COMPAS RECIDIVATE DATA

Predicting recidivate offenders with decision trees

Jinxi Luo

Table of Contents

Executive Summary

In this report, decision trees are utilized to predict general offender recvidisum using the COMPAS data set , which includes positive and negative re-offense cases after screening arrest. Python was used for data pre-processing with decision tree modelling done in SAS. Parameter tuning showed that the best performance trees did not incorporate categorical variables in their splits. With the best tree trained in this investigation having a balanced classification rate of 68.98%. Decile Score and other numerical covariates where found to be good predictors with higher decile score in general contributing to higher rates of recidivism.

Project Aim and Data Overview

This project aims to predict all recidivation of criminals using the COMPAS decile score and offender covariates. Python is used for feature engineering and data pre-processing with decision tree modelling done in SAS. Three trees will be produced to evaluate the effectiveness of covariates and decile score on predicating recidivism through varying modelling parameters.

Twelve variables (table 1) from the full data set (of 7214 observations) is selected for their relevance with the creation of more variables in feature engineering. A data dictionary of the full list is available in table 15 of appendix.

Table 1: Data dictionary of 12 variables from the COMPAS data set.

Feature	Definition	Format
sex	Gender of offender	Categorical: Male or Female
age	Age of offender at recidivism charge date	Numerical: 18 to 96
age_cat	Offender age divided into three bins	Categorical: "Greater than 45","25-45", "Less than 25"
race	Offender race divided into six factors	Categorical: "African-American", "Asian", "Caucasian", …
juv_fel_count	Total offender juvenile felony count at COMPAS screening	Numerical: 0 to 20
decile_score	The COMPAS evaluation score of offender recidivism risk	Numerical: 1 to 10
juv_misd_count	Total offender juvenile misdemeanour count at COMPAS screening	Numerical: 0 to 13
juv_other_count	Total juvenile crime counts for other offenses at COMPAS screening	Numerical:0 to 17
priors_count	Count of all offender previous crimes (including juvenile counts) at COMPAS screening	Numerical: 0 to 38
c_charge_degree	Charge of offender at COMPAS screening divided into felony or misdemeanour.	Categorical: F or M
c_charge_desc	Description of charge degree divided into 437 factors	Categorical: "Abuse Without Great Harm", "Agg Fleeing and Eluding", …
is_recid	State of offender recidivism divided into yes or no	Categorical: 1 or 0

Data Pre-processing

Recidivism, charge degree and days from arrest have not been included in the analysis as they are a source of data leakage dependent on offender recidivism status. Additionally, features with unknown definitions(days before screening arrest, crime days from COMPAS) and dates are excluded; as they are unusable in their current format.

Feature Cleaning

Of the original 11 predictive features, only crime charge description had 0.4% missing values, which were arbitrarily replaced with the most frequent factor level as the proportion of NAs were quite small. The variable days before jail, created in feature engineering had 4.3% missing values which were replaced with 0. These cases were missing both jail in and out dates, assuming the offender was not sentenced to jail at the COMPAS screening trial. Imputation is used here as it is assumed that missing values are a result of human errors in recording and systematic consequences.

While SAS automatically deals with categorical variables, in python they need to be encoded to work with the classification models and feature selection using L1 regularization. Through trying a range of encoding methods using default decision tree classifier parameters from the sklearn library; with imputation and three numerical variables from feature engineering. It was found that One hot and CatBoost encodings performed the best in terms of AUC (table 2). I will be using CatBoost encoding as it works better for the high cardinality features, reduces computation time in feature selection and overfitting.

Table 2: Comparison of AUC for different categorical encoding methods.

Encoding Method	*AUC*
One hot	0.6283
CatBoost	0.6175
No categorical columns at all (baseline)	0.6118
Target	0.6100
Label	0.5979
Count	0.5625

Feature Engineering

Removal of date and unknown features potentially introduces a loss in explanatory power, this section will create features which makes use of the removed features and introduces additional features that may aid in prediction. Crime offense jail in and jail out dates are used to create the days in jail feature. Days before jail was created from the summation of days from COMPAS and days before screening arrest. These variables are aimed to describe the severity of an offense by length of the punishment in jail and the urgency to which the offender was sentenced on explaining recidivism.

Table 3: Created features from dates.

Feature	Definition	Format
Days in jail	Days spent in jail for the crime with COMPAS screening	Numerical: 0 to 800
Days before jail	Days between going in jail and offender arrest or offense date with the COMPAS screening	Numerical: -300 to 9484

Juv counts was created to summarize the three juvenile crime count features as they are "sparse" in the respect that 75th percentile of data are 0s for each feature. In hopes that the predictive power of Juv counts would be better than the other three features individually. Variable interactions were also investigated on recidivism prediction using the combinations of two categorical variables to hopefully improve model performance (table 4). The full list of these variables can be found in table 16 of appendix.

Table 4: Created features from juvenile crime counts and categorical variable interactions

Feature	Definition	Format
Juv counts	Sum of all juvenile offenses	Numerical: 0 to 21
Categorical variable interactions	Ten categorical variables that are the combinations of two of the original four categorical variables	Categorical: Var1level_Var2level

Feature Selection

A total of 24 features exist after feature engineering. This section aims to reduce the number of features to 11, which is the number of predictive features from the beginning to minimize overfitting. L1 regularization is used to obtain the following 11 best features: (highlighted are the original variables)

1. Age
2. Days before jail
3. Days in jail
4. Decile score
5. Juv counts
6. Priors count
7. Age category and crime charge degree (CB encoded)
8. Age category and crime charge description (CB encoded)
9. Crime charge degree and race (CB encoded)
10. Crime charge description and race (CB encoded)
11. Crime charge description and sex (CB encoded)

Eight of the final features are from feature engineering with new model AUC being 0.6225, an improvement on the CatBoost encoding model (in table 2) with 14 features. Suggesting the selection process was worthwhile and that overfitting from additional features was reduced.

It should be noted that results from python may not also apply to models in SAS, as it was evident that target encoding with the same features had lower misclassification errors than cat boost using default parameters in SAS. However, the selected encoded features had also performed better than the original features and their respective non encoded counter parts. For simplicity, I will use the selected features in their non encoded format in decision tree modelling.

Feature Analysis and Summary

The recidivism feature is binary and nominal with 1 indicating positive recidivism. A bar plot of its distribution shows that about 48% of the observations are positive cases (figure 1). Suggesting that any good classifier that doesn't predict a constant 1 should have misclassification error lower than 52%.

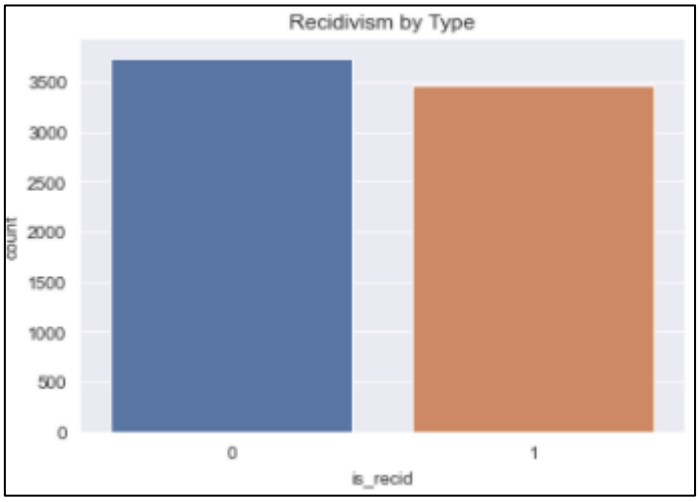

Figure 1: Bar plot of recidivism.

Age is of interval type; its histogram (figure 2) shows that majority of offenders in the data are skewed towards the younger age groups around 25 to 30. Count of recidivism per bin also shows that most positive cases occur in younger age groups suggesting that a classifier may predict younger ages as having a higher chance for recidivism.

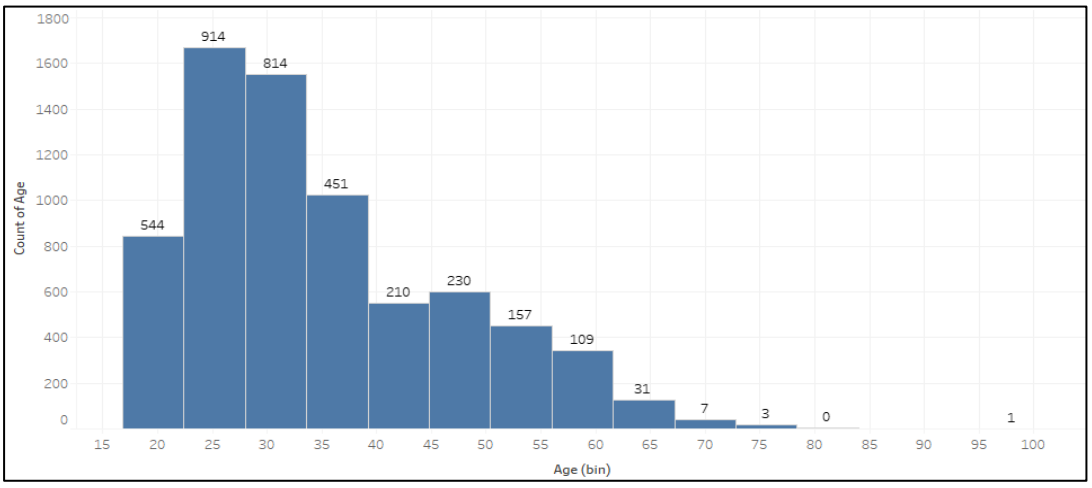

Figure 2: Histogram of age with numbering showing cases of recidivism.

Decile score is nominal and ranges from 1 to 10. Figure 3 of its distribution shows that majority of the scores for offenders are low. As decile score increases the number of offenders with that score also decreases while recidivism count remains almost the same. This may cause offenders with higher decile scores to be classified more often to re-offend.

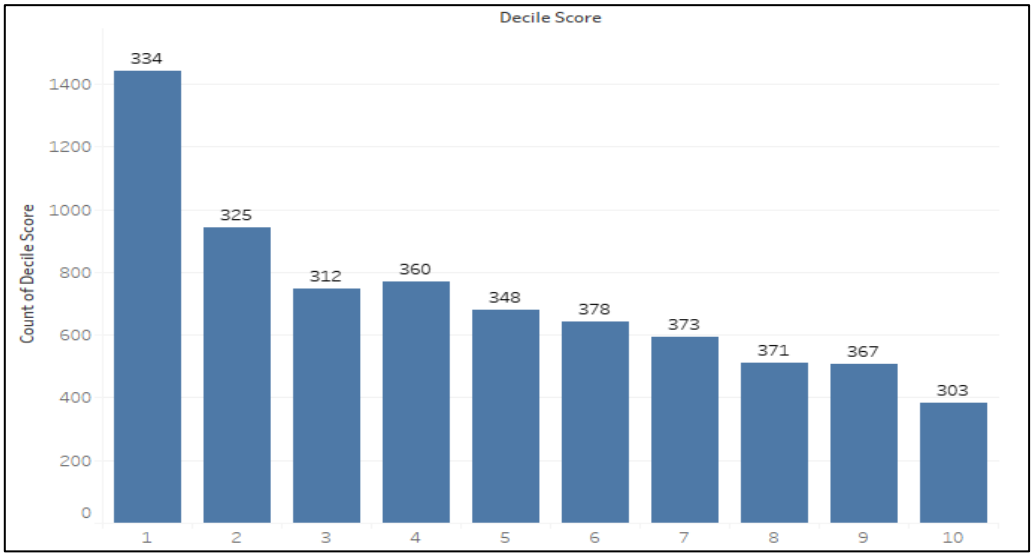

Figure 3: Bar plot of decile score with count of re-offense per bin.

The remaining numerical variables in the data can be summarized in table 5 as they are relatively sparse with most days being 0, suggesting that the features may not be very useful in classification. Most recidivism cases are sentenced to jail on the day of their arrest/offense date and remain for only a day. Negative days indicate that the offender was already in jail during their offense. Juv counts seems to be always less than prior counts suggesting that most offenders had not committed any juvenile offense prior to their screening trial.

Table 5: Summary statistics for the remaining numerical variables.

	Days Before Jail	Days in Jail	Juv Counts	Prior Counts
Mean	40.07	17.93	0.27	3.47
STD	291.85	50.12	0.95	4.88
Min	-300	0	0	0
25%	0	1	0	0
50%	0	1	0	2
75%	0	9	0	5
Max	9484	800	21	38

Scatter plot matrix of the numerical variables doesn't show any multicollinearity except for Juv and prior counts with decile score (figure 4). This is expected from how the score is calculated. For the interaction categorical variables multicollinearity should be irrelevant as new factor levels are created. So, no further feature selection is required.

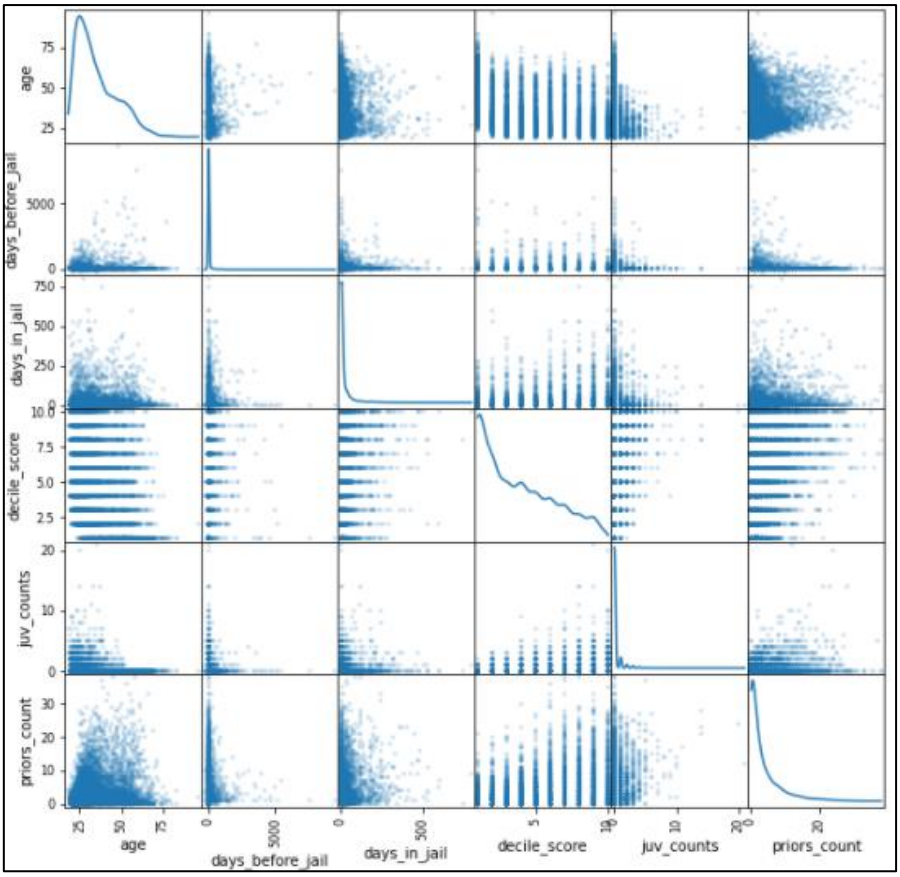

Figure 4: Scatter plot matrix for all numerical variables.

Crime charge degree and race is a nominal variable having a cardinality of 12. Felony committed by African Americans is the most common occurrence with over half of offenders re offending. This may cause classifiers to favour predicting felony by African Americans to re-offend as this high proportion of crime count to recidivism is not seen for the other factor levels (figure 5).

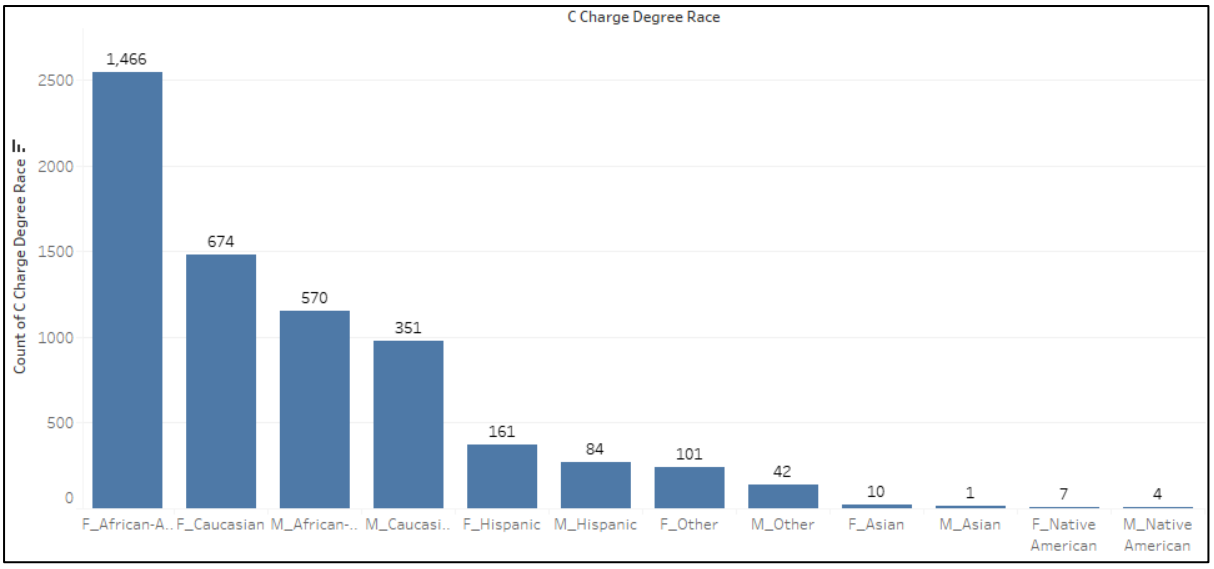

Figure 5: Histogram of crime charge degree and race.

Age category by crime charge degree is a nominal variable having a cardinality of six. Its bar plot (figure 6) supports what was seen in figure 2. felony committed by offenders in age group 25-45 seems to dominate the data so classification may favour the group to re-offend.

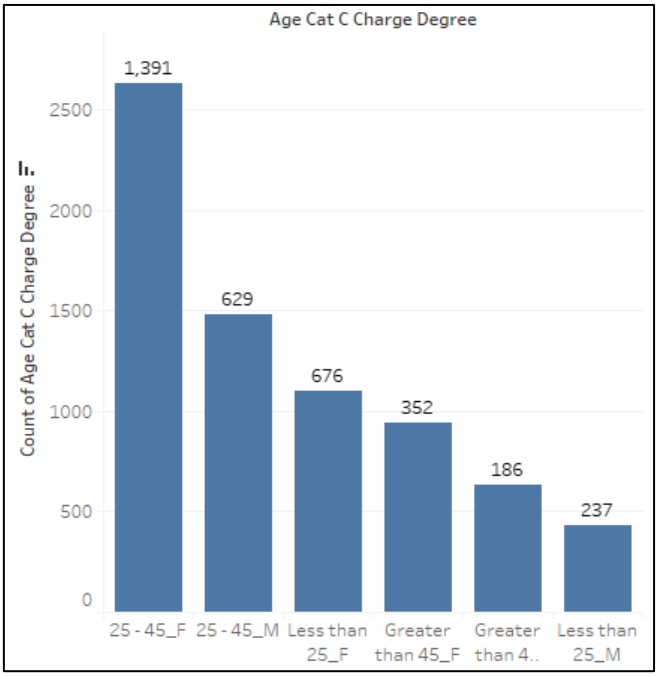

Figure 6: Histogram of Age category and crime charge degree.

The cardinality of the remaining categorical variables is too high to be concisely represented graphically and instead summarized in the table 6 below. It is unclear for now how they may impact classification.

Table 6: Summary of the remaining categorical variables.

	C Charge Desc Race	C Charge Desc Sex	Age Cat C Charge Desc
Proportion Most Common %	8.9	13.0	9.5
Most Common	arrest case no change_Africian-American	arrest case no charge_Male	25-45_Battery
Cardinality	845	578	739

Decision Tree Building in SAS

Three decision trees will be made with the 11 variables from feature selection varying five hyperparameters that are deemed most appropriate to the data set. Shown below with other settings on default unless specified:

When Splitting

- There are only nominal and interval data. Other splitting parameters are set default as no reasonable justification can be made to change them as of now.

1. Nominal target criterion

When Pruning

- These are general parameters that the SAS guide has provided for tree pruning. There is currently no justification to change other similar parameters.

2. Maximum depth
3. Leaf size
4. Number of surrogate rules

The data will be first split into 80/20 for training and validation with random sampling. So that as most data as possible can be used to train a good classifier, assuming that recidivism in training and validation would be equally represented. This decision is made so that if after parameter tuning and the model still performs worse on validation data. Then it may be known that the model is overfitting and that parameter turning was not done well. Subtree assessment measure is also set to misclassification as the target variable is binary. Furthermore, only binary trees are used as variables in the data contain high cardinality; multiple impurity measures will be tested to determine the effectiveness of this decision.

Decision Tree 1

Two of the major pruning parameters are set relatively large from their default to examine if SAS can find a simpler tree given a larger minimum leaf size that was determined based on the counts of categories from figures 5 and 6.

Table 7: Major parameters for decision tree 1.

Parameter	*Value*
Nominal Target Criterion	Default = ProbChisq
Maximum depth	30
Leaf Size	200
Number of surrogate rules	5

Figure 7 shows that SAS was able to find a small tree model with 8 leaf that fitted both training and validation data well. Performance evaluation measures for validation and training are also similar with the model doing better on validation data for four of the six metrics. This suggests that the split of 80/20 split of data was appropriate in training a model that may predict well on future unseen data (table 8).

Feature importance in figure 8 shows mostly numerical features to have very high importance with age_cat_c_crime_charge and crime_charge_race being the only categorical variables in the top 8. They and other numerical variables were however not included in the tree in figure 9, suggesting that better model predictions may be gained by lowering minimum leaf size and increasing surrogate rules.

Table 8: Performance evaluation measures for decision tree 1.

Measure	*Validation*	*Training*
FN	246	990
TN	536	2151
FP	210	846
TP	451	1784
Precision	0.6823	0.678327
Recall	0.647059	0.643115
F-Score	0.664212	0.660252
Kappa	0.366179	0.361531
B.C.R	0.682779	0.680416
Misclassification Error	0.316008	0.318142

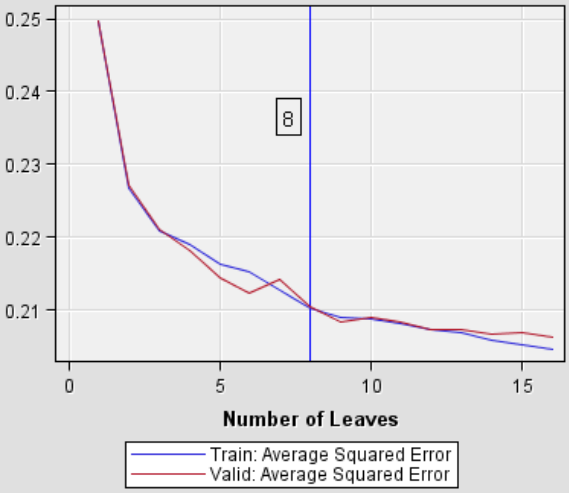

Figure 7: Average squared error vs number of leaves for training and validation data with decision tree 1.

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
decile_score	decile_score	1	3	1.0000	1.0000	1.0000
age_cat_c_...	age_cat_c_...	0	7	0.9732	0.9639	0.9905
priors_count	priors_count	2	2	0.9672	0.9122	0.9431
juv_counts	juv_counts	0	5	0.9329	0.9579	1.0268
days_in_jail	days_in_jail	1	2	0.8344	0.8309	0.9958
c_charge_d...	c_charge_d...	0	3	0.7882	0.7408	0.9398
age	age	3	2	0.5834	0.5904	1.0121
days_befor...	days_befor...	0	3	0.4697	0.4979	1.0599
age_cat_c_...	age_cat_c_...	0	0	0.0000	0.0000	.
c_charge_d...	c_charge_d...	0	0	0.0000	0.0000	.
c_charge_d...	c_charge_d...	0	0	0.0000	0.0000	.

Figure 8: Feature importance table for decision tree 1.

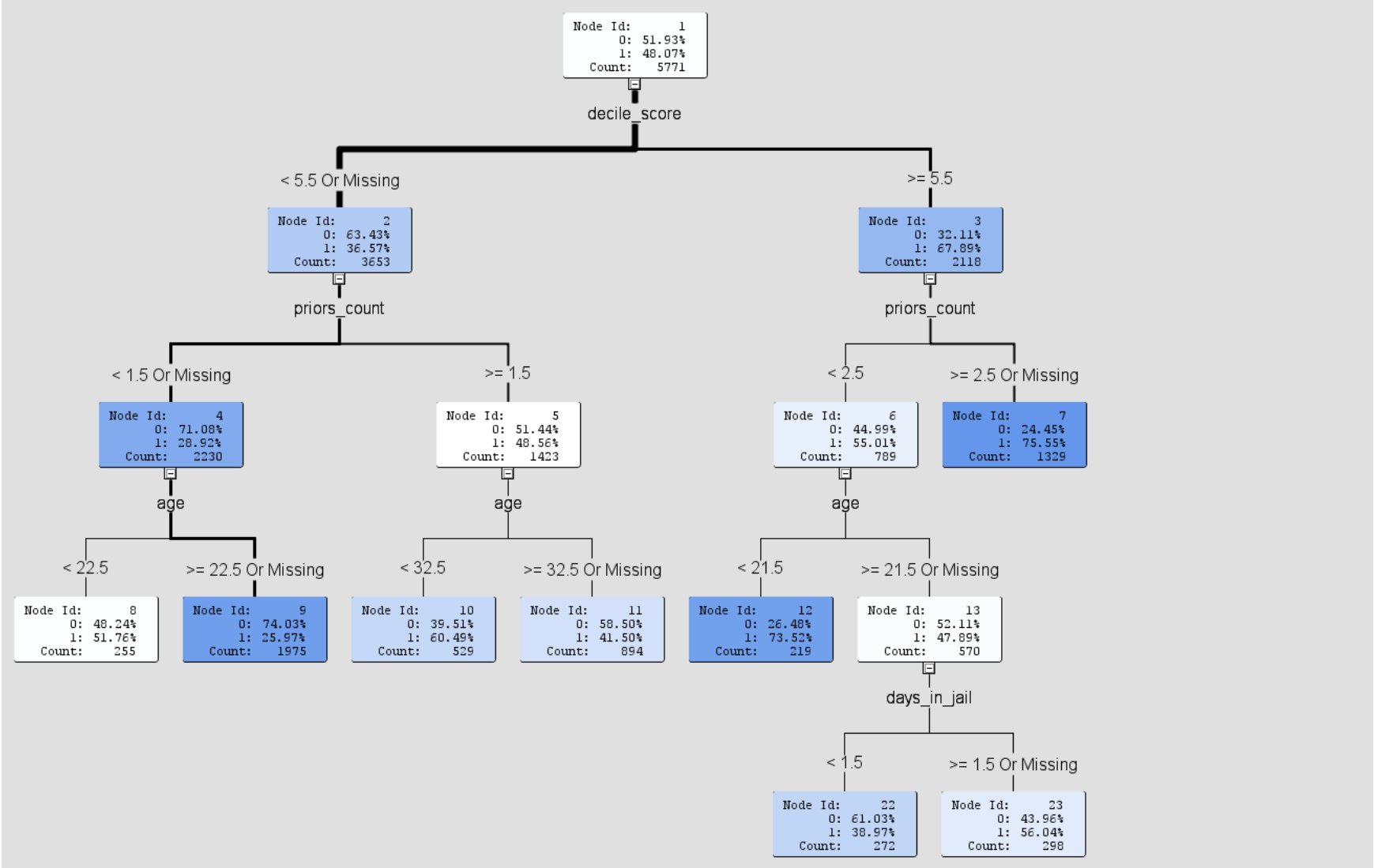

Figure 9: Decision Tree 1

Decision Tree 2

Model pruning parameters are set at default with nominal target criterion set to Gini to examine how node impurity measures would affect the final tree output (table 9). As it is known that Gini and Entropy works well with high cardinality features.

It is clear from table 10 and figure 10 that the model has over fit; with model performing well on the training data only. Figures 11 and 12 also shows this situation with mostly one variable that have the highest feature importance and a deep tree that is split a lot of times by variables that have relatively low importance.

In other experiments, setting the criterion to entropy and keeping model pruning parameters the same as in decision tree 1 yielded similar results. Suggesting that while impurity measures do not cause the tree to prune its categorical variable nodes, using them in prediction could cause more overfitting and complex models.

Table 9: Major parameters for decision tree 2.

Parameter	*Value*
Nominal Target Criterion	Gini
Maximum depth	Default = 6
Leaf Size	Default = 5
Number of surrogate rules	Default = 0

Table 10: Performance evaluation measures for decision tree 2.

Measure	*Validation*	*Training*
FN	253	914
TN	530	2162
FP	216	835
TP	444	1860
Precision	0.672727273	0.690166976
Recall	0.637015782	0.670511896
F-Score	0.654384672	0.680197477
Kappa	0.348078385	0.392315617
B.C.R	0.673735773	0.695949975
Misclassification Error	0.325017	0.303067

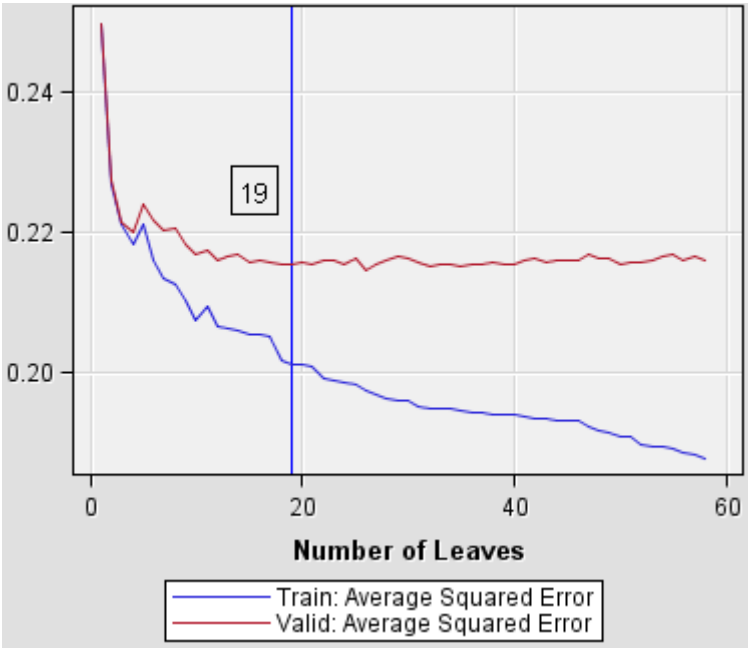

Figure 9: Average squared error vs number of leaves for training and validation data with decision tree 2

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance
decile_score	decile_score	2	1.0000	1.0000
priors_count	priors_count	4	0.6871	0.6146
age_cat_c_...	age_cat_c_...	4	0.5728	0.2944
age	age	3	0.3719	0.3827
days_in_jail	days_in_jail	2	0.2648	0.2520
c_charge_d...	c_charge_d...	1	0.2307	0.0000
c_charge_d...	c_charge_d...	1	0.1710	0.1109
days_befor...	days_befor...	1	0.1250	0.0110
c_charge_d...	c_charge_d...	0	0.0000	0.0000
age_cat_c_...	age_cat_c_...	0	0.0000	0.0000
juv_counts	juv_counts	0	0.0000	0.0000

Figure 10: Feature importance for decision tree 2.

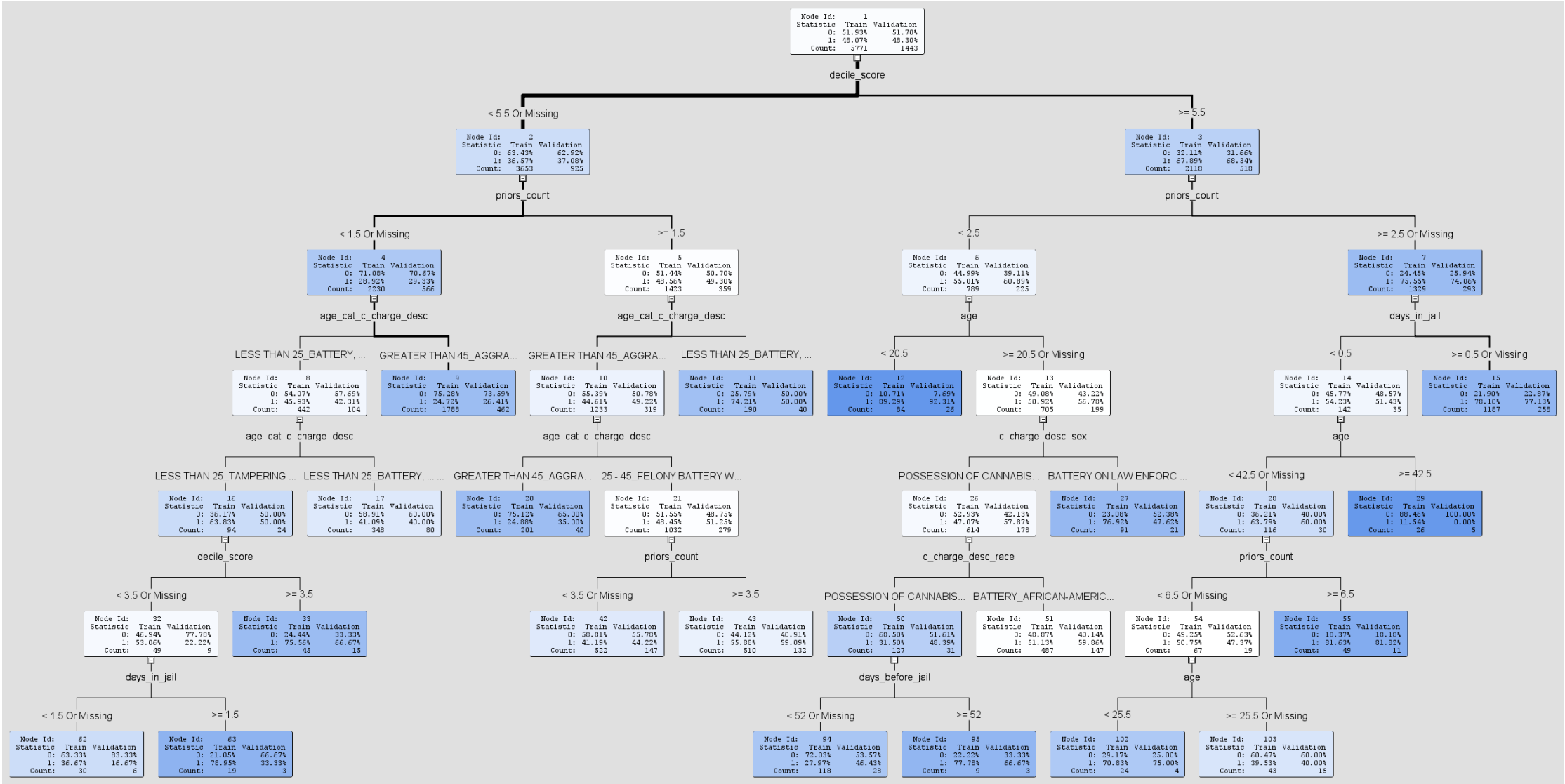

Figure 11: Decision tree 2

Decision Tree 3

The last decision tree parameters are set to be a compromise between the previous trees. Nominal target criterion is set default to prune categorical variable nodes. Leaf size has been reduced and surrogate rules increased to incorporate more of the high importance numerical variables in the decision tree (table 11).

Table 11: Major parameters for decision tree 3.

Parameter	*Value*
Nominal Target Criterion	Default = ProbChisq
Maximum depth	30
Leaf Size	10
Number of surrogate rules	20

Results for the tree are expected and shows performance evaluation results between validation and training data to be close; with lower classification errors than tree 1. However, the model may have underfitted the training data a bit more than decision tree 1 with a more complicated tree (table 12 and figures 13-15).

Table 12: Performance evaluation measures for decision tree 3.

Measure	*Validation*	*Training*
FN	240	973
TN	540	2162
FP	206	835
TP	457	1801
Precision	0.689291101	0.68323217
Recall	0.655667145	0.64924297
F-Score	0.672058824	0.665804067
Kappa	0.380136729	0.371318263
B.C.R	0.689763867	0.685315513
Misclassification Error	0.309078	0.313291

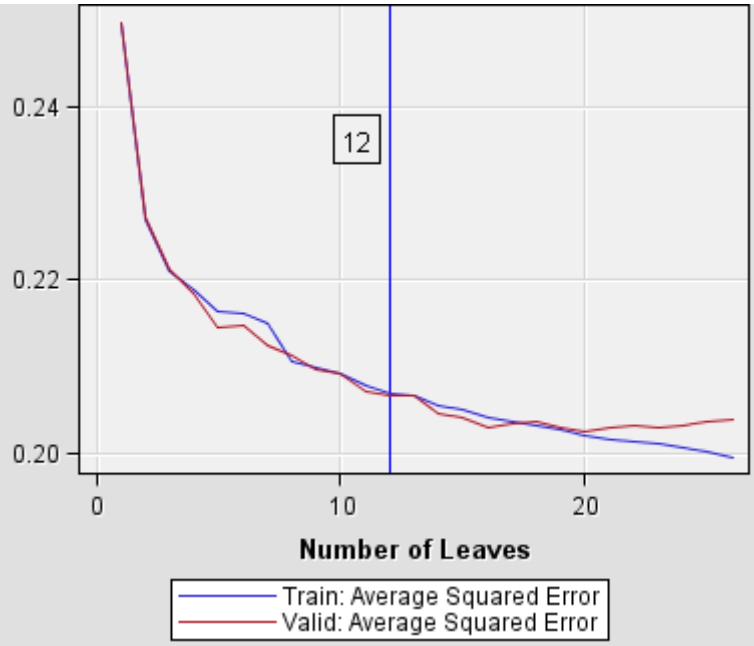

Figure 12: Average squared error vs number of trees for training and validation data with decision tree 3.

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
age	age	5	5	1.0000
decile_score	decile_score	1	3	0.9538
priors_count	priors_count	2	3	0.9368
age_cat_c_...	age_cat_c_...	0	7	0.9212
juv_counts	juv_counts	1	5	0.8961
days_befor...	days_befor...	0	5	0.8209
days_in_jail	days_in_jail	2	2	0.8179
c_charge_d...	c_charge_d...	0	3	0.7508
age_cat_c_...	age_cat_c_...	0	0	0.0000
c_charge_d...	c_charge_d...	0	0	0.0000
c_charge_d...	c_charge_d...	0	0	0.0000

Figure 13: Feature importance for decision tree 3.

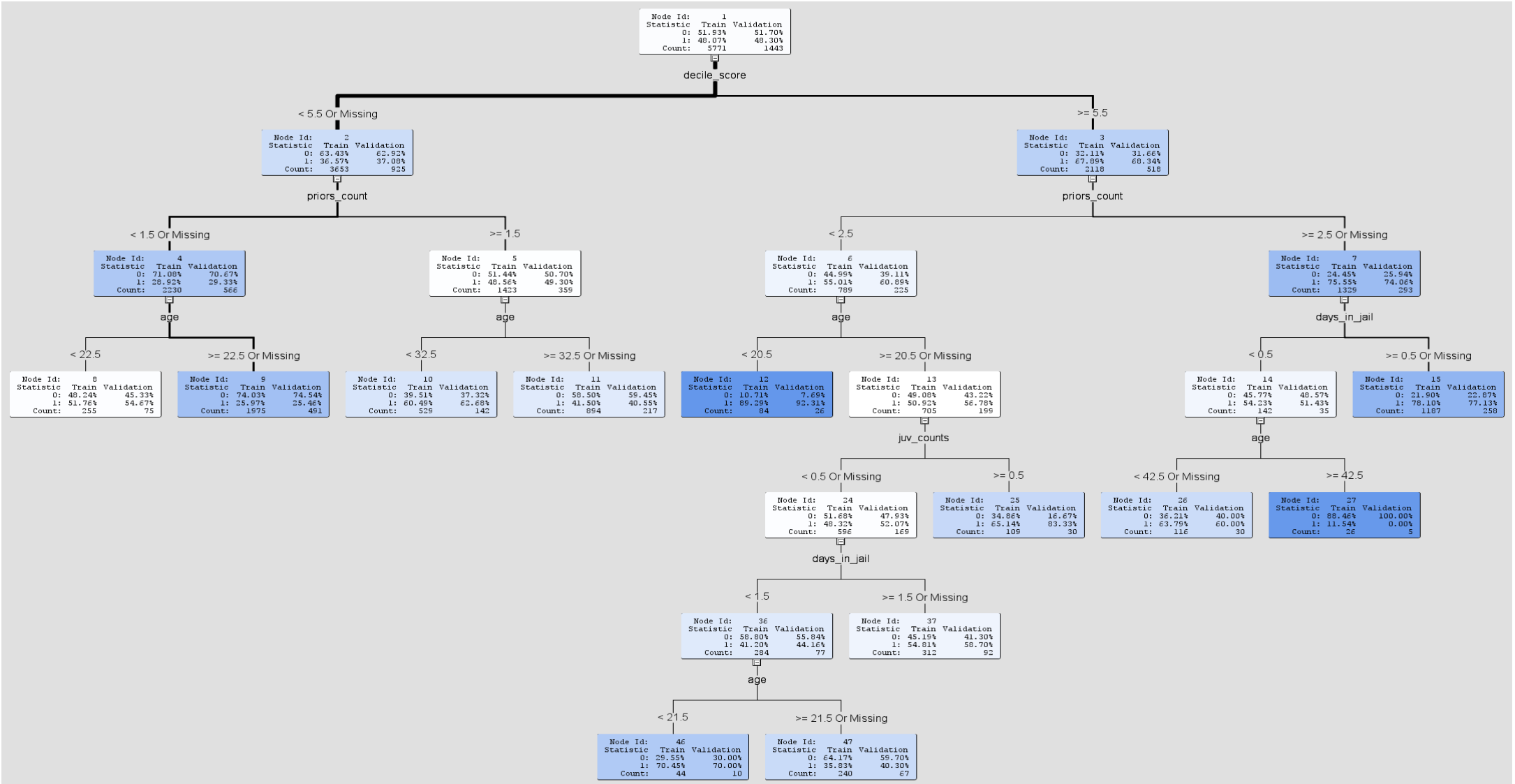

Figure 14: Decision tree 3.

Model Comparison

The main model performance evaluation metric we will be using is the Balanced Classification Rate (table 13). It is chosen as it is able to deal with an imbalanced target variable; as seen in figure 1 where there are more re-offended cases. As it is also of interest to investigate decile score on predicting recidivism, the precision, recall, F1 score and misclassification error metrics are biased towards certain cases in the confusion matrix. Which won't give much of a clear picture of how useful decile score actually is in giving any more information than that could be seen using other covariates.

Cumulative lift and ROC charts will also be used to evaluate the best model as they provide a way to compare the performance of the trees to random guess and predicting recidivism in different percentiles of the data. Validation results will mostly be of focus as it better describes how the trees will perform on unseen data.

Table 13: Balanced classification rate by validation and training data for each tree.

Tree	*Validation*	*Training*
T3	0.689763867	0.685315513
T1	0.682779	0.680416
T2	0.673735773	0.695949975

Keeping in mind that T2 is generally not a good model having overfit the training data and so will not be considered. It can be seen that T3 had the best B.C.R on both validation and training data. This supports the decision to not incorporate categorical variable nodes and that parameter turning was working well. It is also noted that T3 is a more complicated model than T1; and that applying the principle of Occam's razor, T1 should be preferred. However, the ROC plots for T3 had better sensitivity for all false positive rates than T1 in both training and validation (figure 16). T3 also outperformed T1 in the cumulative lift charts, having more area under its line for all depths, supporting T3 to be better at predicting recidivism cases (figure 17).

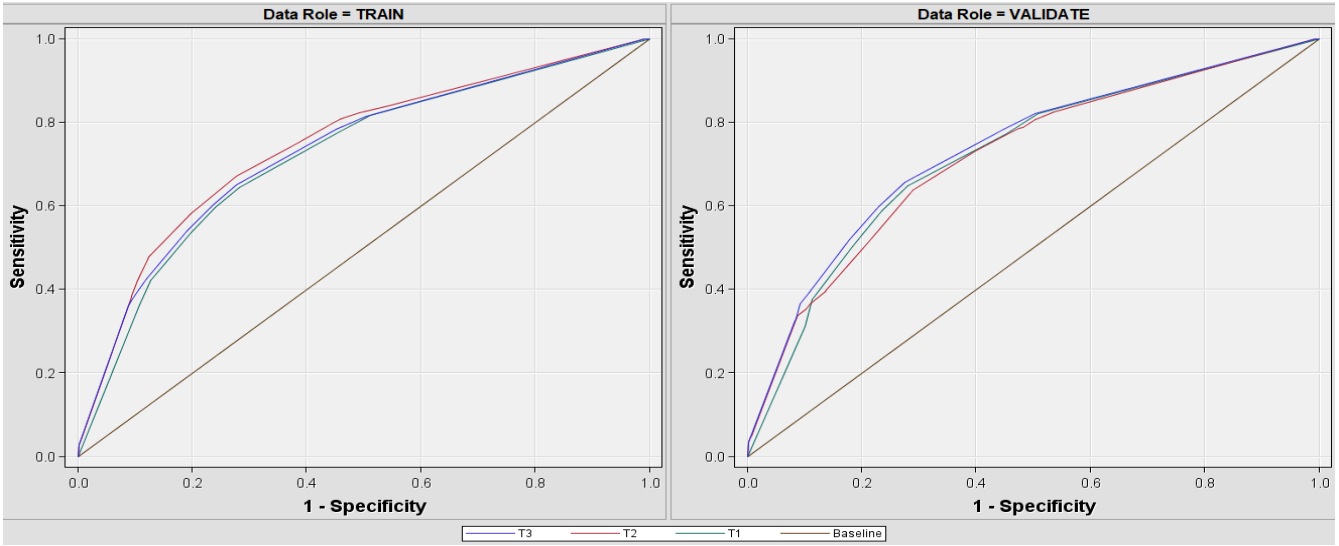

Figure 15: ROC curves of the three decision trees for training and validation data.

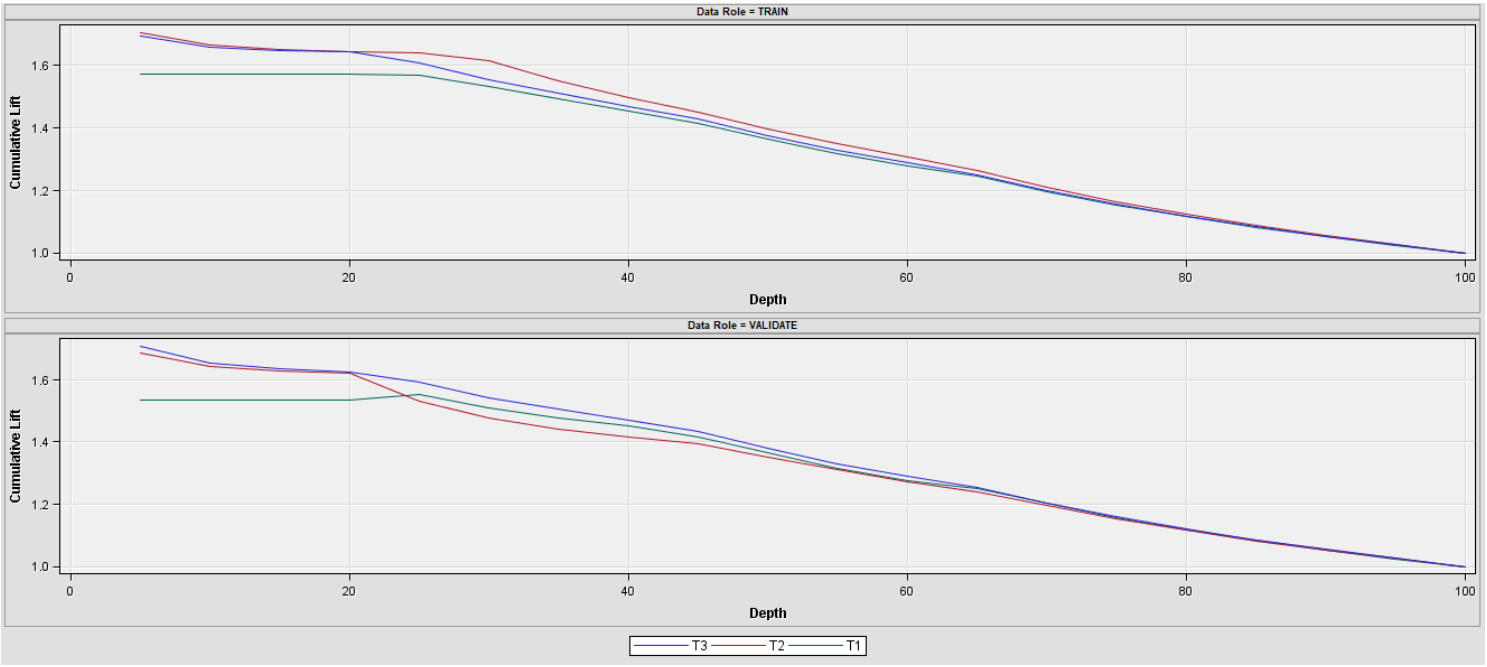

Figure 16: Cumulative lift chart of the three decision trees for training and validation data.

Then, the best model is found to be T3. In addition to discussions on major parameters and results in the previous section. T3 has a depth of 7, 12 leaf nodes and all of its decision nodes are numerical variables listed below in order of importance:

Table 14: Variables used to make decision tree 3.

Importance	*Variable*	*Importance Value*
1	age	1
2	decile_score	0.9538
3	priors_count	0.9368
5	juv_counts	0.8961
7	days_in_jail	0.8179

Decile score appears to have a major contribution to predicting recidivism, but so does the other covariates. These results suggest that simple numerical attributes such as an offender's age, number of previous crimes they have committed during and after juvenile age and how many days they were in jail for are almost as if not more useful than decile score.

It is known for the validation data that offenders are most likely to re-offend if their decile score is over 5. This agrees with observations in figure 3 and with figure 15, it can be seen that this filtering for validation data was able to label 68.34% of reoffenders correctly if they have a decile score over 5. And label 62.92% correctly that did not re-offend with decile score under 5. Comparing these values to the BCR of 68.98% and T3 model precision of 68.93% shows that Decile score by itself can be a valid indicator of recidivism. As it often is also the variable with highest importance in similar tested trees such as T1 and T2.

Appendix

Table 15: Data dictionary of all variables in the COMPAS data set.

Feature	Definition	Format
id	ID of an observation	Numerical: 1 to 11001
sex	Gender of offender	Categorical: Male or Female
age	Age of offender at recidivism charge date	Numerical: 18 to 96
age_cat	Offender age divided into three bins	Categorical: "Greater than 45","25-45", "Less than 25"
race	Offender race divided into six factors	Categorical: "African-American", "Asian", "Caucasian", …
juv_fel_count	Total offender juvenile felony count at COMPAS screening	Numerical: 0 to 20
decile_score	The COMPAS evaluation score of offender recidivism risk	Numerical: 1 to 10
juv_misd_count	Total offender juvenile misdemeanour count at COMPAS screening	Numerical: 0 to 13
juv_other_count	Total juvenile crime counts for other offenses at COMPAS screening	Numerical:0 to 17
priors_count	Count of all offender previous crimes (including juvenile counts) at COMPAS screening	Numerical: 0 to 38
days_b_screening_arrest	Unknown	Numerical: -414 to 1057
c_jail_in	Date of offender arrest where COMPAS was screened	Date: Year or dd/mm/yy hh:mm
c_jail_out	Date offender was released after COMPAS	Date: Year or dd/mm/yy hh:mm
c_offense_date	Date offender committed crime prior to COMPAS	Date: Year or dd/mm/yy
c_arrest_date	Data offender was arrested for their crime prior to COMPAS	Date: Year or dd/mm/yy
c_days_from_compas	Unknown. However, its sum with days_b_screening_arrest provides days between crime arrest/commit date until jail in.	Numerical: 0 to 9485
c_charge_degree	Charge of offender at COMPAS screening divided into felony or misdemeanour.	Categorical: F or M
c_charge_desc	Description of charge degree divided into 437 factors	Categorical: "Abuse Without Great Harm", "Agg Fleeing and Eluding", …
is_recid	State of offender recidivism divided into yes or no	Categorical: 1 or 0
r_charge_degree	Recidivism charge degree divided into 10 factors	Categorical: "(F3)", "(M1)","(F2)", …

r_days_from_arrest	Unknown	

Table 16: Categorical variables created from the combinations of age, race, c_charge_degree and c_charge_desc.

Categorical variable interactions	
age_cat_c_charge_degree	c_charge_degree_c_charge_desc
age_cat_c_charge_desc	c_charge_degree_race
age_cat_race	c_charge_degree_sex
age_cat_sex	c_charge_desc_race
race_sex	c_charge_desc_sex