Predicting the Intensity of Wrongful Convictions with Demographics and Contributing Factors

Eric Geisler

John Carroll University

May 03, 2024

Executive Summary

The United States judicial system is not infallible. Wrongful convictions—when an innocent person is convicted of a crime—almost always result from procedural errors. These systemic wrongdoings, called contributing factors, include falsified forensic evidence, coerced confessions, false accusations, mistaken witness identification, inadequate defense, and misconduct from officials. Organizations carrying on the fight against wrongful convictions are small and poorly funded, meaning they have to be selective in the cases they take up. Thus, having a method of predicting a defendant's sentence and time spent in prison before and during a trial based on demographics (state, race, sex) and the aforementioned contributing factors would enhance the efficiency of these organization's resource allocation; they could predict which defendants will have the worst outcomes. The question remains—is the intensity of wrongful conviction outcomes able to be predicted by these demographic and contributing factors? After constructing a feed forward neural network model and using the factors listed above (9 in total) as inputs to predict both sentence (classification) and time spent in prison (regression), this answer is in doubt. When predicting the classification of sentences, the model had an accuracy score of 55%. A prediction was considered correct if either of the model's top two predicted categories for each case matched the actual sentence. This accuracy is more than 30% greater than chance (22%), but still largely unreliable. Moving onto the regression, the mean average error for the model's predictions for time spent in prison was 6.83. Clearly, using the regression output of the model in its current state is impractical. In terms of field use, this model should only be used as a reference for sentence classification. Going forward, utilizing more general conviction data that do not exclusively contain wrongful convictions would serve more practical importance.

Background

Kenny Phillips and Michael Sutton were released from prison in 2021. Fifteen years earlier, they had been living a normal life, out in the city of Cleveland celebrating Kenny's eighteenth birthday. After accidentally driving between two cars involved in a shootout and, being kids, fleeing from the police out of fear, they were arrested. During trial, the police officers involved in the case lied about the story, framing the teenagers as the perpetrators. The two friends were promptly convicted. Fifteen years gone.

Frighteningly, wrongful convictions are just that easy. At a moment's notice, anyone's life can be stolen from them by the justice system simply for being in the wrong place at the wrong time. Ultimately, the United States justice system is built around a flawed principle: finality over truth. Convicting the right person does not matter as much as simply convicting a person. Thankfully, organizations exist that combat wrongful convictions, the most famous being The Innocence Project. However, these organizations operate with little resources and face the challenge of having to be very selective in what cases they intervene in. Furthermore, these organizations can only intervene post-conviction. If there was a way to predict what cases will result in the worst wrongful conviction outcomes, these organizations could intervene earlier and have a greater effect on mitigating the worst of the problem. At its core, the ability to predict the intensity of wrongful convictions based on demographic information and contributing factors would save lives.

Related Work

Because of the niche nature of wrongful convictions as an area of study, there is not much previous work in this field. Regarding machine learning specifically, there is almost no previous work. All prior research has involved studies comparing cases where innocent defendants were acquitted to wrongful convictions. These studies were used to discover and prove the existence of the aforementioned contributing factors. In other words, all previous work has centered around a very forward-looking, surface level view of the reasons behind wrongful convictions. No previous work has gone backwards—taking the already proven contributing factors and trying to predict the intensity of wrongful convictions. While prior studies have been beneficial in terms of identifying specific areas where policy change is necessary to stop wrongful convictions (coercion of confessions during interviews, for example), no research has been conducted with the purpose of helping organizations that fight wrongful convictions determine which cases are worth their time and resources. In other words, previous research has been centered around providing evidence for policy change in general. Comparatively, this work exists within the current systemic failure and tries to mitigate the worst of wrongful convictions. Thus, there is a more realistic, practical application for this work.

Overview

Again, these smaller organizations simply do not possess the resources to fight all wrongful convictions. Additionally, they are handling most of their cases years after they have already been decided, and the convoluted appeals process requires an immense application of time and resources. With the ability to predict wrongful convictions based solely on case information before a verdict or shortly after a case is finalized, these organizations could become

more efficient, both in terms of selecting the cases with the worst probable outcomes and spending less time and resources per case. Thus, the solution to this problem (outside of any systemic policy change) is to acquire the ability to accurately predict the intensity of wrongful convictions, specifically through the labels of sentence length and time spent in prison. Again, this would result in better efficiency for organizations that fight against wrongful convictions, and allow more people to take back their lives from the corruption of the justice system.

Data

The data used for this research came from The National Registry of Exonerations. This organization tracks exonerations (where a person is released from prison and the justice system admits the case was a wrongful conviction) dating back decades. For each case, the database contains the relevant demographic information and, most notably, all of the contributing factors that occurred. Furthermore, the database also includes sentence length and year of conviction and exoneration, providing the labels for modeling. Essentially, The National Registry of Exonerations provides all possible information that could be useful in determining if predicting the intensity of a wrongful conviction based on basic demographic information and the presence of contributing factors is practical. These variables beget the use of deep learning, as this research is trying to discover patterns between the inputs to predict given outputs.

Methods

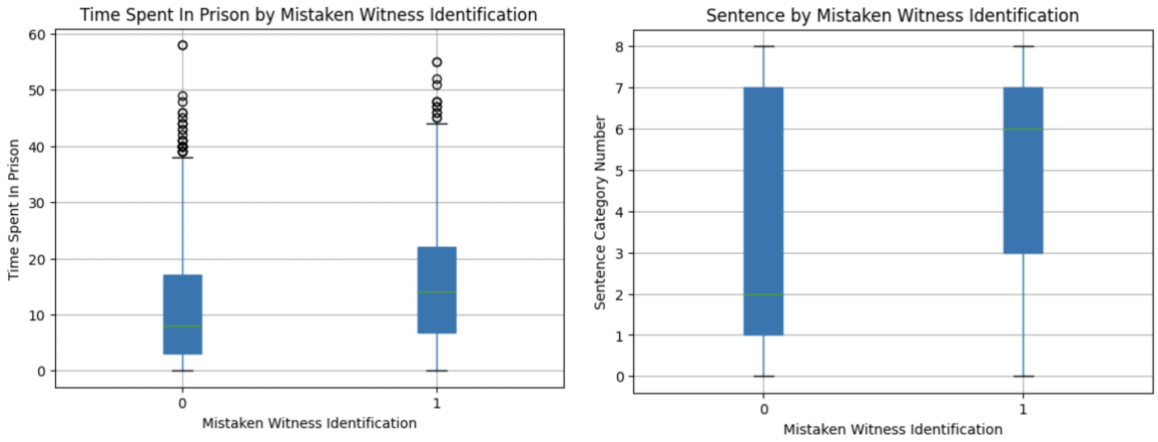
Code can be found here: https://github.com/EricGeisler/WrongfulConvictions

In preparation of building a model, a process of data cleaning and feature engineering was carried out. Firstly, all contributing factor variables (portrayed as text in the original data)

were encoded into new dummy variables, with a value of "1" representing the presence of the specific contributing factor in a given case and a value of "0" representing its absence. This was additionally true for the variable accounting for the sex of an exoneree. The variables representing the state where a conviction occurred and the race of an exoneree were also transformed into numerical columns. In total, nine features were studied, with three involving demographic information and six representing the presence of a contributing factor. The nine features are as follows: "State", "Race", "Sex", "OM_Binary" (Official Misconduct), "FC_Binary" (False Confession), "PFA_Binary" (False Accusation), "ILD_Binary" (Inadequate Defense), "FMFE_Binary" (False Forensic Analysis), and "MWID_Binary" (Mistaken Eyewitness Identification). Additionally, the "State" and "Race" variables were normalized as inputs. Regarding the two labels, cleaning and engineering was also done. For the variable representing the sentence of an eventual exoneree, the original data presented a challenge: there was a vast array of sentences. For example, "Life", "Life without parole", "75 years", and "95 years" were all separate values despite substantively being the same sentence. For simplicity, the original variable was pared down to only nine buckets based on this substantive similarity: "Death", "Possible Life", "More Than 40 Years", "More Than 30 Years", "More Than 20 Years", "More Than 10 Years", "Less Than 10 Years", "Less Than 1 Year", and "Other". The buckets were then transformed into a numerical variable ("Sentence_Category_Number") of zero through eight, with eight corresponding to "Death" and the rest following in the above order. Obviously, this bucketization is far from perfect. Inherently, this format weakened the predictive ability to detect lower sentences (specifically less than ten years) and extremely high sentences described exclusively in years, such as the "95 Years" example above. The creation of the label involving time spent in prison was much simpler. This was created by subtracting the conviction

year from the exoneration year. To eliminate the right skewness, the natural log of these values was taken, with the final label being named “Time_Spent_In_Prison_log”. Finally, a new dataframe was created with only the prospective inputs and labels, and missing data was removed.

Before building a model, each feature was analyzed on a cursory level against each of the labels (the non-log version of time spent in prison). Since race, location, and sex have been long proven to affect convictions, these will not be included in this deeper analysis—only the contributing factors. For example, the presence of mistaken witness identification in a case resulted in a median time spent in prison of about five years higher and a median sentence category of four levels more extreme. The plots are shown below:



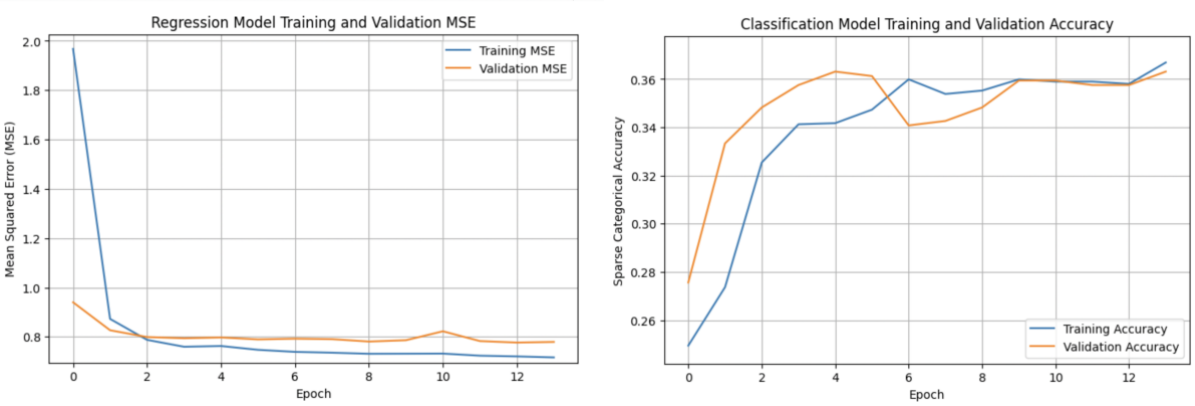
All variables were deemed to have a significant enough effect on a cursory level to validate their inclusion as an input into the model. Specifically, mistaken witness identification, false forensics, false confessions, and official misconduct appeared to have the most significant effects on the target labels.

Moving onto the model, a dense neural network was utilized, as the research goal centers around discovering patterns and connections between the inputs in order to predict the labels. Specifically, a feed-forward neural network was built, meaning the data gets passed directly from

layer to layer. As this research required both a regression and classification output, this model type was chosen. The TensorFlow machine learning package was utilized to build the model. The regression side of the model contains three layers, two hidden dense layers (ReLU activation) and an output layer (Linear activation). The layers have twenty-five units, twelve units, and one unit, respectively. The classification side of the model also contains three layers, two hidden dense layers (ReLU activation) and an output layer (Softmax activation). The layers have twenty units, fifteen units, and nine units, respectively. The model contains fourteen epochs, a batch size of 100, and a learning rate of .01. The model was run with a split of 80% and 20% for both the validation and test data. The aforementioned nine features served as the inputs, predicting "Time_Spent_In_Prison_log" and "Sentence_Category_Number".

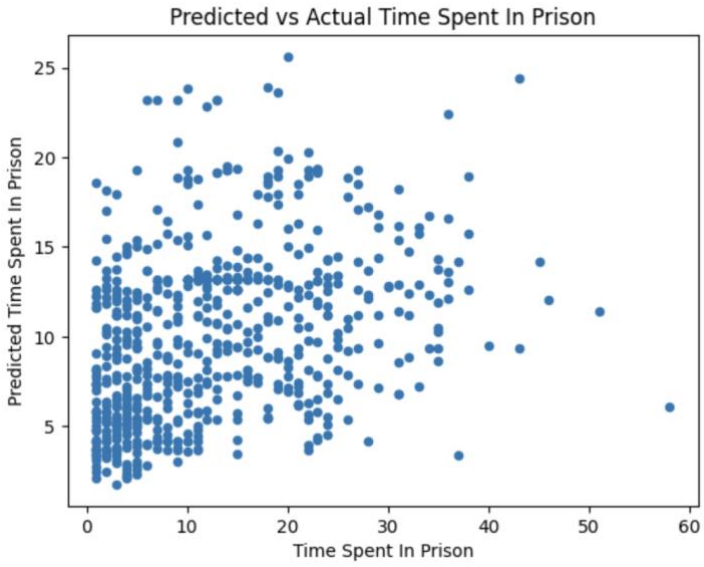
Evaluation

The performance of the model left a lot to be desired. The model seemed to learn well between the training and validation data, seen below:

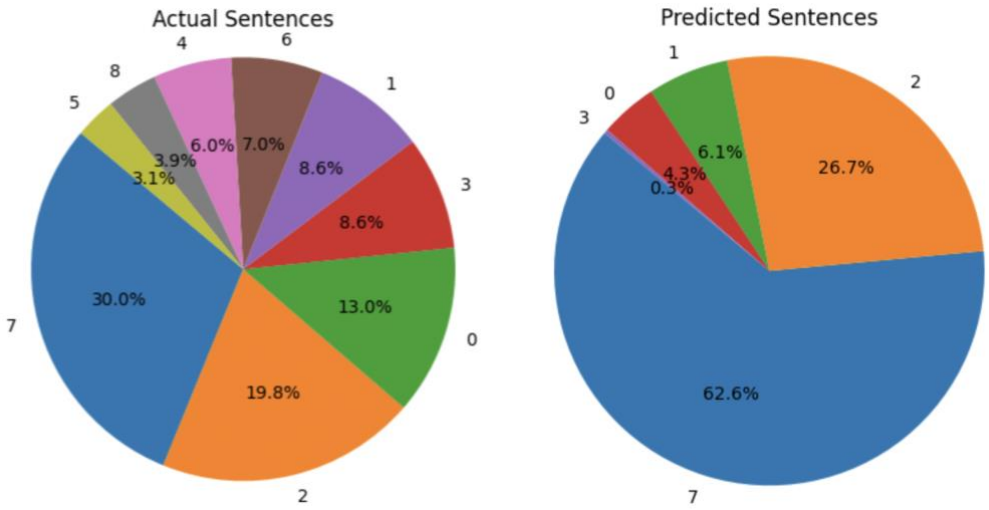


From the validation data, the model produced a root mean squared error (corrected from taking the natural log) of 2.34 and a classification accuracy of around 37%. However, the end results were not impressive, reflected in the model's predictions on the test data. With the test data, the

model produced a corrected mean absolute error of 6.83 for the predictions on time spent in prison. This lack of accuracy is demonstrated by the scatter plot of predicted versus actual time spent in prison below:



Regarding the classification output, the model performed much better with an accuracy of 55%. Again, a prediction was deemed to be correct if either of the model's top two predicted sentence buckets matched the actual sentence bucket. Making a correct prediction at random would resemble a 2/9 chance, or 22%. Thus, the model performed around 33% better than random chance. However, the model's array of classification predictions was not good, seen below:



Evidently, the model predicted "Possible Life" (category seven) far too often, and struggled to predict any of the buckets involving years. While the classification accuracy was significantly better than random chance, the model as a whole still did not perform well.

Discussion

Upon analyzing the results, predicting the intensity of wrongful convictions in terms of time spent in prison and sentence type using basic demographic information and contributing factors appears to be impractical, at least with this data. The model would provide little value for an organization fighting wrongful convictions trying to maximize the efficiency of its allocation of resources. While the classification output could be a good starting point to predict the seriousness of a case as it's unfolding, it is nothing more than that—a starting point. Making any decisions solely based on the model would be unreasonable and risky.

Field Trial

Designing a field trial for wrongful convictions is challenging, as there exists a general lack of data on the subject. Thus, it would be necessary to collaborate with these organizations,

such as The Innocence Project, to access as much data as possible. In order to test the model, a sample of ongoing cases that have been flagged by these organizations could be gathered. These cases have a distinct possibility of ending in a (presumed) wrongful conviction. More specifically, these cases would have to possess some obvious contributing factor at play. For example, a defendant might be insisting that their confession in an interview was coerced, or that officers abused them during interrogation. Once this sample has been gathered, say 100 cases as an example, the model can be put to use. The demographic information (state, race, sex) of the defendant and the alleged contributing factor(s) at play can be used as inputs. Then, the model predictions can be compared directly to the actual outcomes of the trials.

Obviously, this design would present some issues. Firstly, time spent in prison would have to be tracked into the future (for potentially decades), which would require diligence and a large amount of resources. Also, the timing of finding and following cases as they are happening would be difficult. Again, this is why a partnership with wrongful conviction organizations would be necessary. Also, alleged contributing factors are not necessarily guaranteed to be publicly known before a trial verdict. Thus, complete data would be hard to acquire. However, with ideal data acquisition, this design would be an excellent way of testing the model in real world practice.

Limitations

The greatest limitation with this model is the practical timing that would apply to a real world scenario. As previously stated, not all wrongful convictions are flagged before they occur. The stories of contributing factors in cases sometimes do not come out until years or even decades after convictions. Thus, the model would only be most useful for cases in which

contributing factors are well documented before the verdict of a trial. That way, both predictions could actually serve a practical importance, as organizations could devote more resources to a case as it unfolds because of this. For example, if the model discovers a pattern that people of a given race and a given sex typically receive life sentences when there is a mistaken eyewitness identification, these cases could be especially targeted in the future.

Additionally, there were clear limitations with the labels being used. Bucketizing the sentence lengths was not seamless. Since many of the original sentences were substantively the same but had different descriptions, making buckets that encompassed all the sentences perfectly was not possible. This likely altered the true distribution of sentence lengths, and potentially affected the prediction accuracy for the classification component of the model. Regarding time spent in prison, this label inherently under-represents current data, as people who have been wrongfully convicted but not yet exonerated are not represented in the data.

Future Work

Regarding future work, there are two potential avenues. Firstly, the model efficiency can be optimized to a greater extent. While all the inputs appeared to have a significant effect on the labels, a deeper analysis can be conducted on all the variables individually to determine their actual significance. If certain variables do not have any real predictive power, they can be removed to increase the simplicity and efficiency of the model. Secondly, the model can be developed with broader data. Through further research, more demographic and contributing factors that possibly have an effect on the intensity of wrongful convictions can be added to the model. Additionally, the classification aspect of the model can be used on more recent data, apart

from the limited regression label involving time spent in prison. Focusing on the classification output, the model can potentially be run on data with a cleaner bucketization of sentence length.

Proposal

Going forward, individual contributing factors must be analyzed on a deeper level. Ultimately, being able to predict the intensity of possible wrongful convictions only provides a temporary solution in terms of helping small organizations target specific cases. In an ideal world, these organizations fighting wrongful convictions would not exist in general. As of now, there is little research in the area besides surface level examinations of what contributing factors actually exist. By using machine learning techniques such as regression models, the effects of these contributing factors can be researched further. Thus, policy change initiatives could become more realistic by targeting specific areas that are found to be especially troublesome within the context of the overall issue, such as fixing the causes of false forensic analysis. Additionally, more data needs to be gathered on wrongful convictions. If more data existed that combined wrongful convictions with cases of innocent defendants who were not convicted, more machine learning techniques could be utilized to better understand this complex issue.

Appendix

Labels:

- "Time_Spent_In_Prison_log": the natural log of the difference between the exoneration year and the conviction year for each case
- "Sentence_Category_Number": the sentence length for each case, bucketized by substantive similarity, and converted to a numeric value

Features:

- “State”
- “Race”
- “Sex”
- “MWID_Binary”
- “ILD_Binary”
- “PFA_Binary”
- “FMFE_Binary”
- “OM_Binary”
- “FC_Binary”

Hyperparamters:

- Regression:
    - Hidden Layer 1:
        - 25 Units
        - ReLU activation
    - Hidden Layer 2:
        - 12 Units
        - ReLU activation
    - Output Layer :
        - 1 Unit
        - Linear activation
- Classification:
    - Hidden Layer 1:

        - 20 Units
        - ReLU activation
    - Hidden Layer 2:
        - 15 Units
        - ReLU activation
    - Output Layer :
        - 9 Units
        - Softmax activation
- Learning Rate: .01
- Epochs: 14
- Batch Size: 100

Validation and Test Data:

- 80% / 20% splits