Document Title:	NIJ Report, Team VT-ISE
Author(s):	Kwok Tsui, Tai-Jung Chen, Mengqi Shen
Document Number:	305047
Date Received:	July 2022
Award Number:	NIJ Recidivism Forecasting Challenge Winning Paper

NIJ report

(**Team VT-ISE**: Kwok Tsui, Tai-Jung Chen, Mengqi Shen)

- **Were variables added to the data set? If so, detail the variables.**

  There were no new variables added to the dataset.

- **What variables were constructed? How were the variables constructed?**

  5 categorical variables were constructed to the dataset, based on the binary variables from prior arrest, prior conviction, prior revocation, conditions, and violations variable groups (the original binary variables were deleted if the new constructed categorical variables were used).

  For example, with the Prior_Arrest variable, there are two binary prioer_arrest-related variables: Prior_Arrest_Episodes_DVCharges and Prior_Arrest_Episodes_GunCharges. So we construct a single four-level categorical variable to replace the original one . FF represents both Prior_Arrest_Episodes_DVCharges and Prior_Arrest_Episodes_GunCharges variables equal to F. TF represents Prior_Arrest_Episodes_DVCharges is T, and Prior_Arrest_Episodes_GunCharges is F. FT represents Prior_Arrest_Episodes_DVCharges is F, and Prior_Arrest_Episodes_GunCharges is T. TT represents Prior_Arrest_Episodes_DVCharges is T, and Prior_Arrest_Episodes_GunCharges is T. FF, TF, TF, TT were treated as 0,1,2,3 respectively in the dataset.

- **Did you try other models? Were they close in performance? Not at all close?**

  We have considered the models from both machine learning algorithm-level and data-analytics level.

  For the male group, from the machine learning-level, we tried decision tree (year 1 2 &3), random forest (year 1 2 &3), logistic regression (year 1 2 &3), artificial neural network (ANN) (year 1 2 &3) and Adaboost decision trees (year 1 & 2). The decision tree, random forest, logistic regression and ANN have similar performance, but random forest performs slightly better in year 2 in general.

- **What type of model was used?**

  For the male year two recidivism prediction, the machine learning algorithm was random forest. The parameters were searched and optimized based on brier_score via grid cv function. Year 1 positive cases are kept in training and but dropped in validation set. The prediction is 3-level categorical as scenario 5. The predictive class 0 cases in validation set are separated from predicted class 1 and class 2 cases. For class 0 cases, the brier score is calculated by scenario 3. For class 1 and class cases the brier score is calculated by rescaling the sum of the probabilities of class 1 and class 2 to 1.

For the female forecasting, we used Logistic regression with elastic net, eXtreme Gradient Boosting (XGBoost), Random Forest, Support Vector Machine (SVM) as our candidate model. XGBoost was chosen as the final model. XGBoost has been widely recognized by the machine learning community.

In particular, for both male and female modeling, we developed a framework based on the special characteristic of the NIJ dataset. The special characteristic of the dataset is that we are aware that the testing dataset in the second year does not contain any first-year recidivism cases. Therefore, we labeled the cases in the training data into three categories: (A) the subject recidivated in the first year, (B) the subject recidivated in the second year, (C) the subject did not recidivate in both first and second year. We first trained a model to do a three-class classification. We split the training dataset into training and validation in a ratio of 6:4. We then eliminated all the first year recidivism cases in the validation set to let it be analogous to the real testing set. We trained the model by using the training set and validated it using the validation set. Since we know that the validation set does not contain any subjects that recidivate in the first year, so if the model classified the subject to be (A) the subject recidivated in the first year, it must be a wrong classification. We then extracted these cases and retrained a new model which may perform better than the 3-class model, i.e., fitting a two-class classification model using only second year data (recidivate in second year: 1; did not recidivate in second year: 0). In short, we use a three-class classifier as a filter to extract the cases that are misclassified, then train another model to deal with these misclassified cases to improve the overall classification performance, and then combine the results from the two models.

- **What other evaluation metrics should have been considered/used for this Challenge? For example, using false negatives in the penalty function.**

  Due to the imbalance of two classes (the overall positive rate of recidivism for year 1 to year 3 are around 0.2 or 0.3), the model tends to predict most cases as negative, which means the model will perform badly on the real positive cases. That is, there will be less false positive in the prediction, which will actually result to a good Fair and Accurate metrics. If false negative is important, maybe a new classification metric should be considered by weighting the sensitivity and specificity differently. Currently, Brier score is already a continuous version of the standard classification accuracy by weighting the sensitivity and specificity equally.

- **Did the fact that the fairness penalty only considered false positives affect your submission?**

  The fact that the fairness penalty only considered false positives did not affect the choice of models and algorithms as both the Brier score and fairness penalty are consistent under the unbalanced classes in the data (as explained in the previous bullet item).

- **Are there practical/applied findings that could help the field based on your work? If yes, what are they?**

  In the NIJ challenge, we develop a framework based on the special characteristic of the NIJ dataset, which may be helpful for the field. The special characteristic of the dataset is that we are aware that the testing dataset in the second year does not contain any first-year recidivism cases. Therefore, we labeled the cases in the training data into three categories: (A) the subject recidivated in the first year, (B) the subject recidivated in the second year, (C) the subject did not recidivate in both first and second year. We first trained a model to do a three-class classification. We split the training dataset into training and validation in a ratio of 6:4. We then eliminated all the first year recidivism cases in the validation set to let it be analogous to the real testing set. We trained the model by using the training set and validated it using the validation set. Since we know that the validation set does not contain any subjects that recidivate in the first year, so if the model classified the subject to be (A) the subject recidivated in the first year, it must be a wrong classification. We then extracted these cases and retrained a new model which may perform better than the 3-class model, i.e., fitting a two-class classification model using only second year data (recidivate in second year: 1; did not recidivate in second year: 0). In short, we use a three-class classifier as a filter to extract the cases that are misclassified, then train another model to deal with these misclassified cases to improve the overall classification performance, and then combine the results from the two models.

- **What should NIJ have considered changing (other than metrics) to improve this Challenge?**

  There were only two participants in the student group. Thus, NIJ should encourage more students to participate in the challenge. The fairness penalty might not be a good evaluation metrics for the final results due to the inconsistency with the Brier score and classification accuracy.

- **For future Challenges, what should NIJ consider changing to improve Challenges? For example, more/less time, different topics, or data issues (missing data)?**

  In terms of missing data, “Gang Affiliated” was missed for all the female subjects. Although we apply imputation algorithm to those subjects, it will be helpful to have more information for those variables containing nan values.

  More time will be better. However, given there were three stages, the time duration of two weeks for each stage seemed to be reasonable.

- **Which variables were statistically significant?**

  For the male dataset, the top five important variables are Percent_Days_Employed, Jobs_Per_Year, Avg_Days_per_DrugTest, Residence_PUMA, Supervision_Risk_Score_First.

  For the female dataset, we selected eXtreme Gradient Boosting (XGBoost) as the final model. The top 5 important features are: 'Percent_Days_Employed', 'Gang_Affiliated', 'Prior_Arrest_Episodes_PPViolationCharges', and 'Jobs_Per_Year'.

- **What variables were not statistically significant? How was this handled? For example, were they dropped from the overall model?**

  For the male group. Employment_Exempt, Race, Gender are the three least significant variables.

  For the female group, The top 3 unimportant features are: 'Prior_Conviction_Episodes_Drug', 'Dependents', and 'Prior_Conviction_Episodes_Felony'. Note that since the nature of the XGBoost model is a tree model, the unimportant features will receive a relatively low coefficient, but it won' t automatically eliminate those unimportant features.

Acknowledgement: We would like to thank our collaborators from the Department of Industrial, Manufacturing, and Systems Engineering and COSMOS at the University of Texas-Arlington, Dr, Tory Chen, Dr. Chen Kan, Dr. Yuan and their students, for helpful discussions and sharing with processed data and information.