National Institute of Justice's Recidivism Forecasting Challenge

Murray Miron for SRLLC

December 8, 2021

Abstract

The winners of the National Institute of Justice's Recidivism Forecasting Challenge were announced August 16th, 2021; SRLLC[1] was one such winner. This paper provides an overview of the methodology employed by SRLLC while developing the program that generated the winning submission, as well as covering the quantitative importance of features within the dataset. Sources of error are also noted.

Contents

List of Tables

[1]Software Research LLC is owned by Murray Miron.

1 Introduction

At the outset of the project, SRLLC decided to develop a universally-applicable translation layer that would convert the contents of any Pandas dataframe into a form suitable for supervised training with a Tensorflow artificial neural network. This library was used for the first year scores that SRLLC received a cash prize for submitting; the majority of the time spent on this project was devoted to the development of said library, and very little was done with the specific dataset provided by NIJ. The contents of this paper should be treated appropriately.

2 Materials

The solution was developed within a Linux desktop environment using an x86-64 CPU and an NVIDIA GeForce GTX 1080 Ti video card. All source code was written in Python. Python modules that were used, but not written by SRLLC, were restricted to Pandas, numpy, and Tensorflow (NVIDIA CUDA version).

3 Methodology

Our numbers are the result of an estimation procedure which was employed after the close of the contest. The estimation process consisted of training an artificial neural network – one with the same architecture employed to generate our submissions – while iteratively excluding a single feature column from the input via setting each individual column's weights equal to 0. Validation scores improved only minimally after 200 training epochs in the average case, and so that was chosen arbitrarily as a sufficient amount of time to devote to each individual feature; this resulted in the entire estimation process requiring 9600 training epochs (200 epochs times 48 features). The relative importance of each feature was then determined by the training loss and the validation loss, separately, when said feature column was excluded during training (see tables).

Note that the scores listed are actually the mean squared error loss when predicting over either the validation (holdout) set, or the training set. Arguably, the training set provides the best indication of feature importance: whether the model generalizes is irrelevant for our purposes here, making the importance of a feature when predicting over the training set a legitimate metric.

The mean squared error formula we used reduces to be equivalent to the Brier score that was used by the contest organizers. Since lower scores signify better predictive performance (a zero would be perfect prediction), a higher number in the table shown corresponds to the labeled feature column having greater predictive importance. That is, if an important column is excluded from the input to the artificial neural network, then the model will predict less accurately and the loss will be higher.

4 Sources of error

During the challenge, after the first submission phase, SRLLC updated its Tensorflow libraries to version 2.5. This proved to be a disastrous mistake, as changes that silently broke the first input layer of our artificial neural network escaped our notice until after the end of the third year's submission phase. All results reported in this paper were generated with a post-contest evaluation process after correcting for the error.

It should also be noted that there are well known determinism issues with the Tensorflow framework, i.e. it is notoriously difficult (if not impossible) to guarantee absolute repeatability when using a GPU with Tensorflow.

4.1 Controlling for statistically significant variation

The Tensorflow global random seed was predetermined and held fixed prior to each feature's training cycle, the specific value of which is irrelevant herein.

To further control for random variation, we provided all feature columns as input to the artificial neural network when testing each individual feature column for its importance. Excluding a column would necessitate that the artificial neural network's architecture itself were modified to accomodate for that difference, and so instead of excluding each feature of interest by leaving it out, we simply set the column weight for said column to zero: this causes the network to multiply the value by 0, effectively ignoring it entirely without altering the architecture of the network, providing consistency when testing individual features for their relative importance in prediction accuracy.

4.1.1 Why is the seed value relevant?

Pseudo random number generators (i.e. software generators) employ mathematical formulas that create a sequence of values which appear random, but that are not random. A "seed value" is used to initialize the formula, and subsequent requests to generate a random number then simply return the next value in the sequence. If a specific pseudo random number generator (e.g. the Mersenne Twister PRNG) is initialized with the same seed value, the "random" sequence it produces will be identical every time. A common method of providing highly random values is to initialize a PRNG with the current time of day (assuming sub-second precision).

It is also common practice to use random values for the trainable parameters of an artificial neural network when one is initialized for training: these are the very same values that are modified during training to improve the accuracy of the network's predictions. As such, unless the training process is allowed to continue until maximum accuracy is reached for a given network architecture (often referred to as convergence), the random values used to initialize the network can have a dramatic influence on prediction accuracy after an arbitrary number of training epochs. This is the reason the random seed value was held constant.

5 Conclusions

Any interpretation of our numbers is left to the reader, but as stated previously, a higher number in the table corresponds to a higher importance of that feature column; i.e, the tables are listed with the *least important* features at the top, and the *most important* ones at the bottom.

We recommend considering only the training set numbers (table 1). This is, once again, due to the fact that – for our purposes herein – it is irrelevant whether our model generalizes without a given feature being available to it.

Table 1: Feature importance by score, over the *training* set

Feature	**Raw Mean Squared Error**
Race	0.00045499851694330
Gender	0.0004681837745010
Prior Revocations Parole	0.08780288696289062
Prior Conviction Episodes Misd	0.08787801116704941
Supervision Level First	0.08789748698472977
Prior Arrest Episodes Drug	0.08790124207735062
Prior Arrest Episodes PPViolationCharges	0.0879332646727562
DrugTests Other Positive	0.08800055831670761
Percent Days Employed	0.08805891871452332
Prior Conviction Episodes GunCharges	0.08814740926027298
Program UnexcusedAbsences	0.08834480494260788
Prior Arrest Episodes Misd	0.08850335329771042
Prior Conviction Episodes Drug	0.08944132179021835
Prison Years	0.0896102711558342
Gang Affiliated	0.08978767693042755
DrugTests Meth Positive	0.08982886373996735
Education Level	0.08994127064943314
Prior Arrest Episodes DVCharges	0.08994424343109131
Prior Conviction Episodes PPViolationCharges	0.0992644727230072
Prior Conviction Episodes Prop	0.09926749765872955
Supervision Risk Score First	0.09926806390285492
Condition Other	0.09926890581846237
Prior Arrest Episodes Felony	0.09926915913820267
Prison Offense	0.0992695763707161
Violations FailToReport	0.09926994144916534
Delinquency Reports	0.09927039593458176
Age at Release	0.09927110373973846
Prior Conviction Episodes Felony	0.09927186369895935
Condition MH SA	0.09927219152450562
Violations Instruction	0.09927263110876083
DrugTests Cocaine Positive	0.099272720515728
Residence Changes	0.09927274286746979
Residence PUMA	0.09927349537611008
Prior Conviction Episodes Viol	0.09927353262901306
Prior Arrest Episodes GunCharges	0.09927403926849365
Prior Arrest Episodes Property	0.09927469491958618
Program Attendances	0.09927473962306976
Dependents	0.09927501529455185
Violations MoveWithoutPermission	0.09927652776241302
Prior Arrest Episodes Violent	0.09927673637866974
DrugTests THC Positive	0.09927673637866974
Violations ElectronicMonitoring	0.09927763789892197
Prior Conviction Episodes DomesticViolenceCharges	0.09927768260240555
Prior Revocations Probation	0.09928061068058014
Jobs Per Year	0.09928195178508759
Avg Days per DrugTest	0.09928666055202484
Employment Exempt	0.09928678721189499
Condition Cog Ed	0.09929008781909943

Table 2: Feature importance by score, over the *validation* (holdout) set

Feature	Raw Mean Squared Error
Gang Affiliated	0.0879855528473854
Prior Arrest Episodes Violent	0.0879948660731315
Supervision Level First	0.0879963934421539
Violations ElectronicMonitoring	0.0879968702793121
Condition Cog Ed	0.088003322482109
Dependents	0.0880052596330642
Prior Conviction Episodes PPViolationCharges	0.0880066379904747
DrugTests THC Positive	0.0880077630281448
Prior Conviction Episodes Felony	0.0880099833011627
Prior Arrest Episodes Felony	0.0880128219723701
Supervision Risk Score First	0.0880137160420417
Prior Arrest Episodes Drug	0.0880161523818969
Prior Arrest Episodes GunCharges	0.0880161821842193
Prior Revocations Probation	0.088022381067276
DrugTests Other Positive	0.0880224034190177
Education Level	0.0880282893776893
Prison Years	0.0880302488803863
Prior Conviction Episodes Drug	0.088031381368637
Prior Conviction Episodes GunCharges	0.0880320891737937
Prior Conviction Episodes Prop	0.0880328640341758
Prison Offense	0.0880397260189056
Jobs Per Year	0.0880424976348877
DrugTests Cocaine Positive	0.0880450457334518
Prior Arrest Episodes Property	0.0880452841520309
Prior Arrest Episodes PPViolationCharges	0.0880637615919113
Condition Other	0.0880684703588485
Percent Days Employed	0.0880735144019126
Delinquency Reports	0.0880742520093917
Residence Changes	0.088076576590538
Prior Arrest Episodes DVCharges	0.0880781039595604
Violations MoveWithoutPermission	0.088090181350708
Age at Release	0.0880902260541915
Prior Revocations Parole	0.0880912691354751
Violations FailToReport	0.0880935713648796
Residence PUMA	0.0880966112017631
Avg Days per DrugTest	0.0881231278181076
Program Attendances	0.0881478860974311
Employment Exempt	0.0881558507680893
DrugTests Meth Positive	0.0881626829504966
Prior Conviction Episodes DomesticViolenceCharges	0.0881685316562652
Prior Conviction Episodes Viol	0.0881853029131889
Program UnexcusedAbsences	0.0882080793380737
Prior Arrest Episodes Misd	0.0883002653717994
Race	0.0883330702781677
Violations Instruction	0.0884382799267768
Prior Conviction Episodes Misd	0.0885362476110458
Condition MH SA	0.0889153108000755
Gender	0.9126847386360168