Deep Interpretable Criminal Charge Prediction and Algorithmic Bias

Abdul Rafae Khan* [1] **Jia Xu*** [1] **Peter Varsanyi** [2] **Rachit Pabreja** [2]

Abstract

While predictive policing has become increasingly common in assisting with decisions in the criminal justice system, the use of these results is still controversial. Some software based on deep learning lacks accuracy (e.g., in F-1), and many decision processes are not transparent causing doubt about decision bias, such as perceived racial, age, and gender disparities. This paper addresses bias issues with post-hoc explanations to provide a trustable prediction of whether a person will receive future criminal charges given one's previous criminal records by learning temporal behavior patterns over twenty years. Bi-LSTM relieves the vanishing gradient problem, and attentional mechanisms allows learning and interpretation of feature importance. Our approach shows consistent and reliable prediction precision and recall on a real-life dataset. Our analysis of the importance of each input feature shows the critical causal impact on decision-making, suggesting that criminal histories are statistically significant factors, while identifiers, such as race, gender, and age, are not. Finally, our algorithm indicates that a suspect tends to gradually rather than suddenly increase crime severity level over time.

1. Introduction

Automatic crime prediction using deep learning algorithms has played an increasingly critical role in criminal justice systems and crime prevention efforts. Previous work, such as (Luo et al., 2017), shows that deep learning can assist decision-making by processing a very high volume of data that are difficult for human researchers to analyze efficiently. Despite the success of deep learning, it is still challenging to convince the general public to trust our algorithms to assist a judge's decisions due to a lack of interpretability and sufficient accuracy. In particular, criticism that bias exists for identifiers such as race and age points (Skeem & Lowenkamp, 2016) toward the need for a fair, transparent, and explainable crime prediction method.

Our work presents a criminal charge prediction problem and introduces a framework that leads to an accurate decision with post-hoc interpretability. A criminal charge is a formal accusation made by a governmental authority asserting that an individual has committed a crime. There are three primary classifications of criminal offenses: felonies (we call these level 1 crimes), including terrorism, murder, kidnapping, treason, elder abuse; misdemeanors (level 2), such as arson, extortion, threat, bribery, larceny; and infractions (level 3), the least severe crimes, including damaging property, burglary, smuggling, obscenity.

We aim to understand individuals' criminal behavior over twenty years and focus our study on whether there are patterns in a criminal's charge record over time. These charge records include a person's prior arrests, type of crime or crimes committed, prior convictions, and personal attributes of race, gender, and age. We look specifically at whether a person will receive a charge in the next two years and if yes, its level given the criminal history in the previous eighteen years and personal information. We also analyze each input feature's importance for prediction results as well as the changes in an individual's crime levels over time.

The first technical challenge in our study is the heavily *imbalanced data*. A fully connected feed-forward neural network can give very low recall and F-measure scores despite high precision. The negative samples are much fewer than the positive samples with the predictor tending to predict everything positive to minimize error. We apply an RNN along with the Bi-LSTM to capture temporal patterns of a suspect's criminal behavior over time, an approach that brings both high precision and high recall.

The second challenge is the explanation for the causality inference of our predictions. We analyze feature importance using values from attention weights learned from the attentional Bi-LSTMs. The weights learned while training and the node connections activated while predicting will indicate the importance of the input features concerning the decision reached during classification. Our results show that, contrary to the belief that a machine learning algorithm is

*Equal contribution [1]Department of Computer Science, Stevens Institute of Technology [2]Rutgers University. Correspondence to: Jia Xu <jxu70@stevens.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

biased, our algorithms learned that criminal charge histories are essential for prediction and that personal data such as race and age are minimally useful. Our algorithms therefore do not need to explicitly rely on personal information such as race, gender, and age that can cause biased decisions. Our result, however, does not indicate whether there is bias in the human raw dataset, a question that is outside the scope of this study. We summarize our three major contributions below:

1. We propose four research questions on an individual's future criminal charge prediction and introduce Bi-LSTM with attention to alleviate the imbalanced data problem and to interpret feature importance for decision-making.

2. Interestingly, our results show that criminal history plays a significant role in predicting a person's future crime and crime level, while personal identifiers, gender, race, and age, do not.

3. Our models indicate that a person is more likely to commit the same or a similar level of crime in the near future than a very different one.

In the following context, we will first describe our task and dataset, introduce our models, then demonstrate experimental results and interpret our models. Finally, we will discuss related work and conclude.

2. Task Description and Dataset

Dataset The real-life crime prediction data we use contains the criminal charge record in Newark, New Jersey, from 1997 to 2017, with $17,335$ suspect individuals. Each suspect is provided with her/his Personal ID, GenderCode(Female, Male), race, and a list of Bookings. Each booking has a list of sentences, where each sentence has the age of committing the crime and the individual criminal charge history, such as the NCIC crime code, NCIC category code, and the crime level. The NCIC category code includes the list of ["ASSL","TO", "DAD", "LARC", "FO", "PP", "BURG", "SV", "DP", "OP"]. The input features also include the number of bookings, age average, number of crimes at level 1, 2, 3, and so on. We take the first 18 years as the training set and the last two years as the test set. A suspect with at least two bookings is one sample. The label for each sample is the crime level (or whether it is a specific crime level) during the last booking.

Task Our goal is to predict a person's crime or crime level (1 for most severe and 3 least severe) given her/his personal information and previous criminal records. We propose a multitude of tasks addressed by Question (Q)1/2/3/4:

Gender/Age/Race	Crime level 1		Crime level 2		Crime level 3	
	Yes	No	Yes	No	Yes	No
All	6.7	93.3	3.5	96.5	8.7	91.3
Men	6	94	2	98	13	87
Women	3.8	96.2	0.9	99.1	12.3	87.7
≤ 20	16.9	83.1	7.2	92.8	12.8	87.2
21-30	7.8	92.2	4.3	95.7	8.1	91.9
31-50	5.9	94.1	2.9	97.1	9.1	90.9
>50	2.8	97.2	1.4	98.6	7.2	92.8
W	6.8	93.2	3.6	96.4	9.2	90.8
A	9.8	90.2	2.1	97.9	8.1	91.9
U	2.6	97.4	2.1	97.9	5.1	94.9
B	8.6	91.4	3.6	96.4	6.9	93.1
I	3.1	96.9	0.8	99.2	6.2	93.8

Table 1. Highly imbalanced data: most bookings are labeled as non-crime. We measure the percentage of the crime- and -non-crime labeled sample number [in %], respectively. The statistics is calculated on the whole dataset, and on three different data classifications of the gender, age, and race, respectively.

Is a suspect going to commit a crime at level 1/2/3/any?

Essentially we train four different systems for each of the above four questions, viewed as a binary classification task. We observe from Table 1 that this data is highly imbalanced, with most of the instances having the "No crime" label. This resembles a real-life scenario, where a suspect stops committing a crime more often than continuing to commit crimes. It is technically challenging to handle imbalanced data. A fully connected feed-forward neural network does not work in our experiments since it may classify all samples as negative. For example, suppose there are only 2 positive samples among 100 samples. In that case, the precision is still 98%, although nothing is classified.

In the following context, we will first describe our task and dataset, introduce our models, then demonstrate experimental results and interpret our models. Finally, we will discuss related work and conclude.

3. Methods

Learning Temporal Patterns with Sequential Models To solve the imbalanced data problem, we observe that the criminal charges of an individual follow temporal patterns. A suspect likely repeats a similar behavior, such as committing a crime with the same level or after an equal amount of time (i.e. somewhat periodically). This kind of temporary pattern can be learned with the whole dataset's statistics and shared among different individuals. We introduce a Recurrent Neural Network (RNN) to capture the sequential patterns of individual criminal charge records along one's lifetime. To reduce the vanishing gradient problem, we add LSTMs (Hochreiter & Schmidhuber, 1997) in the activation functions, which greatly reduced the data imbalance problem. As a data preparation step, we had a list of booking for each person. Each booking has an age that shows the person's age when he/she gets the sentence. We find that the maximum number of bookings for a person of different ages (sorted ages from young ages to old ones) is 13. We

decided to create sequential data with 12 parts, and we use the last booking as a label of the dataset.

Interpreting Feature Importance with Attentional Mechanisms *The second challenge is that our models lack interpretability.* We provide a certain degree of post-interpretability using the attention mechanism (Vaswani et al., 2017). In particular, the attention mechanism allows to focus on, or pay attention to, different parts of the input sequence to generate a prediction. The idea of an attention model is to consider all the hidden states of the LSTM and compute their linear combination to produce a new hidden state. This new hidden state is used for the classification instead of just the last hidden state. The attention-based model does not only in better learning to improve the accuracy but also provides a method of post-hoc interpretability. In particular, the attention scores obtained while predicting will indicate the importance of the input features towards the decision.

To obtain the feature importance: First, we check the impact of all the input features on our model predictions by setting the window size to be the total number of input features. After calculating the embedding of each feature, we compute their relative weights. Then, we apply softmax overall features to obtain probabilities of the predictions. Each input is associated with a probability value, and the sum of all probabilities should be 1.

4. Experiments

Setup Our model (Winata et al., 2018) is based on long short-term memory (**LSTM**) network and Bi-LSTM with attentions (Hochreiter & Schmidhuber, 1997) with two hidden layers, where each hidden layer has ten nodes with or without attention layers. The inputs are the features, and the outputs are fed to a linear layer mapping to the target classes. The dropout parameter is set to 0.1. The parameters were updated using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 and random initialization.

Model	L	Acc.	Prec.	Recall	F1
LSTM	1	94.1	92.0	94.1	92.3
	2	96.4	93.2	96.4	94.8
	3	92.3	91.2	92.3	91.5
	Any	87.5	86.8	87.5	87.0
Bi-LSTM	1	94.5	93.0	94.5	93.1
with	2	96.5	93.2	96.5	94.8
attention	3	91.4	87.0	91.4	87.6
	Any	90.6	90.5	90.6	90.5

Table 2. LSTM and Bi-LSTM with attention results for crime level prediction. L: the crime level.

Prediction Results Table 2 shows the accuracy scores on the test data using different models on four research questions in Section 2. We introduce sequential models, including LSTM, Bi-LSTM with and without attention. We achieve both high precision and high recall overcoming the data imbalance problem by considering the criminal charge records along the previous eighteen years and learning the common patterns across samples.

Interpreting Results Our attention model provides some interpretations for its predictions, suggesting that the prediction relies on historical criminal records rather than personal data such as age, race, and gender. *Feature weights are learned as a mean to measure the "Causality" of model decisions in our algorithm.* We visualize softmax probabilities for randomly selected test samples. See Figure 1. We see that surprisingly, although there are many input features, only a few of them have the highest impact. More importantly, as shown in Figure 1, we find that personal information such as race contributes very little towards the prediction. Instead, the features of an individual's criminal history greatly impact the prediction, such as "time since last crime" and "variance of the time gap between successive crimes". This indicates the personal information features do not explicitly contribute to our algorithmic decisions. Note that our work only focuses on the feature importance explicitly used in our model. Discussions on any bias made by humans in the previous charge decisions (Brantingham et al., 2018; Arnold et al., 2018) are out of the scope of this paper.

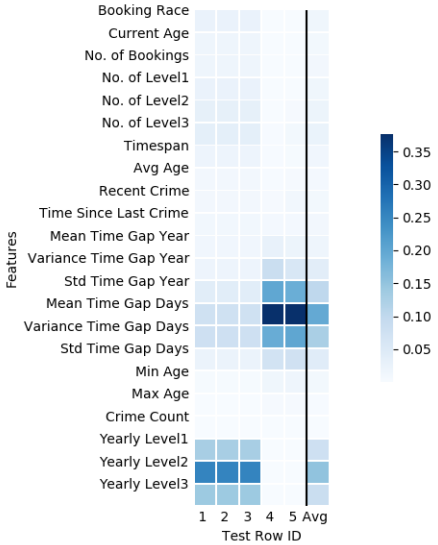

Figure 1. Each column shows the relative weights of different features for one random test samples. The higher, the more important. No. of levels values are the average values across all years.

The correlation and "causality" are different, and we will

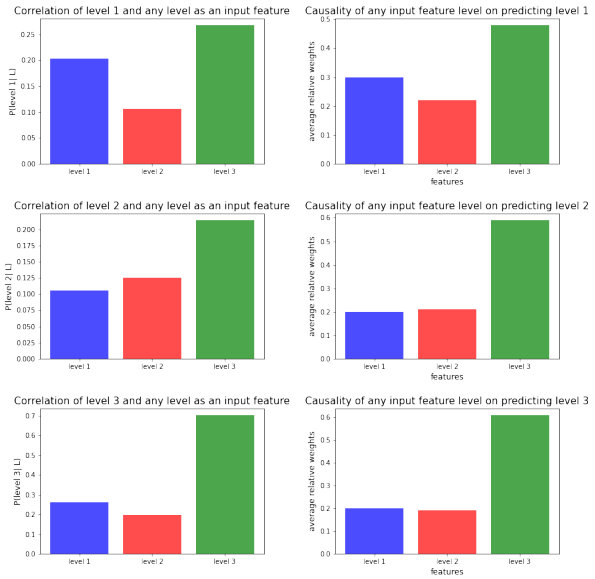

Figure 2. Crime level transitions. Left: $P(L|level)$; Right: Causality of level to L.

show a case study on an individual's crime level development. In our second analysis, we study the transitions of a suspect from one crime level to another. We find that crime level features are essential to predict a new crime level. More precisely, we illustrate whether a lower level crime record indicates a higher level of crime, or vice versa, as shown in Figure 2. The diagrams on the left show the probability of crime level 1, 2, and 3 given the different crime levels committed in history. We use the relative frequency to compute the probabilities. The diagrams on the right show the importance of our machine learning models' features on the data to predict a certain crime level. We consider it as the "causality" of our algorithmic decisions. We can observe a discrepancy between the correlation on the left and the causality on the right. The "causality" diagram shows that level L crime depends more on level $L+1$ or $L-1$ than the correlation diagram suggests. For example, level 2 shows greater importance in predicting level 3 in the "causality" diagram than in the correlation diagram. To predict level 2, both level 1 and level 3 are more important in the "causality" diagram than in the correlation diagram. Finally, for level 3, level 2 shows a more significant impact in the "causality" diagram than in the correlation diagram. Therefore, our algorithmic "causality" results show a stronger indication that a suspect is likely to commit a crime with the same or similar level again than a much higher or lower level.

5. Related Work

There has been an increasing trend in criminal justice to leverage machine learning for high-stakes prediction applications that deeply impact human lives. (Yu et al., 2011) uses Support Vectors Machine (SVM) model, 2-layered feed forward neural networkand Naive Bayes model for crime forecasting. (Stec & Klabjan, 2018) and (Stalidis et al., 2018) use feed forward, convolutional and recurrent-convolutional networks, (Luo et al., 2017) uses recurrent based networks with attention mechanism to analyze law articles. Our model directly predicts the criminal charge based on the criminal history records and interprets our algorithms' decisive factors. (Dressel & Farid, 2021) courts across the United States are using computer software to predict whether a person will commit a crime, the results of which are incorporated into bail and sentencing decisions. It is imperative that such tools be accurate and fair, but critics have charged that the software can be racially biased, favoring white defendants over Black defendants. We evaluate the claim that computer software is more accurate and fairer than people tasked with making similar decisions. We also evaluate and explain the presence of racial bias in these predictive algorithms.

6. Conclusion

Our work introduce a trustable criminal charge prediction method with high precision and high recall as well as post-interpretability. We show that the use of deep learning can be a part of the criminal justice assistant system as long as model transparency and accuracy are taken care of. Perhaps most importantly, we draw attention to the erroneous assumption that social features such as race, gender, and age are statistically insignificant to influence our model prediction, even though data may introduce bias.

Acknowledgement

We appreciate the National Science Foundation (NSF) Award No. 1747728 and the National Science Foundation of China (NSFC) Award No. 61672524 to fund this research. We thank Kovid Inc. to provide the dataset within the NSF project, the initial input of Subhadarshi Panda and the comments of Feng Gu, Mengyan Dai, and Periklis Papakonstantinou. Finally, we would like to give our appreciation to the support of the Google Cloud Research Program.

References

Arnold, D., Dobbie, W., and Yang, C. S. Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4): 1885–1932, 2018.

Brantingham, P. J., Valasik, M., and Mohler, G. O. Does predictive policing lead to biased arrests? results from a randomized controlled trial. *Statistics and public policy*,

5(1):1–6, 2018.

Dressel, J. and Farid, H. The dangers of risk prediction in the criminal justice system. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, 2021.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Luo, B., Feng, Y., Xu, J., Zhang, X., and Zhao, D. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2727–2736, 2017.

Skeem, J. L. and Lowenkamp, C. T. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4):680–712, 2016.

Stalidis, P., Semertzidis, T., and Daras, P. Examining deep learning architectures for crime classification and prediction. *arXiv preprint arXiv:1812.00602*, 2018.

Stec, A. and Klabjan, D. Forecasting crime with deep learning. *arXiv preprint arXiv:1806.01486*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Winata, G. I., Kampman, O. P., and Fung, P. Attention-based lstm for psychological stress detection from spoken language using distant supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6204–6208, 2018.

Yu, C., Ward, M. W., Morabito, M., and Ding, W. Crime forecasting using data mining techniques. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 779–786, Dec 2011. doi: 10.1109/ICDMW.2011.56.