A Statistical Approach On Prediction Of Offender Group Reconvinction

Pritam Mishra
ID 35390350

***Abstract*— The criminal courts in England and Wales may request the probation service to submit pre-sentence reports which are considered by magistrates and judges before making their sentencing decision. Pre-sentence reports must include an assessment of the risk of re-offending and the risk of harm to the public which the convicted offender presents. This project provides a statistical aid to such risk assessment. The scope of this project is to assess criminology data of an offender group and devise a comprehensive measure to determine re-offending score for any offender in future. In this project several advanced statistical tools including but not limited to Neural Network, Logistic Regression, Tree based models, Ensemble Systems have been applied to ensure high prediction accuracy of re-conviction score.**

I. INTRODUCTION

On November, 1996 the Home Office launched the offender group re-conviction scale (OGRS), a statistical risk score for use by the probation service in England and Wales. The score estimates, as far as is possible from the limited information that it uses, the probability that a convicted offender will be re-convicted at least once within the subsequent period of 2 years. . The main aim of the score is to provide background information for probation officers in their writing of pre-sentence reports, these reports being designed to inform judges and magistrates when deciding what sentence might be appropriate for each individual offender[1]. The initial version of OGRS was based on a logistic regression analysis of data on a large sample of offenders who have been convicted in the recent past on a subsequent 2-year history of re-convictions traced through official records. It continued to ORGS3 which improved the overall performance of ORGS in terms of prediction time, accuracy based on a fewer simple risk factors[2].

Although, the prediction accuracy was determined by both ORGS and ORGS3 based on static risk factors like age, gender or other criminal history; the accuracy could further be improved if dynamic risk factors or secondary risk factors are analyzed which is introduced in ORGS4[3]. The scope of this project is to obtain high accuracy to predict re-conviction of offenders which could provide further insights into primary factors responsible for re-conviction based on the static variables from the data-set. In this project, several statistical tools have been implemented to pre-process the data and further compare multiple models based on their performance on validation set and average cross validation sets. Finally the best model has been identified with optimal hyper-parameters to be tested on the test data-set. It is likely that this model could provide even better accuracy if secondary risk factors are assessed for prediction.

II. METHODOLOGY

This project utilizes the offender group re-conviction data-set, which contains 21 columns and 3449 rows of data; with OINUM as the offender identification number, several features referring to past convictions and RECONV as the target variable or offender re-convicted or not before 1999, for training and validation purposes. A further test data-set has been assessed on the best trained and validated model with optimal hyper-parameters. The methodology implemented in this project includes, exploratory data analysis, pre-processing and training several models with distinct hyper-parameters and finally choosing the optimal model with hyper-parameters based on the performance on the validation set or validation sets(for k-fold cross-validation). The training data-set has been split into training and validation data-set as 70 percent and 30 percent respectively for this project.

A. Exploratory Data Analysis

The exploratory data analysis section helps to provide some key insights based on the observation on the training data-set. The relationship between the target variable RECONV with other static variables has been studied and explored in this section.

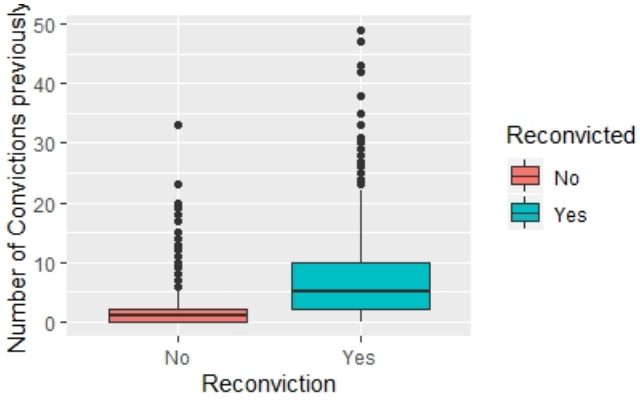

Figure 1: Reconviction vs Number of Convictions previously

From Figure 1, it is clear that, there is a strong relationship between target variable RECONV and NUMCONV(number of previous convictions). It can be observed that people with several years of conviction previously are more likely to be re-convicted again based on the data.

Also, re-conviction has a strong relationship with type of principal offense committed (TARGOFF), this could vary from robbery, burglary to theft and sexual offense. This can be analysed from the following figure that, offenders who committed robbery, burglary and other offenses are very

likely to be re-convicted compared to sexual offense, criminal damage. Also, it is quite undecisive to say if the offender will be re-convicted for crimes like Fraud and forgery, violence against person, theft.

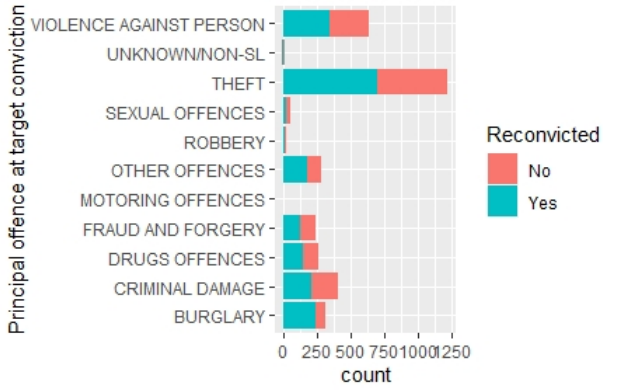

Figure 2: Reconviction vs Principal Offense

Further, it can be observed re-conviction relies on whether the offender had a custodial sentence or not (CUST); this can be observed in the following figure.It can be concluded, people with custodial sentence are more prone to re-conviction while for community sentence it has 50 percent chance of getting re-convicted hence nothing can be concluded.

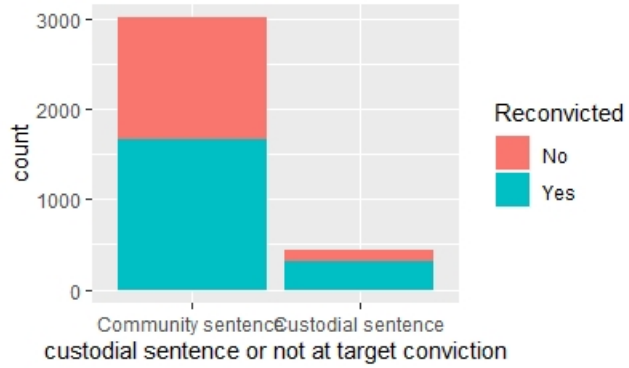

Figure 3: Community Sentence or Custodial vs Reconviction

It can be observed from figure 4, number of previous youth custody sentences has a strong relationship with re-conviction, as we see offenders with more youth custody sentences are very likely to be reconvicted.

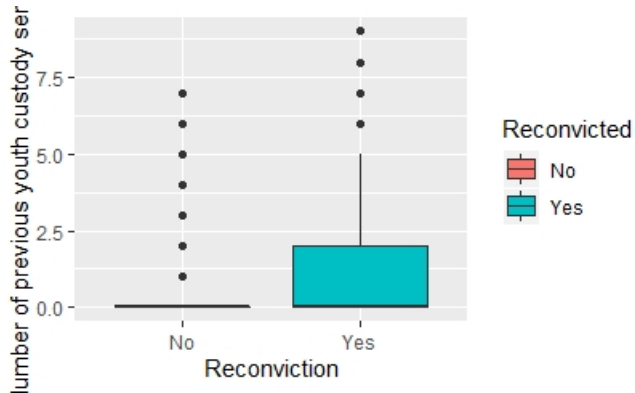

Figure 4: Number of youth custody sentence vs Reconviction

B. Pre-Processing

The re-conviction data-set contains both numerical and categorical features, therefore some distance based supervised algorithm requires all of the variables into numeric rather than both together. Ideally, any tree-based model works best with both numeric and categorical data. In this project, categorical features have been encoded into numeric data using one-hot encoding in order to apply the data on a neural network. Furthermore, the encoded data has been scaled so that it can produce optimal accuracy for neural network.

C. Checking Class Imbalance

The target variable RECONV has been analysed if there is any class imbalance on the data-set. Based on the observation, we can conclude there is no major or significant class imbalance(major class 1.3 times the minor class) present on this data-set, therefore accuracy will not provide any misleading results for class imbalance, although repeated cross-validation has been performed in this project.

```
summary(data$RECONV)
summary(data$RECONV)[2]/summary(data$RECONV)[1]
```

```
  No  Yes
1470 1979
     Yes
1.346259
```

D. Modeling

1) Decision Tree: In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Decision Tree models are created using 2 steps: Induction and Pruning. Induction is where we actually build the tree i.e set all of the hierarchical decision boundaries based on our data. Because of the nature of training, decision trees they can be prone to major overfitting. Pruning is the process of removing the unnecessary structure from a decision tree, effectively reducing the complexity to combat overfitting with the added bonus of making it even easier to interpret.

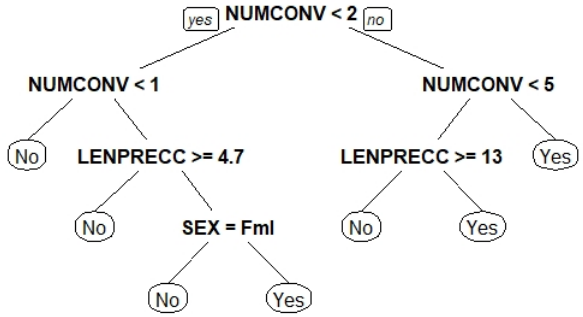

Figure 5: Decision Tree on reconviction datas-set

In this project, repeated cross-validation has been implemented to train a decision tree using package caret and

further rpart2 library has been used to manually adjust the hyper parameters of the decision tree for optimal tuning of hyper-parameters. Finally these models have been compared to determine the model with optimal performance.

2) Multi layer Perceptron: This is one of the advanced machine learning algorithms that has been implemented in this project for optimal performance on the given data-set. In this project several variations of multi layer perceptron has been simulated and cross-validated to get the average accuracy of the model after repeated cross-validation. For MLP, one hidden layer has been implemented for one model with Softmax activation function and learning rate 0.01, while in another model two layers were implemented with varying number of units(tried and tested for best results); LEAKY RELU and softmax were used as activation functions for the layers respectively. The learning rate has been devised as 0.05 for the second model for best results. The number of epochs has been simulated several times to provide the optimal value for that model.

Furthermore, dropout has been implemented in both models as part of the regularisation, which reduces the chance of overfitting the model and improves further accuracy as well.

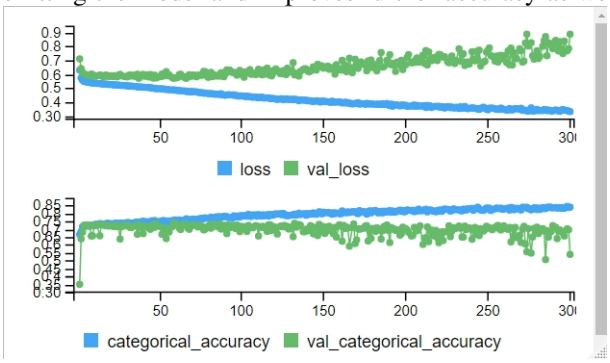

Figure 6: MLP- Training vs Validation Loss and Accuracy

3) Logistic Regression: Logistic Regression measures the relationship between the target variable (labels) and the one or more independent variables (features), by estimating probabilities using it's underlying logistic function.These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the sigmoid function. The sigmoid function takes a any number as input and map it into a value between the range of 0 and 1. The value in between 0 and 1 can then be transformed into either 0 or 1 using a threshold classifier.

$$f(x) = \frac{1}{1+e^x} \tag{1}$$

4) Random Forest: Random Forest is one of the most common ensemble systems that perform really well on any data-set. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features.

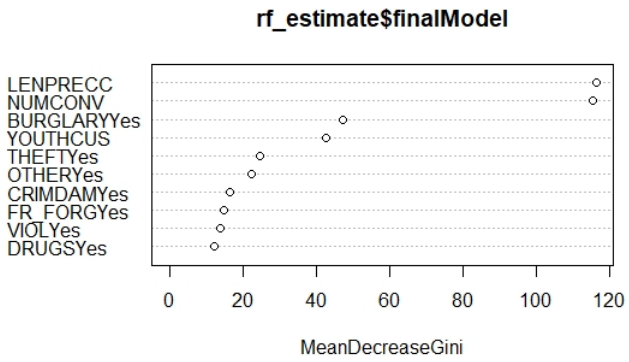

Figure 7: Random Forest: Feature Importance

III. Evaluation and Discussion

The models that were discussed previously were cross-validated or validated using train and validation split. The final model has been selected based on the highest accuracy found on a model with optimal hyper-parameters. The comparison of these models with accuracy metrics can be shown with the following table.

Model Name	Hyper-Parameters	Accuracy
MLP I	Layer=1,Dropout rate=0.2,$lr = 0.01, activation = softmax$	70.6
MLP II	Layer=2,$activation1 = leaky - relu, activation2 = softmax, lr = 0.05$	61.7
Logistic Regression	$cutpoint = 0.5, k - fold = 5$	73.5
Decision Tree I	$k - fold = 5, repeats = 10, maxdepth = (1to10)$	73.98
Decision Tree II	$minsplit = 3, cp = 0, maxdepth = 5$	72.82
Decision Tree II Pruned	$minsplit = 3, cp = 0.009, maxdepth = 5$	74.18
Random Forest	$k - fold = 5, repeats = 10, mtry = c(1, 2, 3)$	73.5

We can conclude, even though some of the advanced models were expected to yield very high accuracy, that was not the case in real data. In fact, it's really surprising how close the accuracy of a very simple yet effective model Logistic Regression has performed. This could be one of the reasons why logistic regression were used in ORGS for several years in it's iterations. Although, Random Forest should have had the highest accuracy among all the other models, as it makes decisions based on a voting scheme of hundreds of decision trees. This could be the accuracy metric based on a random state which probably did not perform better than the pruned decision tree. Nevertheless, all of the supervised algorithm with optimal hyper-parameters

have performed really well considering the data-set. Based on our observation, we can conclude almost every algorithm is performing in a similar way to produce highest accuray even though decision tree II with pruning produced the highest accuracy in this case. Therefore, in this case on that random state our best model would be decision tree II pruned. However, a more balanced more reliable decision making algorithm would be Random Forest. If a recommendation needs to be made to the ministry of justice for determining outcome of a sentence, random forest could be a more reliable model for an unknown data-set. The other advantages of random forest includes displaying the most important features based on mean gini index.

IV. Conclusion

This could be concluded that, although most of the supervised models implemented in this project did not yield a very high accuracy. The accuracy could be further improved if dynamic variables or secondary risk factors such as marital status, housing status, number of children could be assessed in this project[3].

References

[1] "The offender group reconviction scale: a statistical reconviction score for use by probation officers)," Copas and Marshall, 1996.

[2] The revised Offender Group Reconviction Scale .Howard, Francis, Soothill and Humphreys (2009).

[3] A compendium of research and analysis on the Offender Assessment System. Moore, (2016).