Decision Trees in Risk Modelling

Gerard D. Ompad

Question:

Data Monetization:

- The practice of trading personal data.
- Production of new products and services.
- How social media sites earn profit.
- Other firms also conduct this practice.
- Embedded in the “Terms & Agreement” button.

- https://searchcio.techtarget.com/definition/data-monetization
- https://sloanreview.mit.edu/article/demystifying-data-monetization/
- Bataineh et al., 2016, Monetizing Personal Data: A Two-Sided Market Approach, Procedia Computer Science 83 (2016) 472 – 479

- https://www.slideshare.net/DataReportal/digital-2019-philippines-january-2019-v01 (Slide 15)

ICT Data:

- Loading behavior: frequency of load, amount of load, how often do you load (days between).
- Mobile service consumption: what kind of telco service do you use, how frequent and how often do you use it? How much would you spend for a service? How frequent do you text and call?
- Internet Consumption: How frequent do you use your mobile internet? How much do you consume (MB).

- Bataineh et al., 2016, Monetizing Personal Data: A Two-Sided Market Approach, Procedia Computer Science 83 (2016) 472 – 479

Issues with Data Monetization

- Trading of personal data have an impact on data privacy.
- Exposing of sensitive data like names, addresses and medical conditions that puts an individual's morale in jeopardy, affects working and living conditions, and social relationships.
- Users needs to have more control on data privacy, once the users perceived that their data is being exposed, they have a lower customer relationship quality with the service providers.
- Companies needs to have a Good Data Governance Officer.

- Beierle et al., 2019 Context Data Categories and Privacy Model for Mobile Data Collection Apps Procedia Computer Science Volume 134, 2018, Pages 18-25
- Mpinganjira & Maduku, 2019, Ethics of mobile behavioral advertising: Antecedents and outcomes of perceived ethical value of advertised brands Journal of Business Research Volume 95, February 2019, Pages 464-478

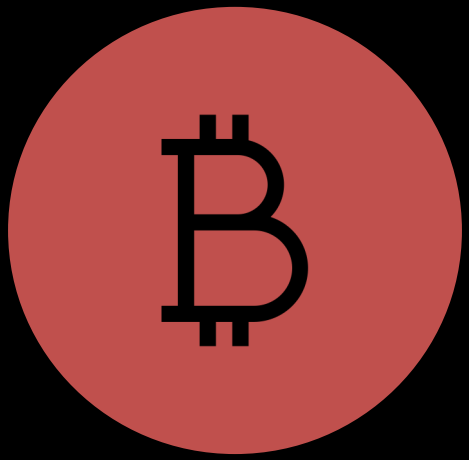
The Good Side

Data monetization have assisted micro-financial institutions in lending money to the unbanked.

Providing secondary data to assess risk.

Determine good borrowers.

Case Study:



GIVEN AN UNBANKED INDIVIDUAL CAN WE USE THEIR MOBILE DATA TO DETERMINE IF THEY ARE A GOOD OR A RISKY BORROWER, PROVIDED THEY HAVE ALLOWED THEIR MOBILE DATA TO BE ***SOLD*** TO THE MFI.

Oversimplified Approach

OBTAIN HISTORICAL DATA OF GOOD AND BAD BORROWERS.

PERFORM EXPLORATORY DATA ANALYSIS.

OBTAIN EACH PERSON'S TELCO DATA.

SPLIT THE DATA INTO TRAINING AND TESTING.

TRAIN A ML MODEL ON THE TRAINING DATA.

TEST THE ML MODEL ON THE TESTING DATA AND DETERMINE THE CONFUSION MATRIX.

USE K-FOLD VALIDATION TO DETERMINE THE ROC-AUC

Oversimplified Approach

OBTAIN HISTORICAL DATA OF GOOD AND BAD BORROWERS.

PERFORM EXPLORATORY DATA ANALYSIS.

OBTAIN EACH PERSON'S TELCO DATA.

SPLIT THE DATA INTO TRAINING AND TESTING.

TRAIN A ML MODEL ON THE TRAINING DATA.

TEST THE ML MODEL ON THE TESTING DATA AND DETERMINE THE CONFUSION MATRIX.

USE K-FOLD VALIDATION TO DETERMINE THE ROC-AUC

Financial Data					Mobile Data (Recent Month)			
sex	age	est_monthly_income	est_monthly_expenses	Payer Type	Amount Load	Days Between Top Up	Frequency of Load Loan	Amount of Load Loan
F	31	12600	8600	B	1701	1.705882353	0	0
F	32	17500	12000	B	0		0	0
F	44	18000	12000	B	0		0	0
F	36	18000	15000	B	171	0	0	0
F	29	13500	9500	B	486	0	0	0
F	34	16500	14000	B	0		0	0

The Historical Dataset

- 9 members per group
- One member will solve one column.

Oversimplified Approach

OBTAIN HISTORICAL DATA OF GOOD AND BAD BORROWERS.

PERFORM EXPLORATORY DATA ANALYSIS.

OBTAIN EACH PERSON'S TELCO DATA.

SPLIT THE DATA INTO TRAINING AND TESTING.

TRAIN A ML MODEL ON THE TRAINING DATA.

TEST THE ML MODEL ON THE TESTING DATA AND DETERMINE THE CONFUSION MATRIX.

USE K-FOLD VALIDATION TO DETERMINE THE ROC-AUC

Exploratory Data Analysis

Method

- Determine the basic statistics of the data
- Mean, median and mode (Discrete data)
- Mean and standard deviation (Continuous Data)
- Frequency & Proportion (Categorical data)
- Maximum & Minimum Value

Activity

You are provided with a masked dataset. Within your group perform the exploratory data analysis by:

Calculating the Mean:

$$\bar{x} = \sum \frac{x}{n}$$

Determine the Median:

The middle observation when the sample measurements are ordered according to magnitude (when n is odd), or the average of the two middle observations (when n is even).

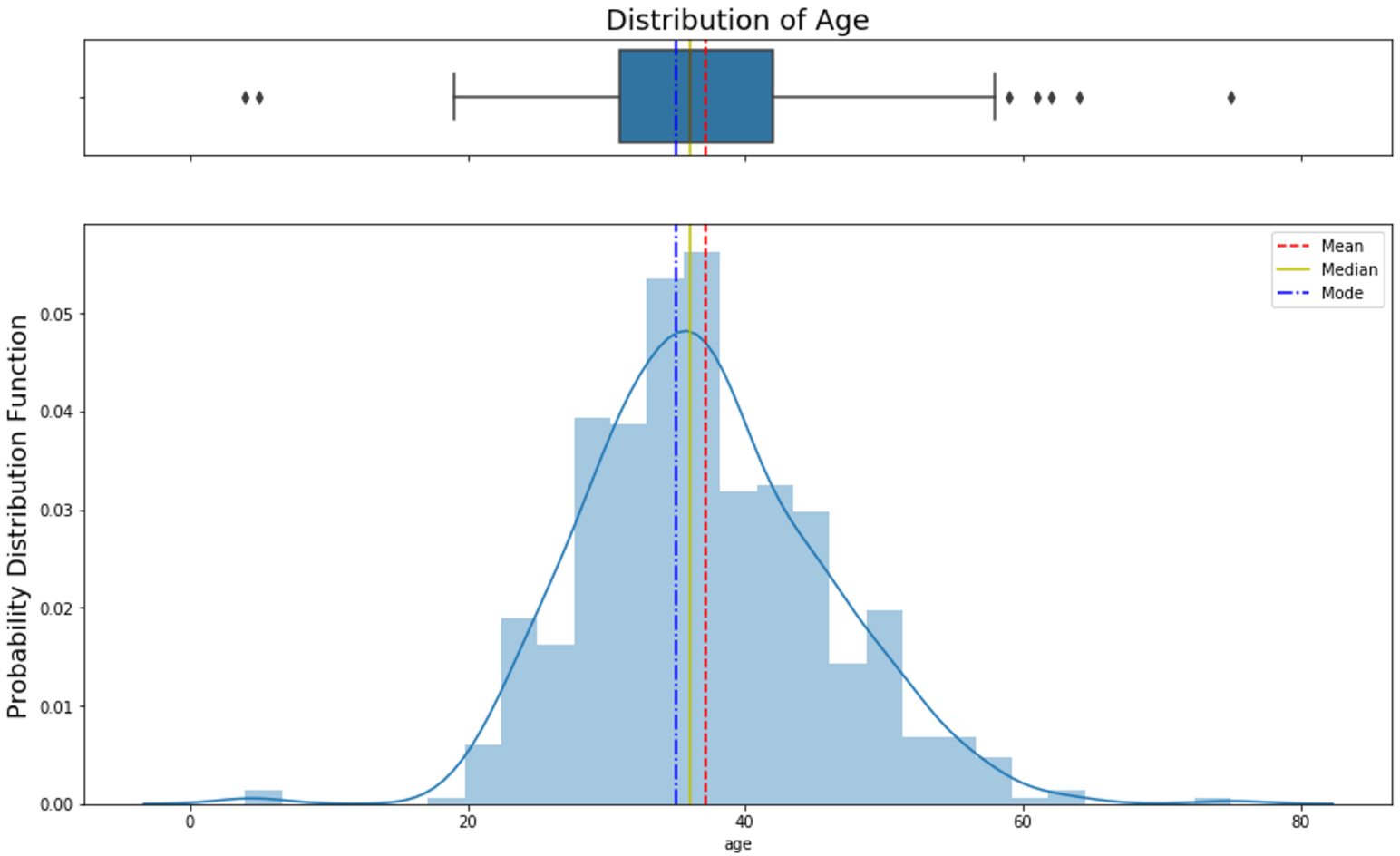
Determine the Mode:

The most frequent observation.

The Maximum & Minimum:

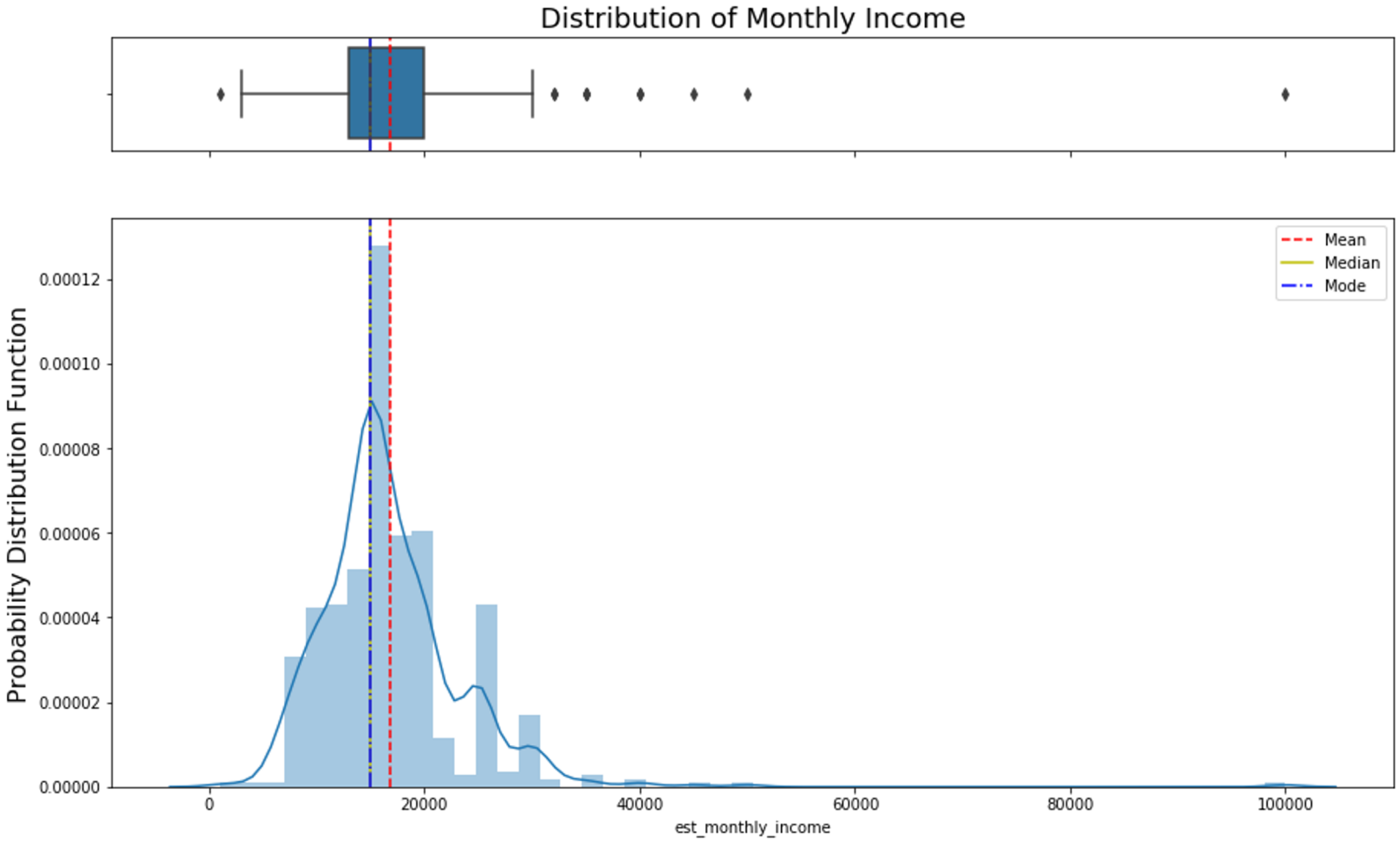
The largest and smallest observations (respectively)

Age



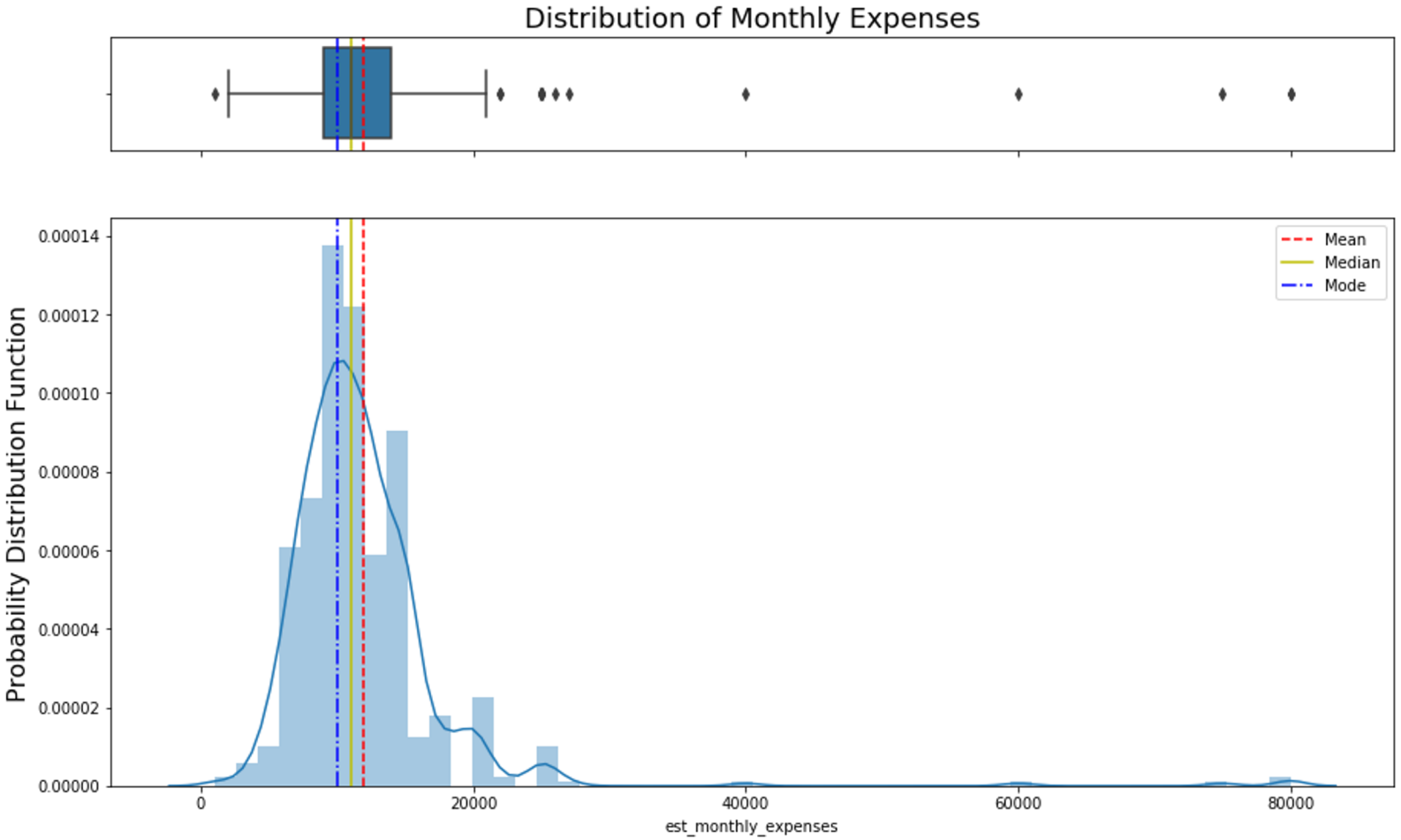
The average age is 37 years old.

Monthly Income



On average, users earn approximately 15,000 - 17,000.

Monthly Expenses



On average, users consume approximately 11,000 - 12,000.

Oversimplified Approach

OBTAIN HISTORICAL DATA OF GOOD AND BAD BORROWERS.

PERFORM EXPLORATORY DATA ANALYSIS.

OBTAIN EACH PERSON'S TELCO DATA.

SPLIT THE DATA INTO TRAINING AND TESTING.

TRAIN A ML MODEL ON THE TRAINING DATA.

TEST THE ML MODEL ON THE TESTING DATA AND DETERMINE THE CONFUSION MATRIX.

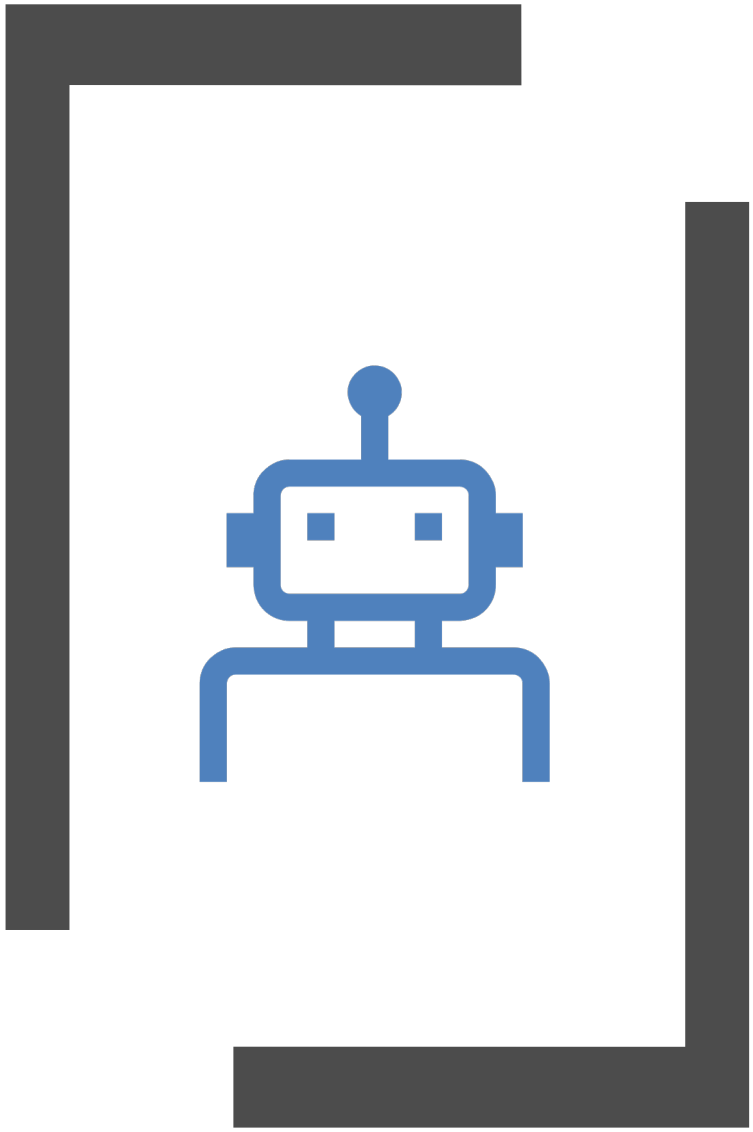
USE K-FOLD VALIDATION TO DETERMINE THE ROC-AUC

Financial Data					Mobile Data (Recent Month)			
sex	age	est_monthly_income	est_monthly_expenses	Payer Type	Amount Load	Days Between Top Up	Frequency of Load Loan	Amount of Load Loan
F	31	12600	8600	B	1701	1.705882353	0	0
F	32	17500	12000	B	0		0	0
F	44	18000	12000	B	0		0	0
F	36	18000	15000	B	171	0	0	0
F	29	13500	9500	B	486	0	0	0
F	34	16500	14000	B	0		0	0

The Historical Dataset

- Telco data on the right hand side.

Data Splitting



- Separate the good payers from the bad payers in the sample.
- Determine the percentage of the good payer the bad payers.
- Find an arbitrary percentage of the original sample set to be used as testing.
- Randomly obtain sub-samples such that the ratio is preserved.

Example:

1. Original sample size: 100
2. 60 good payers (60% good payers)
3. 40 bad payers (40% bad payers)
4. Percentage to be used as testing data: 10% (or n=10)
5. 10 people consisting of 6 good payers (60% good) and 4 bad payers (40% bad)

Oversimplified Approach

OBTAIN HISTORICAL DATA OF GOOD AND BAD BORROWERS.

PERFORM EXPLORATORY DATA ANALYSIS.

OBTAIN EACH PERSON'S TELCO DATA.

SPLIT THE DATA INTO TRAINING AND TESTING.

TRAIN A ML MODEL ON THE TRAINING DATA.

TEST THE ML MODEL ON THE TESTING DATA AND DETERMINE THE CONFUSION MATRIX.

USE K-FOLD VALIDATION TO DETERMINE THE ROC-AUC

Classification Problem

Classification algorithms are a form of supervised machine learning method.

The a set of predictor variables is a pre-defined category.

The goal is to “classify” new datasets with high performance.

Decision Tree Algorithm

The simplest yet the most widely used classification algorithm.

Root node

Internal Node

Terminal Node

ID3 Algorithm

Selecting the Best Split

1. Consider a category of class *i.*

2. Let t denote the attribute.

3. The fraction of the data that belongs to class *i* , for any given attribute is denoted by:

$$\Pr(i|t)$$

4. The best split can be determined by the measure of impurity:

$$Gini = 1 - \sum_{i} \Pr(i|t)^2$$

Selecting the Best Split

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
5	No	Divorced	95K	Yes
8	No	Single	85K	Yes
10	No	Single	90K	Yes

binary categorical continuous class

Determine the homogeneity of the Target Variable:

Default Borrower:

# of Yes = 3

# of No =

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
6	No	Married	60K	No
7	Yes	Divorced	220K	No
9	No	Married	75K	No

Determine the homogeneity of the Target Variable:

Default Borrower:

# of Yes = 3

# of No = 7

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Determine the homogeneity of the Target Variable:

Default Borrower:

# of Yes = 3

# of No = 7

$$Gini = 1 - \sum_i \Pr(i|t)^2$$

$$Gini = 1 - \left[\left(\frac{3}{10}\right)^2 + \left(\frac{7}{10}\right)^2\right]$$

$$Gini = 1 - \left[\frac{29}{50}\right]$$

$$Gini = \frac{21}{50} = 0.42$$

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	125K	No
4	Yes	Married	120K	No
7	Yes	Divorced	220K	No

Determine the homogeneity in each of the attributes:

Home Owner = Yes

# of Defaulted Borrower Yes = 0

# of Defaulted Borrower No = 3

$$Gini = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = 1$$

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
5	No	Divorced	95K	Yes
8	No	Single	85K	Yes
10	No	Single	90K	Yes

Determine the homogeneity in each of the attributes:

Home Owner = Yes

# of Defaulted Borrower Yes = 0

# of Defaulted Borrower No = 3

$$Gini = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = 1$$

Home Owner = No

# of Defaulted Borrower Yes = 3

# of Defaulted Borrower No =

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
2	No	Married	100K	No
3	No	Single	70K	No
6	No	Married	60K	No
9	No	Married	75K	No

binary categorical continuous class

Determine the homogeneity in each of the attributes:

Home Owner = Yes

# of Defaulted Borrower Yes = 0

# of Defaulted Borrower No = 3

$$Gini = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = 1$$

Home Owner = No

# of Defaulted Borrower Yes = 3

# of Defaulted Borrower No = 4

$$Gini = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0.48$$

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Determine the homogeneity in each of the attributes:

Home Owner = Yes

# of Defaulted Borrower Yes = 0

# of Defaulted Borrower No = 3

$$Gini = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right] = 1$$

Home Owner = No

# of Defaulted Borrower Yes = 3

# of Defaulted Borrower No = 4

$$Gini = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0.48$$

Total Gini for Home Owner:

$$Gini = \sum_i \frac{n_i}{n_{total}} \Pr(i|t)^2$$

$$Gini = \left(\frac{0+3}{10} x\ 0\right) + \left(\frac{3+4}{10} x 0.48\right)$$

$$Gini = 0.34$$

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
8	No	Single	85K	Yes
10	No	Single	90K	Yes

binary categorical continuous class

Determine the homogeneity in each of the attributes:

Marital Status = Single

# of Defaulted Borrower Yes = 2

# of Defaulted Borrower No =

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	binary	categorical	continuous	class
1	Yes	Single	125K	No
3	No	Single	70K	No

Determine the homogeneity in each of the attributes:

Marital Status = Single

# of Defaulted Borrower Yes = 2

# of Defaulted Borrower No = 2

$$Gini = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$$

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
2	No	Married	100K	No
4	Yes	Married	120K	No
6	No	Married	60K	No
9	No	Married	75K	No

Determine the homogeneity in each of the attributes:

Marital Status = Single

# of Defaulted Borrower Yes = 2

# of Defaulted Borrower No = 2

$$Gini = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$$

Marital Status = Married

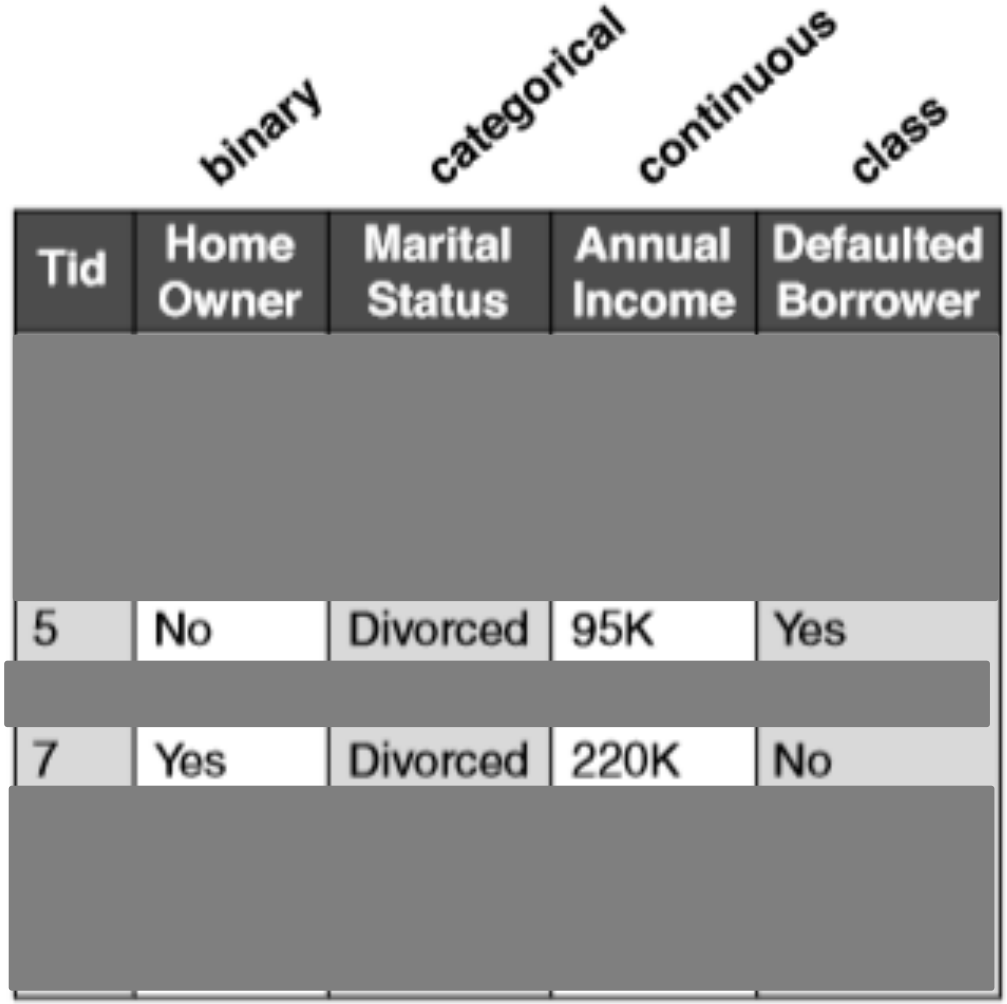
# of Defaulted Borrower Yes = 0

# of Defaulted Borrower No = 4

$$Gini = 1 - \left[\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2\right] = 0$$

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split



Determine the homogeneity in each of the attributes:

Marital Status = Single

# of Defaulted Borrower Yes = 2

# of Defaulted Borrower No = 2

$$Gini = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$$

Marital Status = Married

# of Defaulted Borrower Yes = 0

# of Defaulted Borrower No = 4

$$Gini = 1 - \left[\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2\right] = 0$$

Marital Status = Divorced

# of Defaulted Borrower Yes = 1

# of Defaulted Borrower No = 1

$$Gini = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$$

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Selecting the Best Split

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Determine the homogeneity in each of the attributes:

Marital Status = Single

# of Defaulted Borrower Yes = 2

# of Defaulted Borrower No = 2

$$Gini = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$$

Marital Status = Married

# of Defaulted Borrower Yes = 0

# of Defaulted Borrower No = 4

$$Gini = 1 - \left[\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2\right] = 0$$

Marital Status = Divorced

# of Defaulted Borrower Yes = 1

# of Defaulted Borrower No = 1

$$Gini = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$$

Total Gini for Marital Status:

$$Gini = \left(\frac{2+2}{10} x\ 0.5\right) + \left(\frac{0+4}{10} x0\right) + \left(\frac{1+1}{10} x0.5\right)$$

$$Gini = 0.30$$

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Home Owner	Marital Status	Annual Income	Defaulted Borrower
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

	No	No	No	Yes	Yes	Yes	No	No	No	No
	60	70	75	85	90	95	100	120	125	220
	55	65	72	80	87	92	97	110	122	172	230
	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >
Yes	0 3	0 3	0 3	0 3	1 2	2 1	3 0	3 0	3 0	3 0	3 0
No	0 7	1 6	2 5	3 4	3 4	3 4	3 4	4 3	5 2	6 1	7 0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	*0.300*	0.343	0.375	0.400	0.420

Selecting the Best Split

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Determine the homogeneity in each of the attributes:

Annual Income <= 97,000.00

# of Defaulted Borrower Yes = 4

# of Defaulted Borrower No = 3

$$Gini = 1 - \left[\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right] = \frac{24}{49}$$

Annual Income <= 97,000.00

# of Defaulted Borrower Yes = 0

# of Defaulted Borrower No = 4

$$Gini = 1 - \left[\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2\right] = 0$$

Total Gini for Annual Income:

$$Gini = \left(\frac{4+3}{10} x \frac{24}{49}\right) + \left(\frac{3+4}{10} x 0.0\right)$$

$$Gini = 0.3429$$

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

The Gain

- Compares the degree of the impurity between the target variable and each of the attribute variable

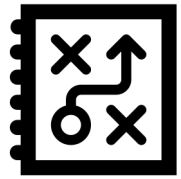
- $Gain = Gini\ of\ Target\ Variable - Gini\ of\ the\ Attribute\ Variable$

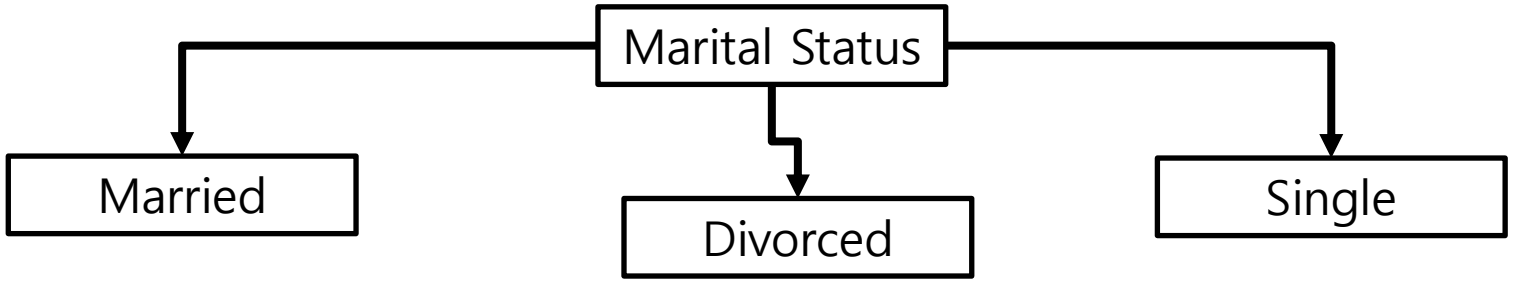
- Gain for Home Owners:
- $Gain = 0.42 - 0.34 = 0.08$

- Gain for Marital Status:
- $Gain = 0.42 - 0.30 = 0.12$

- Gain for Annual Income:
- $Gain = 0.42 - 0.3429 = 0.0071$

- Choose the attribute with the largest Gain as the root node.

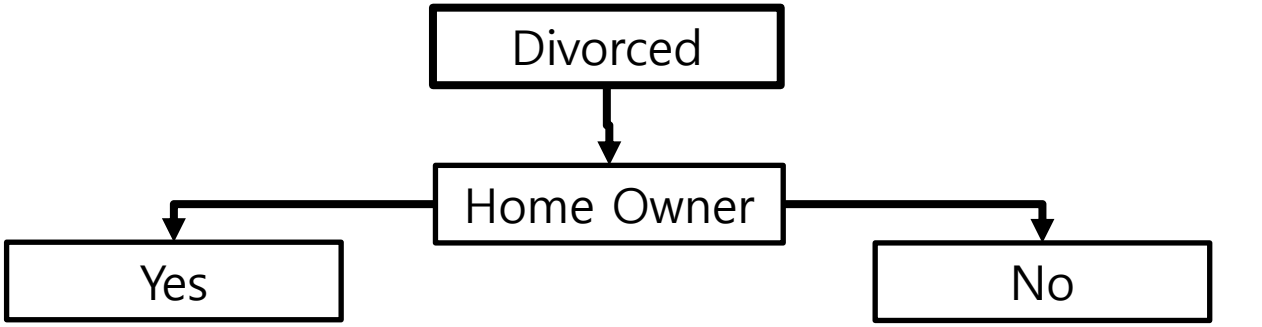




DEFAULT BORROWER	
YES	NO
0	4

DEFAULT BORROWER	
YES	NO
1	1

DEFAULT BORROWER	
YES	NO
2	2

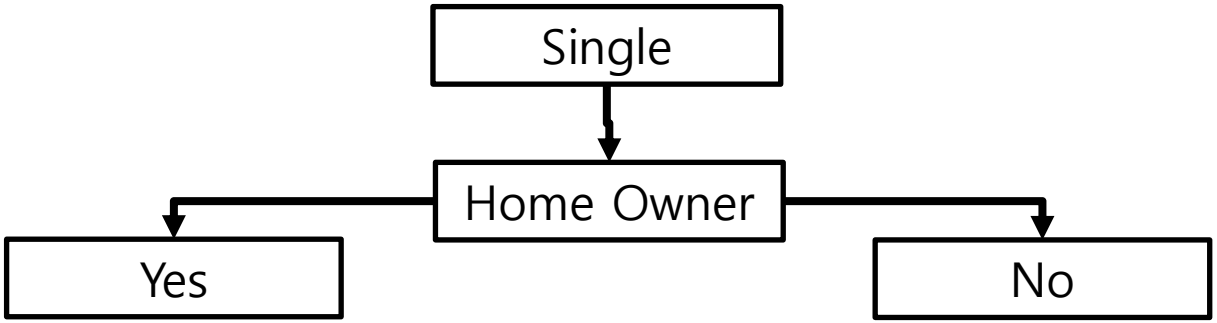


DEFAULT BORROWER	
YES	NO
0	1

DEFAULT BORROWER	
YES	NO
1	0

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Divorced	95K	Yes
Yes	Divorced	220K	No

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013



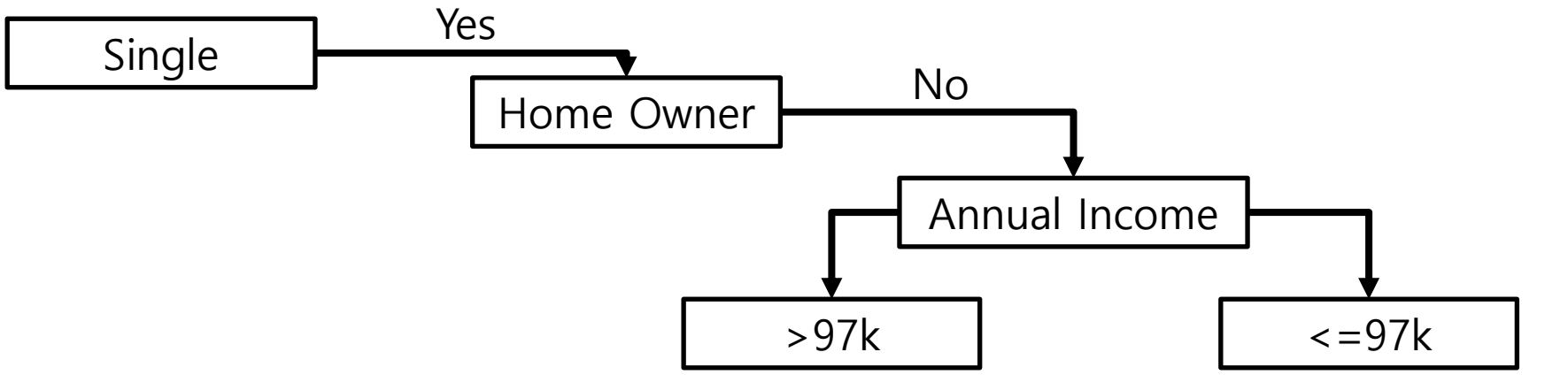
DEFAULT BORROWER	
YES	NO
0	1

DEFAULT BORROWER	
YES	NO
2	1

Home Owner	Marital Status	Annual Income	Defaulted Borrower
Yes	Single	125K	No

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Single	70K	No
No	Single	85K	Yes
No	Single	90K	Yes

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013



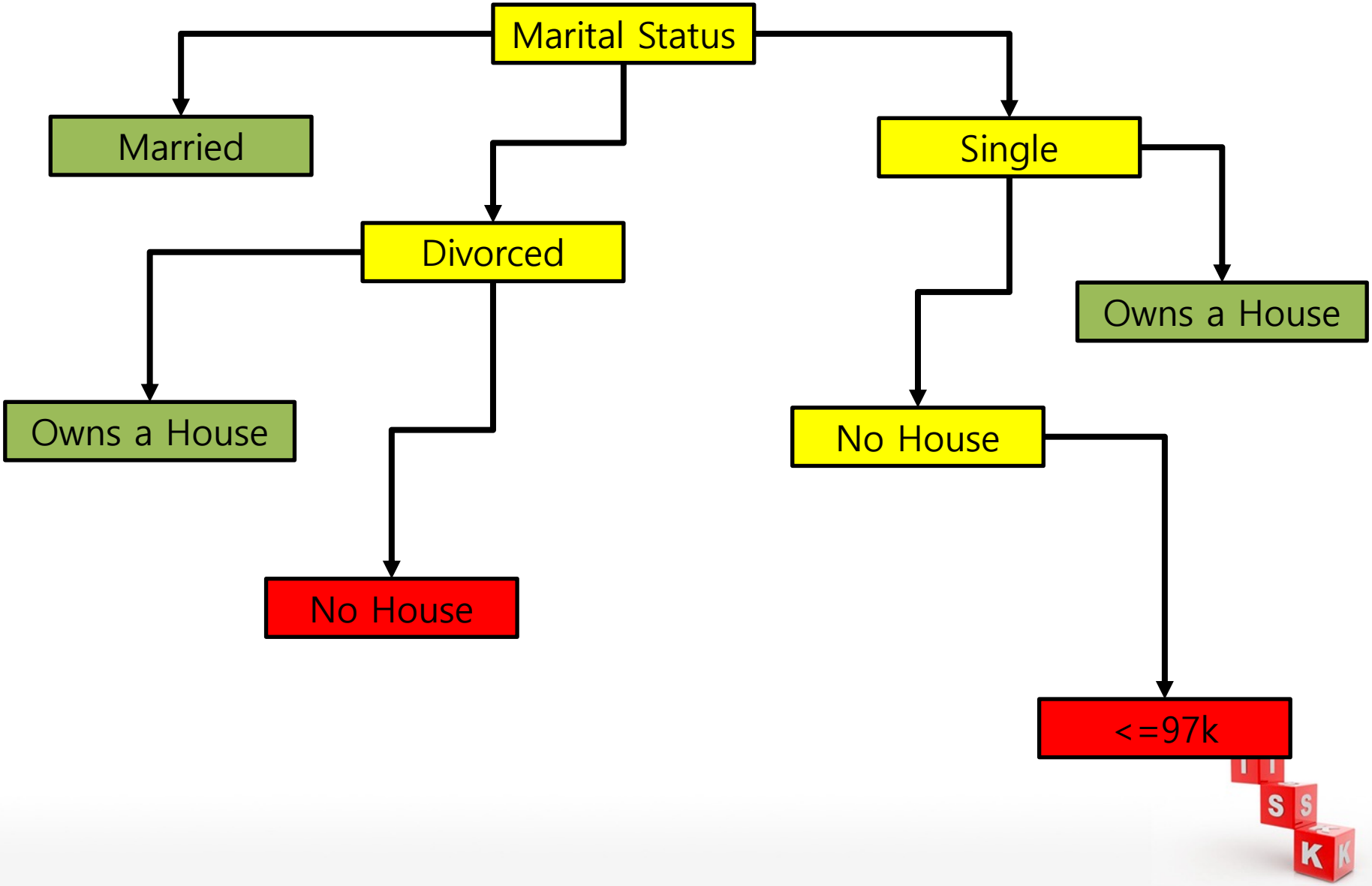
DEFAULT BORROWER	
YES	NO
0	0

DEFAULT BORROWER	
YES	NO
2	1

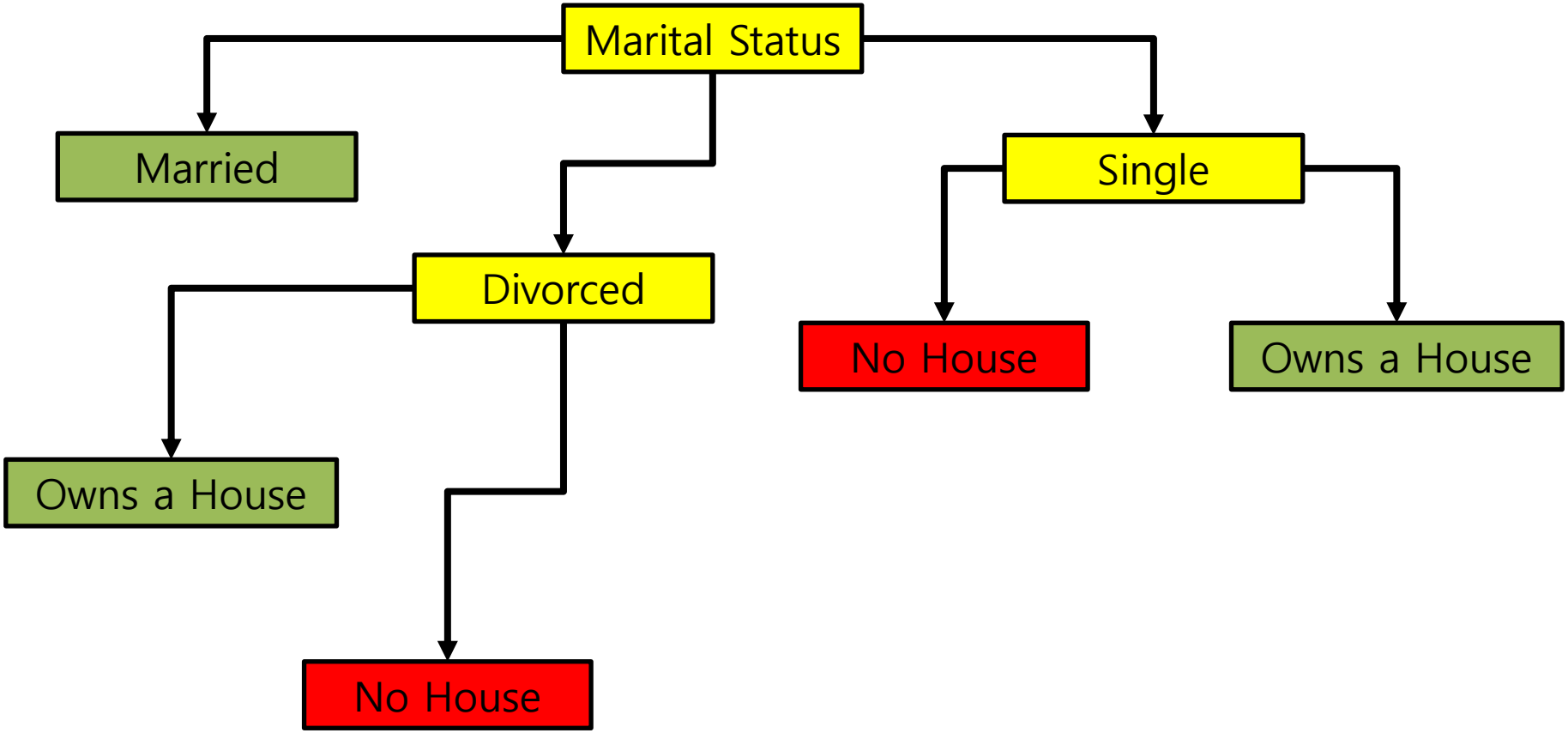
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Single	70K	No
No	Single	85K	Yes
No	Single	90K	Yes

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

The Decision Tree



The Decision Tree



Activity: Using The Training Data

Calculate the Gini of the Target Variable:

$$Gini = 1 - \sum_i \Pr(i|t)^2$$

Calculate the Gini of the Attribute Variable:

$$Gini = \sum_i \frac{n_i}{n_{total}} \Pr(i|t)^2$$

Calculate the Gain:

$$Gain = Gini\ of\ Target\ Variable - Gini\ of\ the\ Attribute\ Variable$$

Find the Root Node

Generate the Tree

Oversimplified Approach

OBTAIN HISTORICAL DATA OF GOOD AND BAD BORROWERS.

PERFORM EXPLORATORY DATA ANALYSIS.

OBTAIN EACH PERSON'S TELCO DATA.

SPLIT THE DATA INTO TRAINING AND TESTING.

TRAIN A ML MODEL ON THE TRAINING DATA.

TEST THE ML MODEL ON THE TESTING DATA AND DETERMINE THE CONFUSION MATRIX.

USE K-FOLD VALIDATION TO DETERMINE THE ROC-AUC

Classification Performance Metrics

$$Accuracy = \frac{Number\ of\ Targets\ Correctly\ Classified}{Total\ Number\ of\ Classes}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Example

Home Owner	Marital Status	Annual Income	Defaulted	Predicted	
12	Single	125K	No	No	True Negative
No	Married	100K	No	Yes	False Positive
No	Single	70K	No	No	True Negative
Yes	Married	120K	No	No	True Negative
No	Divorced	95K	Yes	Yes	True Positive
No	Married	60K	No	Yes	False Positive
Yes	Divorced	220K	No	Yes	False Positive
No	Single	85K	Yes	No	False Negative
No	Married	75K	No	No	True Negative
No	Single	90K	Yes	No	False Negative

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Constructing a Confusion Matrix

Defaulted	Predicted	
No	No	True Negative
No	Yes	False Positive
No	No	True Negative
No	No	True Negative
Yes	Yes	True Positive
No	Yes	False Positive
No	Yes	False Positive
Yes	No	False Negative
No	No	True Negative
Yes	No	False Negative

Actual	Yes		
	No		
		Yes	No
	Predicted		

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Constructing a Confusion Matrix

Defaulted	Predicted	
No	No	True Negative
No	Yes	False Positive
No	No	True Negative
No	No	True Negative
Yes	**Yes**	**True Positive**
No	Yes	False Positive
No	Yes	False Positive
Yes	No	False Negative
No	No	True Negative
Yes	No	False Negative

Actual	Yes	1	
		Yes	
	Predicted		

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Constructing a Confusion Matrix

Defaulted	Predicted	
No	No	True Negative
No	**Yes**	**False Positive**
No	No	True Negative
No	No	True Negative
Yes	Yes	True Positive
No	**Yes**	**False Positive**
No	**Yes**	**False Positive**
Yes	No	False Negative
No	No	True Negative
Yes	No	False Negative

Actual	Yes	1	
	No	**3**	
		Yes	No
	Predicted		

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Constructing a Confusion Matrix

Defaulted	Predicted	
No	No	True Negative
No	Yes	False Positive
No	No	True Negative
No	No	True Negative
Yes	Yes	True Positive
No	Yes	False Positive
No	Yes	False Positive
Yes	**No**	**False Negative**
No	No	True Negative
Yes	**No**	**False Negative**

Actual	**Yes**	1	**2**
	No	3	
		Yes	**No**
	Predicted		

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Constructing a Confusion Matrix

Defaulted	Predicted	
No	No	True Negative
No	Yes	False Positive
No	No	True Negative
No	No	True Negative
Yes	Yes	True Positive
No	Yes	False Positive
No	Yes	False Positive
Yes	No	False Negative
No	No	True Negative
Yes	No	False Negative

Actual	Yes	1	2
	No	3	**4**
		Yes	No
	Predicted		

Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Constructing a Confusion Matrix

Defaulted	Predicted	
No	No	True Negative
No	Yes	False Positive
No	No	True Negative
No	No	True Negative
Yes	Yes	True Positive
No	Yes	False Positive
No	Yes	False Positive
Yes	No	False Negative
No	No	True Negative
Yes	No	False Negative

Actual	Yes	1	2
	No	3	4
		Yes	No
	Predicted		

$$Accuracy = \frac{1+4}{10} = 0.5$$

$$Sensitivity = \frac{1}{1+2} = 0.3333$$

$$Specificity = \frac{4}{3+4} = 0.5714$$

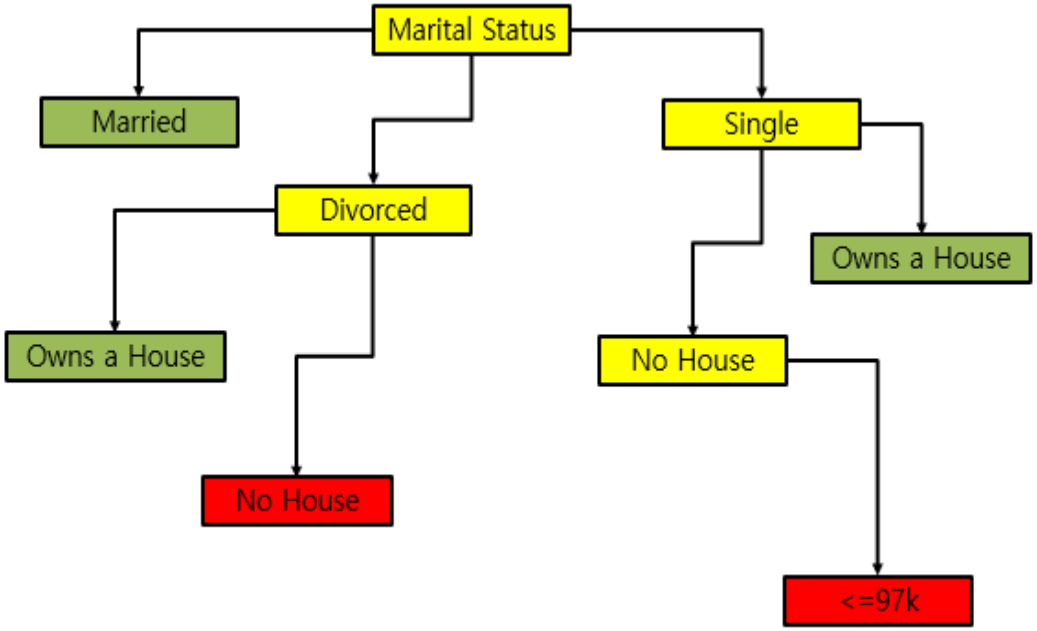
Data taken from Introduction to Data Mining by Pan-Ning Tan, Pearsons Publication 2013

Activity: Performance Determination

Using the testing dataset and the tree you constructed:

1. Classify the test data.
2. Construct a confusion matrix
3. Calculate the Accuracy, Sensitivity & Specifity

Home Owner	Marital Status	Annual Income	Predicted Default
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	Yes
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	Yes
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes



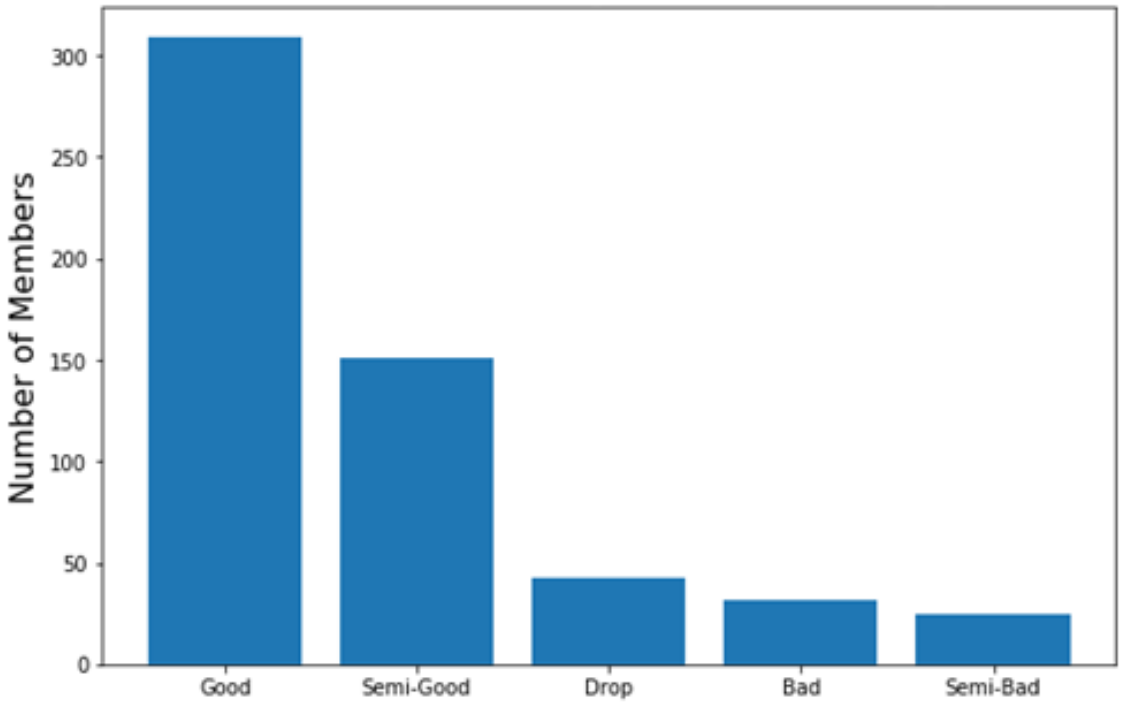
Case Study:

- Target Institutions: Micro-Finance Institution targeting unbanked farmers.
- Target Market: Unbanked farmers.
- Experiment: Prove that telco data can be used to determine good paying behavior, in the absence of financial data.

Model | Benchmark

Proportion of Chance Criterion (PCC) is equivalent to predicting the users by random chance.

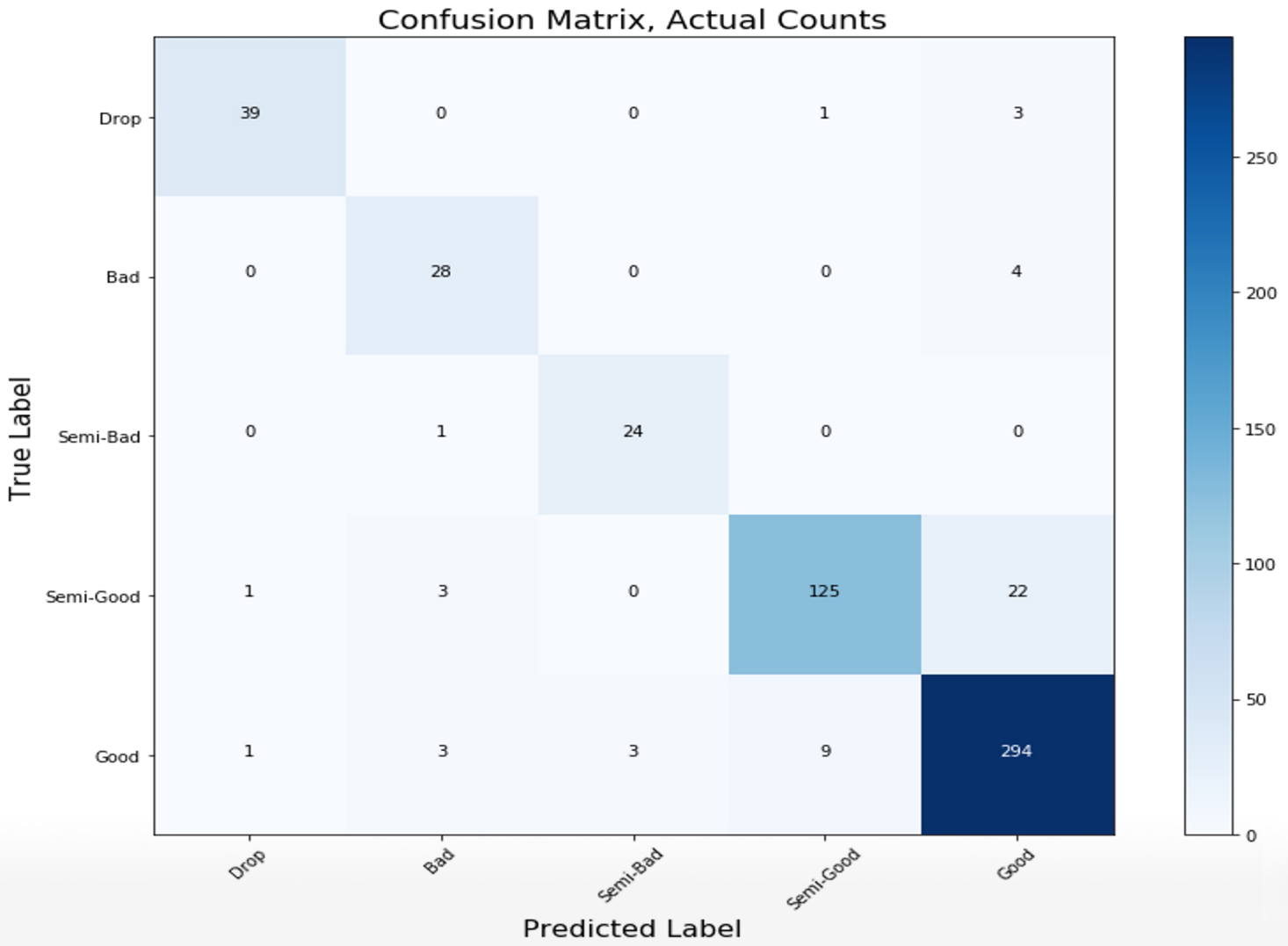
The minimum accuracy the classifier must have is 48.54% to be considered significant, based on PCC * 1.25.



Real World Dataset:

Random Forest (depth=14, n_estimators=100, criterion='gini')

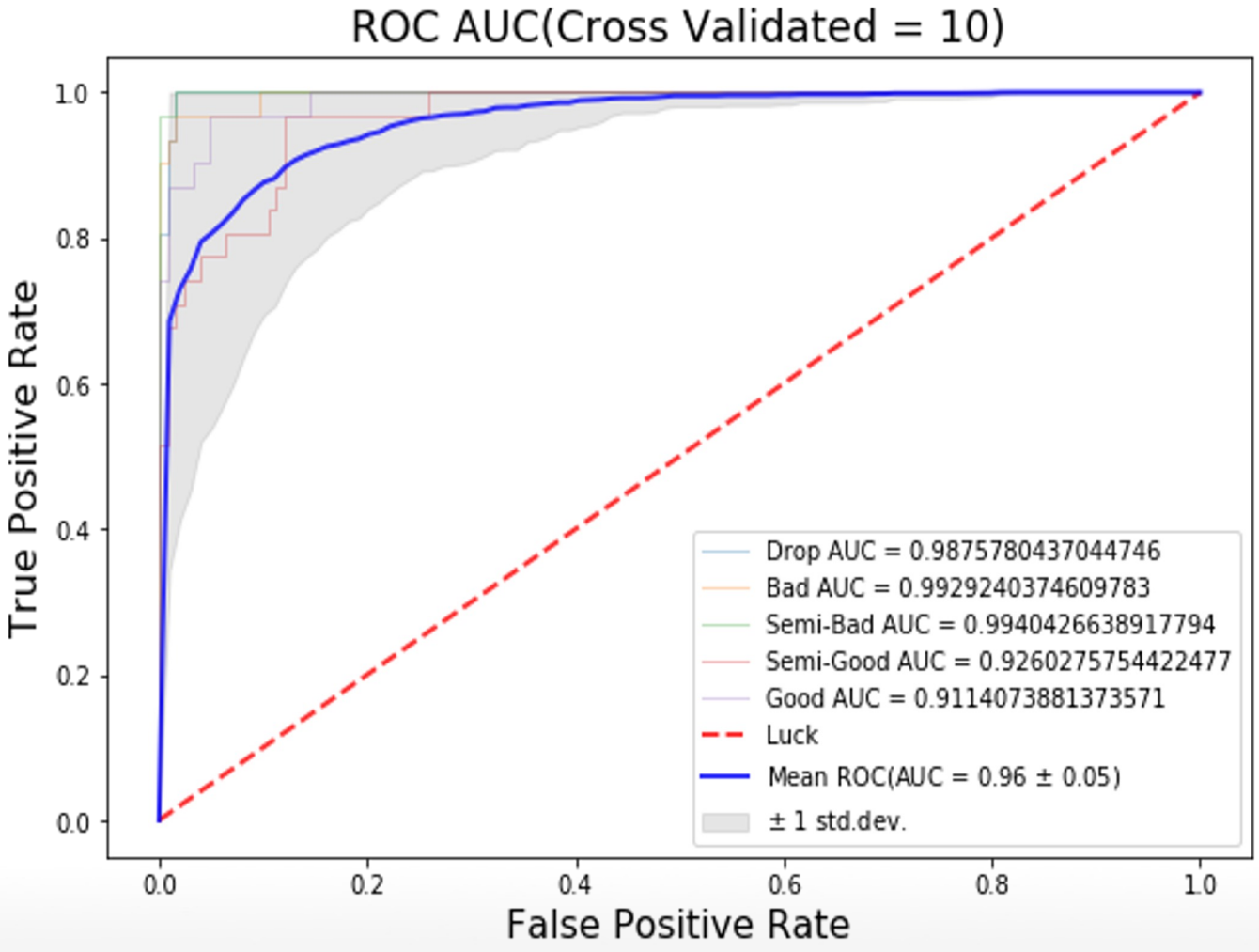
Model Accuracy: **90.9%**



Real World Dataset:

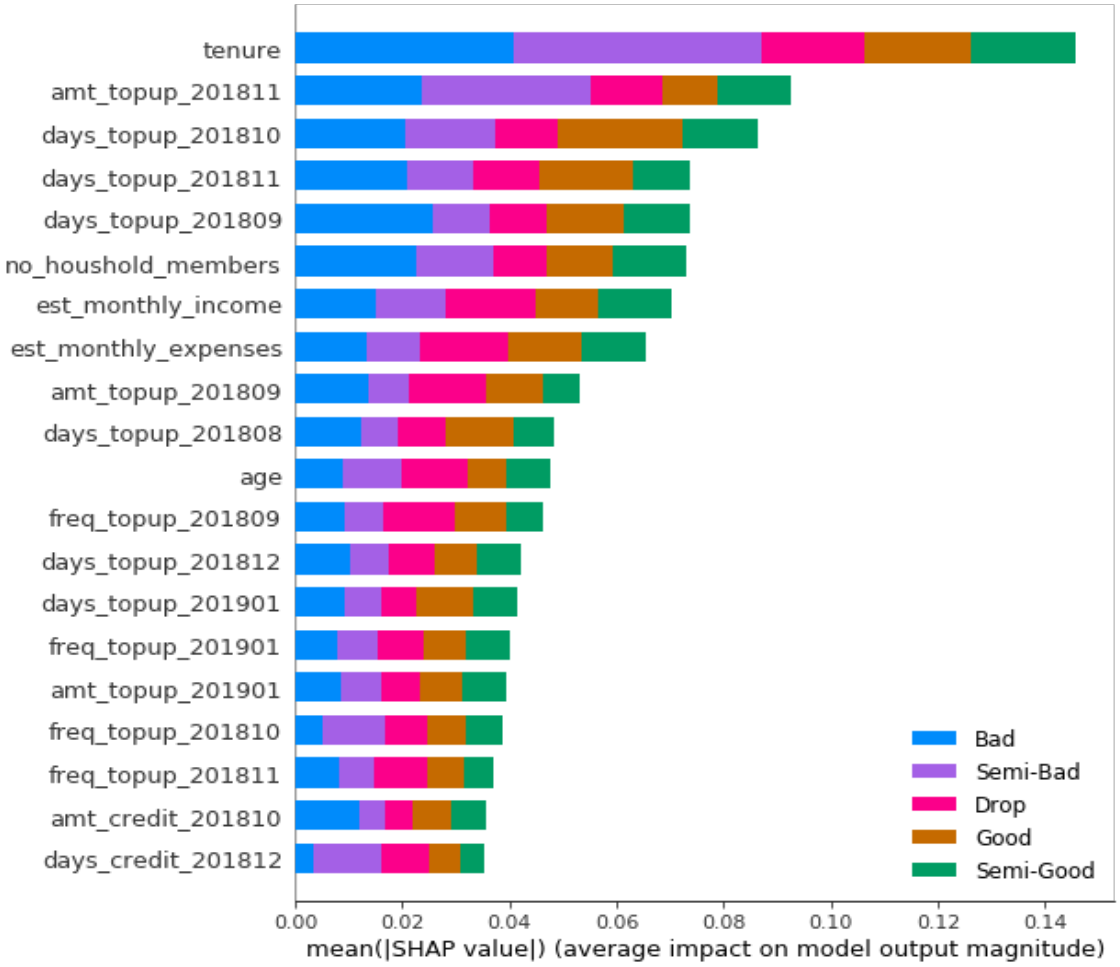
Random Forest (depth=14, n_estimators=100, criterion='gini')

ROC-AUC: **96%**



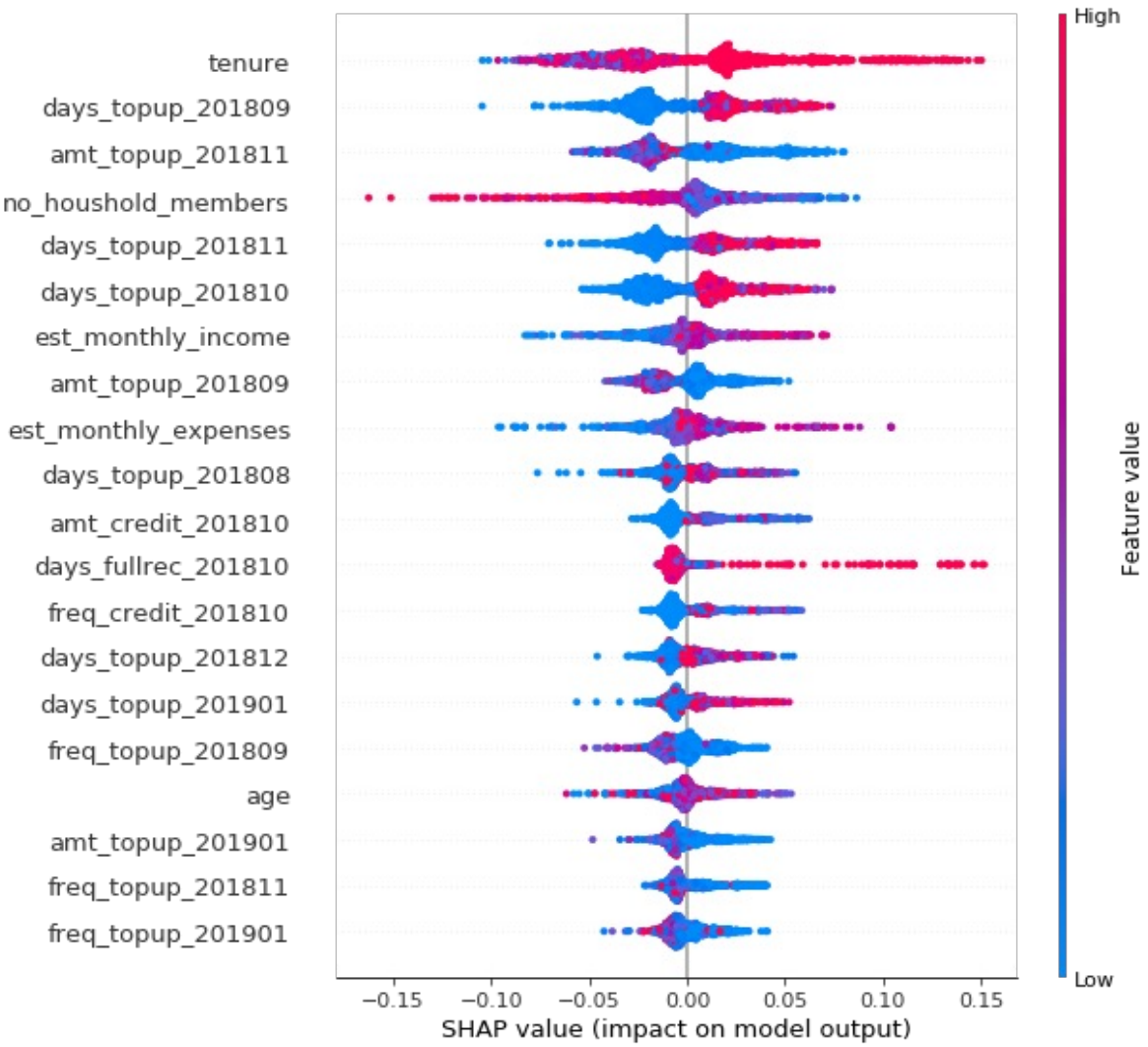
Feature Importance

- Overall, the most important feature is **tenure months.**
- Majority of the most important features are from **mobile behavior**.



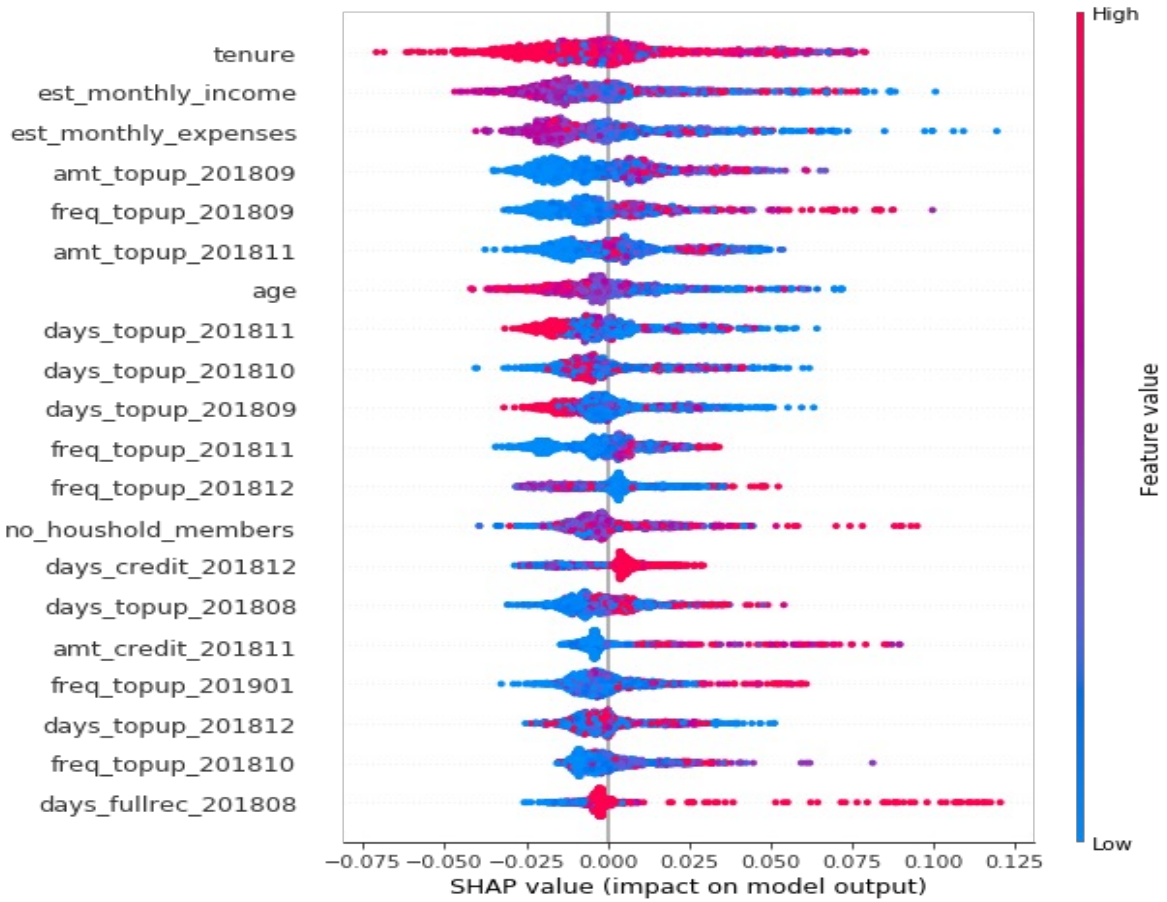
Insights | Bad Borrowers

- Higher **number of days between top-up** increases the risk of becoming bad borrowers.



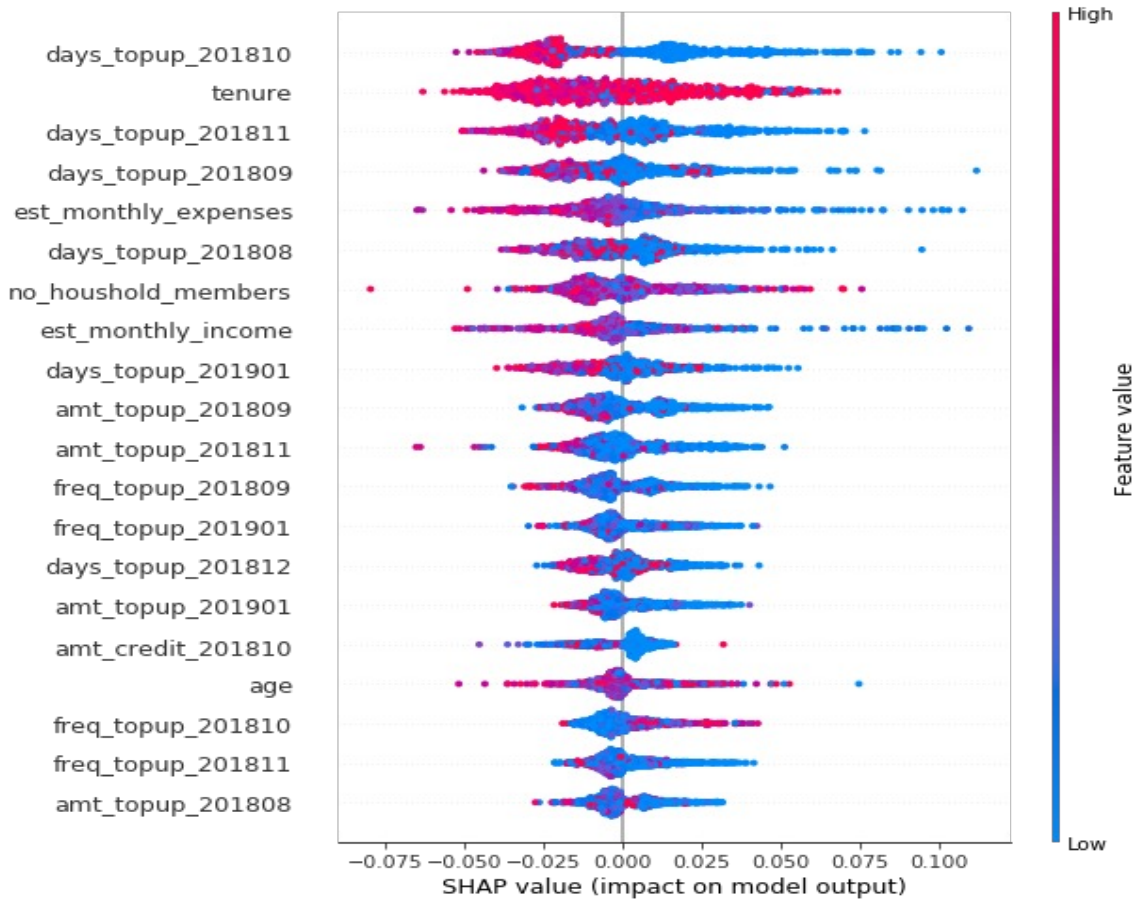
Insights | Dropped Borrowers

Higher **age** decreases the risk of becoming dropped borrowers.



Insights | Good Borrowers

- Lower **number of days between top-up** increases the risk of becoming good borrowers.



Thank You

QUESTIONS?