# Skynet is Alive and Well: Leveraging a Neural Net to Predict Felon Recidivism

Dylan Hanson (Student Category), *Vanderbilt University*

## I. Introduction

The 2021 National Institute of Justice's (NIJ) "Recidivism forecasting challenge" emphasized the growing importance of using predictive deep learning models to help better understand the factors that cause recidivism. According to the NIJ's provided information, 'revolving door' prison sentences cost around 9.3 billion dollars annually[1]. Thus, a better understanding of these factors can assist corrections officers and parole officers in preventing recidivism, potentially saving billions in taxpayer money. Many current models also suffer from racial bias, and minimizing this bias is an essential element of the challenge. This submission details the implementation of a deep learning model, specifically a neural net, to help predict recidivism and assist corrections officers in identifying high-risk individuals and thus helping to prevent recidivism.

## II. Variables

The model was trained on the data provided by the NIJ. This dataset included roughly 26000 persons from Georgia on parole through the years 2013 to 2015. The model was trained on the

[1] Council of State Governments (2019). Confined and Costly: How Supervision Violations are Filling Prisons and Burdening Budgets. Washington, DC: Council of State Governments.

70% of information provided as “training” data and applied on the remaining 30% of “testing” data for project submission.

Depending on the nature of the underlying variables, they were encoded differently as to be processed by the model correctly. Numeric columns were left unchanged. Boolean values, such as “gang affiliation” were changed from being encoded as true-false to 1-0 respectively. Finally, categorical variables, such as “supervision level” were assigned numeric values. For instance, the supervision levels of ‘standard’, ‘high’, and ‘specialized’ were assigned 0, 1, and 2 respectively. This was preferred to one-hot encoding due to the limited computing power available. Following this, buckets were generated for a subset of variables using the keras buckets feature, specifically grouping convictions, arrests, violations, and drug test results. As the variable ranges following this process were similar, it was deemed unnecessary to standardize the variables. Due to the relatively small number of input variables to the model, it was similarly deemed unnecessary to conduct a robust principal component analysis (PCA).

No additional variables were added to the dataset.

# III. Model

As mentioned previously, a classification sequential model was constructed for the challenge. This model was built in the python language, relying heavily on the keras API. Keras allows for the construction of deep learning models, having been built upon the machine learning platform TensorFlow. Keras’s high-level interface allows for the full abilities of TensorFlow to be exploited by individuals with a low barrier to entry, requiring relatively little background knowledge of computer science.

The sequential model implemented is the simplest model that can be constructed with keras. The sequential model is a linear stack of keras layers. In keras, layer objects are used to build up a neural network for deep learning. Each layer is generated with a variable number of neurons and a specific activation function. These activation functions determine how each node of the neural net processes the information provided to it. Following training, layers also store information

regarding the weights of the different inputs that each node receives. These inputs are obtained from higher levels of the neural network, beginning with the input variables but modified to be highly abstract predictors by each successive layer[2].

To begin, an arbitrary sequential model was created by taking an arbitrary number of layers, nodes, and arbitrary activation functions. This model was then trained based upon the training data provided by the NIJ. The data was split into a test/train split using a 20-80 ratio. The training data was then further split into validation/training, once again based on a 20-80 ratio. In this method, the model was trained using the validation/training split, with the algorithm optimizing for accuracy using the validation/training set. Finally, model performance was analyzed using the testing data, the model being optimized for classification accuracy.

Additional models were then generated by varying the number of layers, nodes on each layer, and each layer's activation function. As with all machine learning projects, this is more of a deliberate 'art' than it is a science. Through a process of trial and error, an optimal combination of the above hyperparameters was determined; the final model selected was the one with the highest accuracy. This optimized model was then applied to the remaining 30% of data that the competition used to evaluate the model and select winners.

The final deep learning model was able to predict recidivism with a 67.73% accuracy, at the 0.5 threshold.

# IV. Discussion

Despite its accuracy in generating predictions, the neural network model used is unable to provide us with information about which variables are more relevant than others in predicting recidivism. This is because the computations applied by each layer of the neural network take into account multiple variables from the prior layer of the neural network, causing the aforementioned abstraction from the provided input layers. Unfortunately, this makes the

[2] Chollet, François et al. "Keras." https://keras.io. (2015).

underlying logic used by the model something of a 'black box'; this is common among many deep learning methods.

With this in mind, it is easy to see how the very concept of 'statistical significance' is inapplicable to a deep learning model: it is impossible to assign any meaningful measure of statistical significance when the impact of each variable is unknown. If the model 'decided' in its computations that a variable was not important in predicting recidivism, those nodes processing that variable would have lower weights assigned to them.

As the neural network was optimized for accuracy and no other metric, the fact that the fairness penalty only affected false positives did not affect the submission. The 0.5 threshold, while usually being the standard for many deep learning projects- literally being "better than a coin flip"- could potentially have led to an excess of false positives, affecting the fairness penalty. An alternative would be to use a range of thresholds (e.g. from 0.5 to 0.9) to compare results and determine an optimum with a minimum of false positives, due to the innate meaning of a false positive result. Alternatively, the model could have been optimized for kappa instead of accuracy. Cohen's Kappa statistic takes into account class imbalances[3], which could potentially reduce the biases from low socioeconomic status individuals.

In the real world, usage of this model (or similar models) by parole officers could allow them to determine which individuals that they are supervising are the most likely to reoffend. Knowledge of these predictions would allow officers to redirect more resources to helping rehabilitate these individuals, with the intent to prevent them from being arrested again.

# V. Future Considerations

Further model development could allow for higher accuracies in predicting recidivism. The implementation of a Recursive Feature Elimination (RFE) algorithm could allow one to determine which variables are most important in determining recidivism. RFE algorithms operate by iteratively generating deep learning models with different subsets of variables, thus

[3] McHugh, Mary L. "Interrater reliability: the kappa statistic." *Biochemia medica* vol. 22,3 (2012): 276-82.

determining the most impactful variables in determining the outcome from provided data. Such information could assist policymakers and parole officers in trying to reduce the risk of certain individuals with "high-risk" characteristics from reoffending.

Implementation of other forms of deep learning models, such as tree-based models (e.g. random forests), gradient boosting, or Bayesian models could prove to yield more robust results but would be more computationally intensive. Bayesian models in particular could prove very useful: while they are highly computationally intense, they are based on conditional probabilities. This means that a given Bayesian model can generate predictions based on partial information and is capable of accepting new information to update predictions. This would be of great assistance to parole officers who can actively update information in real-time about the individuals they are supervising to help determine resource allocation to their supervisees.

However, for the sake of this competition, the implementation of a tree-based model would not suffice as many tree-based models are incapable of generating probability estimates for a given observation, but rather classify observations right off the bat. Similarly, the computational intensity of a Bayesian model is too high a barrier to entry for a student project.

Further, the variables being considered in the dataset do not necessarily take into account the "root causes" of recidivism, but rather symptoms of deeper systematic issues. For instance, the category of "drug tests" may highlight that those who test positive for narcotics are more likely to reoffend, but it does not address the underlying reasons as to why an individual on parole may choose to partake in drug use.

Future challenges may benefit from focusing more on the impact of rehabilitation programs for formerly incarcerated individuals rather than simply factors that may or may not correlate with recidivism.

# VI. Conclusion

It is self-evident that the usage of deep learning tools has immense potential in assisting parole officers and policymakers in the rehabilitation of formerly incarcerated members of society. As proof of this, the neural network created as an entry for this project was able to predict

recidivism with a 67.73% accuracy, despite being a relatively simple model. More sophisticated models that consider the underlying variable distributions could easily prove to yield more robust predictions.

However, innate socioeconomic inequalities can cause models that are trained off available data to be skewed towards overpredicting the risk of recidivism among those of low socioeconomic status. As such, any models generated must be done so carefully as to prevent any possible racial bias. The false positive score associated with scoring this challenge is one step in the right direction, but it is important to keep in mind that some variables may still be implicitly biased.

# VII. Acknowledgments

Thank you to Aakash Manapat, a fellow undergraduate student at Vanderbilt. Without his guidance in model creation and feedback in writing this report, my submission to the NIJ's Recidivism forecasting challenge would not have been possible.

# VIII. Relevant Literature

1. Council of State Governments (2019). *Confined and Costly: How Supervision Violations are Filling Prisons and Burdening Budgets*. Washington, DC: Council of State Governments.
2. Chollet, François et al. "Keras." https://keras.io. (2015).
3. McHugh, Mary L. "Interrater reliability: the kappa statistic." *Biochemia medica* vol. 22,3 (2012): 276-82.

# IX. Appendix

Link to code repository:

https://github.com/jovialis/nij-recidivism-challenge