RECIDIVISM

PREDICTION OF PAROLE VIOLATION

CONTENTS

INTRODUCTION

The last four decades in the United States have been characterized by a dramatic increase in the number of people incarcerated in jails and prisons. Since the early 1980s, the rate of incarceration has more than quintupled, with over 2.2 million individuals currently behind bars (Bureau of Justice Statistics, 2016)[1]. Although almost all of these people maintain their freedom upon release, over two-thirds of former prisoners commit new crimes which leads them back into the system. As such, the prison institution is often referred to as a 'revolving door'. ***Recidivism*** is the tendency of a convicted criminal to reoffend. It "is measured by criminal acts that result in re-arrest, re-conviction or return to prison with or without a new sentence during a three-year period following the prisoner's release."[2]

In many criminal justice systems around the world, prisoners deemed not to be a threat to society are released from prison under the parole system prior to completing their sentence. Parole is designed to function as a critical surveillance and rehabilitative mechanism for offenders transitioning from prison to the community. Upon exiting prison, individuals under parole are given strict conditions to which they must adhere, such as regular visits with their assigned parole officer, refraining from committing new crimes, getting employed, remaining drug-free, etc…They are still considered to be serving their sentence while on parole, and can be returned to prison if the terms of their parole are violated. Parole boards are responsible for identifying which inmates are good candidates for release on parole. They aim to release inmates who will not commit additional crimes after release. Failure to accurately predict the risk of recidivism may lead to harmful consequences if a dangerous individual is put back into society.

The goal of this project is two-fold: first, I endeavor to build and validate a model that predicts if an inmate will violate the terms of his or her parole. Such a model could be useful to a parole board when deciding to approve or deny an application for parole. Secondly, I intend to help parole boards identify which key variables make a parolee more likely to relapse. For the analysis, I will use a subset of data from the U.S 2004 National Corrections Reporting Program, a nationwide census of parole releases that occurred during 2004. The subset of data contains observations of 675 parolees who either successfully completed or violated the terms of their parole in 2004. As inputs, the model will take various demographic attributes such as age, race, and gender, as well as variables such as crime committed, and length of time served in prison. The output will be a binary response variable predicting the likelihood of parole violation (1 meaning 'violated' and 0 meaning 'not violated'). Since the target variable is categorical, I will use classification algorithms to build a robust predictive model. I will explore the use of Tree-Based algorithms as well as Logistic Regression. Cross-validation methods will subsequently be used to test the model performances and help arrive at the model with the optimal predictive power, measured by the lowest error rate of the classification prediction. All analysis will be performed using R Studio.

[1] Siege, Jonah l Aaron. "Prisoner Reentry, Parole Violations, and the Persistence of the Surveillance State." University of Michigan, U.S. Department of Justice, 2014, pp. 1–4.

[2] Office of Justice Programs, U.S. "Recidivism." National Institute of Justice, 2014, nij.ojp.gov/topics/corrections/recidivism.

DATA DESCRIPTION

The subset of data derived from the U.S 2004 National Corrections Reporting Program contains observations of 675 parolees who either successfully completed or violated their terms of parole in 2004. This subset limits our scope to parolees who served no more than 6 months in prison and whose maximum sentence did not exceed 18 months.

Response Variable

- *violator*: 1 = the parolee violated the parole, and 0 = the parolee completed the parole without violation.

Predictors

- *male*: 1= the parolee is male, 0 = the parolee is female
- *race*: 1 = the parolee is white, 2 = the parolee is non-white
- *age*: the parolee's age (in years) when he or she was released from prison
- *state*: a code for the parolee's state. 2 = Kentucky, 3 = Louisiana, 4 = Virginia, and 1 = any 'Other' state. The three states were selected due to having a high representation in the original dataset.
- *time.served*: the number of months the parolee served in prison (limited by the inclusion criteria which does not exceed 6 months).
- *max.sentence*: the maximum sentence length for all charges, in months (limited by the inclusion criteria which does not exceed 18 months).
- *multiple.offenses*: 1 = the parolee was incarcerated for multiple offenses, 0 = the parolee was incarcerated for just one offense.
- *crime*: a code for the parolee's main crime leading to incarceration. 2 = larceny, 3 = drug-related crime, 4 = driving-related crime, and 1 = any 'Other' crime.

I began my analysis by examining the distribution of each predictor in the dataset, and visualizing the relationship between the predictors and the target variable. When observing the relationships, I ensured that the categorical variables were recognized by R as categorical, by transforming them using the *as.factor* function. [See Appendix for detailed descriptions on variable distributions, as well as visualizations].

MODEL SELECTION & METHODOLOGY

My target variable is categorical; therefore, I will use a classification model to predict the likelihood of a parolee violating his or her parole.

SIMPLE DECISION TREE

I first considered a tree-based model. Decision trees are very popular and work well with classification tasks because they can handle categorical predictors easily. Tree models are computationally simple and quick to fit, even for large problems. Additionally, they can handle non-linear interactions effectively. Moreover, the terminal nodes of the model naturally cluster the data into homogenous groups.

Using my understanding of each predictor's relationship on the target variable, I opted not to exclude any of the variables. This is because Decision Trees also play a key role in feature selection when building the model. Trees are able to rank variables by importance and use the most important features for the optimal model.

The goal of a Decision Tree is to split the branches at the right internal nodes in order to minimize the Residual Sum of Squares.

$$RSS = \sum_{j} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

To minimize RSS, we use an algorithm called Recursive Binary Splitting. Trees have a tendency to overfit if we let them get fully grown. If the tree is too big, the lower "branches" tend to model noise in the data. Therefore, to build the optimal tree and obtain the best predictions, we prune the tree by using a complexity parameter (cp). CP controls the growth of the decision tree and selects the optimal size. To find the optimal CP value, the usual paradigm is to start with a very complex tree (small cp of 0.0001) and "prune" back unnecessary splits. Then, using the 'out-of-bag' cross-validation technique, we study the out-of-sample performance of our tree model to decide how complex we want to make the tree. By plotting the out-of-performance error as a function of CP (Figure 18), we note that the lowest error is obtained when CP is 0.017. Thus, this is our optimal CP and we use this value to build our optimal Decision Tree.

Results & Interpretation

Our optimal tree (Figure 19) has 4 internal nodes and 5 terminal nodes. By observing the tree, we can see the feature selection mechanism at play. Only 4 out of the 7 predictors are taken into account when building the optimal tree. The most important factor influencing the *violator* output variable is {*state; 1,2,4*} because this is the variable that the tree considers in first internal node. Among parolees from state 3 (Louisiana), the next most important criterion of parole violation is {*multiple.offenses, 0*}. Among parolees from Louisiana with multiple offenses, the third most important variable is {*max.sentence, 13*}. Lastly, among this subset of parolees with a maximum sentence of less than 13 months, the fourth most important feature is {*time.served, 2.6*}.

88% of observations ended up in terminal node 1. In this node, we predict that parolees who are from state 1,2,4 will not violate their parole. 4% of observations ended up in the 2nd terminal node. In this node, we predict that parolees from Louisiana who only have one offense will likely not violate their probation. Looking at the 3rd terminal node, we can see that 4% of observations fell into this node. Here, we predict that parolees from Louisiana who were convicted for multiple offences, had a maximum sentence of less than 13 months and served more than 2.6 months are also likely not to violate their parole.

On the other hand, the last two terminal nodes classified parolees as likely to violate their parole. The fourth terminal node captures 3% of observations and predicts that parolees from Louisiana, who have committed multiple offenses, have a maximum sentence less than 13 months, and have served less than or equal to 2.6 months of their sentence are likely to re-offend. Finally, the last terminal node captures 1% of observations and predicts that parolees from Louisiana, with multiple offences, and a max sentence of more than 13 months are also likely to re-offend.

Model Accuracy

To measure the performance of this model, we calculate the error rate by multiplying the *root node error* by the *xerror* of the last split level (these values are given in R via the classification summary output)

$$Error\ rate\ of\ simple\ tree = 0.11556 \times 1.0256 = 0.1185$$

This means that about 11.85% of all observations are misclassified. In other words, if the parole board uses this model, 11.85% of parolees who are deemed fit to be released under parole, might end up re-offending. Although this error might seem small, in the case of criminal activity, this is quite significant as it could be a matter of life and death. Therefore, it is essential to find a way to increase the accuracy of our model and decreasing the error rate.

RANDOM FOREST

Simple Decision Trees tend to suffer from high variance, leading to overfitting and poor model performance. In order to fix this, I chose to use the Random Forest Algorithm. Random Forest is a type of ensemble machine learning algorithm called "bootstrap aggregation" or bagging. Bagging is a method used to increase the predictive power of Decision Trees. It works by selecting a bootstrapped sample of *n* observations with replacements and fits a tree on each sample. Then, it merges the trees together by taking the average response from each observation across all tress that that observation appears in. In addition to this bagging process, Random Forest only uses a subset, *m,* of the predictors, *p,* when building each tree (where $m = \sqrt{p}$). The benefit of a bagged forest is that it reduces the model complexity and the variance we might get from a simple tree, given that a tree is highly dependent on the data it is fed. As such, a Forest is more representative of the entire dataset and leads to more accurate and stable predictions. Another advantage of using Random Forest is the fact that Simple Trees could have a higher level of instability; a little change in data can result in a large change in the tree's structure. However, with a Random Forest, if we change the data a little, the individual trees may change but the forest is relatively stable because it is an average combination of many trees.

Model Accuracy

After running the Random Forest algorithm, the out-of-bag error was 11.70%. This means that 11.70% of observations were misclassified. Although not a large improvement from the Simple Decision Tree, it still shows that Random Forest produces a model with a higher prediction accuracy. It is important to note that one disadvantage of the Random Forest algorithm is that we lose interpretability of the tree. Unlike a simple tree, we are unable to plot the tree as we had done in Figure 19.

In order to further improve the accuracy of a tree, the Boosting algorithm can be used. Boosting is similar to bagging, with two key additional features: 1) it assigns different weights to different observations, so some appear in the new datasets more than others; and 2) the trees are built and trained sequentially and take into account the previous tree's performance, which helps in the generalizability of the model. I also performed the boosting algorithm on my dataset, however, the error rate was 13.33%. In this case, Random Forest performed better than the boosting algorithm.

LOGISTIC REGRESSION

When working on a predictive task, it is always wise to test several different models in order to identify which works best and which is most accurate. Logistics regression is one of the most popular classification algorithms particularly when the target variable is binary.

In Logistic Regression, we use the logistic function to compute probabilities:

$$\rho(X) = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$

Logistic regression uses Maximum likelihood estimation. It seeks coefficient values (β_i) that try to maximize *L* and give a probability ρ(x) that most closely resembles observed data:

$$L(b_0 , b_1) = \prod_{i:y=1} p(x) \prod_{i:y=0} (1 - p(x_i))$$

It is important to note that estimating exponential functions is computationally intensive. Therefore, the Logistic Regression model transforms the $\rho(X)$ function to a linear one, which estimates the *odds* (ln(odds)) rather than the probability:

$$\ln\left(\frac{p(X)}{1 - p(X)}\right) = b_0 + b_1 X$$

When building the model, I did not exclude any predictors. I split the data into training (80%) and test (20%) sets in order to perform cross-validation after the model is built.

Results & Interpretation

The results (Table 3) show that:

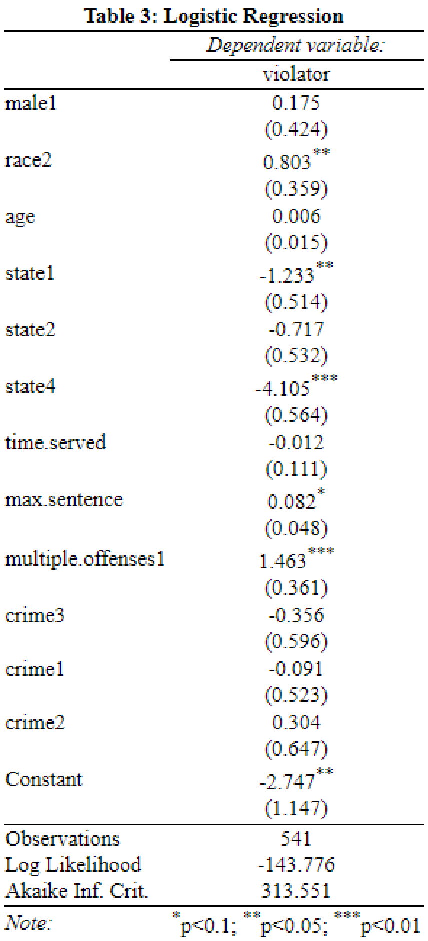


Table 3: Logistic Regression

	Dependent variable:
	violator
male1	0.175
	(0.424)
race2	0.803^{**}
	(0.359)
age	0.006
	(0.015)
state1	-1.233^{**}
	(0.514)
state2	-0.717
	(0.532)
state4	-4.105^{***}
	(0.564)
time.served	-0.012
	(0.111)
max.sentence	0.082^{*}
	(0.048)
multiple.offenses1	1.463^{***}
	(0.361)
crime3	-0.356
	(0.596)
crime1	-0.091
	(0.523)
crime2	0.304
	(0.647)
Constant	-2.747^{**}
	(1.147)
Observations	541
Log Likelihood	-143.776
Akaike Inf. Crit.	313.551
Note:	*p<0.1; **p<0.05; ***p<0.01

- In comparison to a female parolee, a male parolee has 0.18 times higher odds of violating their parole
- In comparison to a White parolee, the odds of a non-White parolee violating their parole is 0.803 times higher.
- An increase in age by 1 year increases the odds of parole violation by 0.006
- In comparison to a parolee from state 3 (Louisiana), a parolee from state 1 ('Other') has 1.23 times lower odds of violating their parole, a parolee from state 2 (Kentucky) has 0.717 times lower odds of violating their parole, and a parolee from state 4 (Virginia) has 4.11 times lower odds of parole violation.
- An increase in time served by 1 month decreases the odds of parole violation by 0.012
- An increase in maximum sentence by 1 month increases the odds of parole violation by 0.082
- In comparison to a parolee with only one multiple offense, a parolee with multiple offenses has 1.43 times higher odds of being a violator of his or her parole
- In comparison to a parolee who was convicted of a driving-related crime, a drug-related criminal has 0.36 times lower odds of violating their parole, a parolee who committed a crime in the 'Other' category has 0.091 times lower odds of violating their parole, but a larceny criminal has 0.30 times higher odds of parole violation.

Model Accuracy

The R-Squared value of this Logistic Regression model is 0.334. Using cross-validation, the error rate is 10.45%, which makes this the best performing model in this analysis.

Interestingly, I build another Logistic Regression model using only the 4 variables that the Decision Tree deemed most important (*state, multiple offenses, time served* and *max sentence)*, and the model performed worse. The R-Squared value was 0.315 and the error rate was 11.94%. Therefore, features that work well for one model, may not be the best for another model and trial-and-error is important.

FEATURE SELECTION

In addition to creating a robust classification model, the second goal of this analysis is to help the parole board identify which key variables greatly influence a parolee's propensity to violate his or her parole.

DECISION TREE

As seen above, tree-based algorithms have a great feature which aids in identifying the importance of each variable. Rather than relying on the visual interpretation of the importance from Figure 19, I calculated the statistical importance values to justify the previous findings from above.

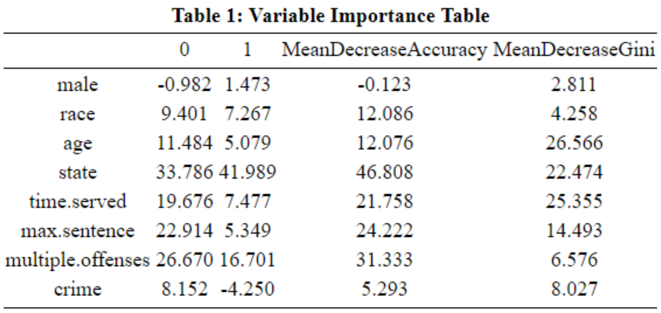


Table 1: Variable Importance Table

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
male	-0.982	1.473	-0.123	2.811
race	9.401	7.267	12.086	4.258
age	11.484	5.079	12.076	26.566
state	33.786	41.989	46.808	22.474
time.served	19.676	7.477	21.758	25.355
max.sentence	22.914	5.349	24.222	14.493
multiple.offenses	26.670	16.701	31.333	6.576
crime	8.152	-4.250	5.293	8.027

Table 1 shows the different coefficients attributed to each variable which pertains to their level of importance. Figure 20 plots these values. *State* is the most important variable because if removed from the model, this will have the largest effect on the accuracy – it will decrease the model's accuracy by 46.81%. *Multiple offenses* is the second most important variable, and if removed, it will decrease the model accuracy by 31.33%. Third most important feature is *max.sentence* and if removed, the model's accuracy will decrease by 24.22%. The fourth most important variable is *time served* and if excluded from the model, the accuracy will decrease by 21.76%.

PRINCIPAL COMPONENT ANALYSIS (PCA)

In order to confirm that these are indeed the most important variables, we can also use the Principal Component Analysis (PCA) technique. PCA is an unsupervised learning technique which allows us to visualize highly complicated data, simply, by reducing the data's complexity. The idea is that, although the observations in the dataset lie in a *p*-dimensional space (i.e, *p* predictors), not all dimensions are equally significant. PCA produces a low-dimensional representation of the data that captures as much information and variability as possible.

When we analyze the first two principal components in the dataset, we see in Table 2 that in the first component, the variable that accounts for the most variability in the data is *state* (which has a loading coefficient of -0.59), followed by *multiple.offenses* (-0.53). In the second component, *max.sentence* (0.73) and *time.served* (0.46) are the variables that account for the most variability. This mirrors the findings from the importance plot derived from the Decision Tree, although not entirely. Using PCA, we see that *race* has a big influence on the first component (-0.43), but race is not denoted as an important feature by the Decision Tree. We also notice race's influence when we plot the two principal components using *autoplot()* (Figure 21). We observe that the variables *state, multiple offenses* and *race* are more directed towards parole violators (orange datapoints), compared to the other variables. This implies that they are reliable indicators that a parolee will violate his or her parole. Another caveat to note is that PC1 explains about 26% of the overall variance, and PC2 explains about 22% of the variance, so with both PC1 and PC2, we explain about 48% of the variance (Figure 22).

Therefore, when we combine the insights from both the importance plots and the PCA analysis, we can conclude that parole boards should perhaps weigh these key factors more than others: *state, multiple offenses, race, time served* and *max sentence.*

CLASSIFICATION/PREDICTIONS & CONCLUSIONS

Based on the analysis, it is possible to conclude that the odds of a parolee violating his or her parole is higher when he or she is from Louisiana, has committed multiple offenses, is non-White, has a long maximum sentence length, and has served a short time.

Given our models, we are now able to make predictions. Consider a prisoner who is male, of non-White race, age 35 years at prison release, from the state of Kentucky, served 4 months, has a maximum sentence of 12 months, did not commit multiple offenses, and committed a drug-related crime.

- Using the Simple Tree model, this inmate is given a 0.069 likelihood of violating his parole.
- Using the Random Forest model, the inmate is given a 0.040 likelihood of violating his parole.
- Using the Logistic Regression Model, this inmate's odds of violating his parole is 0.15. By applying the following equation: $probability = \frac{odds}{(1+odds)}$, the odds equate to a probability of 0.133.

Across all 3 models, this prisoner is classified as *0 (not violated).* Therefore, the parole board will likely agree to release him on parole. If asked to choose only one model to use, I believe that Logistic Regression, which has the lowest error rate and produced the highest probability value for the above prediction, would be best. The Logistic Regression model would lead to a higher parole application denial rate. This could mean refusing parole to potentially dangerous re-offenders, who might have otherwise been released under the guidance of the Tree models due to their lower probabilities. As is often said, "it is better to be safe than sorry."

Managerial Implications

It is well known that high parole recidivism rate is one of the major contributing factors to the growing U.S. prison population. The role of a parole board is to ensure that prisoners who are released into society do not pose a risk to the general public. As such, it is particularly important that they have the right set of data and predictive models to enable them to make confident decisions.

Through this analysis, a few key insights have been uncovered which can be very beneficial to the judicial system. First, *state* is denoted as the most important factor influencing the odds of a parolee violating his or her parole. More specifically, the state of Louisiana has a significant relationship with the target variable. The justice and rehabilitation systems in Louisiana could use this information to begin an investigation into why parolees in their state are more likely to return to prison. Perhaps they could look into strengthening the support the state provides to ex-convicts upon their re-entry into society. Studies have shown that support with employment, mental and physical health, housing, and community cohesion greatly decrease the likelihood of parole violations[3]. In conjunction with the aforementioned insight, *multiple offenses* was the second most important factor in predicting parole violation. Strengthening the rehabilitation and post-imprisonment resources for people on parole will consequently help decrease the likelihood of parolees committing another offense.

Additionally, during the analysis, I noticed a relationship whereby the longer the time served in prison, the lower the odds of violating parole. As such, parole boards should perhaps refrain from releasing prisoners too early during their prison sentence. This insight could serve to persuade the correctional system in the US to further invest in ameliorating the rehabilitation systems within prisons. In other words, the longer the prisoner serves in prison, the more time they have to take advantage of the in-prison rehabilitation services, and the better equipped they will be when released on parole. Instead of rehabilitation services being separate entities in correctional institutions and parole, the two should mix and become one process. Norway has one of the lowest recidivism rates in the world at 20%, while the USA has one of the highest rates at 60%[4]. Prisons in Norway and the Norwegian criminal justice system focus on restorative justice and rehabilitating prisoners rather than punishment, deterrence, and marginalizing suspects. Therefore, increasing the presence and quality of pre-release services may be beneficial, particularly in states with high recidivism, such as Louisiana, and especially for non-White parolees who face a higher rate of relapse.

Limitations

There were a few limitations to this analysis. Firstly, the size of the data was fairly small. When building a robust predictive/classification model, it is best advised to have a large training dataset, and 675 observations is too low. Although the results obtained seem reasonable, it will be necessary to conduct the analysis with a larger dataset in order to extend the generalizability of the results, increase accuracy, and increase the parole board's confidence in the predicted outputs.

Furthermore, this dataset looks at the number of parolees who violated their parole in 2014 only. Recidivism measures the reconviction rate within a three-year period, post-release. Therefore, ideally, the target variable used to train the models should have captured information on parolees who either violated or did not violate their parole within a 3-year period, rather than just in an isolated year.

Additionally, as previously stated, the dataset used to build out models is a subset of a larger original census from the U.S 2004 National Corrections Reporting Program. The variables included in this

[3] Robinson, Jacqueline A. (2005) "Relationship Between Parole and Recidivism in the Criminal Justice System, *McNair Scholars Journal:* Vol. 9: Iss. 1, Article 12.

[4] Sterbenz, Christina. "Why Norway's Prison System Is so Successful." Business Insider, Business Insider, 11 Dec. 2014, www.businessinsider.com/why-norways-prison-system-is-so-successful-2014-12?op=1.

subset of data are somewhat limiting. There are a few variables that I believe should have been included which would have a large influence on the target variable – such as level of education, status of employment, homelessness, presence of mental illness and presence of a physical disability. Several studies have demonstrated a link between these variables and recidivism; therefore, the inclusion of these variables might have improved the accuracy and predictive power of our models.

Lastly, the categories created for the *state* and *crime* variables seem too few. By grouping several diverse crimes under 'Others', it prevents the model from finding meaningful relationships between different types of crimes and parole violation. For instance, it can be assumed that burglars and rapists would have different propensities to re-commit a crime, because they both fall under 'Others'. If the categories for such variables were extended further, perhaps, we would have seen the importance level of the *crime* variable increase in the models.

Despite these limitations, this analysis is incredibly relevant to a great issue facing America today: *incarceration*. It is important that such studies are done to identify factors that influence the ever-increasing rates of incarceration and recidivism in order to improve the rehabilitation services given to ex-convicts, and help parolees remain out of prison. Ultimately, the loss of an adult to the prison system is the loss of a contributing member to America's economy and society.

APPENDIX

Appendix A: Exploration of Variables

Figure 1 exhibits the distribution of **age** which is largely right-skewed. The minimum age in the dataset is 18.40 years and the oldest parolee is 67 years old. The median age is 33.70 and the mean is 34.51. Figure 2 shows the relationship between *age* and *violator*. More violators fall in the younger age range, while as the parolee gets older, they are more likely to successfully complete their parole without violation. For numerical predictors, I am able to run a logistic regression model to measure the significance of the variable's relationship with the categorical response variable. For *age* and *violator*, the p-value of 0.908 suggests that this relationship is not statistically significant, as it falls below the 0.05 threshold.

Next, I examined **time served**. The distribution of this variable is left-skewed, as observed in Figure 3. The mean is 4.20 months and there are four outliers on the left tail. Figure 4 shows the relationship between *time served* and *violator*. Parolees more often violate their parole when they have served a longer time. The p-value from the logistic regression of these two variables (0.011) suggests that this relationship is statistically significant.

The distribution for **max sentence** (Figure 5) seems normal with a mean and mode of 13 months. There are five outliers on the left tail. Figure 6 shows that as length of sentence increases, so does likelihood of parole violation. The p-value from the logistic regression of these two variables (1.94e-05) suggests that this relationship is significant.

Next, I looked at the **male** variable. Most parolees in the dataset (81%) identify as male (Figure 7). I created a mosaic plot to observe the relationship between the categorical features and the target variables. By observing Figure 8, the relationship between *male* and *violator* does not seem to be significant because the number of violators remains fairly similar between male and female.

When looking at the **state** variable (Figure 9), we see that most parolees are from state 4 (Virginia), followed by 'Other', third being state 2 (Kentucky), and last being state 3 (Louisiana). Figure 10 demonstrates that there seems to be a significant relationship between Louisiana and the target variable. Parolees from Louisiana are more likely to violate their parole than parolees from the other three state categories.

When exploring the **race** variable (Figure 11), we see that most parolees (58%) identify as White. Figure 12 shows that there is a relationship between the variables: more non-White parolees tend to violate their parole than their White counterparts.

I then looked at **crime** (Figure 13). Most of the crimes committed fall into the 'Other' category, followed by drug-related crime, then larceny, and lastly driving-related crime. By looking at the mosaic plot (Figure 14), we can see that parolees convicted for driving-related crimes tend to violate their parole less than parolees who were convicted for crimes in the other three categories.

Finally, I explored the **multiple offenses** variable (Figure 15). There was a larger frequency of parolees who had committed multiple offenses than those who committed one offense. Looking at Figure 16, we can see that parolees with multiple offenses tend to be more likely to violate their parole than parolees with only one offense.

Next, I constructed a correlation matrix for all variables in the dataset to ensure that the predictors are not influenced by each other. It is important to eliminate collinearity between variables because this could lead to double-counting and instability of my model results. As observed in Figure 17, no two variables had a correlation coefficient higher than +/- 0.50 (which was the correlation between *state* and *multiple.offenses*). This is well below the ~0.80 - 0.85 threshold which would suggest a collinear relationship between variables. Therefore, collinearity is not present in the dataset.

Although outliers were detected in two of the numerical predictors (*time.served* and *max.sentence*), I chose not to remove outliers from any of the variables. Estimates of predictive power are more accurate by retaining outliers as removing them would merely be ignoring data that the model does not fit well and would reduce the generalizability and ability of the model to predict accurately in a test-setting. Moreover, the dataset is fairly small, and it would not be wise to further decrease the size.

Figure 1: **Distribution of *Age* Variable**

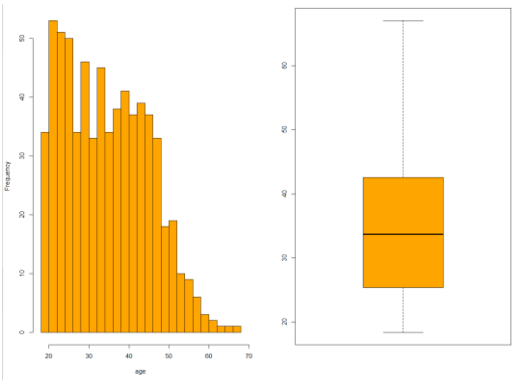


Figure 2: **Relationship between *Age* and *Violator***

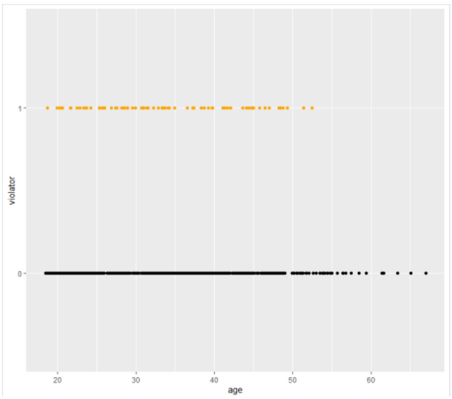


Figure 3: **Distribution of *Time Served* Variable**

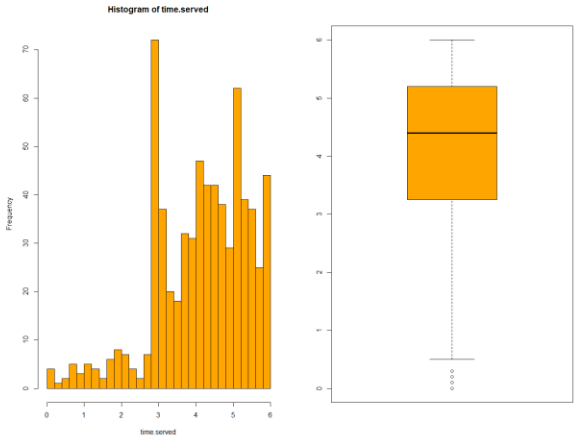


Figure 4: **Relationship between *Time Served* and *Violator***

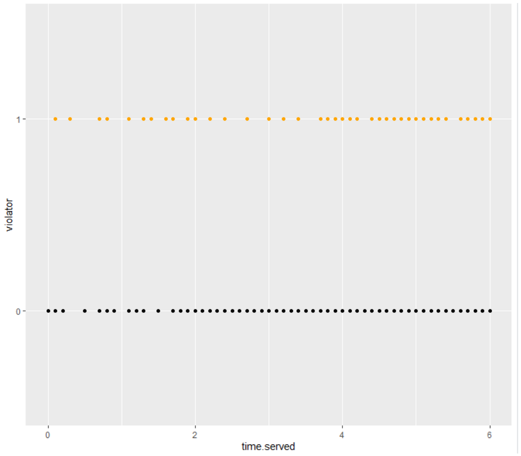


Figure 5: **Distribution of *Max Sentence* Variable**

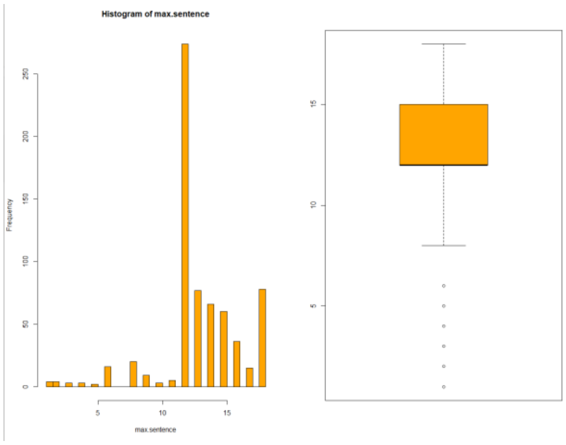


Figure 6: **Relationship between *Max Sentence* and *Violator***

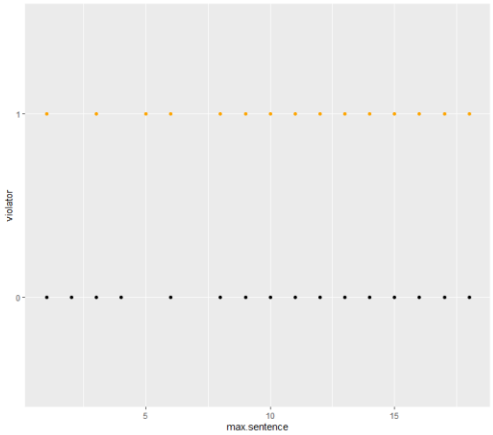


Figure 7: Distribution of *Male* Variable

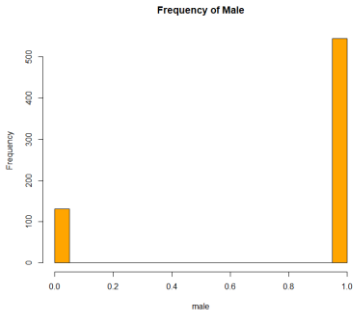


Figure 8: Relationship between *Male* and *Violator*

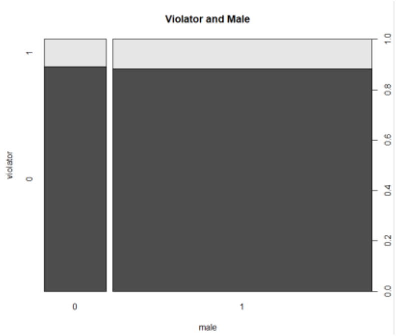


Figure 9: Distribution of *State* Variable

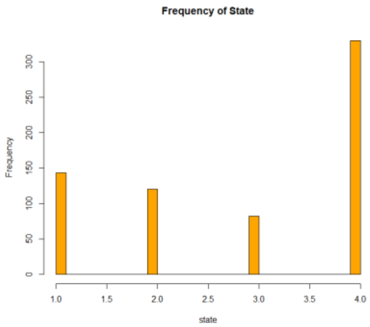


Figure 10: Relationship between *State* and *Violator*

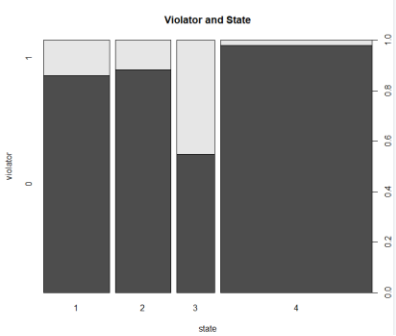


Figure 11: Distribution of *Race* Variable

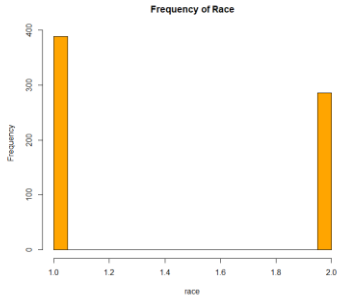


Figure 12: Relationship between *Race* and *Violator*

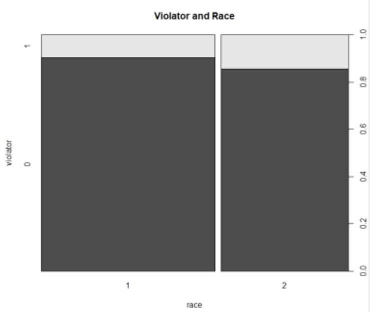


Figure 13: Distribution of *Crime* Variable

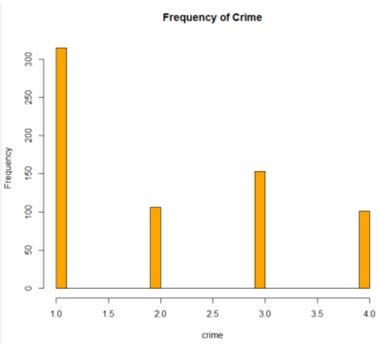


Figure 14: Relationship between *Crime* and *Violator*

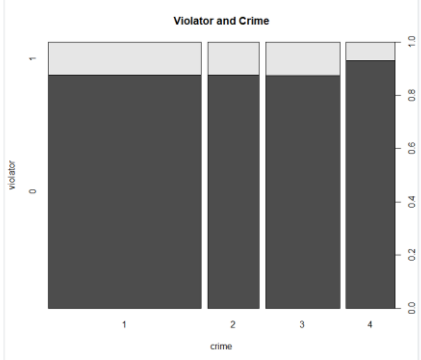


Figure 15: Distribution of *Multiple Offenses* Variable

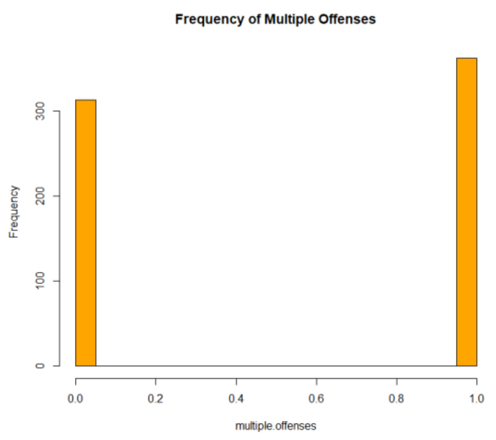


Figure 16: Relationship between *Multiple Offenses* and *Violator*

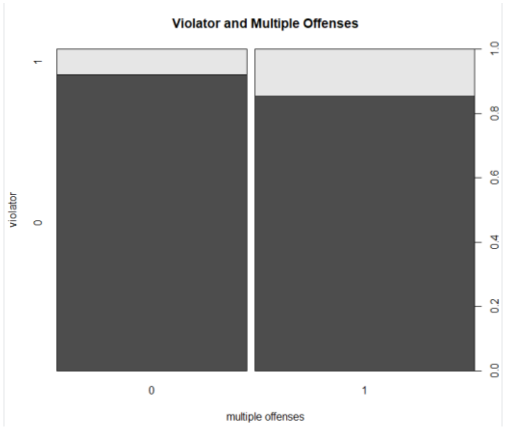


Figure 17: Correlation Matrix of all predictors

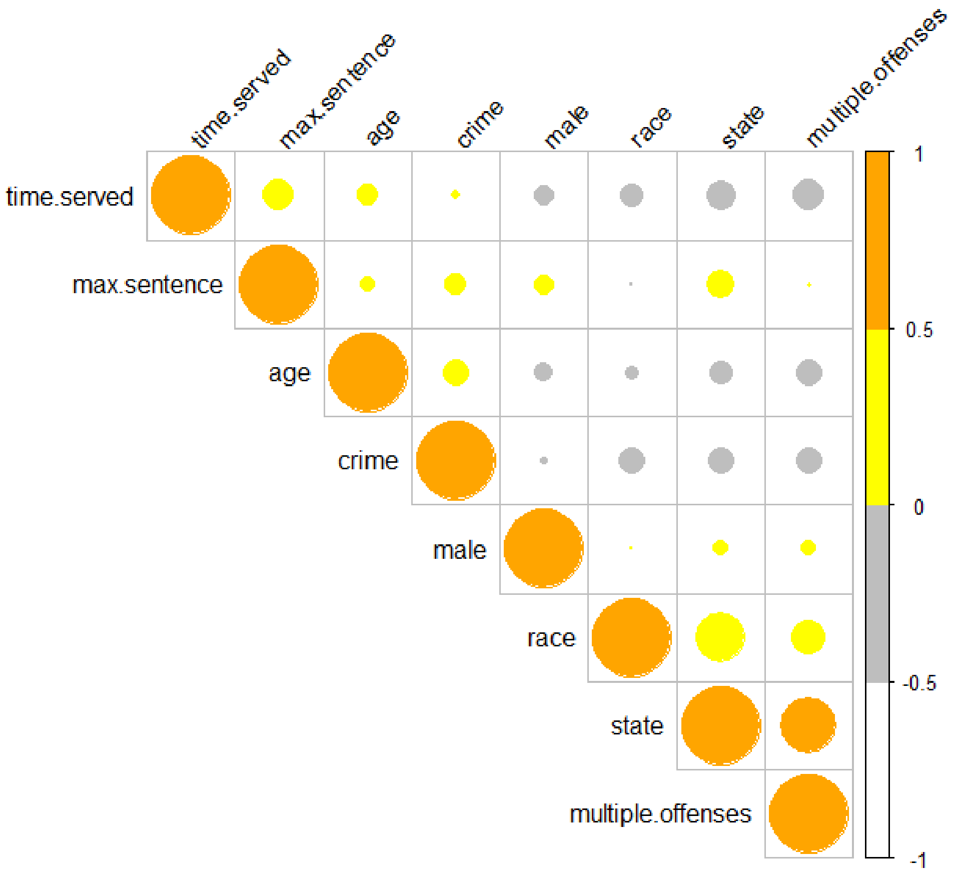


Figure 18: Optimal CP Value

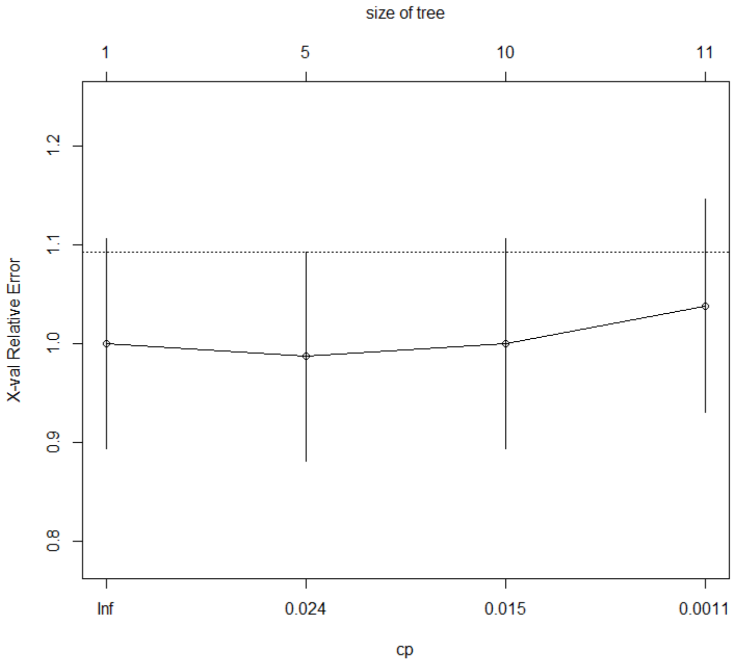


Figure 19: Simple Decision Tree

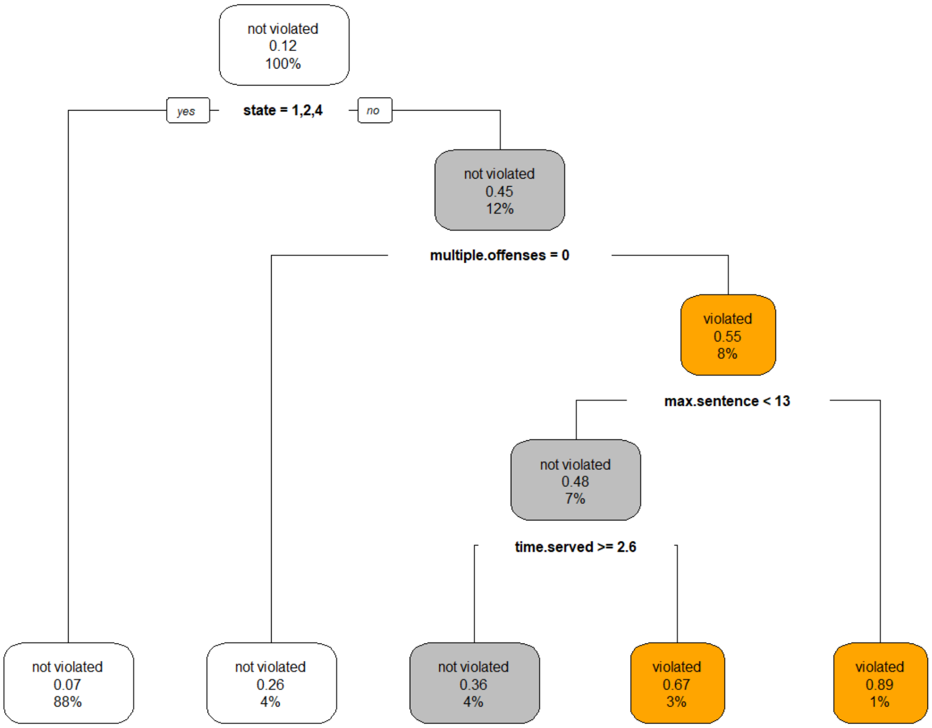


Figure 20: Variable Importance Plot

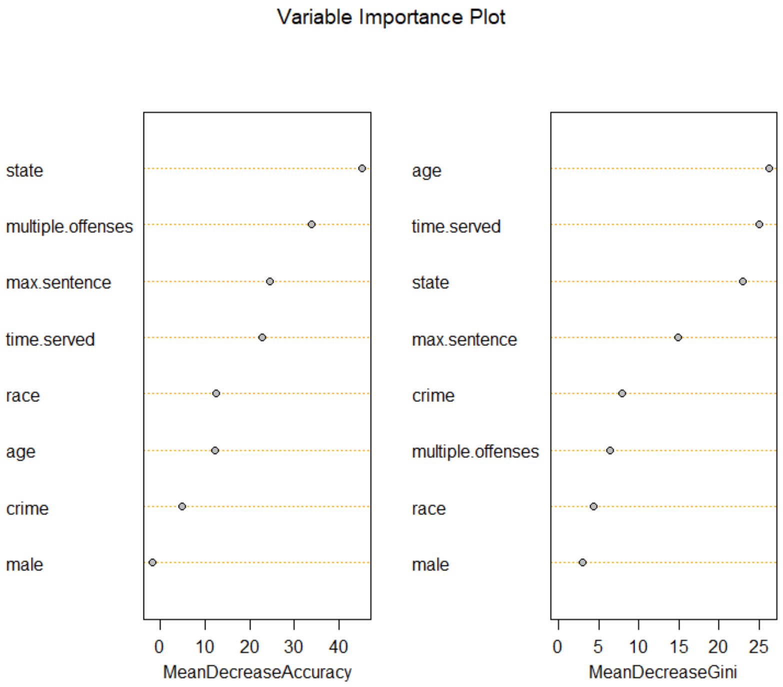
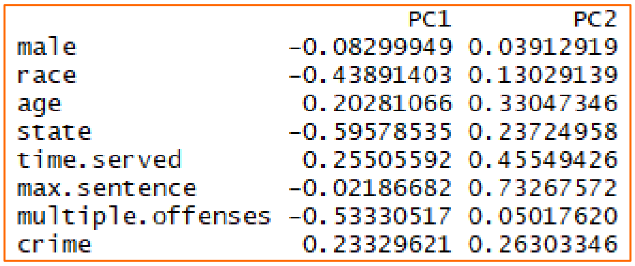


Table 2: Principal Component Analysis Loadings



```
                          PC1        PC2
male              -0.08299949 0.03912919
race              -0.43891403 0.13029139
age                0.20281066 0.33047346
state             -0.59578535 0.23724958
time.served        0.25505592 0.45549426
max.sentence      -0.02186682 0.73267572
multiple.offenses -0.53330517 0.05017620
crime              0.23329621 0.26303346
```

Figure 21: Autoplot of PC1 and PC2

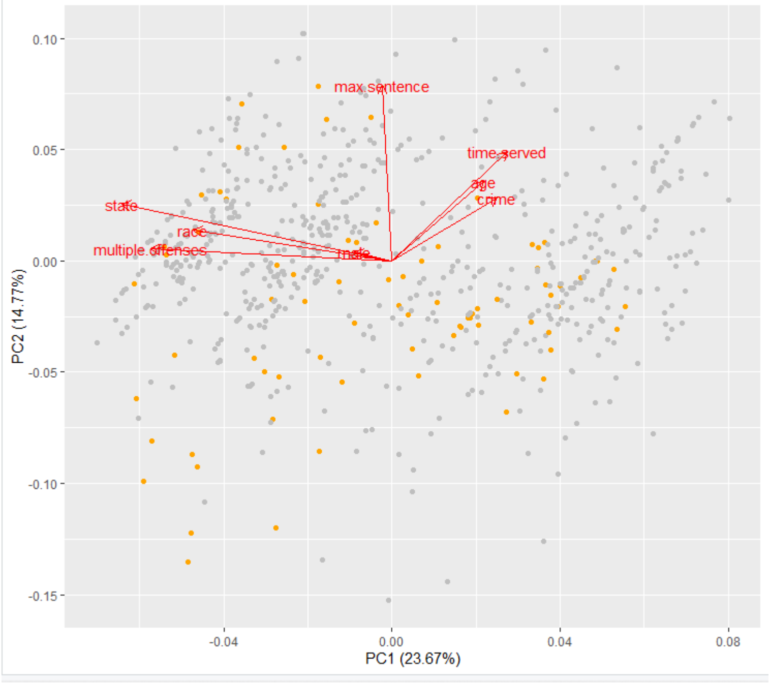
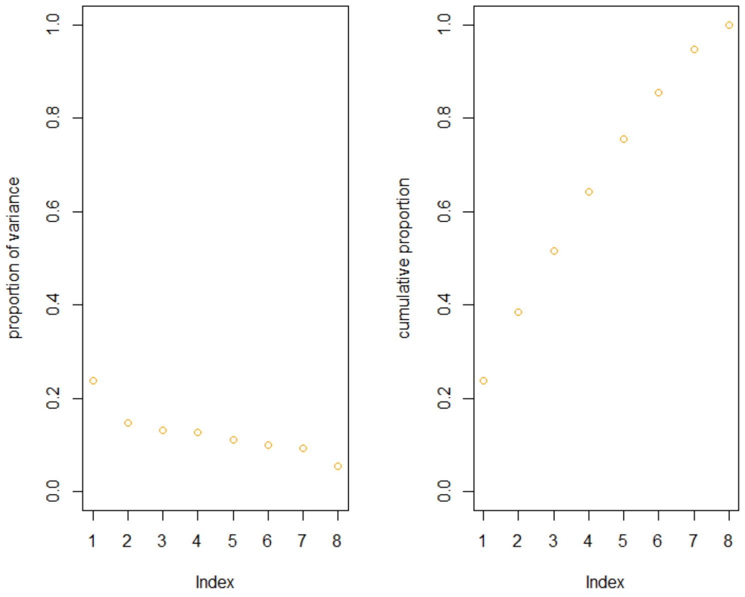


Figure 22: PCA – Proportion of Variances


proportion of variance
Index
cumulative proportion
Index