Analysis of Bank Loan Data

Katie Adamson, Brian Bruno, Colin McCunney, Brandon Rank

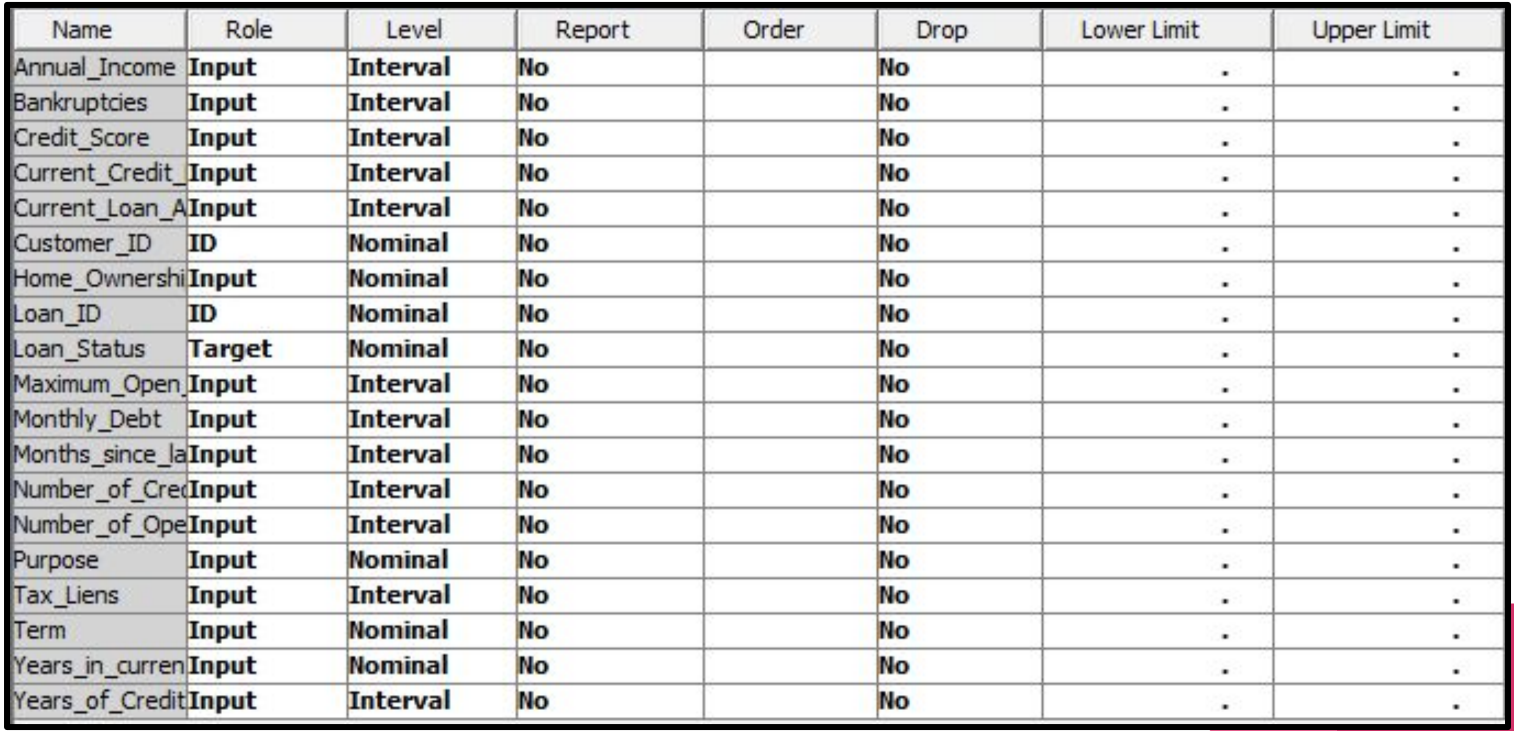
Table of Contents

- Project Motivation
- Variable Overview
- Data Explanation
- Data Preparation
- Models
- Conclusion

Project Motivation

- Examine dataset: bank loans and corresponding customers
- Long-term success in banking: How to evaluate quality of potential customers?
  - *How to determine the likelihood that an applicant will default on their loan?*
- Availability of data and software innovations changes application decision & risk calculation: human experience → data analysis
- **How can different models perform on the dataset we found? What does this tell us about methods that businesses might use to improve operations?**

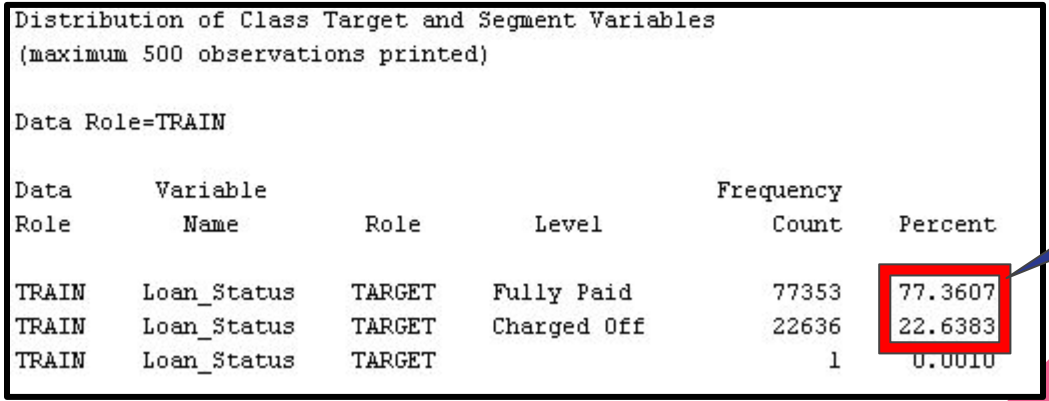
Variable Overview



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Annual_Income	Input	Interval	No		No	.	.
Bankruptcies	Input	Interval	No		No	.	.
Credit_Score	Input	Interval	No		No	.	.
Current_Credit_	Input	Interval	No		No	.	.
Current_Loan_A	Input	Interval	No		No	.	.
Customer_ID	ID	Nominal	No		No	.	.
Home_Ownershi	Input	Nominal	No		No	.	.
Loan_ID	ID	Nominal	No		No	.	.
Loan_Status	Target	Nominal	No		No	.	.
Maximum_Open_	Input	Interval	No		No	.	.
Monthly_Debt	Input	Interval	No		No	.	.
Months_since_la	Input	Interval	No		No	.	.
Number_of_Cre	Input	Interval	No		No	.	.
Number_of_Ope	Input	Interval	No		No	.	.
Purpose	Input	Nominal	No		No	.	.
Tax_Liens	Input	Interval	No		No	.	.
Term	Input	Nominal	No		No	.	.
Years_in_curren	Input	Nominal	No		No	.	.
Years_of_Credit	Input	Interval	No		No	.	.

Data Explanation

- 99,990 records, 19,000 missing values
  - Model performance concerns
- 19 variables: 2 identifiers, both interval (12) and nominal (5) values

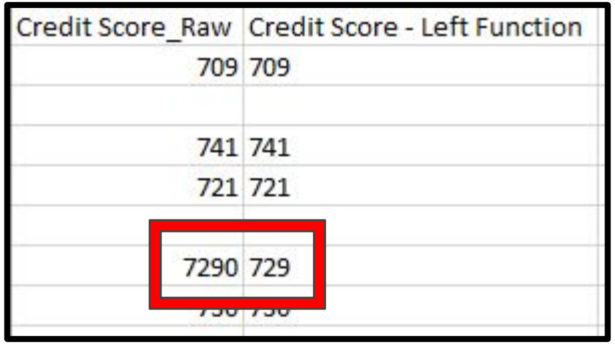


```
Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data      Variable                                    Frequency
Role        Name         Role          Level              Count      Percent

TRAIN     Loan_Status    TARGET      Fully Paid           77353      77.3607
TRAIN     Loan_Status    TARGET      Charged Off          22636      22.6383
TRAIN     Loan_Status    TARGET                               1       0.0010
```

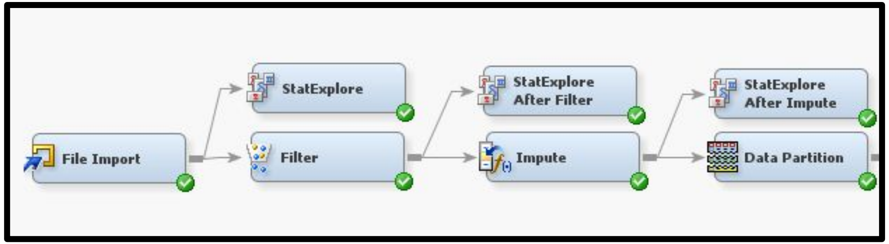
Baseline model comparison statistic

Data Preparation

- ❖ Extraneous record removal
- ❖ Credit Score cleaning
- ❖ Replacement of “NA” and “n/a” values
  - ➢ Allows SAS to properly interpret missing values
- ❖ SAS filtering
  - ➢ Removal of only outliers
- ❖ Imputation
  - ➢ Mean for interval variables
  - ➢ Count for nominal variables
- ❖ Data Partition, 70/30

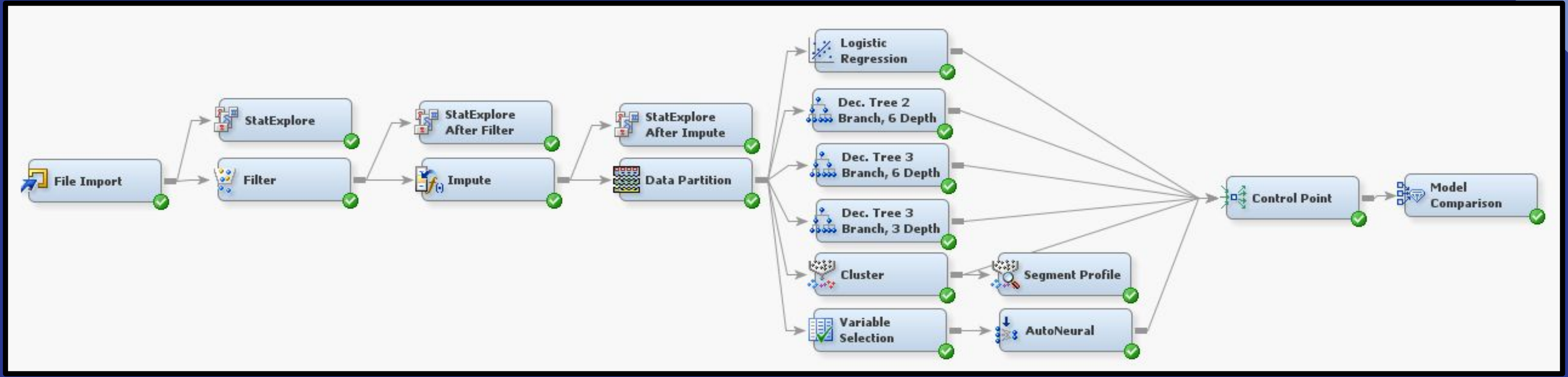


Credit Score_Raw	Credit Score - Left Function
709	709
741	741
721	721
7290	729
730	730



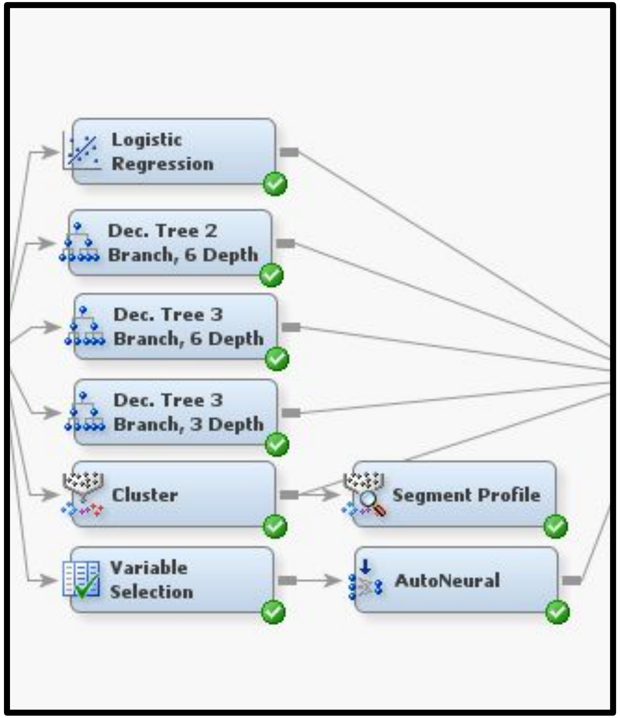
Model Selection and Implementation

SAS EM Diagram



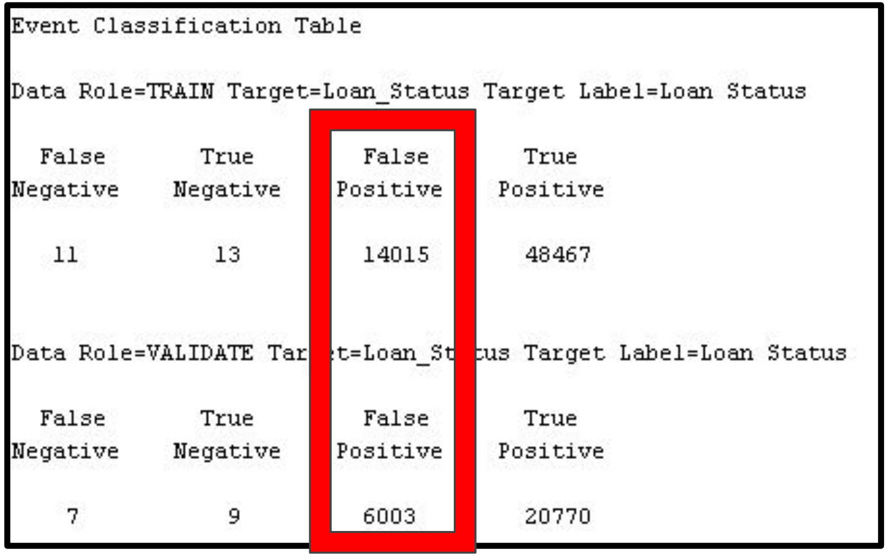
Models

- Logistic regression
- 3 decision tree variations
- Clustering analysis and segment profile
- Variable selection and auto neural network


Logistic Regression
Dec. Tree 2 Branch, 6 Depth
Dec. Tree 3 Branch, 6 Depth
Dec. Tree 3 Branch, 3 Depth
Cluster
Segment Profile
Variable Selection
AutoNeural

Logistic Regression Analysis

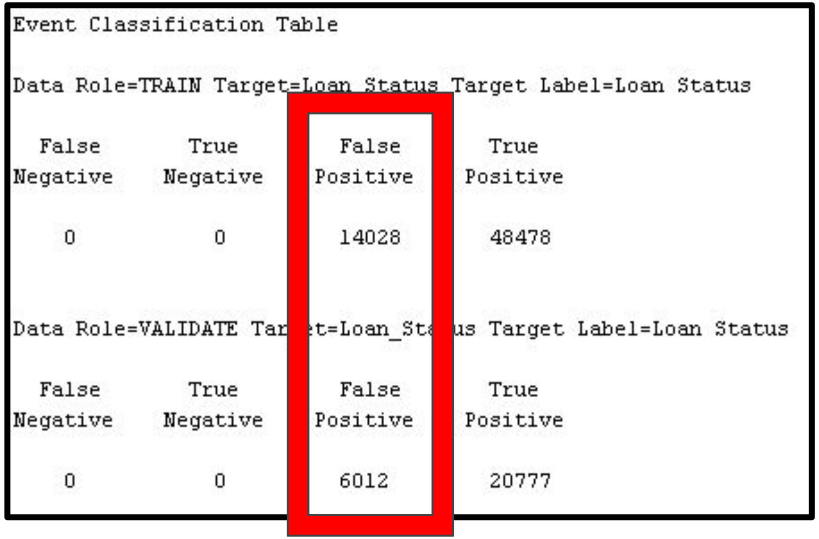
- Default SAS settings
- Training Set
  - Misclassification Rate - 22.4382%
  - ASE - 16.263%
- Validation
  - Misclassification Rate - 22.4346%
  - ASE - 16.3053%
- High number of false positives



```
Event Classification Table

Data Role=TRAIN Target=Loan_Status Target Label=Loan Status

  False       True       False       True
Negative    Negative    Positive    Positive

   11          13         14015       48467


Data Role=VALIDATE Tar t=Loan_St tus Target Label=Loan Status

  False       True       False       True
Negative    Negative    Positive    Positive

    7           9          6003       20770
```

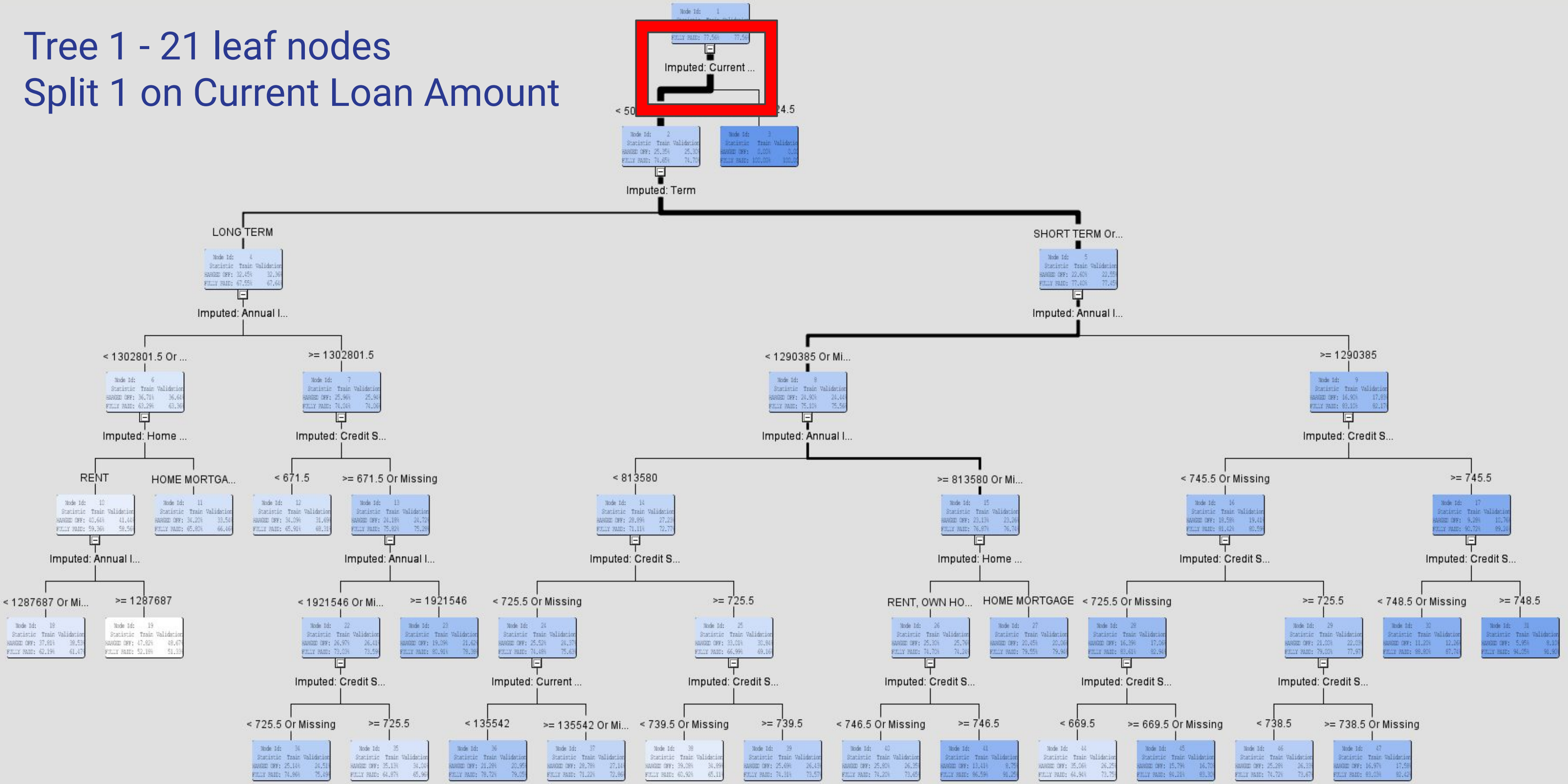
Decision Tree #1 - Maximum Branch of 2, Depth of 6

- Default tree settings except:
  - Assessment measure of ASE
- Training set
  - Misclassification Rate - 22.4414%
  - ASE - 16.1942%
- Validation set
  - Misclassification Rate - 22.442%
  - ASE - 16.2928%
- Once again, high false positive rate



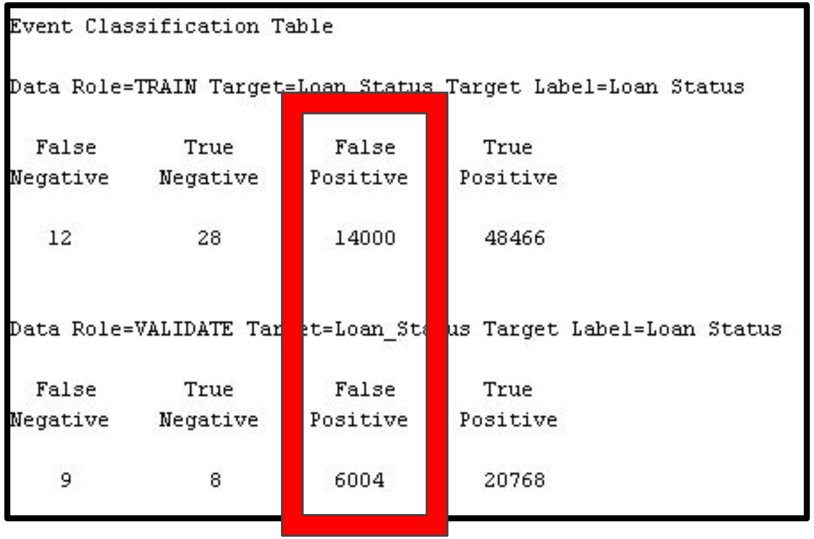
```
Event Classification Table

Data Role=TRAIN Target=Loan_Status Target Label=Loan Status

  False        True        False        True
Negative     Negative     Positive     Positive

    0            0          14028        48478


Data Role=VALIDATE Target=Loan_Status Target Label=Loan Status

  False        True        False        True
Negative     Negative     Positive     Positive

    0            0           6012        20777
```

Tree 1 - 21 leaf nodes
Split 1 on Current Loan Amount



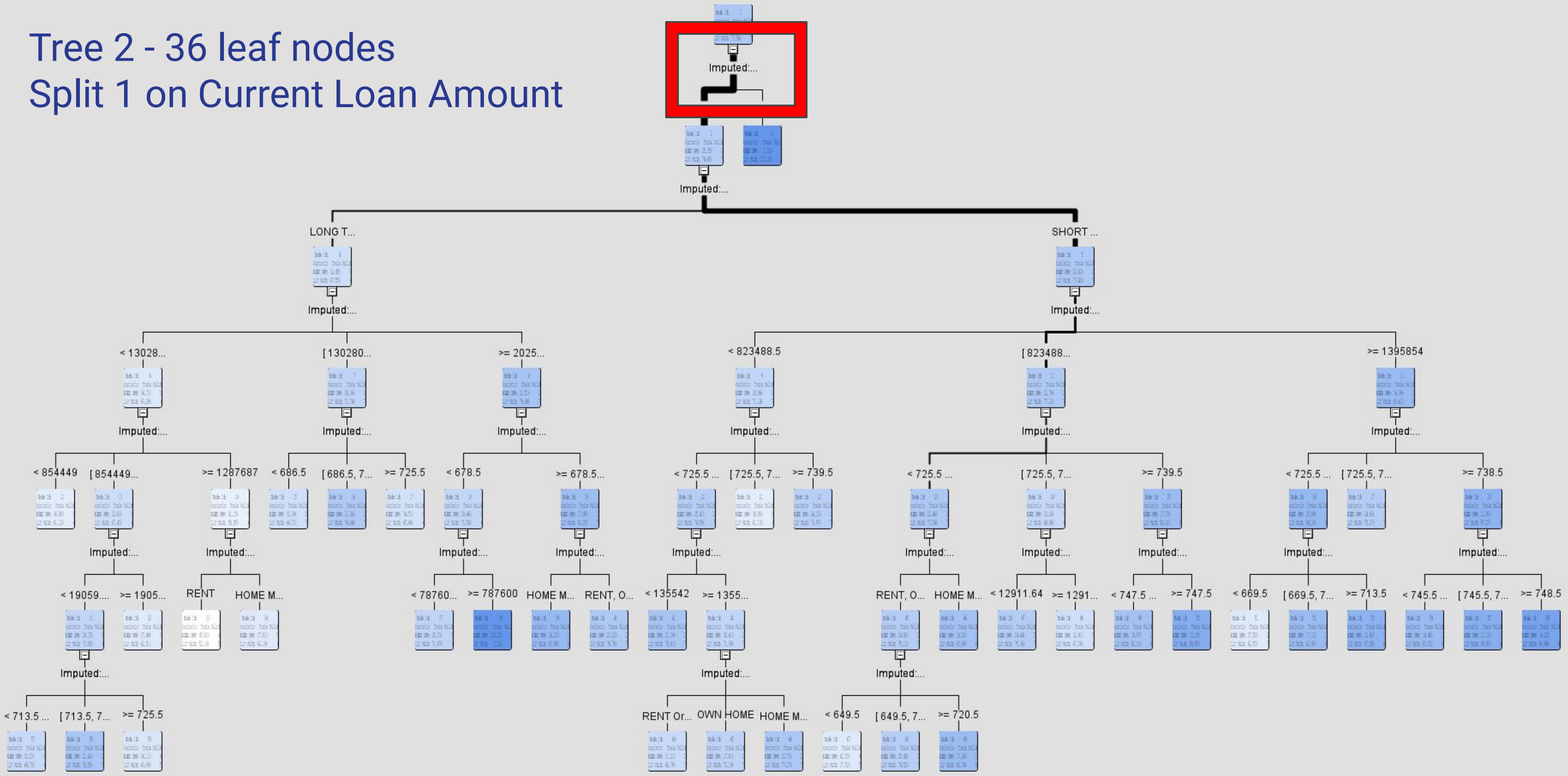
Decision Tree #2 - Maximum Branch of 3, Depth of 6

- Default tree settings except:
  - Assessment measure of ASE
  - Max. branch set to 3
- Training set
  - Misclassification Rate - 22.4158%
  - ASE - 16.0913%
- Validation set
  - Misclassification Rate - 22.4458%
  - ASE - 16.2289%
- High false positive rate



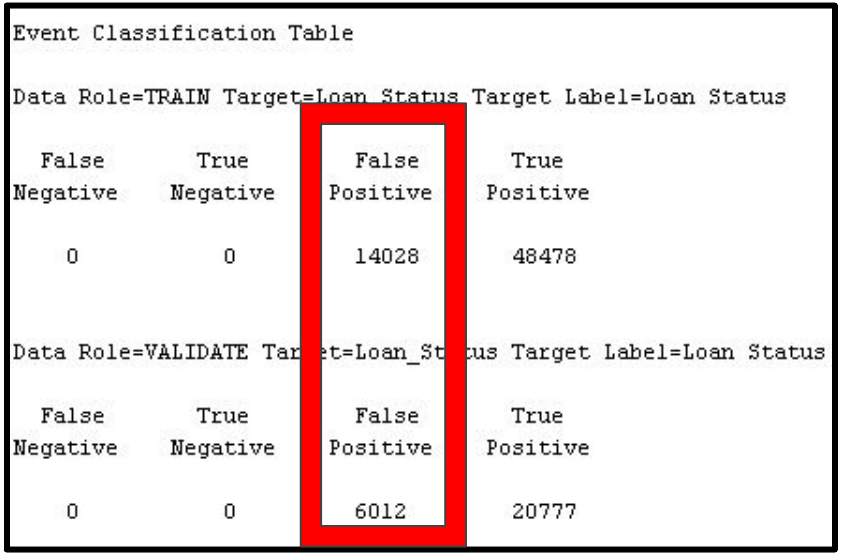
```
Event Classification Table

Data Role=TRAIN Target=Loan_Status Target Label=Loan Status

  False       True        False       True
Negative    Negative    Positive    Positive

   12          28        14000       48466


Data Role=VALIDATE Target=Loan_Status Target Label=Loan Status

  False       True        False       True
Negative    Negative    Positive    Positive

    9           8         6004       20768
```

Tree 2 - 36 leaf nodes
Split 1 on Current Loan Amount



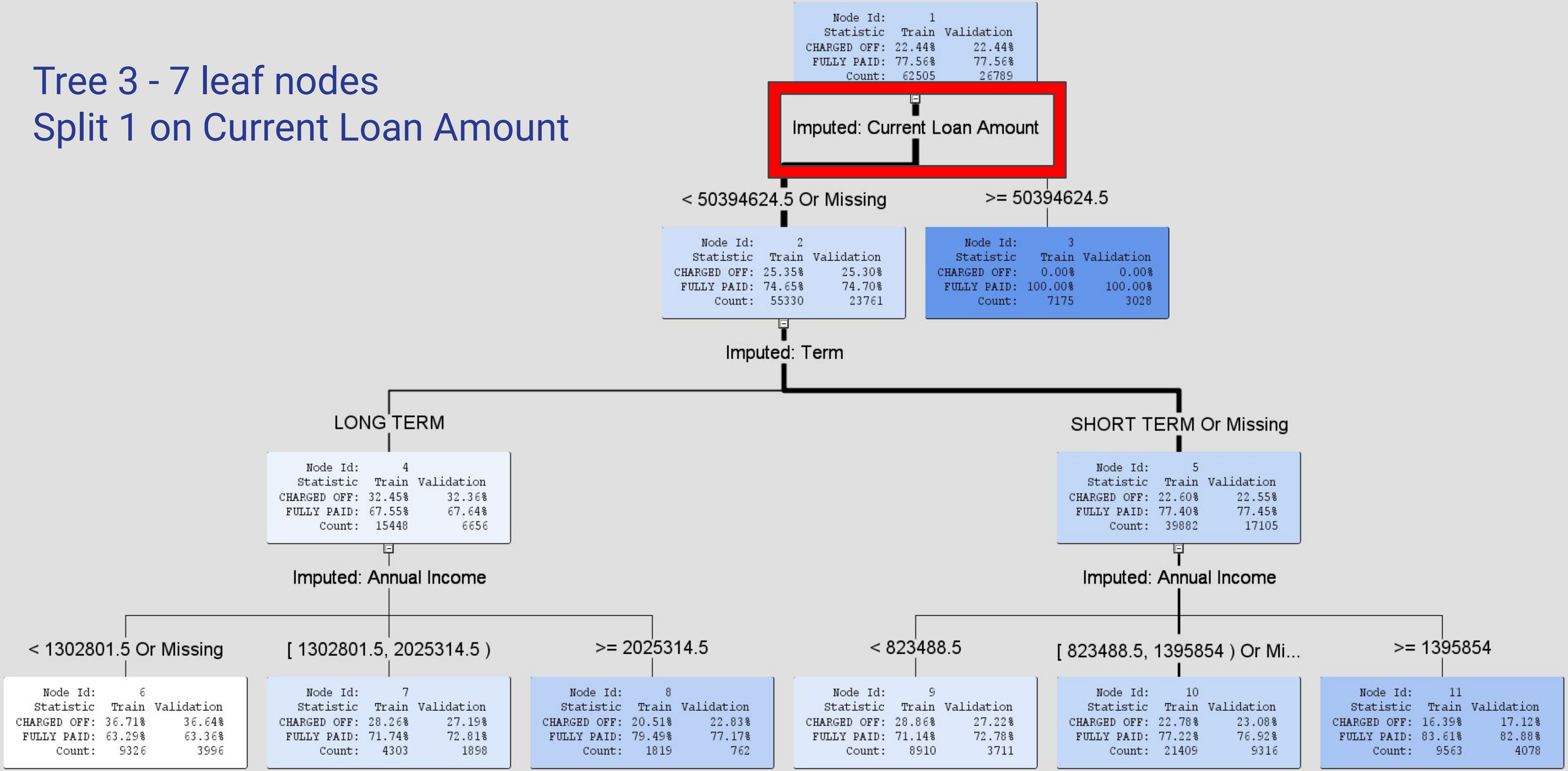
Decision Tree #3 - Maximum Branch of 3, Depth of 3

- Default tree settings except:
  - Assessment measure of ASE
  - Max. branch set to 3
  - Max. depth set to 3
- Training set
  - Misclassification Rate - 22.4158%
  - ASE - 16.0913%
- Validation set
  - Misclassification Rate - 22.4458%
  - ASE - 16.2289%
- High false positive rate



```
Event Classification Table

Data Role=TRAIN Target=Loan_Status Target Label=Loan Status

  False        True        False        True
Negative     Negative     Positive     Positive

    0            0          14028        48478


Data Role=VALIDATE Target=Loan_Status Target Label=Loan Status

  False        True        False        True
Negative     Negative     Positive     Positive

    0            0           6012        20777
```

Tree 3 - 7 leaf nodes
Split 1 on Current Loan Amount



```
         Node Id:       1
       Statistic   Train  Validation
    CHARGED OFF:  22.44%      22.44%
     FULLY PAID:  77.56%      77.56%
          Count:   62505       26789
```

Imputed: Current Loan Amount

< 50394624.5 Or Missing

```
         Node Id:       2
       Statistic   Train  Validation
    CHARGED OFF:  25.35%      25.30%
     FULLY PAID:  74.65%      74.70%
          Count:   55330       23761
```

>= 50394624.5

```
         Node Id:       3
       Statistic   Train  Validation
    CHARGED OFF:   0.00%       0.00%
     FULLY PAID: 100.00%     100.00%
          Count:    7175        3028
```

Imputed: Term

LONG TERM

```
         Node Id:       4
       Statistic   Train  Validation
    CHARGED OFF:  32.45%      32.36%
     FULLY PAID:  67.55%      67.64%
          Count:   15448        6656
```

SHORT TERM Or Missing

```
         Node Id:       5
       Statistic   Train  Validation
    CHARGED OFF:  22.60%      22.55%
     FULLY PAID:  77.40%      77.45%
          Count:   39882       17105
```

Imputed: Annual Income

< 1302801.5 Or Missing

```
         Node Id:       6
       Statistic   Train  Validation
    CHARGED OFF:  36.71%      36.64%
     FULLY PAID:  63.29%      63.36%
          Count:    9326        3996
```

[1302801.5, 2025314.5)

```
         Node Id:       7
       Statistic   Train  Validation
    CHARGED OFF:  28.26%      27.19%
     FULLY PAID:  71.74%      72.81%
          Count:    4303        1898
```

>= 2025314.5

```
         Node Id:       8
       Statistic   Train  Validation
    CHARGED OFF:  20.51%      22.83%
     FULLY PAID:  79.49%      77.17%
          Count:    1819         762
```

Imputed: Annual Income

< 823488.5

```
         Node Id:       9
       Statistic   Train  Validation
    CHARGED OFF:  28.86%      27.22%
     FULLY PAID:  71.14%      72.78%
          Count:    8910        3711
```

[823488.5, 1395854) Or Mi...

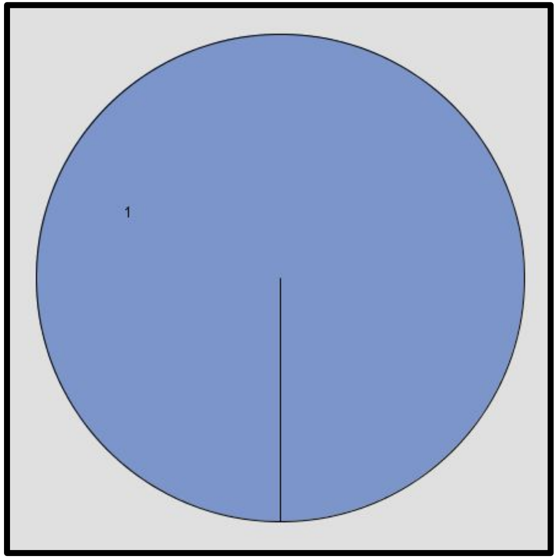
```
         Node Id:      10
       Statistic   Train  Validation
    CHARGED OFF:  22.78%      23.08%
     FULLY PAID:  77.22%      76.92%
          Count:   21409        9316
```

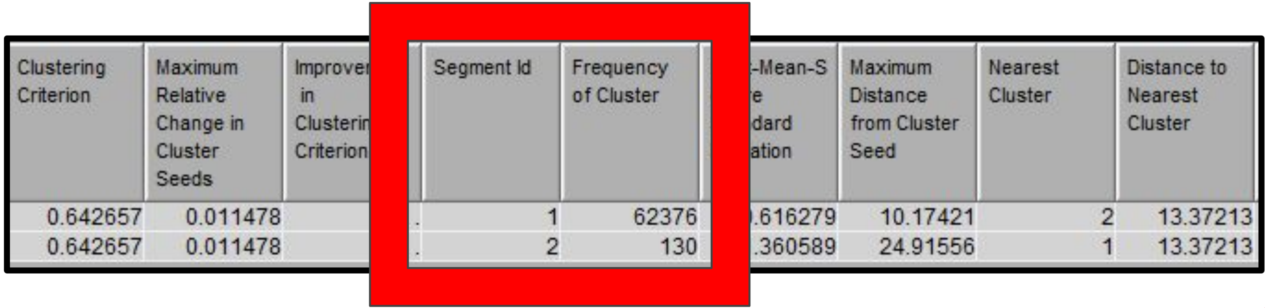
>= 1395854

```
         Node Id:      11
       Statistic   Train  Validation
    CHARGED OFF:  16.39%      17.12%
     FULLY PAID:  83.61%      82.88%
          Count:    9563        4078
```

Clustering Analysis

- ❖ Default SAS settings except:
  - ➢ Ward cluster method used
- ❖ 2 clusters found with extreme frequency inequality
- ❖ Experimented with various settings
  - ➢ No change found
  - ➢ Changed clustering method
  - ➢ Adjusted minimum allowed clusters



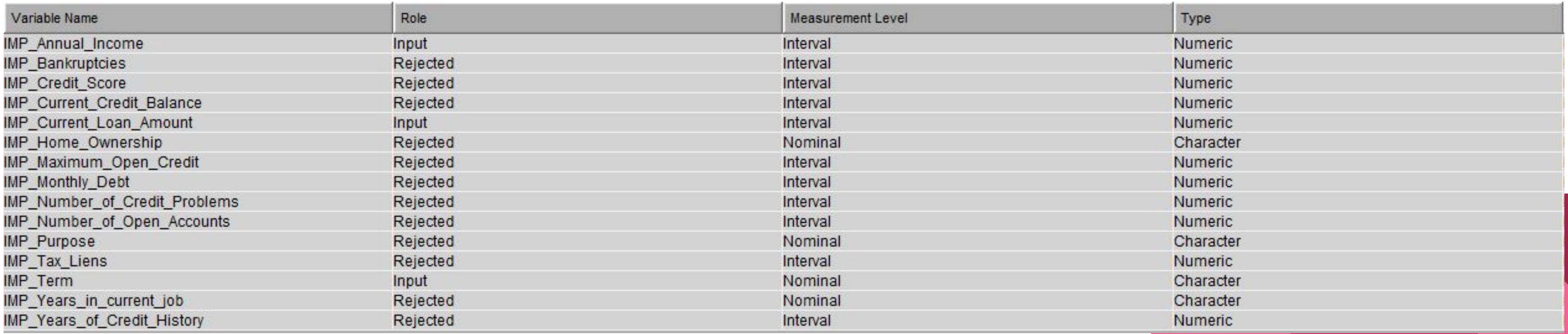


Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improver in Clusterin Criterion		Segment Id	Frequency of Cluster	-Mean-S e dard ation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
0.642657	0.011478		.	1	62376	.616279	10.17421	2	13.37213
0.642657	0.011478		.	2	130	.360589	24.91556	1	13.37213

Auto Neural Network and Variable Selection

Variable Selection

- Default settings
- Done to reduce computational requirements
- Reduced number of variables down to 4 (R^2 evaluation)

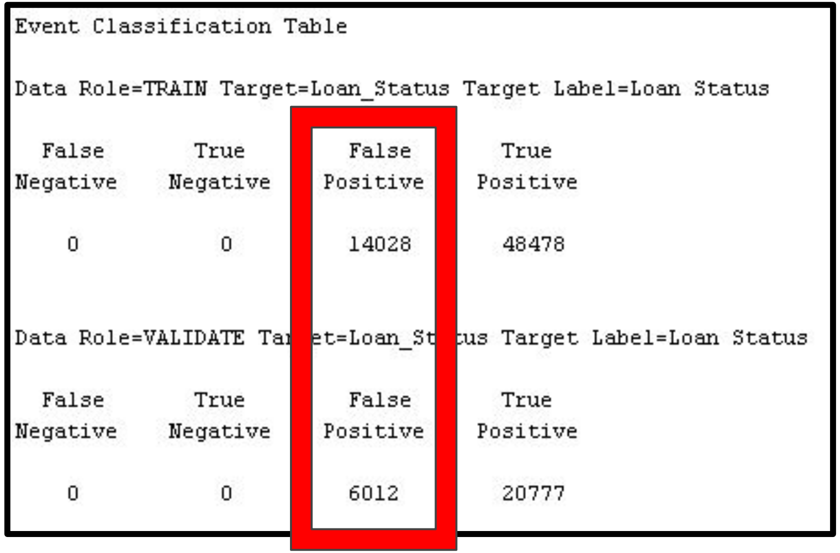


Variable Name	Role	Measurement Level	Type
IMP_Annual_Income	Input	Interval	Numeric
IMP_Bankruptcies	Rejected	Interval	Numeric
IMP_Credit_Score	Rejected	Interval	Numeric
IMP_Current_Credit_Balance	Rejected	Interval	Numeric
IMP_Current_Loan_Amount	Input	Interval	Numeric
IMP_Home_Ownership	Rejected	Nominal	Character
IMP_Maximum_Open_Credit	Rejected	Interval	Numeric
IMP_Monthly_Debt	Rejected	Interval	Numeric
IMP_Number_of_Credit_Problems	Rejected	Interval	Numeric
IMP_Number_of_Open_Accounts	Rejected	Interval	Numeric
IMP_Purpose	Rejected	Nominal	Character
IMP_Tax_Liens	Rejected	Interval	Numeric
IMP_Term	Input	Nominal	Character
IMP_Years_in_current_job	Rejected	Nominal	Character
IMP_Years_of_Credit_History	Rejected	Interval	Numeric

Auto Neural Network and Variable Selection

Auto Neural Network

- Changed number of hidden units to 3
- Training Set
  - Misclassification Rate - 22.4414%
  - ASE - 17.4052%
- Validation Set
  - Misclassification Rate - 22.442%
  - ASE - 17.4056%
- False positive rate

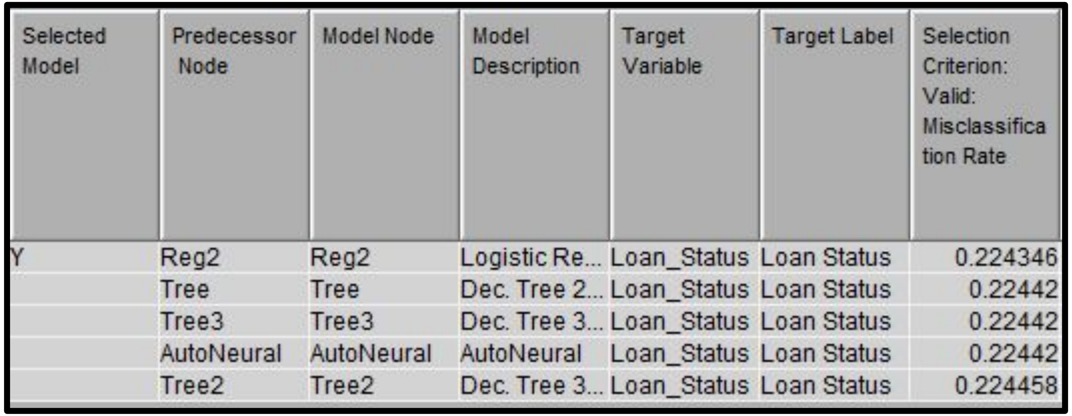


```
Event Classification Table

Data Role=TRAIN Target=Loan_Status Target Label=Loan Status

  False       True       False       True
Negative   Negative   Positive   Positive

     0          0       14028       48478

Data Role=VALIDATE Target=Loan_Status Target Label=Loan Status

  False       True       False       True
Negative   Negative   Positive   Positive

     0          0        6012       20777
```

Model Comparison and Conclusion

Model Comparison

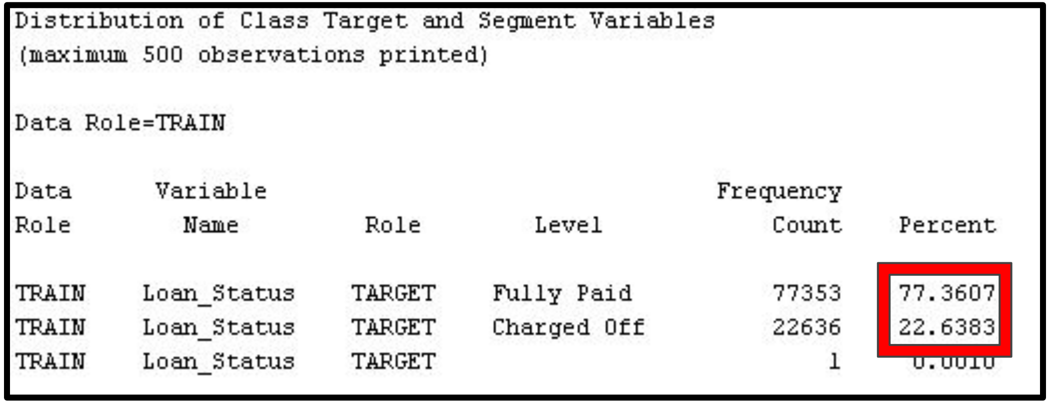
- Logistic regression was the chosen model
  - 0.74% better misclassification rate as compared to other models
- Ranked order below



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Reg2	Reg2	Logistic Re...	Loan_Status	Loan Status	0.224346
	Tree	Tree	Dec. Tree 2...	Loan_Status	Loan Status	0.22442
	Tree3	Tree3	Dec. Tree 3...	Loan_Status	Loan Status	0.22442
	AutoNeural	AutoNeural	AutoNeural	Loan_Status	Loan Status	0.22442
	Tree2	Tree2	Dec. Tree 3...	Loan_Status	Loan Status	0.224458

Conclusion

- Baseline model statistics
  - 22.6383% misclassification rate expected
- Logistic regression statistics ★
  - 22.4346% misclassification rate
  - There is a small improvement in our overall performance by using the logistic regression



```
Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data      Variable                                      Frequency
Role        Name         Role        Level                  Count    Percent

TRAIN     Loan_Status    TARGET     Fully Paid              77353    77.3607
TRAIN     Loan_Status    TARGET     Charged Off             22636    22.6383
TRAIN     Loan_Status    TARGET                                 1     0.0010
```

Limitations and Considerations

- Attempted PCA analysis prior to model running
  - Overall effect was negligible, and led some models to perform slightly worse
- Large number of missing values
  - Although imputed, having a complete dataset would be preferred
- Inequality in class target variable
  - Possible solutions include technique known as oversampling
- Oversampling
  - Not conducted here
  - Draws a greater number of sample records from the class considered to be a “rare event”

References

Arafa, A. (2020, August 8). Bank Loans. Retrieved March 2022, from https://www.kaggle.com/code/abdelrahmanarafa/bankloans34.

Wang, R., Lee, N., &amp; Wei, Y. (n.d.). A Case Study: Improve Classification of Rare Events with SAS Enterprise Miner. SAS Support Papers. Retrieved April 2022, from https://support.sas.com/resources/papers/proceedings15/3282-2015.pdf