variables which act as proxies to them, so we are not losing valuable information by removing them. We also decided to exclude "vandcomm" and "poorhou" since they are both proxies to poor living conditions, which is already reflected by the variable "rubbcomm". Keeping them in the model caused some spurious correlations between other variables, and removing them solved this issue.

We also plotted boxplots to identify potential correlations between any of the continuous and categorical covariates. We spotted possible correlations between some of the independent variables, but upon further check, we decided to not drop any of the variables as the correlations that exist were mostly weak, and simply dropping them might cost us some valuable information. Next, through the Goodman Kruskal association measure, we spotted high correlations between some of the categorical covariates (there is also an instance of perfect correlation) and following the same reasoning and procedure as before, we dropped "dftdrive", "weekday", "mobwh", "vehowner", and "reverw".

Finally, before running ML models on the data, we looked at some basic plots to identify variables that could potentially correlate well with the four response variables. One example would be Figure 4 - here, we could conjecture that for household crime, there seems to be a difference in the

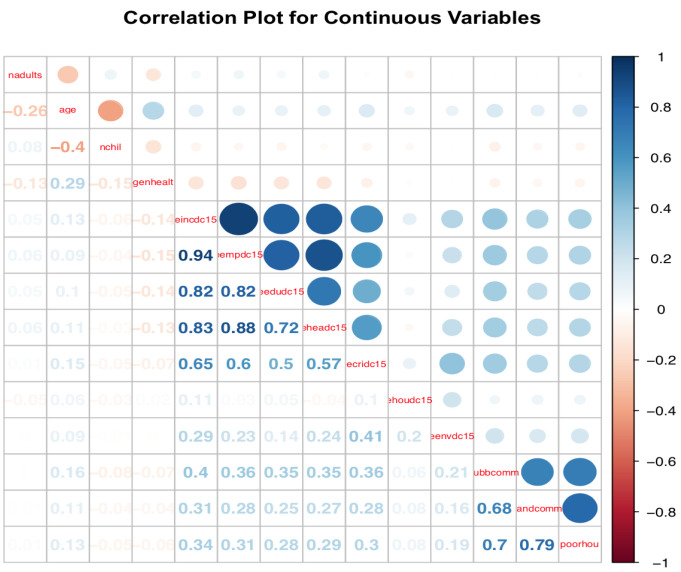

Figure 3: Correlation Plot for Continuous Variables

mean “age”, “eincdc15”, “ecridc15”, and “eenvdc15” reported between victims and non-victims, with victims on average being younger and more impoverished in terms of income, crime exposure and living environment. Thus, we would expect these variables to be included in our final model for household crime (although the final decision still depends on the significance of these mean differences).

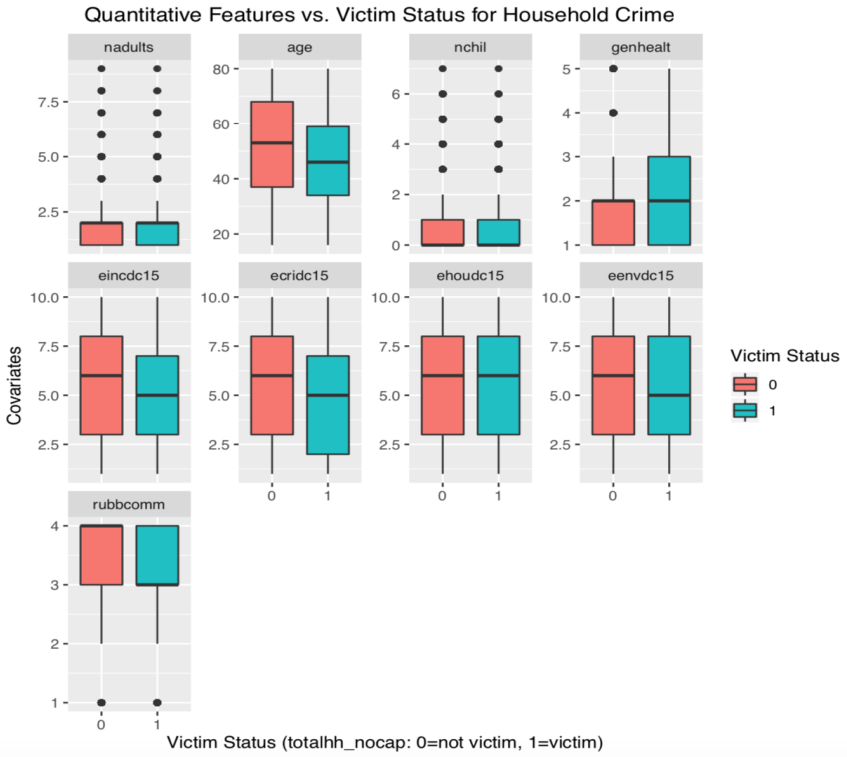

Figure 4: Correlations between Quantitative Features and Victim Status for Household Crime

2.4.2 ML Classifiers, Variable Selection, and Dealing with Imbalanced Data

After EDA, we started building ML models for each type of crime. In this section, we will only report the results for the two best-performing ML methods explored, namely Logistic Regression and Random Forest. Next, to deal with the data imbalance problem, we also explored the Upsampling method across these two ML methods.

As explained in Section 2.3.3, we also performed variable selection for our Logistic Regression models using LASSO, and only included variables that are significant at 5% significance level. Figure 5 shows the ROC curves for our four best models for household crime, namely Logistic Regression with LASSO using original sample and the upsampled data, as well as Random Forest using original sample and upsampled data. In this figure, we can see that the Logistic Regression model (with LASSO) using the original sample outperformed all the other models in terms of AUC score. Table 4 displays the AUC scores for the four models explored for each type of crime - it is clear that across all types of crime, Logistic Regression with LASSO outperformed Random Forest regardless of the sample type, thus the former is selected as our final type of classifier for all types of crime.

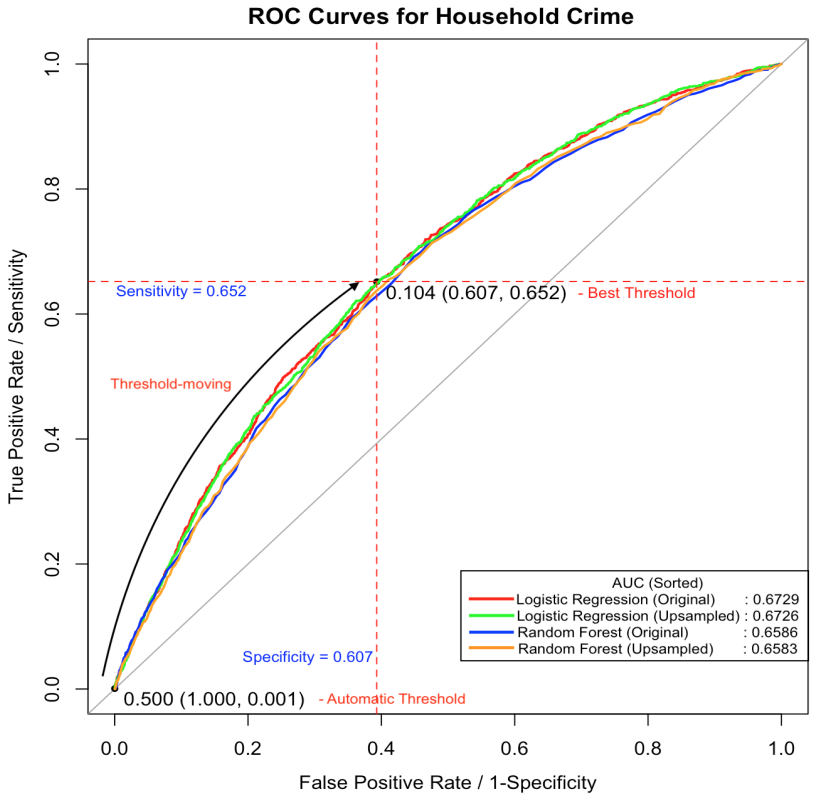

Figure 5: ROC Curves for Different Classification & Sampling Methods for Household Crime

Next, although we could observe marginal improvements in terms of the AUC scores for the Logistic Regression models for personal crime and threats (Table 4), we decided to still select the Logistic Regression models using the original sample. This is because for both types of crime, the upsampled models selected almost twice the number of variables compared to the corresponding

models built using the original sample (25 vs. 13 for personal crime and 29 vs. 15 for threats). This is expected as one of the main disadvantages of the upsampling method is the risk of overfitting, especially when observations from the minority class had to be duplicated many times to match the number of observations in the majority class. Similar scenario is observed for the other two types of crime, whereby the upsampled models recorded much higher complexities in terms of variables included (31 vs. 21 for household crime and 29 vs. 10 for domestic & acquaintance violence). Additionally, the upsampled models performed poorer than the models built using the original sample for these two crimes in terms of AUC scores (Table 4). Thus, our final decision is to select the Logistic Regression model with LASSO using the original sample for all four types of crime.

		Area Under ROC Curve (AUC)			
		Household Crime	Personal Crime	Threats	Domestic & Acquaintance Violence
Logistic Regression	Original Sample	0.6729	0.6766	0.7250	0.7415
	Upsample	0.6726	0.6812	0.7252	0.7396
Random Forest	Original Sample	0.6586	0.6595	0.6815	0.7270
	Upsample	0.6583	0.6539	0.6969	0.7059

Table 4: AUC Values for Different Classification & Sampling Methods

Since we rejected the upsampling method, we still have to deal with the data imbalance problem - thus, we applied "threshold-moving" on our final Logistic Regression models. We used a built-in function "print.thres" from the pROC package in R to select the "best" threshold, i.e. the threshold with the highest sum sensitivity + specificity. Figure 5 shows the threshold-moving process on the final Logistic Regression model for household crime - the function we used moved our probability threshold from the usual 0.5 to 0.104, giving us a better trade-off between sensitivity and specificity (the original 0.5 threshold correctly (wrongly) classified almost all non-victims (victims), which is what we intend to avoid). The moved threshold gave a better sensitivity-specificity trade-off (0.652 vs. 0.607).

Likewise, we applied threshold-moving to all of our chosen models across the four types of crime - the "best" thresholds are displayed in Figure 6. From the ROC curves, we see that among all types of crime, our model for domestic & acquaintance violence performed the best, followed by our models for threats, personal crime and household crime respectively. Additionally, for domestic & acquaintance violence and threats, our "best" thresholds managed to surpass the 0.7 target that we set (Section 2.3.4) for sensitivity, implying that our models performed well as victim screener

tests for these two crimes.

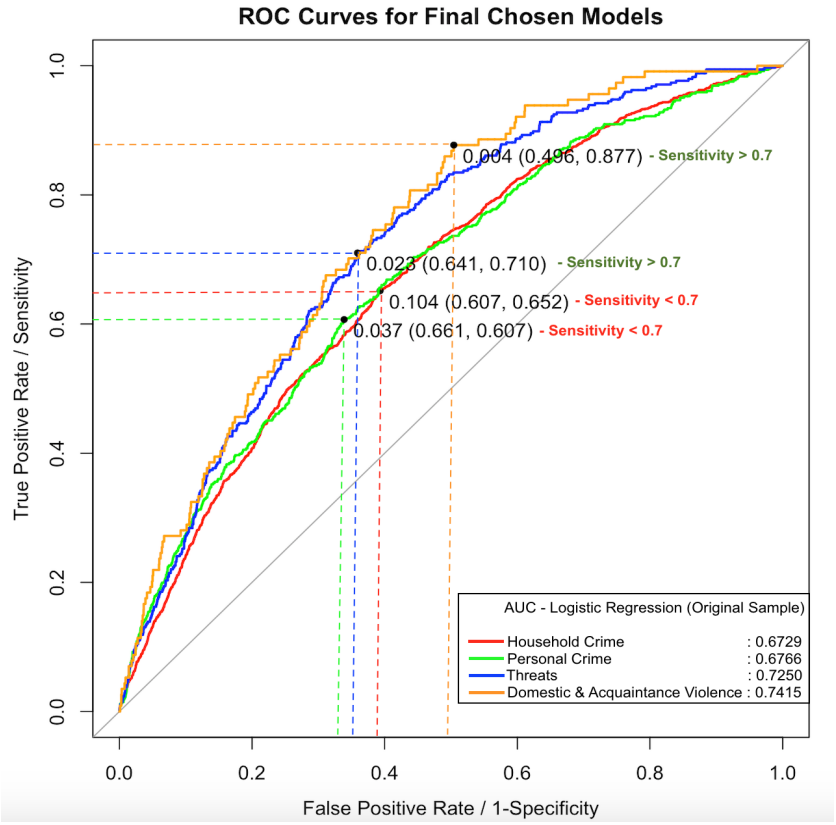

Figure 6: ROC Curves for Final Baseline Models with the 'Best' Thresholds for Four Types of Crime

However, for household crime and personal crime, the "best" thresholds that gave the best trade-off between sensitivity and specificity were not able to surpass the 0.7 target for sensitivity, albeit not being too far from it. Of course, the threshold can be moved manually to achieve a sensitivity score of 0.7, but that would not give the best trade-off between the two measures. Given more training data, we might be able to improve our models for these two crimes, thus able to achieve the best trade-off between sensitivity and specificity while surpassing the sensitivity target of 0.7.

Below we include the final Logistic Regression equations across the 4 types of crime; These only include significant variables, and for the purpose of space efficiency, in this report, we only include coefficients of levels that are significant for any categorical variable selected in the final models - they can be compared to their corresponding baseline levels as described in Table 1. Please refer to Table 1 for the interpretation of each variable and their levels (for categorical variables).

Note: *The coefficients below are rounded to 2 significant figures. The numbers attached to the right of the variable names refer to the levels that are significant for those categorical variables, in comparison to their respective baseline levels (emboldened levels in Table 1).*

Household Crime:

$$\text{logit}\,[Pr(\text{totalhh\_nocap=1}|\boldsymbol{x})] = -2.56 - 0.0038\mathbf{age} + 0.089\mathbf{nchil} + 0.074\mathbf{genhealt} + 0.20\mathbf{onsdisab1} + 0.37\mathbf{relig39} + 0.14\mathbf{educat22} + 0.49\mathbf{managhh21} + 0.20\mathbf{managhh22} - 0.26\mathbf{rural3} - 0.24\mathbf{gor1} - 0.18\mathbf{gor2} - 0.15\mathbf{gor9} - 0.44\mathbf{gor10} - 0.018\mathbf{eincdc15} - 0.062\mathbf{ecridc15} - 0.026\mathbf{eenvdc15} + 0.11\mathbf{acctyp3} + 0.25\mathbf{acctyp4} + 0.22\mathbf{acctyp7} - 0.10\mathbf{rnssec52} + 0.10\mathbf{rnssec53} - 0.28\mathbf{rnssec56} + 0.18\mathbf{livharm1a2} + 0.11\mathbf{livharm1a3} + 0.27\mathbf{livharm1a4} + 0.26\mathbf{livharm1a5} - 0.22\mathbf{rlstweek10} + 0.30\mathbf{bikowner} + 0.53\mathbf{numcar31} + 0.71\mathbf{numcar32} + 1.1\mathbf{numcar33} + 0.43\mathbf{antisocial2} + 0.76\mathbf{antisocial3} + 0.65\mathbf{antisocial4} - 0.13\mathbf{mobproc2} - 0.25\mathbf{mobproc3} - 0.09\mathbf{homesecurity4}$$

Personal Crime:

$$\text{logit}[Pr(\text{totperls\_nocap} = 1|\boldsymbol{x})] = -3.16 - 0.20\mathbf{sex2} + 0.21\mathbf{pubclub2} + 0.10\mathbf{genhealt} + 0.40\mathbf{onsdisab1} + 0.49\mathbf{managhh21} + 0.30\mathbf{managhh22} - 0.77\mathbf{gor1} - 0.27\mathbf{gor2} - 0.21\mathbf{gor9} - 0.25\mathbf{gor10} + 0.22\mathbf{acctyp3} + 0.40\mathbf{acctyp5} + 0.26\mathbf{acctyp6} - 0.32\mathbf{nsethgrp} + 0.20\mathbf{livharm1a2} + 0.37\mathbf{livharm1a3} + 0.62\mathbf{livharm1a4} + 0.65\mathbf{livharm1a5} + 0.42\mathbf{livharm1a6} + 0.97\mathbf{rlstweek2} - 0.27\mathbf{rlstweek7} - 0.33\mathbf{rlstweek10} + 0.76\mathbf{pubclub5} + 0.45\mathbf{antisocial2} + 0.84\mathbf{antisocial3} + 0.77\mathbf{antisocial4} - 0.20\mathbf{mobproc2} - 0.29\mathbf{mobproc3} - 0.02\mathbf{age} + 0.39\mathbf{pubclub3} + 0.59\mathbf{pubclub4}$$

Threats:

$$\text{logit}[Pr(\text{threats\_nocap=1}|\boldsymbol{x})] = -5.13 + 0.32\mathbf{sex2} + 0.27\mathbf{yrsarea4} + 0.12\mathbf{genhealt} + 0.46\mathbf{onsdisab1} + 0.72\mathbf{relig35} + 0.61\mathbf{relig38} + 0.60\mathbf{educat21} + 0.52\mathbf{educat22} + 0.40\mathbf{managhh21} + 0.26\mathbf{managhh22} + 0.39\mathbf{gor6} - 0.18\mathbf{acctyp2} + 0.24\mathbf{livharm1a3} + 0.43\mathbf{livharm1a4} + 0.40\mathbf{livharm1a5} + 1.0\mathbf{antisocial4} + 0.39\mathbf{rlstweek3} - 0.37\mathbf{rlstweek10} + 0.20\mathbf{bikowner} + 0.61\mathbf{antisocial2} + 0.88\mathbf{antisocial3} - 0.60\mathbf{gor1} - 0.27\mathbf{mobproc2} - 0.41\mathbf{mobproc3} + 0.15\mathbf{pubclub2} + 0.62\mathbf{pubclub5}$$

Domestic & Acquaintance Violence:

$$\text{logit}[Pr(\text{dom\_acq\_nocap} = 1|\boldsymbol{x})] = -5.74 - 0.022\mathbf{age} + 0.14\mathbf{genhealt} + 0.47\mathbf{onsdisab1} + 0.61\mathbf{gor2} + 0.77\mathbf{managhh21} + 0.24\mathbf{managhh22} + 1.1\mathbf{gor3} + 1.3\mathbf{gor4} + 0.76\mathbf{gor5} + 0.95\mathbf{gor6} + 1.0\mathbf{gor8} + 0.83\mathbf{gor9} + 0.91\mathbf{gor10} + 0.55\mathbf{livharm1a2} + 0.64\mathbf{livharm1a3} + 1.3\mathbf{livharm1a4} + 0.99\mathbf{livharm1a}5 + 0.35\mathbf{tenharm2} - 1.5\mathbf{rlstweek10} + 0.47\mathbf{antisocial2} + 1.0\mathbf{antisocial3} + 0.72\mathbf{antisocial4} - 1.4\mathbf{nationality1} - 2.1\mathbf{nationality4}$$

In this report, we will only focus on highlighting some main and interesting findings across the four types of crime. However, in general, the coefficients can be interpreted this way: positive

(negative) coefficients reflect positive (negative) association with the response variable, a larger magnitude implies larger effect of that variable, and the exponents of the coefficients highlight the amount of increase (positive) or decrease (negative) in the odds of being a victim for that specific type of crime, when the amount for that variable increases by one unit (continuous variable) or when the level changes from baseline to that level (categorical variable), keeping all else constant. There are altogether 21, 13, 15, and 10 significant variables for the four types of crime respectively.

Here, we present some interesting findings. There is a regional effect across all 4 types of crime, but the effect varies across them; For crimes other than domestic & acquaintance violence, the susceptibility to crime in a few regions (refer to Table 1 and equations above for more details) decreases in general compared to the baseline region, i.e. London, but for domestic and acquaintance violence, every other geographic location (except for North East) recorded higher risk than London. With every model antisocial was a significant variable, and people are usually more at risk the greater the severity of antisocial behavior they witness. Pubclub is a significant factor for only threats and personal crime, implying that visiting pubs/ clubs is associated with a higher exposure to experiencing personal crime/ threats, but not necessarily household/ domestic crimes.

Having more cars or owning a bike increases the propensity to experience household crime, but not other types of crime (except threats, where owning a bike increases the risk for that crime). Where age is significant, it has a negative association with crime victimisation, i.e. older people are less at risk. As expected, having worse health conditions/ a mental or physical disability/ not being able to find £100 to meet unexpected expenses (a proxy of financial condition) are all associated with higher risks of experiencing all four types of crime. Interestingly for domestic & acquaintance violence, people who are single, divorced, separated, or cohabiting are more at risk of domestic and acquaintance violence than married people, and it is the only type of crime whereby the nationality of a person matters in determining his/ her susceptibility to crime, with Europeans (excluding UK nationals)/ Middle Easterners/ Asians being at lower risk to facing this type of crime compared to UK nationals.

2.4.3 ML Fairness via Post-processing – Equalised Odds, Equal Opportunity

As highlighted in Sections 2.3.5 and 2.3.6, the Logistic Regression models that we built for the four types of crime did not take into account any fairness constraints. As we aim to not disadvantage any groups belonging to the two protected classes $\{z_i\}_{i=1}^{2}$ i.e. sex and gender when identifying those who are susceptible to experiencing crime, our next step was to implement methods from the ML Fairness umbrella to minimise any gaps between our four subgroups namely White males, Non-white males, White females, and Non-white females in terms of their (primarily) FNRs and

FPRs (whenever possible).

Before that, we examined our existing models to confirm that indeed, our models discriminate between the four subgroups. The disparities between the ROC curves for the different protected groups as observed in Figures 7 - 10 (for the four types of crime) confirm that our models are not treating each group fairly, with some groups having better classification performance, and some others much worse. An interesting observation - our classifiers are consistently most unfair to the Non-white males across all types of crime.

Still referring to Figures 7 - 10, we found that in general, our classifiers performed better for the Whites compared to the Non-whites, except in the model for threats (Figure 9), where our classifier performed the best for the Non-white females. Although the much smaller number of observations in the two minority groups namely Non-white males and Non-white females did result in their ROC curves taking more of a stepwise form, we cannot say for sure that this is also the reason behind poor performance, as there are instances (household crime and threats) where the Non-white females recorded better performances compared to the Whites.

After establishing that our models are not treating the four groups fairly, we implemented our first method of fairness called the post-processing method, whereby we still used our unfair models (from this point onwards we will call these unfair models as our baseline models), but we tweaked the probability threshold selection process at the end. As explained in Section 2.3.6, there are two techniques under this post-processing method, i.e. Equalised Odds and Equal Opportunity (Hardt et al, 2016). The former ensures the gaps between all groups in terms of FNR and FPR are minimised as much as possible, whereas the latter only ensures similar FNRs but not necessarily FPRs across all groups.

Figures 7 - 10 highlight exactly the process of selecting the best thresholds that maximised fairness under both Equalised Odds and Equal Opportunity techniques, for each type of crime. We first identified the "best" threshold for the worst-performing group, which in our case was the Non-white males for all types of crime, and used this as the baseline threshold across both techniques. This is because for Equalised Odds, this would be one of the best options to select from the convex hull (Section 2.3.6), and by choosing a point on the worst-performing group, we can avoid performing randomisation for the group with the worst performance.

Since Equal Opportunity works by setting the same sensitivity level across all groups, we decided on the sensitivity level by looking at the sensitivity corresponding to the "best" threshold for the worst-performing group (horizontal dotted lines in Figures 7 - 10) - this would at least give the best trade-off between sensitivity and specificity for the worst-performing group, and of course better

specificity levels for the other groups (the vertical dotted lines), given the same sensitivity. This would ensure fairness in terms of FNR without sacrificing the FPRs too much, especially for the worst-performing group.

For Equalised Odds predictors, we performed randomisation as described in Section 2.3.6 between the two thresholds for each group other than the Non-White males as displayed in the zoomed in images on Figures 7 - 10 before conducting predictions on the test set. This is done so that the FPRs and FNRs for the 3 other groups are similar to the ones for the Non-white males, satisfying Separation. Across the four types of crime and both Equalised Odds and Equal Opportunity techniques, we were able to achieve sensitivities of over 70% for each group (except for Domestic & Acquaintance Violence where the lowest sensitivity achieved per group was 66.7% - still very close to the 70% mark). This implies that our models are doing well since we would only be classifying at most 30% of victims wrongly for each group, which is a good level of FNR (Section 2.3.4).

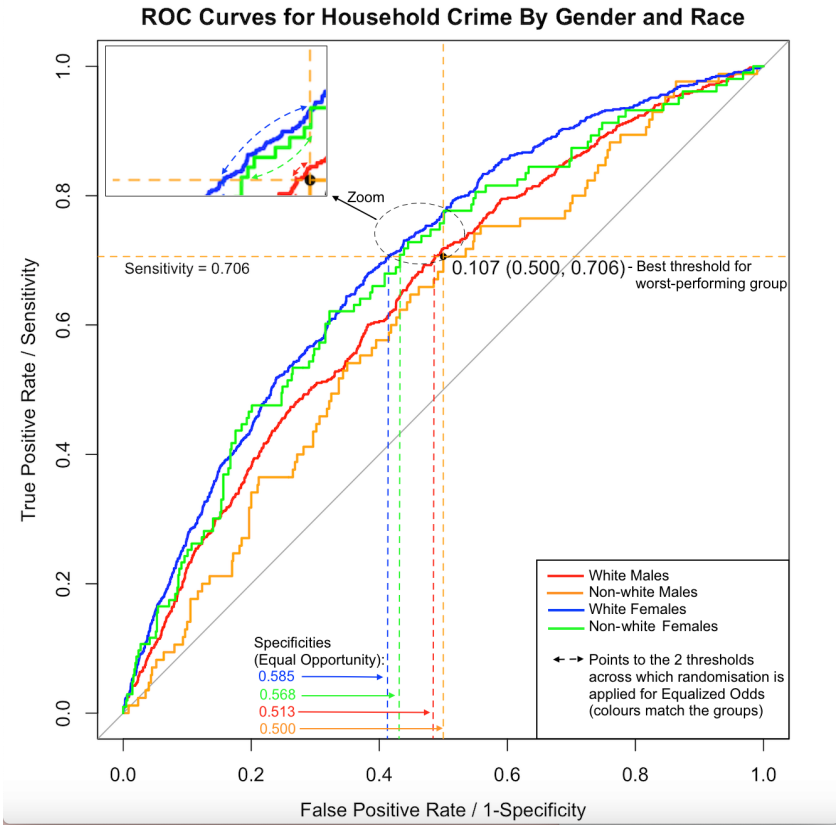

Figure 7: Household Crime: ROC Curves showing the construction of Equalised Odds and Equal Opportunity predictors. The zoomed-in image highlights the threshold randomisation process to obtain the Equalised Odds predictor, whereas the points of intersection between the vertical dotted lines and the horizontal line correspond to the thresholds used to construct the Equal Opportunity predictor. The sensitivity and specificity corresponding to the "best" threshold for the worst-performing group are the base targets that we try to achieve from the randomisation process for the Equalised Odds predictor, and the same sensitivity level is also selected as the base target to achieve across all 4 groups.

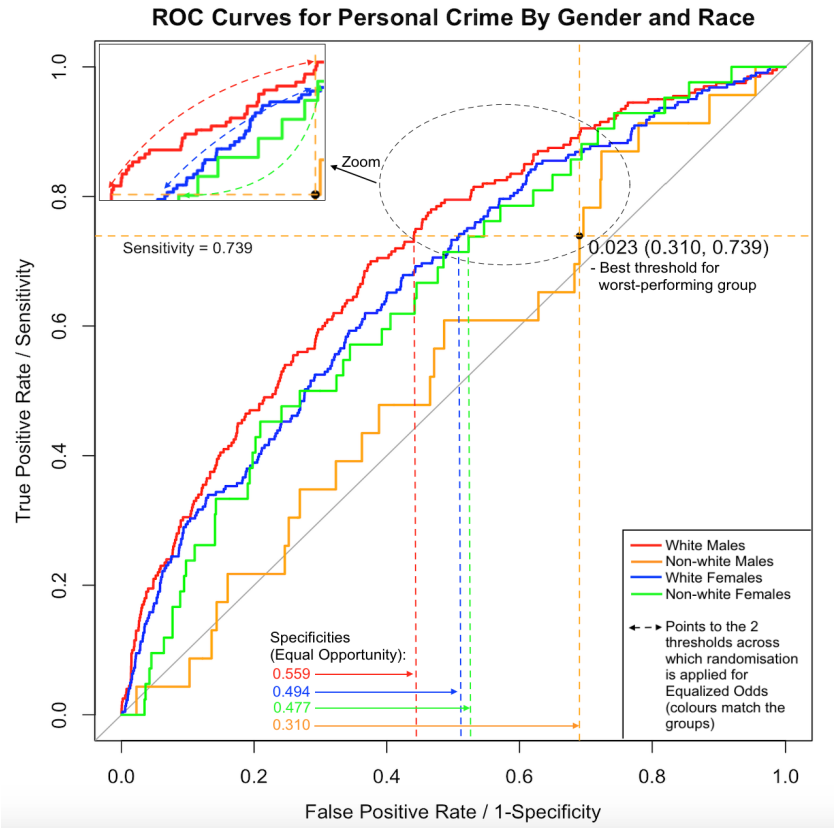

Figure 8: Personal Crime: ROC Curves showing the construction of Equalised Odds and Equal Opportunity predictors.

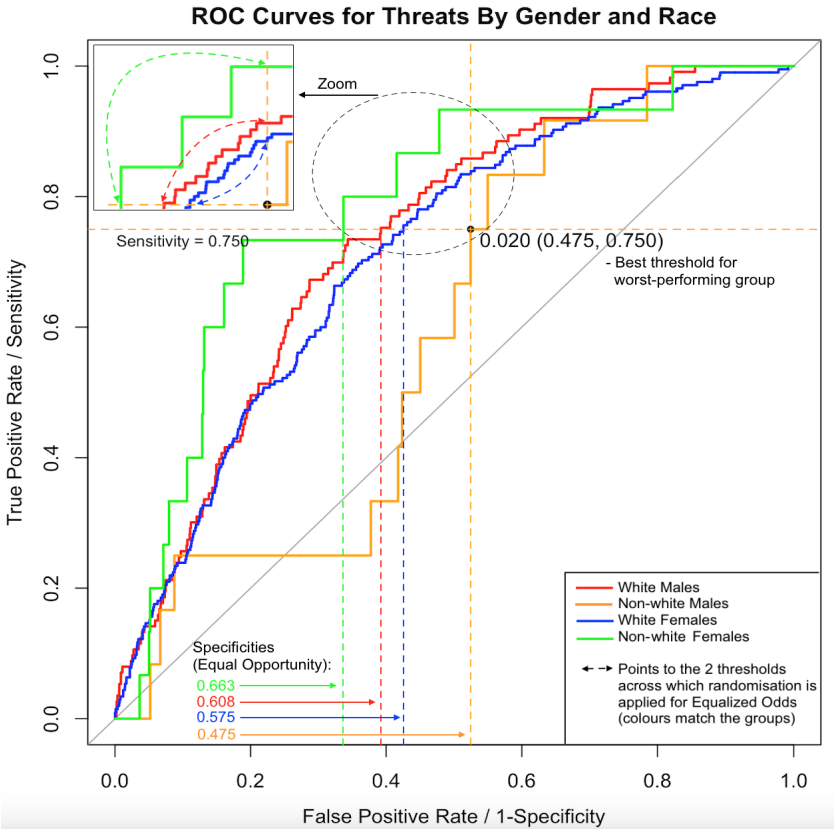

Figure 9: Threats: ROC Curves showing the construction of Equalised Odds and Equal Opportunity predictors.

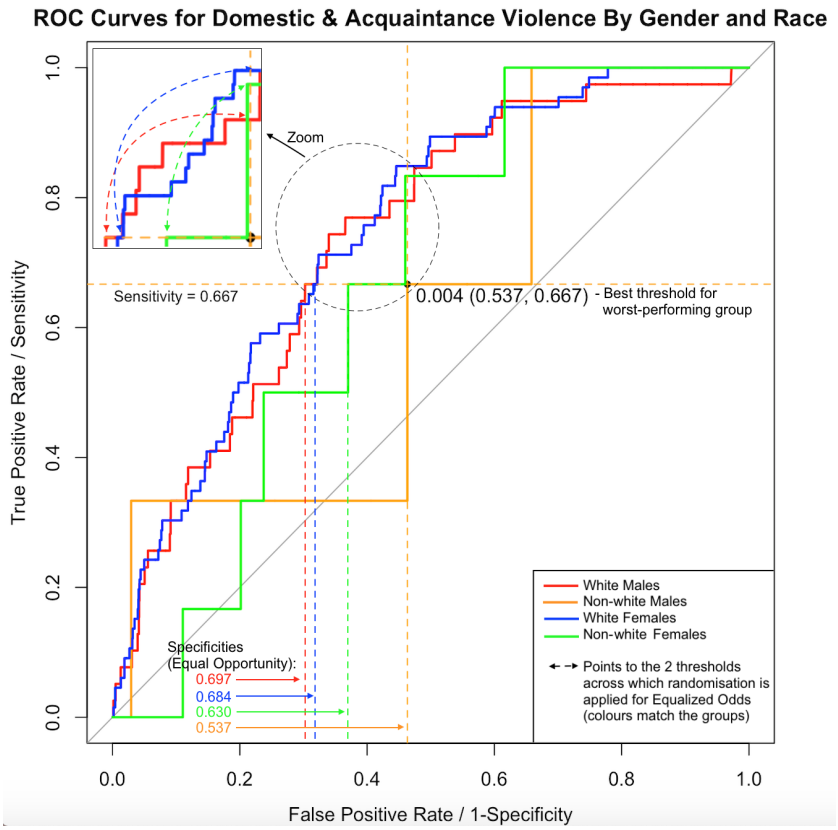

Figure 10: Domestic & Acquaintance Violence: ROC Curves showing the construction of Equalised Odds and Equal Opportunity predictors.

As expected, given the highly imbalanced dataset across all crimes, the high levels of sensitivity were achieved at some expense of specificity; Although we were able to mostly achieve above 50% specificities across all groups for all types of crime, there were instances where the specificities recorded were lower than 50%, going as low as 31% for the Non-white males under personal crime (Figure 8). Even so, 50% is not the best target, and if we have access to more data on the minorities (Non-whites), we could build better predictors with better trade-offs between sensitivity and specificity; However, under the current circumstances, the results we achieved are sufficient for the predictors to fulfill their roles as screening tests (high sensitivity with not too much focus on specificity).

2.4.4 ML Fairness via Optimisation during Training

Next, we trained Logistic Regression classifiers without both, disparate mistreatment and disparate treatment simultaneously on our real world dataset by adding fairness constraints during the training process. Misclassification can be measured in different ways, we aim to eliminate disparate mistreatment in terms of FPR and FNR specifically.

Disparate mistreatment in our case is represented by D_{FPR} and D_{FNR}, which indicate the biggest absolute difference in FPR and FNR respectively, between any pair from the four subgroups. The closer the values of D_{FPR} and D_{FNR} to zero, the lower the degree of disparate mistreatment. Dis-

parate treatment is removed simultaneously by not including the two sensitive attributes while training the models. This method offers high flexibility in choosing the trade-off between accuracy and fairness by selecting different thresholds for $c \in \mathbb{R}^+$ as in Eqn.(9). The smaller the value of c, the higher the fairness of our model and when $c = 0$, we get the fairest classifier. In our case, by setting $c = 0$, all values of D_{FPR} and D_{FNR} obtained are around 0.1.

2.4.5 Performance Comparison between Different Classifiers

In this section, we aim to compare the performance of our unconstrained (unfair) Logistic Regression models for each type of crime to their corresponding fair models (generated via both, post-processing and optimisation during training). Besides, we also aim to compare and contrast the pros and cons of both fairness techniques we implemented - the first technique operates by post-processing the outcomes of an unfair classifier, and the second technique trains classifiers with fairness constraints.

To achieve this second aim, we compare the results obtained from the Equalised Odds classifier to the classifier generated by optimising both FNR and FPR constraints during the training process, as they essentially apply the same constraints, but via different techniques. On the other hand, we compare the Equal Opportunity classifier to the classifier generated by optimising only FNR constraint during the training process, as the former only requires that the same sensitivity value (i.e. TPR = 1 - FNR) is set across all subgroups under consideration. Although the models we built for the four types of crime are essentially different and so, their performances are not directly comparable, we can attribute a part of the decline in performance (Table 5) for some of the crime types (domestic & acquaintance violence) to the much smaller victim to non-victim ratio.

As shown in Table 5, the unconstrained baseline models for all crimes recorded larger $D_{FNR}s$ than $D_{FPR}s$, implying that the unfairness across the four subgroups is more severe in terms of FNR than FPR. In the baseline model, threats recorded a noticeable D_{FNR} value, which is 0.683. The classifier yielded FNRs of 0.363, 0.750, 0.239 and 0.067 respectively for the White males, Non-white males, White females, and Non-white females, implying that the baseline victim classifier brought on a huge disadvantage to the Non-white males - about 75% of actual victims would be misclassified as non-victims (as opposed to only about 7% for the Non-white females). The baseline models for crimes other than threats recorded D_{FNR} values of around 0.20. Moreover, the D_{FPR} values for the four types of crime are 0.150, 0.081, 0.227 and 0.044 (respectively, following the order in Table 5 for baseline models). These results show the existence of disparate mistreatment in terms of both FPR and FNR in our unconstrained Logistic Regression models.

For the Equal Opportunity predictors (Table 5), since the same FNR is set for all subgroups (or almost equal when it's not possible to set the same FNR), all $D_{FNR}s$ decreased dramatically from

the high values in the unfair baseline models to nearly zero; The D_{FNR} in domestic crime even dropped to exactly zero. However, for personal crime and domestic & acquaintance violence, the $D_{FPR}s$ for the Equal Opportunity predictors are slightly higher than in their baseline models. Moreover, for the Equalised Odds predictors, both values of D_{FPR} and D_{FNR} are lower than the corresponding values in the unconstrained baseline models for all types of crime. These results, as highlighted in Table 5, suggest that both Equalised Odds and Equal Opportunity techniques are effective in eliminating unfairness in terms of FPR and FNR, often entailing a small cost in terms of accuracy.

For the optimisation during training method, the values of $D_{FPR}s$ and $D_{FNR}s$ are always smaller than those from the unconstrained baseline models, but slightly larger than those from the corresponding post-processing techniques. The reason for not achieving absolute fairness (i.e. $D_{FPR}s$ and $D_{FNR}s$ are not equal to zero) via this optimisation during training method might be that the size of our dataset is not large enough. According to Zafar et al (2017a), having a relatively small number of respondents would hinder a robust estimate of the misclassification covariance.

Among all methods, the unfair baseline models have the highest accuracies for most crimes except Domestic & Acquaintance Violence (this crime has the biggest data imbalance, we conjecture this may have something to do with the unusual results), but the models discriminate against people from certain groups, treating them unfairly. Applying post-processing and optimisation during training methods effectively removed the disparate mistreatment, but often with a small cost in terms of accuracy. However, as mentioned in Section 2.3.4, accuracy is not our chosen performance metric, since our dataset is largely imbalanced - rather, we focus on achieving high sensitivities / low FNRs as we are building screening tests for identifying victims and our focus is to be able to classify most victims correctly without sacrificing FPRs too much. In Table 5, comparing the fair models to the unfair baseline models, the former always recorded lower FNRs (except in Domestic & Acquaintance Violence), and if we compare between the two fairness techniques, post-processing always recorded lower FNRs compared to the optimisation during training method, suggesting that the fair models achieved via post-processing techniques performed the best in terms of our chosen performance metric.

To summarise, the optimisation during training method is appealing as it can simultaneously remove disparate mistreatment and disparate treatment - something that the post-processing technique is not able to achieve, because although we did not include the sensitive attributes in the training process, we still had to access them at the time of threshold-moving and decision-making. In addition, this method also systematically achieved slightly better accuracies compared to the corresponding post-processing methods. However, the post-processing method is easier to im-

plement as it operates on the outcomes of the classifiers without requiring retraining the models which can be time-consuming especially for very large amounts of data. Besides, the post-processing method also achieved the smallest overall FNRs, $D_{FPR}s$ and $D_{FNR}s$, implying that this method not only performed the best in terms of our chosen performance metrics, but it also produced the fairest outcome. Given a bigger dataset, the optimisation during training method might work better, as highlighted in the original paper by Zafar et al (2017a), but the cost of retraining models may still be an issue.

		FNR Constraint				**Both constraints**			
		Acc	FNR	D_{FPR}	D_{FNR}	Acc	FNR	D_{FPR}	D_{FNR}
Household Crime	Baseline (Unfair)	0.612	0.348	0.150	0.176	0.612	0.348	0.150	0.176
	Equalised Odds	-	-	-	-	0.552	0.283	0.058	0.019
	Equal Opportunity	0.566	0.293	0.085	0.003	-	-	-	-
	Opt during Training	0.603	0.310	0.097	0.108	0.581	0.298	0.099	0.102
Personal Crime	Baseline (Unfair)	0.659	0.393	0.081	0.222	0.659	0.393	0.081	0.222
	Equalised Odds	-	-	-	-	0.396	0.199	0.075	0.131
	Equal Opportunity	0.519	0.261	0.250	0.002	-	-	-	-
	Opt during Training	0.571	0.299	0.122	0.123	0.433	0.283	0.113	0.120
Threats	Baseline (Unfair)	0.643	0.290	0.227	0.683	0.643	0.290	0.227	0.683
	Equalised Odds	-	-	-	-	0.549	0.220	0.099	0.050
	Equal Opportunity	0.593	0.246	0.187	0.050	-	-	-	-
	Opt during Training	0.590	0.246	0.130	0.147	0.564	0.252	0.123	0.153
Domestic & Acquaintance Violence	Baseline (Unfair)	0.499	0.123	0.044	0.227	0.499	0.123	0.044	0.227
	Equalised Odds	-	-	-	-	0.576	0.290	0.040	0.106
	Equal Opportunity	0.679	0.333	0.161	0.000	-	-	-	-
	Opt during Training	0.633	0.305	0.160	0.143	0.590	0.291	0.159	0.139

Table 5: This table contains the accuracy, overall FNR, D_{FPR}, and D_{FNR} results for the four types of crime for the unconstrained baseline model, as well as the fair models implemented via post-processing and optimisation during training. The post-processing technique includes the Equalised Odds predictor and the Equal Opportunity predictor. The Equalised Odds predictor considers both FNR and FPR and so is comparable to the classifier optimised with both constraints during the training process (to the right of the bold vertical line). The Equal Opportunity predictor considers only the FNR constraint, and so is comparable to the classifier optimised with only FNR constraint during the training process (to the left of the bold vertical line). The unfair baseline model is unconstrained, so the numbers are the same to the left and right of the bold vertical line.

3 Conclusion

3.1 Main Findings

In this report, we strived to develop a holistic framework for developing a fair screening test to identify individuals susceptible to experiencing four different types of crime, namely household crime, personal crime, threats, and domestic & acquaintance violence. We largely drew upon the concept of EpiCrim, i.e. addressing crime victimisation from the public health perspective by treating crime like a social disease, and used statistical tools to develop a fair solution to the problem of victim detection so that those susceptible to experiencing crime are correctly identified from the society, and provided the much needed support to curb victimisation and revictimisation.

We began constructing our framework by identifying significant factors that are highly correlated with crime victimisation using multiple aspects of variable selection, i.e. the relevance and statistical significance of the attributes, as well as identifying highly correlated independent variables to remove confounding variables, resulting in a robust variable selection process. We managed to reduce the number of variables from as high as 2,962 to only 21, 13, 15, and 10 respectively across the four types of crime (refer to Section 2.4.2 for the detailed breakdown of the significant variables for each crime). This way, to use our models, the end user will not need to conduct an extensive data acquisition process like in CSEW but instead, should there be a need to conduct a survey, the reduced number of variables will result in much less questions to be asked and so, reduce the monetary and time cost of data collection.

Next, in line with the main theme of the report which is to incorporate fairness into the classifier-building process, we also examined our selected models (after variable selection) for unfairness towards any the four subgroups that we have, namely the White males, Non-white males, White females, and Non-white females. We emphasised on whether or not the FNRs between the four subgroups are vastly different, thus disadvantaging any of the subgroups when it comes to receiving the support they deserve as individuals who are susceptible to crime. From our analysis, we found that all of our classifiers across the four types of crime are discriminative towards the different subgroups, with all of them being most unfair to the Non-white males. Thus, we established a solid case for the use of ML Fairness methods in order to eradicate the unfairness embedded in our baseline Logistic Regression models.

Under the ML Fairness umbrella, we considered the three fundamental observational criteria (Barocas, Hardt and Narayanan, 2019) namely Independence, Separation, and Sufficiency and decided that Separation fits the purpose of our study (achieve equal FPRs and FNRs across all groups belonging to each protected class) and the dataset that we have. We then implemented two tech-

niques to achieve Separation, i.e. post-processing and optimisation during training. Under the post-processing technique, we considered two methods, namely Equalised Odds and Equal Opportunity. Correspondingly, under the optimisation during training method, we implemented two models for each crime, one with both FPR and FNR constraints, and another one with only FNR constraint, to compare with Equalised Odds and Equal Opportunity respectively.

We found that across the crimes in general, fair models recorded much lower values of FNRs, $D_{FNR}s$ and $D_{FPR}s$ compared to the corresponding unconstrained baseline models. However, the latter registered higher accuracy values compared to the former, as expected since there is a trade-off between accuracy and fairness. Nonetheless, in our study, achieving a sufficiently low overall FNR should be prioritised over achieving an overall high accuracy, since our dataset is highly imbalanced. Although prioritising to classify most victims correctly will be traded off with a higher misclassification rate for the non-victims - the police force may need to spend more in support provision as there will also be individuals wrongly classified as victims - in line with our EpiCrim approach to this issue, it is more morally acceptable than to let individuals susceptible to crime to not receive the support that they deserve.

After deciding that indeed, the fair models performed better than the unfair models in terms of our chosen performance metrics, we then compared the performances between the two fairness methods. The fair classifiers achieved via optimisation during the training process recorded slightly higher accuracies compared to the corresponding fair classifiers achieved via the post-processing method, and they were also able to remove disparate treatment simultaneously alongside disparate mistreatment. However, in terms of our main performance metrics, i.e. overall FNR, D_{FNR} and D_{FPR} values, the post-processing method outperformed the optimisation during training method, implying a better and fairer decision-making process between the four subgroups. In addition, the post-processing technique can be applied after any classifiers without having to retrain them, making the fairness implementation process less invasive, easier to implement, and cost-savvy.

Therefore, our final chosen models across the four crimes are the fair Logistic Regression models achieved via post-processing techniques. Choosing between the two post-processing techniques, i.e. Equalised Odds and Equal Opportunity is a trade-off between unfairness in terms of FPR and overall accuracy - the former includes constraints on both, FPR and FNR, whereas the latter only has constraint on FNR; Thus, the former would have smaller $D_{FPR}s$ compared to the latter, but lower overall accuracy. Therefore, the final decision depends on the aim of the end users of our models - whether they are willing to sacrifice some accuracy to gain fairness in terms of FPR, or vice versa.

3.2 Guidelines for End Users

Building on our main findings presented in Section 3.1, here we include guidelines to access and use the models and results presented throughout Section 2.4. Since our final chosen models are the fair models implemented via post-processing techniques, we are first going to outline the steps that need to be taken to use our post-processed models, followed by some guidelines to use our codes to implement fairness via the optimisation during training technique, should the end users be interested in that method as well.

To use any one of our post-processed or optimised during training models for predicting individuals who are susceptible to the four types of crime, the end users must first ensure that they have access to all the significant attributes (refer to Section 2.4.2) for the individuals that they are trying to classify into potential victims of crime or not; Since our models are trained on individuals sampled in Britain, they are only generalisable to the population in Britain. Please refer to our GitHub page as linked for our codes.

Specifically for the post-processing methods, once the details of the individuals are fed into the models via the final equations in Section 2.4.2 and the probabilities of the individuals being susceptible to crime are calculated, to classify them into potential victims or not, the end user should compare with the thresholds reported in Table 6. For Equal Opportunity predictors, if the probabilities calculated exceed the thresholds, then classify to 1 (susceptible to crime), else, classify to 0 (not susceptible to crime). For Equalised Odds predictors, if the probabilities calculated are below t_l, then classify to 0, if they are above t_u, classify to 1, and if they are between t_l and t_u, classify to 1 with probability p.

In our report, we focused on two binary sensitive attributes, namely gender and race, but our framework can be easily extended to multiple categorical variables, even those with more than 2 levels, but it is not advisable to have so many subgroups for small datasets with uneven distribution of observations across the subgroups, as this may cause bias in the final results. Additionally, should other end users be interested in our general methodology, the framework we introduced in this report can also be used alongside other datasets by training models following our methodology outlined in Section 2.3, and imitating our codes in the file titled 'Modelling_final.R' uploaded onto our GitHub page.

Next, we provide guidelines to use our fair models implemented via the optimisation during training method. Please refer to the 'Opt_during_training.ipynb' file on our GitHub page for the full code implementation - all of the codes have been well commented, to ease reference. Via this method, we effectively achieved fairness in terms of disparate mistreatment and disparate treatment simultaneously across our four types of crime. We removed disparate mistreatment via 4

		Equalised Odds			Equal Opportunity Threshold
		Randomisation Probability (p)	Lower Threshold (t_l)	Upper Threshold (t_u)	
Household Crime	White Males	0.5000	0.0875	0.0892	0.0892
	Non-white Males	-	0.1066		
	White Females	0.3000	0.0830	0.0964	0.0964
	Non-white Females	0.5000	0.1079	0.1196	0.1196
Personal Crime	White Males	0.7000	0.0168	0.0285	0.0285
	Non-white Males	-	0.0226		
	White Females	0.6000	0.0163	0.0248	0.0248
	Non-white Females	0.6000	0.0216	0.0305	0.0305
Threats	White Males	0.4000	0.0160	0.0205	0.0205
	Non-white Males	-	0.0195		
	White Females	0.4000	0.0165	0.0201	0.0201
	Non-white Females	0.7000	0.0198	0.0400	0.0281
Domestic & Acquaintance Violence	White Males	0.7500	0.0049	0.0085	0.0085
	Non-white Males	-	0.0044		
	White Females	0.7500	0.0051	0.0089	0.0089
	Non-white Females	0.6000	0.0049	0.0066	0.0066

Table 6: The randomisation probabilities p are calculated using the method introduced in Section 2.3.6. As the thresholds for the Non-white males (worst-performing group) are chosen as the baselines for both Equalised Odds and Equal Opportunity predictors, there is no need for randomisation for that group under the Equalised Odds method, and the numbers are the same across both types of predictors.

types of misclassification constraints, i.e. OMR, FNR, FPR, and both FNR and FPR. These constraints can be incorporated into Logistic Regression formulas as convex-concave constraints. The constrained optimisation problem can be solved using recent advances in convex-concave programming (Shen et al, 2016) as elaborated in the file 'Opt_during_training.ipynb'.

Choosing either one of the four measures is up to the end users, depending on the purpose of their study. In our codes, instead of allowing only one sensitive attribute as implemented in the paper by Zafar et al (2017a), we extended the codes to allow for as many binary sensitive attributes as the users would like to add - refer to the codes for a better understanding. Additionally, the trade-off between accuracy and fairness can be adjusted by choosing a different value for the covariance threshold c; A smaller c leads to a fairer classifier, although often at a small cost in terms of accuracy. Setting $c = 0$ would output the fairest model. Our program is designed to generate the following outputs for our dataset: overall accuracy, FPR, and FNR across the four subgroups

(White males, White females, Non-white males, and Non-white females). To calculate our chosen measure of fairness i.e. D_{FNR} and D_{FPR}, the user should calculate the biggest absolute difference in FNR and FPR respectively between any pair of the four subgroups. Our program has also been designed to achieve fairness with respect to disparate treatment simultaneously, by excluding the sensitive attributes at decision-making time.

3.3 Future Research

One of the potential avenues for future research is to extend the fairness procedures we implemented to models other than Logistic Regression. In our case, we only applied fairness procedures to our Logistic Regression model, as it outperformed the rest of the models we explored. However, it would be interesting to observe how the other models would perform under fairness constraints. For post-processing methods including Equalised Odds and Equal Opportunity, the same procedure can be applied to any other classifier as long as we have their ROC curves. However, when it comes to achieving fairness via optimisation during the training process, the method is task-specific. Therefore, in order to train a different fair classification model, the optimisation problem would change and thus, the classifier needs to be modified, implying that most of the codes we wrote would also need to be modified.

Given that we have constructed screening tests with high sensitivities for the 4 types of crime, another potential avenue for future research would be to construct diagnostic tests with higher specificities to filter out the misclassified non-victims. This is especially important so that our end user, i.e. the TVVRU can provide support to a smaller set of true potential victims, thus reducing the cost of support provision. The cost of constructing a diagnostic classifier is expected to be much larger - we would need access to more detailed information on individuals' social relationships, arrest, abuse, and family histories (Glasser et al (2001) suggested strong correlations between being victims and perpetrators of crime) and many other sensitive information in order to build more predictive and accurate diagnostic classifiers for the four types of crime.

Under the optimisation during training method, we only considered constructing fair classifiers without disparate mistreatment in terms of FPR and FNR. Other measures of disparate mistreatment such as FDR and FOR (Table 3) were not considered in this report because of the high computational complexity (these measures were also not implemented in the original paper by Zafar et al (2017a) due to a similar reason). However, it would be good to explore these other measures of misclassification in future work.

Next, our codes for the optimisation during training method are designed for binary sensitive attributes (even Zafar et al (2017a) only catered for binary attributes), and so, work well for our two

binary sensitive attributes (gender - males and females; race - Whites and Non-whites). However, in most scenarios, we would expect to have multi-level categorical sensitive attributes. Even in our case, we grouped race into Whites and Non-whites because of the comparatively small number of individuals other than the Whites being sampled. Should we have access to more individuals from the minority group, we may be able to further break them up into their respective ethnicities, resulting in a multi-level categorical sensitive attribute. Thus, generating a classifier that ensures fairness even with respect to sensitive features having multiple levels would be an interesting avenue for future work.

4 Bibliography

1. Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). Machine Bias. [online] ProPublica. Available at: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [Accessed 9 August 2020].

2. Badrick, T. and Turner, P. (2015). Review and Recommendations for the Component Tests in the Liver Function Test Profile. Indian Journal of Clinical Biochemistry, Vol. 31 (1), pp 21-29.

3. Barocas, S., Hardt, M. and Narayanan, A. (2019). Fairness in Machine Learning. fairmlbook.org.

4. Barocas, S. and Hardt, M. (2017). Fairness In Machine Learning NIPS 2017 Tutorial. [online] Available at: <https://vimeo.com/248490141> [Accessed 9 August 2020].

5. Barocas, S. and Selbst, A.D. (2016). Big data's disparate impact. Calif. L. Rev., Vol. 104, pp 671.

6. Bellis, M., Ashton, K., Hughes, K., Ford, K., Bishop, J. and Paranjothy, S. (2015). Adverse Childhood Experiences And Their Impact On Health-Harming Behaviours In The Welsh Adult Population. Public Health Wales NHS Trust.

7. Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of Machine Learning Research Vol. 81, pp 1–11.

8. Christmas, H. and Srivastava, J. (2019). Public health approaches in policing. Public Health England.

9. Fisher, B. S. (1995). Crime and fear on campus. Annals of the American Academy of Political and Social Science, Vol. 539, pp 85–101.

10. Fox, K., Nobles, M. and Piquero, A. (2009). Gender, crime victimization and fear of crime. Security Journal, Vol. 22(1), pp 24-39.

11. Gibson, C.L., Zhao, J., Lovrich, N.P. and Gaffney, M.J. (2002). Social integration, individual perceptions of collective efficacy, and fear of crime in three cities. Justice Quarterly, Vol. 19(3), pp 537–564.

12. Glasser, M., Kolvin, I., Campbell, D., Glasser, A., Leitch, I., Farrelly, S. (2001). Cycle of child sexual abuse: Links between being a victim and becoming a perpetrator. The British Journal of Psychiatry, Vol. 179 (6), pp 482–494.

13. Hedderman, C. and Hough, M. (1994). Does The Criminal Justice System Treat Men And

Women Differently?. Research and Statistics Dept United Kingdom, p.4.

14. Jennings, W.G., Gover, A.R. and Pudrzynska, D. (2007). Are institutions of higher learning safe? A descriptive study of campus safety issues and self-reported campus victimization among male and female college students. Journal of Criminal Justice Education, Vol. 18(2), pp 191–208.

15. Kim, N., Yasmineh, W., Freier, E., Goldman, A. and Theologides, A. (1977). Value of alkaline phosphatase, 5'-nucleotidase, gamma-glutamyltransferase, and glutamate dehydrogenase activity measurements (single and combined) in serum in diagnosis of metastasis to the liver. Clinical Chemistry, Vol. 23 (11), pp 2034-2038.

16. Lanier, M. M. (2010). Epidemiological Criminology (EpiCrim): Definition and Application. Journal of Theoretical and Philosophical Criminology, Vol. 2 (1), pp 63-103.

17. Lanier, M. M. and Akers, T. A. (2009). "Epidemiological Criminology": Coming Full Circle. Am J Public Health, Vol. 99 (3), pp 397-402.

18. Lum, K. and Isaac, W. (2016). To predict and serve?. Significance, Vol. 13, No. 5, pp 14-19.

19. Mallicoat, S. (2018). Women, Gender, And Crime. SAGE Publishing.

20. Marmot, M. (2010). Fair Society, Healthy Lives. [London]: Marmot Review.

21. Mcmanus, J. (2018). What exactly is a public health approach to crime and disorder reduction? [online] Available at: <https://jimmcmanus.wordpress.com/2018/10/03/what-exactly-is-a-public-health-approach-to-crime-and-disorder-reduction/> [Accessed 9 August 2020].

22. Murphy, K.P. (2012). Machine learning: a probabilistic perspective. MIT press.

23. Office for National Statistics. (2017-2018). Crime Survey For England And Wales (CSEW) Dataset User Guide: Adults Aged 16 And Over. UK Data Archive Study Number 8464.

24. Reiss, A.J. (1981). Victim proneness in repeat victimization by type of crime.

25. Shen, X., Diamond, S., Gu, Y. and Boyd, S. (2016). Disciplined convex-concave programming. 2016 IEEE 55th Conference on Decision and Control (CDC).

26. Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), Vol. 58 (1), pp 267-288.

27. Zafar, M.B., Valera, I., Gomez-Rodriguez, M. and Gummadi, K. (2015). Fairness Constraints: Mechanisms for Fair Classification. AISTATS.

28. Zafar, M.B., Valera, I., Gomez-Rodriguez, M. and Gummadi, K. (2017a). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. WWW.

29. Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., Weller, A. (2017b). From parity to preference-based notions of fairness in classification. Advances in Neural Information Processing Systems, pp 229-239.

30. Zou, H. and Hastie, T. (2005). Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 67 (2), pp 301-320.