Human Centred New York Learning: Evaluating Parole Hearings of New York City

Lucas G. S. Félix
lucasgsfelix@gmail.com
Department of Computer Science - DCC
Belo Horizonte, Minas Gerais

Flávio Diniz Figueiredo
flaviodf@dcc.ufmg.br
Department of Computer Science - DCC
Belo Horizonte, Minas Gerais

ACM Reference Format:
Lucas G. S. Félix and Flávio Diniz Figueiredo. 2018. Human Centred New York Learning: Evaluating Parole Hearings of New York City. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Currently, we live in a digital era were every person and device is constantly connected. This people and gadget integration is proportionate by the expansion and popularization of internet in recent years [22], generating each day more and more bytes of data. Hereof, estimates that until the year of 2020, 40 trillions gigabytes of data will be produced [1], been companies that work with information management, as Google, Facebook, Microsoft and Apple, worth billions of dollars. Thus, data driven analysis has been common thing to researches and companies, not limited to those. Some of then committed to the task of transform data in information more simpler and efficient, given the challenge that is organize, treat and find the best way to extract information of massive amount of data.

There are several cases where the apply of data has become a success, bringing insights and helping optimize customers services. One example in users everyday is Google with his search engine. Google become each day a well oiled machine, with more and more precise recommendations according to users need, maximizing customer satisfaction in each one of Googles products [2]. Another sign of data assisting in evolution is the so-called smart cities. Several big cities has been doing major investments in data analysis to make better quality-of-life aspects [9, 16].

Nevertheless, not only from online data lives the man. Many researches, governmental and companies offices has spreadsheets and databases with not open data. This data can be well applied and give companies advantage over it competitors and knowledge about it costumers. Also, it can be accurately used by governmental offices to measure and efficient direct resources. However, this datasets can have many sensitive features that when poor employed can lead to misinformation, and even discrimination. Thus, as success cases, there are many examples where data has been poorly applied. One major example is Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a software used as support tool to courts in U.S. counties [8]. Based on a set of questions, *COMPAS* measures the probability of a person relapse in his/hers laws infractions. Recently, *COMPAS* has been recently widely criticized for it bias over people skin color [3]. Bias problems occurred in ML models are due poor input and data treatment, as stated in [23], "bias in, bias out". Moreover, historical data is usually biased due societies behavior, however, with evolution of thought, bad behaviors and bias are no longer accepted.

To study and assist soft problems when using Machine Learning (ML) algorithms a new venue has rise, so-called Human Centred Machine Learning (HCML). This new field efforts to study bias, fairness, justice, interpretability in ML models. As pointed in [17], "Examining machine learning from a human-centered perspective includes explicitly recognising this human work, as well as reframing machine learning workflows based on situated human working practices, and exploring the co-adaptation of humans and systems". Thus, give a better understanding to ML models, using it as a assisting tool, not universal truth.

Given this facts, in this work we use HCML techniques to make a study over the Parole Hearings in New York City (NYC). Parole is a early release of a criminal given that he/she agrees to certain conditions. Only in the state of New York, over $10,000$ parole eligible prisoners has freedom denied each year. And year this criminals cost about $\$60,000$ to be held in prison, while even more is spend to treat ill prisoners. Besides that, the parole process in unclear, not been transparent the parole commissioners decision, most of them argue that a individual verdict is only based in his crime [7]. Thus, **we want to investigate if other features, more specifically sensitive and geo-spatial attributes, can interfere in the decision of the parole commissioner**.

More specifically, our goal is addressed to the following research questions (RQs):

(1) It is possible to predict the change commissioner decision ?
(2) Which features are most important to define the sentence of a criminal ?
(3) Joining different datasets, we can better understand the sentences ?

To address these questions properly we join several different NYC datasets taken from Kaggle [1]. All the datasets were pre-processed and joined, forming two datasets, one with our target feature and one with auxiliary attributes that can assist to properly explain the parole commissioner decision. Data from other cities of the New York state were removed, remaining only NYC. The analyzed portion of the New York state can be seen by the Figure 1.

[1] https://www.kaggle.com

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Jurisdiction		Population	Area	
Borough	County	April 1º, 2010 Census	In square Miles	In square KM
(1) Manhattan	New York	1 585 873	23	59
(2) Brooklyn	Kings	2 504 700	71	183
(3) Queens	Queens	2 230 722	109	283
(4) Bronx	Bronx	1 385 108	42	109
(5) Staten Island	Richmond	468 730	58	151
New York City		8 175 133	303	786
New York State		19 378 102	47 214	122 284

Source: United States Census Bureau

Figure 1: New York Boroughs

To properly answer this questions, we make a study of the datasets using Human Centred Machine Learning tools. More specifically, we address this questions to a *Random Forest/Logistic Regression* models, verifying the interpretability/predictability of the parole commissioners decision. Moreover, we analyze how this models behave over some changes in the dataset.

2 BACKGROUND AND RELATED WORK

In this section, we analyze relevant works to our proposal. Given that our major aim is to investigate if sensitive and geo-spatial features can interfere in the decision of the parole commissioner, we evaluate works that relate to our in 2 ways: (*i*) Human Centred Related and (*ii*) Evaluation of criminal justice using Machine Learning techniques. We highlight that in some works this two sides are merged, turning the work high related with ours.

Currently, machine learning and data analysis have been widely applied in most society fields. From agriculture until traffic ML has been performed aiming to solve problems and facilitate in populations difficulties. Thus, algorithms has now been applied in human venues, and taking decisions accordingly to historical data in persons place. Hence, although computational resources have been assisting in the humans progress, it can carry humans problems, and take it to the decisions made by the algorithms, the so-called bias.

Aiming to solve problems that can be provoked by ML models and directly interfere in a persons life, a Human Centred Machined Learning (HCML) field became a hot topic in research field, turning in a way to step aside from this problems. In this section, we evaluate works related to ours in the sense of criminal justice evaluation using a ML methodology. It is high valued that not necessarily this productions apply HCML techniques, but, for treating the same kind of problem with different perspectives it is interesting to give a literature overview and point failures that this work can assist to solve.

Directly related with this work the proposal [11], evaluates the impact in the use of ML systems to assists parole release decisions in Pennsylvania state, U.S.. The author points, that risk assessments have been use in the U.S. since 1920*s*, and is treat as sensible and routine task. However, recently,controversies have surround this matter, and since them studies have been evaluating risk assessments for a variety of criminal justice decisions. To assists on the task of parole decisions which majorly uses risk assessments, the government of Pennsylvania proposed the use of ML models using a dataset labeled by a criminal justice professor. Th aim of the model was to predict the probability of a inmate relapse in crime after parole release. In their results Random Forest was the model with best performance. Thus, the aim of the paper was to verify how the use of a ML model can interfere in parole release decisions. In their results, the author shows that the forecast had no effect on the overall parole release rate, but alter the mix of released criminals. Thus, distinctions were made of the inmates by the algorithm considering their committed crimes, also, the results shows that a parole release decisions have a bigger probability when the crimes committed is considered non-violent. Nevertheless, this work does not considered any metrics of HCML, although is used sensitive features as race.

The work of [25], uses a dataset of parole hearings of New York City. The authors aim is only to predict the result of parole hearing. To complete the task is used in the proposed methodology a Artificial Neural Network, achieving a accuracy of 76.8%. The authors does not make clear were the dataset was gathered, however, by the features we believe that is the same dataset as ours. Nevertheless, this work differ from our for some reasons: (*i*) Is used a single dataset with a reduced number of features and instances, (*ii*) does not consider HCML metrics and do not discuss it, although using sensitive features as race and gender, (*iii*) does not consider other features outside the original dataset.

In [13], the authors point how the subjective judgment, an approach that on intuition guided by experience, can be widely inaccurate and unclear, even more when applied in risk assessments problems. The work, even tough is a law field paper, points to the problem of historical data training ML models, given that datasets has subjective judgment leading to limited retributivist ideologies, which are a idea of application a sentence aiming only on justice, not in punishment.

The proposal in [26] points towards the problem of finding the best model to predict the chances of criminal recidivism. To perform the work, the authors split the data in three types of recidivism: general, violent and sexual, using in the methodology classic statistics and machine learning models and comparing them. The applied in this study was gathered from the *Dutch Offender's Index*, in spite of using only gender as sensible feature, another important, that can be discriminatory feature is the nationality of the criminal. However, as the works aim is to find the best model to predict recidivism, the author make a deep study over the features, but focus in the models performance.

In [18], the authors presents how automates systems can improve and also mix up in risk assessment assignments, also raising complex legal and ethical questions about adequately capturing a individual specific circumstance. Some view points in the US aim that justice has to be individual, and when insufficiently personalized can aggravate injustice. To mitigate worries, most courts apply risk assessment tools to inform, but not to completely replace, judicial determinations. The author indicates also the problems of using historical to feed ML models, given that populations can change over time making the models outdated. To better evaluate these points, the authors analyze the benefits of personalized risk assessment tools for law enforcement, the data applied is from New York City on gun bearing crimes. The work aim is to assist in police officers in evaluate risk in real time, and it is highlighted for also make a summary of heuristics that could assist to softy the use of ML models by law enforcement. Nevertheless, the model proposed by this work fails on solve racial disparities having some biased results, that could lead to excess of police stop towards some races.

The proposal of [19], analyzes how proposals of fair risk assessments in the literature can be unfair and restrain efforts to reform the criminal justice system. The work suggest some ways that ML models could expand to better work in risk assessment problems as the use of unbiased data. The author criticizes the role of the computer scientist as trying to fix something that should be rebuild from scratch, although recognizes the value of interdisciplinary view points.

The work of [12], discuss fairness in criminal justice risk assessments evaluating the lack of conceptual precision, given the amount of fairness metrics in this field. The authors shows that some of the fairness metrics conflict with one another and with accuracy, although it has a broadly similar intent, but with differs in substantive and technical details. More important, the author points that fairness and accuracy are impossible to maximize at same time, been necessary to consider trade-offs. Bellow, we summary the metrics broach in the work:

- **Overall Accuracy Equality (OAE):** The definition of this metric assumes that true negatives are as desirable as true positives. Thus, the overall procedure accuracy is the same for each class of sensible features. *Example:* Considering a parole problem where a inmate can succeed the parole or not and the sensitive feature is the race R, where we have $r_i \in R$. Hence, OAE aim that the accuracy of $r_k == r_j$, independent of the amount of true negatives or the overall accuracy.
- **Statistical Parity:** The definition of this metric assumes that the marginal distributions of the prediction has to be same for each class of sensible features. *Example:* Considering a parole problem where a inmate can succeed the parole or not and the sensitive feature is the race R, where we have $r_i \in R$. Hence, The amount of parole succeed in r_k should be the same in r_j.
- **Conditional Procedure Accuracy Equality:** The definition of this metric assumes that the conditional procedure accuracy is the same for the sensitive features. Example: Given a target feature $\hat{y}$, where a set $c \in \hat{y}$ and a sensible feature S, and a set of classes $s \in S$, them the conditional accuracy a for each class s has to be the same.
- **Conditional Use Accuracy Equality:** The definition of this metric assumes the probability of correct prediction for each class of sensible feature has to be the same. *Example:* Considering a parole problem where a inmate can succeed the parole or not and the sensitive feature is the race R, where we have $r_i \in R$, the probability of correctness P for each class $r_i \in R$ have to be equal.
- **Treatment Equality:** The definition of this metric assumes that the ratio of false negatives and false positives are the same for all sensitive features. *Example:* Considering a parole problem where a inmate can succeed the parole or not and the sensitive feature is the gender G, where we have $g_i \in G$. Hence, the amount of times that a algorithm is mistaken for a gender g_k has to be same for a gender g_j.
- **Total Fairness:** The definition of this metric achieves all previous presented metrics equally.

The main contributions of this work are a (i) study case of real New York City datasets. (ii) application of HCML techniques aiming that could be use to risk assessment/parole forecast and (iii) extensive literature review pointing the main guidelines to evaluate criminal justice problems.

3 PROPOSED METHODOLOGY

This work consists in a exploratory study that apply data mining and machine learning techniques, aiming to investigate if features, in special sensitive features (as race and gender) and geo-spatial attributes, can interfere in the decision of the parole commissioner in NYC. To do so, we propose the following methodology. We high-value that the main focus of this work is not prediction accuracy, but visualize the principle concepts of HCML and study it over several datasets of New York City.

3.1 Datasets

To perform this work, we utilize several New York city and state datasets, gathered from *Kaggle*. The base dataset of our work is the **Parole Hearings** dataset, which has several features that lead our work as race, gender, committed crimes, as place where the law infractions happen. To assist the proposed analyzes and verify correlated factors that could lead to the decision in the parole hearings was joined with the first dataset 3 other bases. Each dataset has complementary information that can assist to a better understanding of the decision taken by the parole commissioner. Below, we present each one of the datasets as make a brief description of each one of it. And in Table 1, we characterize each one of the used datasets showing it amount of instances and features.

- **Parole Hearings:** This dataset has information about parole interviews in the New York State. Between the features in the set are sensible variables as *Race and Gender*, also present crimes committed, local where it was committed and sentence. The *Parole Hearings* dataset is the main dataset used in our analysis [7].
- **Housing:** This dataset has information about housing in NYC. With this data it is possible to find socio-economic needs in each borough and verify it relation with the crimes that happen [6].

- **Census:** This dataset has about a 5 years estimation census of NYC. Between the features are total population, racial/ethnic demographic information, employment and commuting characteristics. With *Census* data it is possible to infer some neighborhood characteristics that could lead to the final sentence decision [4].
- **Crimes:** This dataset has the happened crimes from 2014 – 2015 in NYC. As the first dataset, it has which crime and where it happen assisting in it study. The main difference from the *Parole Hearings* dataset is that *Crimes* does not contain any sensitive data [5].

Datasets	# of Instances	# of Features
Parole Hearings	43519	46
Housing	3287	41
Census	2167	36
Crimes	1048575	24

Table 1: Characterization of the datasets applied in the proposed methodology

3.2 Pre-Process

In the pre-process step, we work to remove unnecessary information and join all datasets aiming to retain as most data as possible. We high value that this step is vital to achieve results better as possible. As we work with several datasets, our task was hampered by the many features presented, thus, the datasets was joined in different stages. Bellow, we show each step of joining and pre-process the datasets:

(1) In the first step we removed all non New York City related data. Leaving only data from the five boroughs that compose the city. This process was made only in Parole Hearings dataset, all the other dataset had only New York City in it.
(2) Second, we consider relevant and no-relevant features for our study, removing those with we judge would not help in our case. This process was made in all datasets.
(3) In the housing dataset we had several blocks with the same *id*, but with different geo-spacial points (i.e. latitude, longitude). Thus, to get around this issue, we join the blocks making the mean latitude-longitude for all points with the same block *id*..
(4) After, we pass to the stage of joining datasets. To join the **Housing, Census and Crimes** datasets we use the geo-spatial features present in the data. As the latitude and longitude points are not the same and have small variations, was calculated the distance between the points of the datasets using the Equation 1. So, given a housing $h \in H$, where H in the set of all housing data, $cinC$, where C is set of all crimes committed and $ce \in Ce$, where Ce is the set of available Census data. Them, a housing h_i is joined with a crime c_k and a census ce_j, if a the distance between it latitude and longitude points are the minimum as possible.

$$\begin{aligned} earth\_radius &= 6373.0 \\ dist\_lat &= lat_b - lat_a \\ dist\_lon &= lon_b - lon_a \\ a &= \sin(\frac{dist\_lat}{2})^2 + cos(lat_a) \times \cos(lat_b) \times \sin(\frac{dist\_lon}{2})^2 \\ b &= 2 \times \text{atan2}(\sqrt{(a)}, \sqrt{(1-a)}) \\ distance &= earth\_radius \times b \end{aligned} \quad (1)$$

3.3 Methods

Bellow, we present the machine learning models applied in our methodology. Our aim is to predict the parole commissioners decision, and verify what are the most important features to define if a criminal will be released or not:

Decision Trees(DT): This model works by making classification rules organized as tree paths [10]. In a overview, decision trees are acyclic graphs, where the internal nodes are the dataset features, the branch nodes are the result and each leaf node is a class [20].

Decision Trees add nodes in the Tree based on Entropy and Information Gain metrics. Thus, a ranking of more appropriated instances is created based on this metrics [24]. Bellow we define this two metrics:

- **Entropy:** This measure defines the average rate of information produced by the data, given by the Formula:

$$Entropy(S) = -(p_+ \times \log_2 p_+) - (p_- \times \log_2 p_-) \quad (2)$$

  Where p_+ are the positive instances and p_- are the negative instances.
- **Information Gain:** This measures defines the expected entropy loss caused by the instances split. Information Gain is given by the Formula:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Instances(A)} \frac{|S_v|}{S} Entropy(S_v) \quad (3)$$

  Where $Instances(A)$ is the set of possible values to the feature A, and S_v is a subset of S, where A has a value v.

Nevertheless, in our work we opt for a more robust, but yet simple based on the decision trees, the so-called **Random Forest**. This technique train a set of N trees, considering random segments of the dataset. Next, these N trees are joined creating a unique tree [14].

To compare with the RF, we use **Logistic Regression(LogR):** The logistic regression model estimates the probability p of a target feature, calculating how this p is affected by independent variables. The model is given by the bellow Formula 4:

$$y = \frac{L}{1 + e^{-k(x-x_0)}} \quad (4)$$

Where e is the euler natural logarithm, x_0 is the x-value of sigmoids midpoint, L is the curve maximum value and k the logistic growth rate or steepness of the curve [21].

We choose this model for it simple, yet effective premises been models with easy interpretability. In both models, it is possible to

better understand the output, and how the algorithm achieve the answers.

4 RESULTS

In this section, we present and discuss the results achieved with the proposed methodology. First we show the parole hearings the dataset, and the effectiveness of the pre-processing in models results. Second, we evaluate the models behavior when permuting sensible features (as gender, ethnicity and borough). Last, we justify the results achieved in the first dataset, with the combination of several New York databases.

4.1 Pre-Processing and Models Results

In our experimental evaluation the main dataset is the "New York State Parole Hearing", our main task in this dataset is to predicted the parole commissioner decision which can be given by several classes, defined in [2]. In our dataset there are 12 of this classes, we focus in two of then "DENIED" and "GRANTED", due the fuzzy definition of the other classes. In other words, this two labels define if a prisoner have his parole requisition granted or not. In most of the cases this requisitions are denied making the original dataset unbalanced and our task harder. In Figure 2, it is possible see to the number of instances per class in the original dataset before pre-processing, and also how some of the classes are related (as Denied and Not Granted), some of the features have missing information (* or **********).

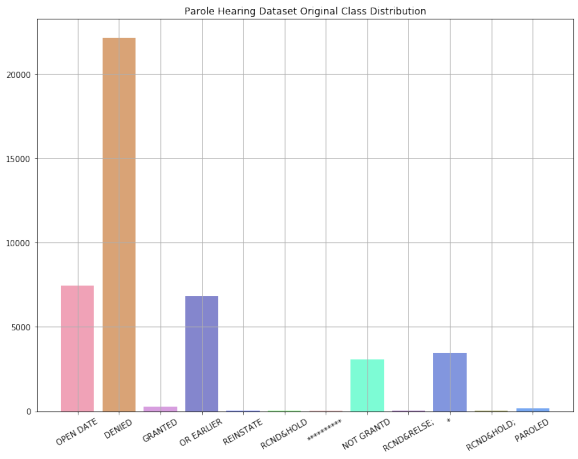

Figure 2: Amount of instances per class in the original dataset

To overcome the problem of unbalanced classes, was removed classes that made no sense (considering we have no law background) and join related classes (i.e. classes which mean the same thing). In the end, as pointed before, we predict only two classes. Finally, was applied a pre-processing technique to over-sample the minority classes and achieve better results with the models. We apply in our methodology SMOTE: synthetic minority over-sampling technique [15], a over-sampling technique based on nearest neighbors (kNN) measuring the euclidean distance between data points in feature space.

To evaluate the models was used a stratified cross fold validation with $k = 10$. The scoring metrics was *Macro-F1* when multi-class and *Binary F1-Score* when having only two labels. This two metrics was selected given that it penalize more when the model fails in more rare classes. Thus, in Table 2, it is presented the achieved results in each model, Random Forest and Logistic Regression, with and without pre-processing. We high-value that the over-sampling steps was not made in the test sets, only in train, given that if we test with generated samples the model would be biased.

Model	# of Classes	# of Instances	Oversampling	Metric	Result
Random Forest	10	12395	No	Macro-F1	0.32
Logistic Regression	10	12395	No	Macro-F1	0.12
Random Forest	10	15268	Yes	Macro-F1	0.32
Logistic Regression	10	15268	Yes	Macro-F1	0.10
Random Forest	2	7906	No	Binary-F1	0.63
Logistic Regression	2	7906	No	Binary-F1	0.51
Random Forest	2	15268	**Yes**	**Binary-F1**	0.62
Logistic Regression	2	15268	Yes	Binary-F1	0.10

Table 2: Result Achieved with the Models

Hence, our first Research Question (RQ1): *It is possible to predict the change commissioner decision ?*, is answered. Thus, we know it is possible to predict if parole commissioner decision, nevertheless, it is a hard question and the used models do not achieve many strong results. The best achieved result was with Random Forest, when predicting only two classes and using a over-sampling technique. However, even with a over-sampling the model still lacks in predicting granted paroles.

In our vision, even applying several over-sample algorithms (SMOTE was the one who show the best result), this technique do not perform well, making generic instances which do not help the classifier do separate each class. We one more time high-value that our aim with this work was not to achieve good prediction results, but evaluate if the dataset is biased. We believe that more strong algorithms as SVM, XGBoost or even Neural techniques can achieve better results.

To answer the Research Question (RQ2): *Which features are most important to define the sentence of a criminal ?*, we use the features importance given by the random forest to evaluate the main features used by the model to define the future of a criminal. We select this model given that the best result was achieved with it, when using two classes and making over-sampling. Following our hypothesis, that sensible features could influence in the models behavior, between the main features are the "Parole Board Interview Type" , "Crime Class" and "Race/Ethnicity". The Table 3, show that the original dataset is mainly balanced when evaluating the number of white and black people, it also possible to see the distribution after the first pre-processing.

Dataset	Amer Ind/Alsk	Asian/Pacific	Black	Hispanic	Other	Unknown	White
Original (New York State)	373	164	12128	5553	563	219	11054
Pre-Processed (New York City)	56	38	4386	2444	120	69	793

Table 3: Amount of instances per race

When evaluating the Table 3, it is possible to notice, that different from the original dataset where the number of instances between

[2] http://www.doccs.ny.gov/calendardatadefinitions.html

black and white are close, in the pre-processed database the amount of afro-descendants is bigger. This is justified by the fact that the original dataset has data from all New York State, while we want to evaluate only the New York City Crimes, which are most committed by people with black ethnicity within our dataset. Thus, this justifies the ethnicity feature between the main features. From the results achieved in this step we make a evaluation present in the next subsection, where the model behavior when changing the sensible features.

4.2 Sensible Features Evaluation

In this subsection we evaluate the model behavior when treating a single instance. The main goal of this subsection it is analyze if our model have a different behavior when treating people with different sensible features (gender, race), class of crime and local of the crime. The gender feature ranges from "male" and "female". The ethnicity ranges from with seven values, as we can see in Table 3. The crime classes ranges from A to E, been a total of 5 classes, where A are most serious crimes (as felony) and E are least serious crimes (as infractions and violations). Lastly, we evaluate if the local where the crime was committed can interfere. Thus, we evaluate if crime happens in one of the five New York City boroughs. To complete this task, we evaluate a set I of instances i, where each instance i as at least one of the described attributes different from other instances in the set. The set I, has two main subsets I_G with granted parole and I_N with denied parole.

In the end we have about 700 combinations which were tested in the Random Forest model. Due the scope of this work, it was not possible to put the results here. However, we high-value that our model is biased, not for the sensitive features, but when choosing a instance $i \in I_N$, the model denies the parole for all combination. The same thing happen when evaluating the set $i \in I_G$, even when the instance was setted with a high class crime (i.e. a felony or rape), the models output was Granted. Thus, we conclude that this features do not directly interfere on the model behavior, even tough "Class Crime" should interfere, in the parole commissioner.

Trying to better understand the parole officer decisions and the achieved results with the models, in the next subsection we evaluate the joined datasets and make comparisons between the two bases.

4.3 Criminal Justice Fairness Metric

To evaluate if the forecast made by the algorithm was considered to be fair, we use the metric *Conditional Procedure Accuracy Equality (CPAE)*, which was described in section 2. This metric assumes that the simple accuracy for each class of sensible features has to be the same. We choose to evaluate our model by this measure due it simplicity and interpretability, been the best suited between all evaluated options.

We analyze gender and race features, evaluating for it if the model is fair. To make this evaluation we train the model with 70 % of the dataset and test with the 30 % left. The results achieved are present in Table 4 and 5.

-	Amer Ind/Alsk	Asian/Pacific	Black	Hispanic	Other	Unknown	White
Accuracy	1	1	0.99	0.98	1	1	0.97

Table 4: Accuracy per class in race

-	Male	Female
Accuracy	0.99	0.954

Table 5: Accuracy per class in gender

Through the bellow tables, it is possible to see that the model are used it is consider to be fair by CPAE metric. We high-value that even tough it is not equal the values of accuracy the results are close, showing good results. Lastly, we point that this metric can evolve (as future work) given weights to labels on unbalanced class problems. Thus, despite the fact that we have good accuracy values, most of the errors happen in minority classes, and our results is mostly biased by the majority classes.

4.4 Joined Dataset

In this subsection we evaluate how joined datasets can help understand the bias on our dataset. As described in section 3, we joined different New York City datasets trying to answer our research questions and evaluate if geo-spatial attributes can interfere in the parole commissioner decision. First we show through Figures 3, 4, the amount of crimes per class in each dataset. It is possible to see that the first dataset has a amount of instances bigger than the joined dataset. Nevertheless, the proportion of crimes per neighborhood is keeped, where in both datasets in possible to see the New York borough as where happen most crimes, and Richmond where happens the least amount. One main advantage of the joined dataset is the location of where crimes happened, latitude and longitude. In Figure 5, it is possible to see the distribution of crimes through New York City. By the Figure it is possible to notice three regions were most crimes happen, New York, Kings and Bronx.

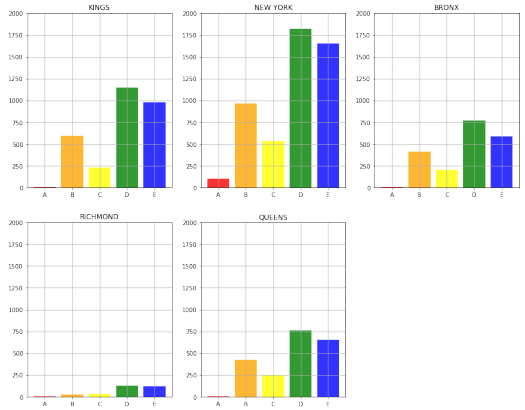

Figure 3: Amount of crimes per class in each neighborhood - Parole Dataset

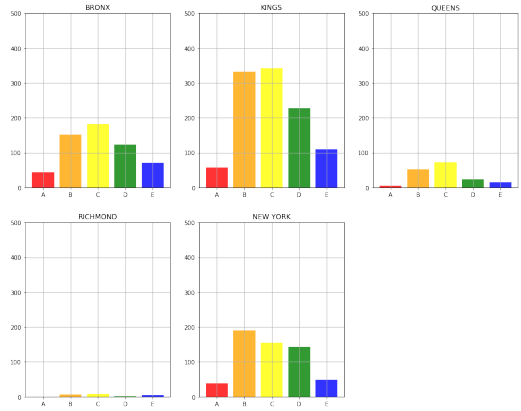

Figure 4: Amount of crimes per class in each neighborhood - Joined Dataset

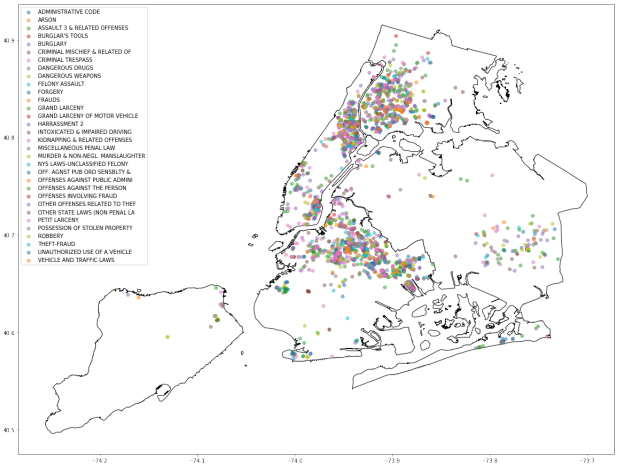

Figure 5: Distribution of crimes through New York City

To make a better evaluation for locality, we analyze how many Granted and Denied parole decisions were taken by county. In Table 6, it is possible to notice that as shown before, most of the counties have more denied decisions. Richmond is high lighted by amount of granted decisions having only one.

	New York	Kings	Queens	Bronx	Richmond
Granted	104	17	35	15	1
Denied	3108	1927	1267	1225	207

Table 6: Parole commissioner decision per county

Trying to better understand we evaluate sensible features for each county, we make a discussion aiming to answer the research question *"Joining different datasets, we can better understand the sentences ?"*. To make so we use census features from the joined dataset. First we high-value that our dataset is composed by the regions present in Figure 5. Second we point that correlation is not causality. Thus, we make a discussion leaving a open discussion, given that the datasets used in this work can be biased and lead to wrong conclusions. In Table 7, we show the census data used to discuss our results.

	New York	Kings	Queens	Bronx	Richmond
Asian Census	0.09	0.05 0.09	0.01	0.01	0.07
Native Census	0.016	0.019	0.019	0.028	0.068
Black Census	0.28	0.52	0.57	0.28	0.35
White Census	0.29	0.17	0.096	0.02	0.20
IncomePercap Census	40730	24695	25003	13583	20916

Table 7: Census data mean

Befitting with the pre-processed dataset (Parole Hearings), the Table 7, shows that most of the people in both of our datasets has a black ethnicity. New York is the county where the most crimes in all classes happen even tough the income per capita is the highest, showing that there is no relation between then. This can be justified by the fact that we are using only the income per capita mean, which is increased by multi-millionaires that live in Manhattan.

Lastly, we high-value that more analysis has to be made in each of the datasets, trying to better understand the parole commissioner's decision. By the end we see that the sensible and geo-spacial features do not made difference in the decision, hence there are other factors that were not consider in either of the datasets that can influence in this decision and can be considered in further works.

5 CONCLUSIONS AND FUTURE WORK

In this work we evaluate the parole hearings dataset, considering the city of New York. Our methodology considered two machine learning algorithms Random Forest and Logistic Regression, to predict the parole commissioner result. Our aim with this work was not to get the best result, but to verify if there was any bias in our model or the dataset. In the original dataset, without any pre-processing, it was possible to notice that the data was balanced between people of black and white ethnicity. However, after the pre-processing a discrepancy was notice, given that the original dataset had data from all New York State, while we evaluate only the city of New York, which has more black people living in the town. The result of the models shows that Random Forest could perform better in the dataset, when using SMOTE for oversample some unbalanced classes. Lastly, we discuss using data from a joined dataset with New York City data from multiple sources. In this discussion, it was not possible to see correlation between income per capita and the number of crimes. Lastly, we high-value that as pointed in some of the papers presented in the background section, risk assessment and criminal justice are naturally biased by the justice system of the United States (can serve to other countries). Thus, the aim has to be not in best performance algorithms and unbiased datasets, but aim to fairness decision by the judges, without considering sensible attributes of persons.

As future work, we aim to explore better the joined dataset, which as many features not used in our discussion. Try new over-sampling techniques, making a better pre-processing and try new interpretative models. Besides that, we aim to evaluate how justice metrics perform in our models, even tough we know that we would lose in statistic sores.

REFERENCES

[1] [n. d.]. BIG DATA BUSINESS, Os grandes e impressionantes números de Big Data. https://www.bigdatabusiness.com.br/os-grandes-e-impressioantes-numeros-de-big-data/. Accessed: 2018-07-21.

[2] [n. d.]. Google Shopping Adds Personalized Homepage and Price Tracking. http://socialbarrel.com/google-shopping-adds-personalized-homepage-and-price-tracking/121570/. Accessed: 2019-10-11.

[3] [n. d.]. Machine Bias Risk Assessments in Criminal Sentencing. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2019-10-14.

[4] [n. d.]. New York City Census. https://www.kaggle.com/muonneutrino/new-york-city-census-data. Accessed: 2019-10-11.

[5] [n. d.]. New York City Crimes. https://www.kaggle.com/adamschroeder/crimes-new-york-city. Accessed: 2019-10-11.

[6] [n. d.]. New York City Housing Units. https://www.kaggle.com/new-york-city/housing-new-york-units. Accessed: 2019-10-11.

[7] [n. d.]. Parole Hearings in New York State. https://www.kaggle.com/parole-hearing-data/parole-hearings-in-new-york-state. Accessed: 2019-10-11.

[8] [n. d.]. Practitioners Guide to COMPAS. https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf. Accessed: 2019-10-11.

[9] [n. d.]. Stories from the World of Municipal Analytics. https://towardsdatascience.com/stories-from-the-world-of-municipal-analytics-d3dc97077682. Accessed: 2019-10-11.

[10] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.

[11] Richard Berk. 2017. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology* 13, 2 (2017), 193–216.

[12] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.

[13] Richard Berk and Jordan Hyatt. 2015. Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter* 27, 4 (2015), 222–228.

[14] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[15] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[16] Michael Fitzgerald. 2016. Data-driven city management: A close look at Amsterdam's smart city initiative. *MIT Sloan Management Review* 57, 4 (2016).

[17] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. 2016. Human-centred machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 3558–3565.

[18] Sharad Goel, Justin M Rao, and Ravi Shroff. 2016. Personalized risk assessments in the criminal justice system. *American Economic Review* 106, 5 (2016), 119–23.

[19] Ben Green. 2018. "Fair" Risk Assessments: A Precarious Approach for Criminal Justice Reform. In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[20] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

[21] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. Vol. 398. John Wiley & Sons.

[22] Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. 2018. A feature-oriented sentiment rating for mobile app reviews. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1909–1918.

[23] Sandra Gabriel Mayson. 2018. Bias In, Bias Out. In *University of Georgia School of Law Legal Studies Research Paper No. 2018-35*. https://ssrn.com/abstract=3257004

[24] S. Russell and P. Norvig. 2004. *Inteligência artificial*. CAMPUS - RJ. https://books.google.com.br/books?id=wBMvAAAACAAJ

[25] Tribhuwan Singh, Yashvardhan Jain, and Vaibhav Kumar. 2017. Predicting parole hearing result using machine learning. In *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)*. IEEE, 1–3.

[26] Nikolaj Tollenaar and PGM Van der Heijden. 2013. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176, 2 (2013), 565–584.