SAN DIEGO STATE UNIVERSITY

Assignment 1
Decision Tree & Knowledge Extraction

Submitted by

Harshal Sanap
822044776

MIS 691: Decision Support System
Jan 13, 2019

Part 1: Building a Decision Tree

Problem Statement:
Create a decision tree for Pau Daunk Bank to assist in the process of loan granting decision making. Assist Ginger and Saurabh in making right decision making with limited amount of information.

Data description:

Size of the data : 33 rows and 8 columns
The attributes are given as below:

First Name
Last Name
Age (Categorized in bins of 0 – 29, 30 – 55, 56 and older)
Loan Type
Ability to Pay
Past Payment Record
Loan Grant
Loan Amount (Categorized in bins of $5,000; $7,500; $10,000.)

1. What did I do?

- In order to build a decision tree, I first copied the data given in the exercise into excel. I created the bins for age using IF formula (=IF(AND(C5>0, C5<=29),"0-29",IF(AND(C5>=30, C5 <55),"30-55","55+"))) in excel and called this new column as age bucket.
- Now in order to build the decision tree, I studied the approach to make it. The basic algorithm for decision tree works as below.
  - Place the best attribute of the dataset as the root node.
  - Split the data set into subsets such that each subset contains data with the same value for an attribute
  - Repeat the above steps until we find a leaf node for each subset.
- Now from above pseudo code we would wonder how to find the root node and how to decide subsets. There are multiple algorithms for the same, in this exercise I have used the approach of information gain also known as ID3.
- ID3 is a top-down decision tree builder that starts from root node and works its way to leaf nodes. It uses the concept of Entropy and Information Gain for creation of the decision tree.
- **Entropy**: It is calculated as the degree of impurity present in a node or the level of uncertainty of a particular class selection in a node. It's value varies from 0 to 1. 0 meaning the node is completely pure or has a single class making it completely certain. The entropy of 1 states that there is 50-50 chance of selecting a particular class thus making it highly uncertain. The nodes with higher entropy generally require more information to describe them and vice versa.

$$\text{Entropy} = -p \log_2 p - q \log_2 q$$

  In the above formula, p and q signifies probability of success and failures respectively for a particular node. Lesser the Entropy the better it is.
- **Information Gain:** It is calculated as the difference between the entropy of the target node and the entropies of the children nodes. Higher the information gain the better it is and the node giving higher information gain is chosen for splitting.

$$\mathbf{Gain(T, X) = Entropy(T) - Entropy\ (T, X)}$$

  In above information gain formula, T is the target node and X is child node attribute.

- **Algorithm:**
  Step 1: Calculate the entropy of Target. In our example that would be Loan Grant.
  Step 2: Calculate entropy for all the attributes other than the target. As for the sub-nodes we have multiple classes, we need to calculate weighted entropy which shall be shown further in the example walkthrough.
  Step 3: Calculate the difference between step 1 and step 2 as information gain. And choose the node that provides highest difference called as the highest information gain as the splitting node.
  Step 4: After splitting, the node would either have entropy 0 or non-0. If it is 0 it means we can not split the node further and it's a leaf node. If it is non-zero, then we can split the node further.
  Step 5: We run the ID3 algorithm again on all non-leaf nodes until we reach leaf nodes for all attributes.

- Let's apply the algorithm to our data set to create the decision tree.

Age Bucket	Loan Type	Ability To Pay	Past Payment Record	Loan Grant	Loan Amount
30-55	luxury	good	slow	yes	$10,000
0-29	necessity	bad	good	no	0
30-55	necessity	good	poor	yes	$7,500
55+	frivolous	good	good	no	0
30-55	necessity	bad	slow	no	0
0-29	luxury	good	poor	no	0
30-55	frivolous	bad	good	yes	$5,000
55+	necessity	bad	good	no	0
30-55	luxury	good	poor	yes	$7,500
0-29	necessity	good	good	yes	$10,000
30-55	frivolous	good	slow	yes	$7,500
55+	luxury	good	slow	no	0
30-55	frivolous	good	poor	yes	$7,500
0-29	necessity	good	slow	yes	$5,000
55+	necessity	good	slow	yes	$5,000
55+	necessity	good	good	yes	$7,500
0-29	necessity	good	poor	no	0
0-29	luxury	good	good	yes	$5,000
0-29	frivolous	good	good	no	0
55+	necessity	good	poor	no	0
55+	necessity	good	slow	yes	$5,000
0-29	luxury	good	slow	no	0
30-55	frivolous	bad	poor	no	0
30-55	luxury	bad	slow	no	0
30-55	luxury	bad	good	yes	$5,000
0-29	necessity	bad	good	no	0
30-55	necessity	good	good	yes	$10,000
30-55	necessity	bad	poor	no	0
55+	luxury	good	good	yes	$5,000
55+	luxury	good	poor	no	0
30-55	necessity	good	good	yes	$10,000
30-55	necessity	good	poor	yes	$7,500
30-55	necessity	bad	good	yes	$5,000

Step 1 : Calculate the entropy for root node, loan grant.

	Yes	No	Total
Loan Grant	18	15	33

Entropy T = (-1 *((18/33)*LOG2(18/33) + (15/33)*LOG2(15/33)) = **0.994030211**

Step 2: Calculate entropies for sub nodes.

	Loan Grant								
Age group	Yes	No	Total						Information Gain
0-29	3	6	9	Entropy X1	0.918295834		Entropy(T,X1-3)	0.901029	0.093001176
30-55	11	4	15	Entropy X2	0.836640742				
55+	4	5	9	Entropy X3	0.99107606				

	Loan Grant								
Loan Type	Yes	No	Total						
frivolous	3	3	6	Entropy X4	1		Entropy(T,X4-6)	0.988367	0.005663457
luxury	5	5	10	Entropy X5	1				
necessity	10	7	17	Entropy X6	0.977417818				

	Loan Grant								
Past Payment Record	Yes	No	Total						
good	9	5	14	Entropy X7	0.940285959		Entropy(T,X7-9)	0.96343	0.030600093
poor	4	6	10	Entropy X8	0.970950594				
slow	5	4	9	Entropy X9	0.99107606				

	Loan Grant								
Ability To Pay	Yes	No	Total						
bad	3	7	10	Entropy X10	0.881290899		Entropy(T,X10-11)	0.916711	0.077318846
good	15	8	23	Entropy X11	0.932111568				

The above calculations are performed in excel. One of the calculation is explained below and can be followed for all of the nodes.

Entropy X1 = -1 * ((3/9)*LOG2(3/9) + (6/9)*LOG2(6/9)) = **0.918295834**

Entropy X2 = -1 * ((11/15)*LOG2(11/15) + (4/15)*LOG2(4/15)) = **0.836640742**

Entropy X3 = -1 * ((4/9)*LOG2(4/9) + (5/9)*LOG2(5/9)) = **0.99107606**

Entropy(T,X1-3) = (9/33) *0.918295834+(15/33)* 0.836640742+(9/33)* 0.99107606 = **0.901029**

Step 3:
Information Gain for attribute age group = Entropy(T) – Entropy(T,X1-3)
= 0.994030211 - 0.901029

= **0.093001176**

As highlighted in the above table, the gain for age group is the highest so we use that as the first split.

Step 4: We have now decided to split the given data on attribute age group. We can see from the above table that different age group are still having impurity meaning they have multiple classes or mixed loan grant

decisions thus we would need to split the node further until we find all the nodes with leaves or single class node.

Step 5: In this step we would split the data further as concluded in above step until we reach leaf nodes for all branches.

Here I would mention how the above algorithm would work in recursion by running it one more time manually. Generally, the above steps are run by sub setting the data further using the nodes found above until we find leaf nodes for all the branches. But for our dataset after the second iteration the further steps were possible by simply observing the data from excel as the volume of data became very low after second iteration. Thus, further breakdown is not shown here and would be seen in the decision trees directly.

So far, we now have divided the dataset based on the age group and now we need to analyze those age groups individually.

Lets see how the second iteration would work:

For Node of age group 0-29:

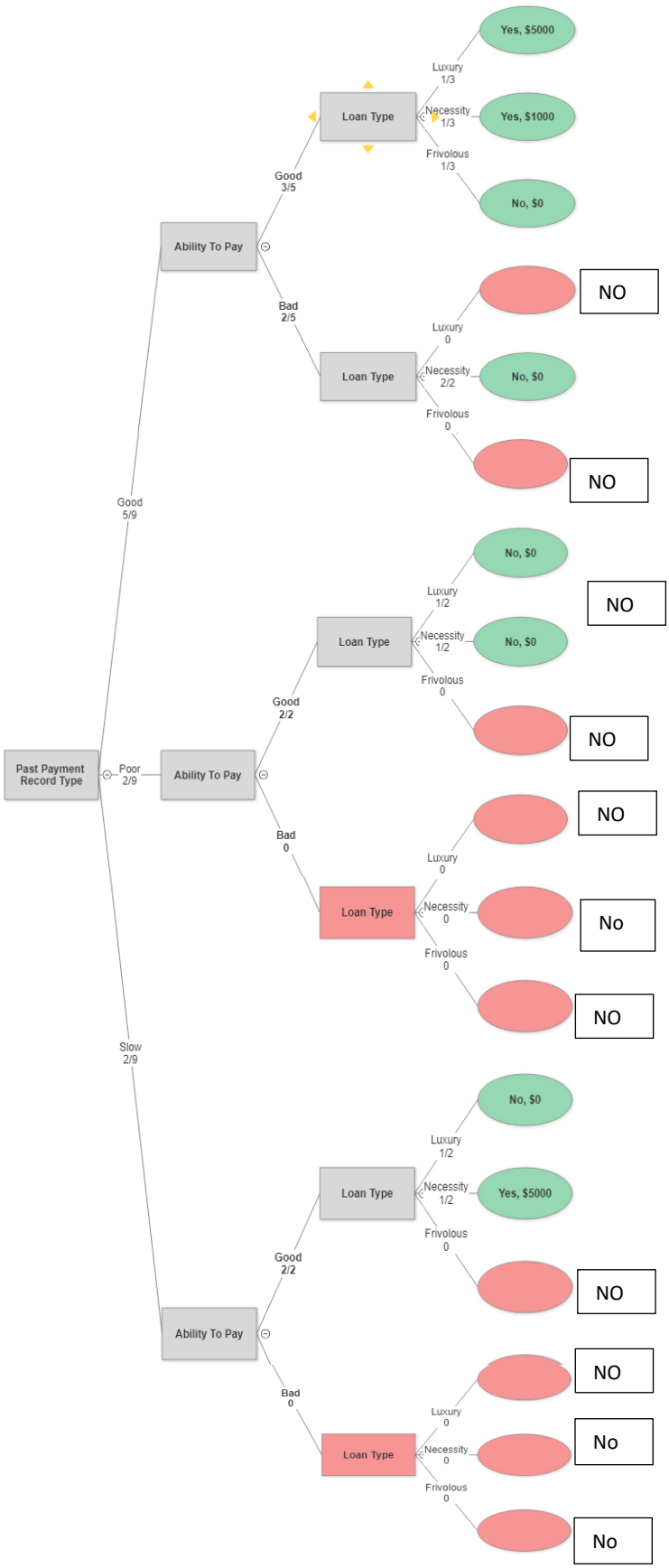
First Name	Last Name	Age	Age Bucket	Loan Type	Ability To	Past Paym	Loan Gran	Loan Amo
Pedro	Serrano	27	0-29	necessity	bad	good	no	0
Paul	Baker	24	0-29	luxury	good	poor	no	0
Karen	Pitman	20	0-29	necessity	good	good	yes	$10,000
Mei Li	Yang	28	0-29	necessity	good	slow	yes	$5,000
Ambika	Arun	29	0-29	necessity	good	poor	no	0
Olesia	Stumm	18	0-29	luxury	good	good	yes	$5,000
Sacha	Mantiol	25	0-29	frivolous	good	good	no	0
Chuang	Peng	28	0-29	luxury	good	slow	no	0
Komatsu	Takumi	25	0-29	necessity	bad	good	no	0

Dataset filtered for only age group 0-29

	Yes	No	Total							
Loan Grant	3	6	9	Entropy T	0.918296					
	Loan Grant									
Loan Type	Yes	No	Total							
frivolous	0	1	1	Entropy X1	0		Entropy(T,X1-3)	0.845516		Information Gain
luxury	1	2	3	Entropy X2	0.918296					0.072780226
necessity	2	3	5	Entropy X3	0.970951					
	Loan Grant									
Past Payment Record	Yes	No	Total							
good	2	3	5	Entropy X4	0.970951		Entropy(T,X4-6)	0.761639		
poor	0	2	2	Entropy X5	0					0.156656615
slow	1	1	2	Entropy X6	1					
	Loan Grant									
Ability To Pay	Yes	No	Total							
bad	0	2	2	Entropy X7	0		Entropy(T,X7-8)	0.766289		
good	3	4	7	Entropy X8	0.985228					0.152007284

Entropy calculation for the above group

From above table we can see that the split node for this dataset would be past payment record as it has highest information gain.



Decision Tree for Age group 0-29.

For calculating the value of loan grant at terminal node, mean response was given if there were multiple observations falling in that same region criteria.

Missing values are filled with specialized imputation. Where the value missing for a particular scenario is filled by using domain knowledge and looking at the mode of scenarios similar to it.

Example of specialized imputation:

For our example the loan grant for loan type frivolous is missing. In such case we would look at other scenarios with loan type as frivolous and as most of the time their loan grant was rejected we assume it will get rejected in our missing observation as well. This approach is followed though out this assignment.

For age group 30-55:

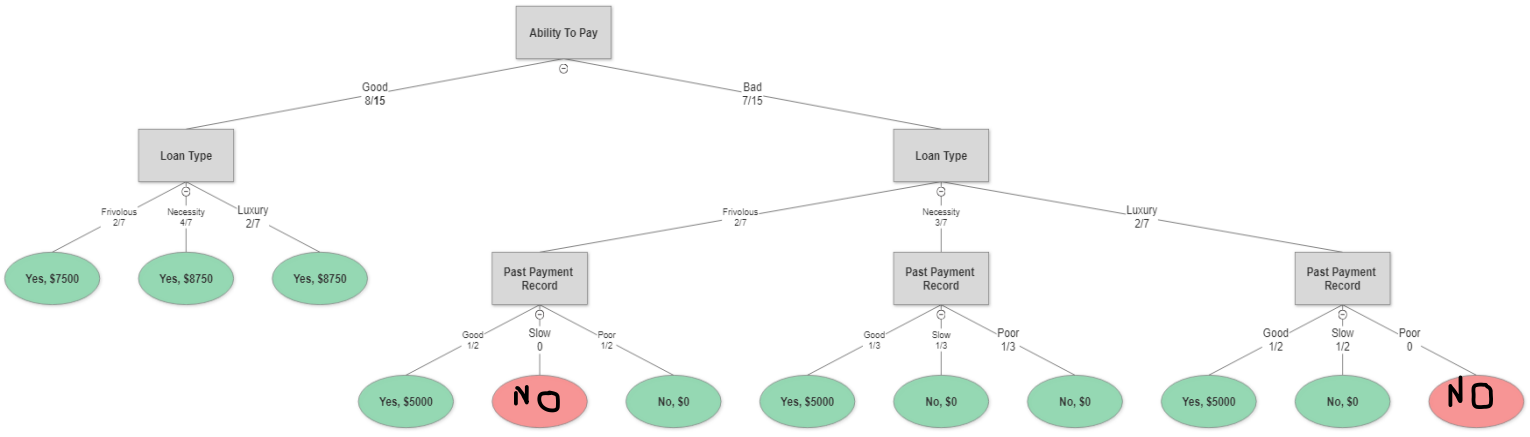
First Name	Last Name	Age	Age Bucke	Loan Type	Ability To	Past Paym	Loan Grant	Loan Amount
Angela	Vargas	35	30-55	luxury	good	slow	yes	$10,000
William	Tyler	46	30-55	necessity	good	poor	yes	$7,500
Sara	Cohen	52	30-55	necessity	bad	slow	no	0
Yolanda	Bassey	45	30-55	luxury	bad	slow	no	0
Joseph	Banks	38	30-55	luxury	good	poor	yes	$7,500
Ann	Swenson	49	30-55	frivolous	good	slow	yes	$7,500
Charlotte	Newman	50	30-55	frivolous	good	poor	yes	$7,500
Anapurna	Shan	34	30-55	frivolous	bad	poor	no	0
Ariel	Sarda	32	30-55	necessity	bad	poor	no	0
Hui	Chang	43	30-55	frivolous	bad	good	yes	$5,000
Bella	Berthog	32	30-55	necessity	good	good	yes	$10,000
Fujiwara	Ayumi	52	30-55	luxury	bad	good	yes	$5,000
Emma	Mane	46	30-55	necessity	good	good	yes	$10,000
Helen	Bacerini	39	30-55	necessity	good	poor	yes	$7,500
Heath	Norson	52	30-55	necessity	bad	good	yes	$5,000

Dataset filtered only for age group 30-55

	Yes	No	Total							
Loan Grant	11	4	15	Entropy T	0.836641					
	Loan Grant									
Loan Type	Yes	No	Total							Information Gain
frivolous	3	1	4	Entropy X1	0.811278		Entropy(T,X1-3)	0.835471		0.001169477
luxury	3	1	4	Entropy X2	0.811278					
necessity	5	2	7	Entropy X3	0.863121					
	Loan Grant									
Past Payment Record	Yes	No	Total							
good	5	0	5	Entropy X4	0		Entropy(T,X4-6)	0.633985		0.202655742
poor	4	2	6	Entropy X5	0.918296					
slow	2	2	4	Entropy X6	1					
	Loan Grant									
Ability To Pay	Yes	No	Total							
bad	3	4	7	Entropy X7	0.985228		Entropy(T,X7-8)	0.459773		0.376867612
good	8	0	8	Entropy X8	0					

Entropy Calculations for the above group

From above table we can conclude that the data set would be further split on ability to pay as it has the highest information gain.



Decision Tree for Age group 30-55.

For calculating the value of loan grant at terminal node, mean response was given if there were multiple observations falling in that same region criteria. The leaf nodes in red signify they are missing and they are filled as No using specialized imputation with mode operation.

For age group 55+:

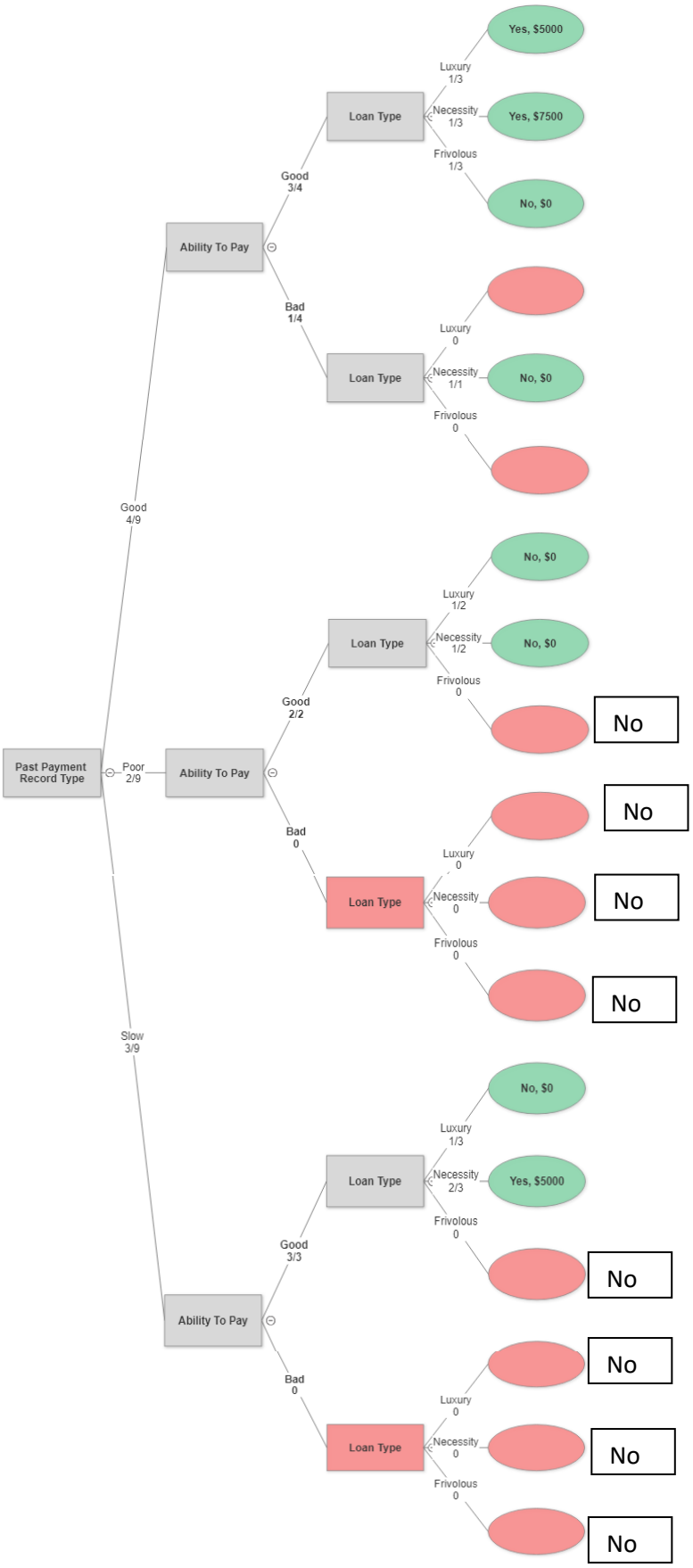
Age Bucket	Loan Type	Ability To	Past Paym	Loan Gran	Loan Amount
55+	frivolous	good	good	no	0
55+	necessity	bad	good	no	0
55+	luxury	good	slow	no	0
55+	necessity	good	slow	yes	$5,000
55+	necessity	good	good	yes	$7,500
55+	necessity	good	poor	no	0
55+	necessity	good	slow	yes	$5,000
55+	luxury	good	good	yes	$5,000
55+	luxury	good	poor	no	0

Dataset filtered by age group 55+

	Yes	No	Total							
Loan Grant	4	5	9	Entropy T	0.991076					
	Loan Grant									
Loan Type	Yes	No	Total							Information Gain
frivolous	0	1	1	Entropy X1	0		Entropy(T,X1-3)	0.845516		0.145560452
luxury	1	2	3	Entropy X2	0.918296					
necessity	3	2	5	Entropy X3	0.970951					
	Loan Grant									
Past Payment Record	Yes	No	Total							
good	2	2	4	Entropy X4	1		Entropy(T,X4-6)	0.750543		0.240533004
poor	0	2	2	Entropy X5	0					
slow	2	1	3	Entropy X6	0.918296					
	Loan Grant									
Ability To Pay	Yes	No	Total							
bad	0	1	1	Entropy X7	0		Entropy(T,X7-8)	0.888889		0.102187171
good	4	4	8	Entropy X8	1					

Entropy calculations for the above group

As shown above we can say that the above group would be split on Past payment record as it gives the highest information gain.



Decision Tree for Age group 55+

For calculating the value of loan grant at terminal node, mean response was given if there were multiple observations falling in that same region criteria.

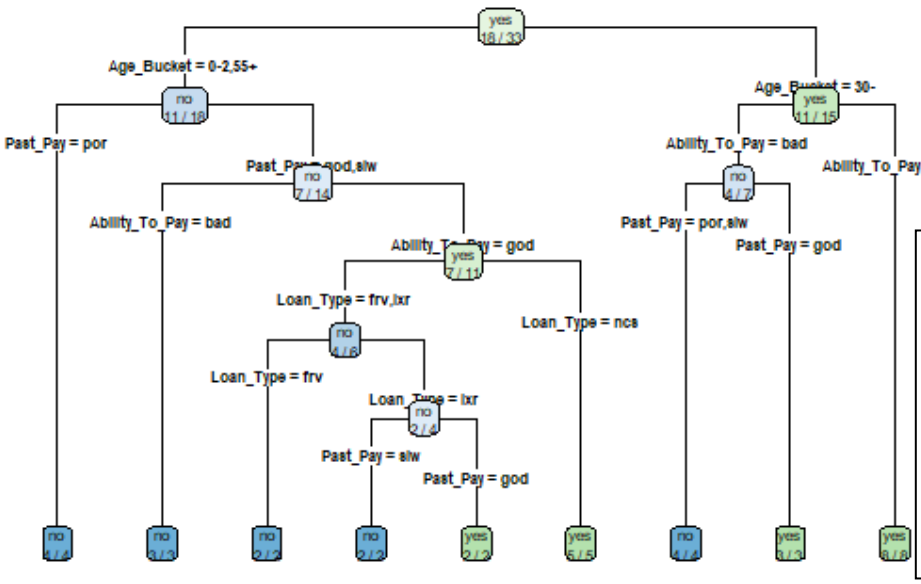
Missing values are filled with specialized imputations as mentioned in above trees.

2. What were the results?

- The final decision tree created as shown above is the outcome of the exercise. The above decision tree shall help in taking better decisions based on the given input in terms of age, ability to pay the loan, loan type and past payment record.
- To get a better idea of the outcome lets take an example. Suppose Saurabh and Ginger gets an applicant with details like (age: 57, past payment record history : Good, Ability to pay: Good and loan Type: Luxury). For this situation we can refer the decision tree with 55+ age group and decide that we can grant the loan up to $5000.
- From the above decision tree we can also summarize that the population in the age group of 30-55 is most likely to get the loan grant.

3. What did I learn?

- Decisions trees are a useful way of creating or establishing heuristics that help in stating a conclusion based on observed data set.
- Decision trees can get complicated based on the count of attributes as well as the data corresponding to them.
- Handling missing data can get complicated at times, a combination of rules and heuristics generally gives best possible result but a good knowledge of respective domain can also play an important role while dealing with missing values.
- The algorithm required for creating a decision tree is straight forward but due to recursive approach it can be time consuming if it is done manually. There are different algorithms used for creating a decision tree and the algorithm of information gain (used in this doc) and Gini index are the most common ones.
- After I built my decision tree manually in excel using the approach of information gain, I also tried plotting the same using statistical tools like R Studio and I found similar results which confirmed my manual decision trees to be correct. The output from R studio can be seen as below:



In the RStudio output, I didn't perform any optimization or didn't work towards making the output more appealing as it was not the purpose of this assignment and I used it only to verify my manual efforts. (This output is only for reference and not meant for grading purposes)

4. How does it relate to class?

- The slides 2-7 from decision support system presentation from night 1 and slides 2-6 from presentation deciding to do a DSS of night 2 helped me understand the basic idea behind how a decision support system is created in general and how it can get complex at times. It also made me realize about the possibility of anticipating missing values as one of the problems faced while creating a decision tree.
- The slides 2-5 from the presentation of Decision Modeling on night 2 helped me understand the basics of decision tree modeling and how the decision trees are designed in general.
- The real life example which was done in class with student 'Kate' where professor tried to predict her nationality based on her inputs related to the kind of snacks she prefers made the idea of how decision tree was working in background got clear.

Part 2

Identify specific rules and heuristics for the following two examples:

Example 1: This is part of a multiple criteria with thresholds, weighting, and scoring. Note that it is not

the only criteria in the model, just the criteria you are addressing

Reliability of service – minimum 98% availability with B/U and recovery (weight 20%)

- 5 - >99.997% availability
- 3 - >99% availability
- 2 - >98% availability

Explanation:

Lets first understand what reliability and availability means.

Reliability of service: It is defined as the probability that a product, service or a system will perform its intended functions adequately for a specified time period while operating in a defined operating environment without failure.

Availability: It represents the probability that a system, product or service is capable of conducting its required function when its called upon, given it is not failed or undergoing a repair action.

Therefore, we can say that availability is a function of reliability.

Now for a year long service, measures for service availability can be estimated as below.

score	availability	Downtime/Year	Calculation
2	98.00%	7.3 Days	365*0.02
3	99.00%	3.65 Days	365*0.01
5	99.997%	15.76 min	365*0.003

Now keeping the above information in mind we can say the set of rules and heuristics for the given company could be as follows.

For the above example we can set the rules and heuristics as below:

Rules:

- Availability of service should be at least 98%
- Backup and recovery option must be present at all times

- Rating of 2 means we can allow only 2% as the total downtime
- Rating of 3 means we can allow only 1% as the total downtime
- Rating of 5 means we can allow only 0.003% as the total downtime in service.

Heuristics:

- Addition of backup and recovery would increase the availability of service
- As reliability of service has 20% weightage, it is an important aspect of decision making.
- Customer is expecting a minimum of 98% availability and lower number for the same can result in loss of customer and revenue for the company and there could be a possibility of getting lower availability rating than 2.

1. What did I do?

- Depending upon the given availability percentages I calculated the downtime values on yearly basis.

availability	Downtime/Year	Calculation
98.00%	7.3 Days	365*0.02
99.00%	3.65 Days	365*0.01
99.997%	15.76 min	365*0.003

- Then I tried to understand the meaning of rating score 2-5:
  Based on the discussion in class and research related to six sigma topics I realized the percentages mentioned here are different levels of sigma. The basic statistical concept of normal distribution is applied here for number of downtimes. And for this normal distribution, different levels of sigma correspond to its deviation away from the mean downtime value and thus the total % probability of that downtime happening. Six sigma being the highest level of perfection which covers almost 99.99966% of service availability, leaving less than 1% for downtimes.

Availability	**Sigma Level**	**Yearly downtime**
99.99966 %	6	1.47 min
99.997 %	5	15.76 min
99 %	3	3.65 days
98 %	2	7.3 days

- Regarding the 20% weightage to reliability, I believe that the reliability is of more importance than other 6 criteria's of Cost of Service vs capacity, Technical Support, Security, Web Development Support, Web Data Collection. If reliability was having equal importance as others then it would have had a weightage of 14.28%

2. What were the results?

- The given company is looking for minimum 98% of downtime and they also have availability of backup and recovery meaning they are trying to aim for almost no downtime in their service.
- As there could be downtime values below 98% but our company needs at least 98% thus we are having a score of 2. The value of 1 would probably be lower than 98% which is not acceptable in our example.
- With backup and recovery however the company can push the availability of service to more than 98% and thus the scores go up correspondingly.

3. What did you learn?

- From the above example I realized that the company is not that big because if we consider a big organization like amazon which operates throughout the year, if they have a 2% downtime then that can cost them millions of dollars in revenue.

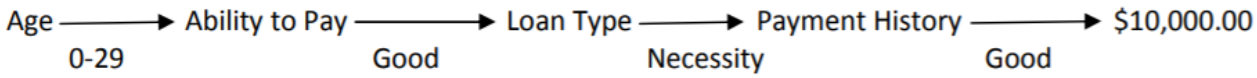
- From the above example I also learnt different aspects of six sigma and why it can be of significance amongst organizations.

4. How does it relate to the material that was covered in the class?

- I learnt about different evaluation criteria from Decision Modeling presentation slide 30 discussed on night 2 and I learnt about a similar problem from the same presentation from step 2 through step 4
- Slides from 37-39 from presentation 'sharda_dss10_ppt_04_jennexmodified' helped me understand the concept of six sigma.

Example 2: This is a branch of a decision tree. Note that this is not the only branch of the decision tree,

just the branch you are addressing



Explanation:

Before talking about the outcome of the above node, lets just explore the meaning of different aspects of the node.

- The ability to pay refers to the capability of an individual to pay off the debt obligations within required time period. Having good ability to pay the loan would mean that person has enough financial capability to repay their loans and thus there is a high chance of their loan getting approved.
- Loan type describes the specific type for which the loan is being sanctioned. It describes purposes like luxury frivolous or necessity towards which the loan is being sanctioned.
- Payment history refers to considering someone's past repayment habit. Good, slow and poor are used as indicators for payment history. Good would state that the person has made payments in time, slow means the person took longer than the provided repayment time period and poor means person has not been regular in making payments.

Rules:

- Loan of $10,000 can be granted to an individual who is in age group of 0-29, with a good ability to pay and with loan type as necessity and having good payment history track.

Heuristics:

- An Individual must be between 0-29 years of age.
- People having good ability to pay and good payment history are eligible for loan. Bank wouldn't take risk on individuals otherwise.
- Loan is generally granted to individuals with loan type as necessity and generally people with need of luxurious loan type having bad ability to pay and bad past payment history would not get their loan approved.

References read:

https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

https://cloud.smartdraw.com

https://en.wikipedia.org/wiki/Decision_tree

http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

https://asq.org/quality-resources/reliability

http://www.serviceperformance.com/the-5-service-dimensions-all-customers-care-about/