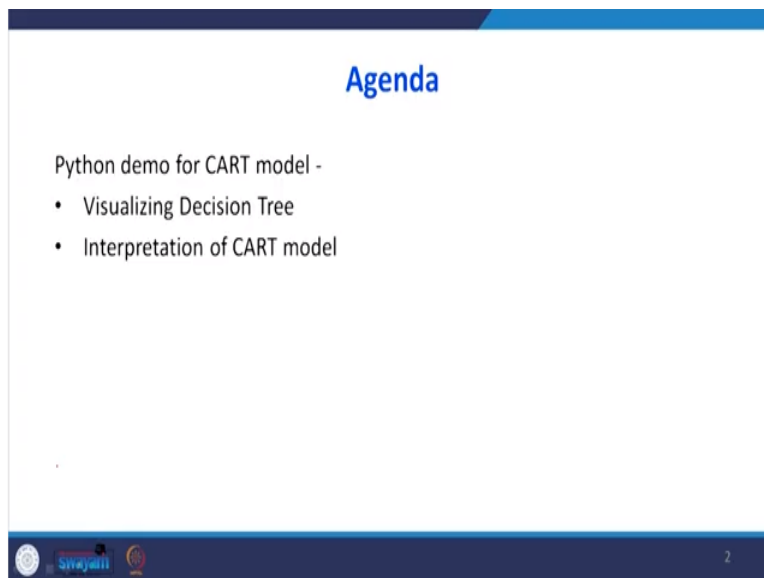


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 60
Classification and Regression Trees (CART) - III

In my previous lecture I have explained how to choose an attribute for the decision tree model. We know that there are three methods one is information gain method; second one is gain ratio method; third one is the Gini index. In previous lecture I have explained by using gain ratio and Gini index how to choose an attributes and I have explained all the procedures.

(Refer Slide Time: 00:49)



In this lecture, we are going to use python with help of python. We are going to construct the CART model then I am going to explain the decision tree then I am going to interpret the output of CART model.

(Refer Slide Time: 01:07)

Example

Problem Description-

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

This was the sample data this is an example say there are variable age, income, student, current rating is the attributes for independent variable the dependent variable is buys_computer. This example is taken from this source Han, Pei and Kamber the title of the book is Data Mining concepts and techniques.

(Refer Slide Time: 01:30)

Import Relevant Libraries and Loading Data File

```

In [1]: 1 import pandas as pd
        2 import numpy as np
        3 import matplotlib.pyplot as plt

In [2]: 1 data = pd.read_excel('CART.xlsx')

In [3]: 1 data

```

Out[3]:

	RID	age	income	student	credit_rating	buys_computer
0	1	youth	high	no	fair	no
1	2	youth	high	no	excellent	no
2	3	middle_aged	high	no	fair	yes
3	4	senior	medium	no	fair	yes
4	5	senior	low	yes	fair	yes
5	6	senior	low	yes	excellent	no
6	7	middle_aged	low	yes	excellent	yes
7	8	youth	medium	no	fair	no
8	9	youth	low	yes	fair	yes
9	10	senior	medium	yes	fair	yes
10	11	youth	medium	yes	excellent	yes
11	12	middle_aged	medium	no	excellent	yes
12	13	middle_aged	high	yes	fair	yes
13	14	senior	medium	no	excellent	no

I have brought this screenshot of first we will import relevant libraries and loading the data file. This file I have copied into the excel so I am importing pandas as pd import numpy as np, import matplotlib.pyplot as plt. I have imported the data then stored any object called data.

(Refer Slide Time: 01:50)

Methods used in Data Encoding

- **LabelEncoder ()**: This method is used to normalize labels. It can also be used to transform non-numerical labels to numerical labels.
- **Fit_transform ()**: This method is used for Fitting label encoder and return encoded labels.

Then there are different methods for encoding the data. The first one is LabelEncoder this method is used to normalize labels it can also be used to transform non-numerical labels into numerical labels. In our examples, the data are non-numerical that we are going to convert into numerical labels. The another function is fit under score transform this method is used for fitting label encoder and return encoded labels.

(Refer Slide Time: 02:23)

Data Encoding Procedure

```
In [4]: 1 import sklearn
        2 from sklearn.preprocessing import LabelEncoder

In [5]: 1 le_age = LabelEncoder()
        2 le_income = LabelEncoder()
        3 le_student = LabelEncoder()
        4 le_credit_rating = LabelEncoder()
        5 le_buys_computer = LabelEncoder()

In [6]: 1 data['age_n'] = le_age.fit_transform(data['age'])
        2 data['income_n'] = le_income.fit_transform(data['income'])
        3 data['student_n'] = le_student.fit_transform(data['student'])
        4 data['credit_rating_n'] = le_credit_rating.fit_transform(data['credit_rating'])
        5 data['buys_computer_n'] = le_credit_rating.fit_transform(data['buys_computer'])
```

So this is a data in coding percentages we import sklearn from sklearn dot preprocessing import LabelEncoder. So le_age LabelEncoder le_income like that for all the variables. Now there are different attributes like age but age, income, student credit rating buys computer this is in the text form. So that I am going to convert into numerical form, but new variable is age under n by

using this function le underscore age dot fit underscore transformation. So that wherever there is a fit_transformation, that is a text data is going to convert into numerical form.

(Refer Slide Time: 03:10)

Data Encoding

```
In [7]: 1 data.head()
```

Out[7]:

	RID	age	income	student	credit_rating	buys_computer	age_n	income_n	student_n	credit_rating_n	buys_computer_n
0	1	youth	high	no	fair	no	2	0	0	1	0
1	2	youth	high	no	excellent	no	2	0	0	0	0
2	3	middle_aged	high	no	fair	yes	0	0	0	1	1
3	4	senior	medium	no	fair	yes	1	2	0	1	1
4	5	senior	low	yes	fair	yes	1	1	1	1	1

Now what is happening is this portion is the text by using transformations I have converted into the numerical form.

(Refer Slide Time: 03:20)

Structuring Dataframe

drop(): This is used to Remove rows or columns by specifying label names and corresponding axis or by specifying directly index or column names.

```
In [8]: 1 data_new = data.drop(['age', 'income', 'student', 'credit_rating', 'buys_computer'], axis='columns')
2 data_new.head()
```

Out[8]:

	RID	age_n	income_n	student_n	credit_rating_n	buys_computer_n
0	1	2	0	0	1	0
1	2	2	0	0	0	0
2	3	0	0	0	1	1
3	4	1	2	0	1	1
4	5	1	1	1	1	1

Then structuring the data frame by using the function drop. This is used to remove rows or columns by specifying label names in corresponding axis or by specifying directly index or column names. So what we are going to do in the previous layers, I told you there is a text data also is there , numerical data also there. Since already we are transforming into numerical that

text portions that I am going to drop it by using this drop function. So after that, this was the my dataset in this there is no text only numerical values. So this data set is going to be taken for building the CART model.

(Refer Slide Time: 03:58)

Independent and Dependent Variables Selection

```
In [9]: 1 feature_cols = ['age_n', 'income_n', 'student_n', 'credit_rating_n']
        2 x = data_new.drop(['buys_computer_n', 'RID'], axis='columns') #input
        3 y = data_new['buys_computer_n'] #target
```

```
In [10]: 1 x.head()
```

Out[10]:

	age_n	income_n	student_n	credit_rating_n
0	2	0	0	1
1	2	0	0	0
2	0	0	0	1
3	1	2	0	1
4	1	1	1	1

```
In [11]: 1 y.head()
```

Out[11]:

0	0
1	0
2	1
3	1
4	1

Name: buys_computer_n, dtype: int32

In the building of the CART model we go to specify what is the independent and dependent variables we know that dependent variable is buys underscore computer. The independent variable is age, income, student credit rating you see that I am using age underscore n that is in the numerical form. In the dependent variables only two options, Yes or No that is buys underscore computer underscore n that has only two levels one is 0 or 1 that is Yes or No.

(Refer Slide Time: 04:31)

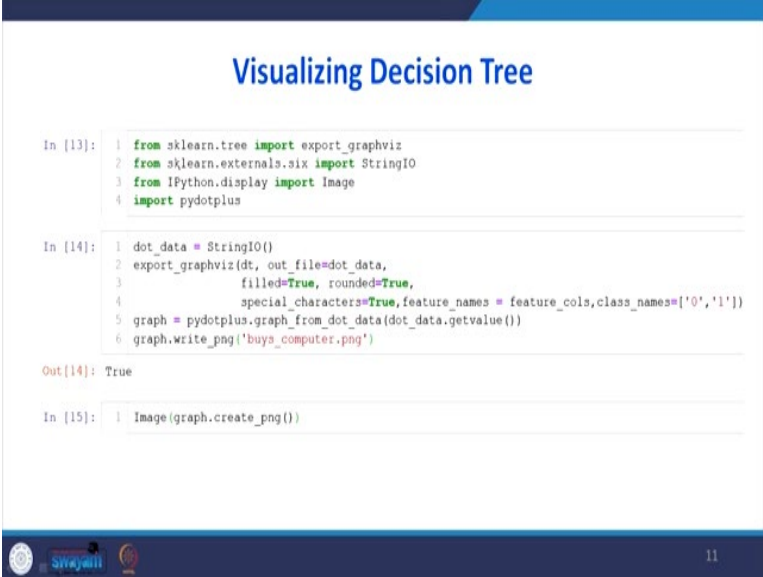
Build the Decision Tree Model without Splitting

```
In [12]: 1 from sklearn.tree import DecisionTreeClassifier
        2 clf = DecisionTreeClassifier()
        3 dt = clf.fit(x,y)
        4 dt
```

Out[12]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best')

Now we are going to build the decision tree model without splitting what is the meaning of without splitting is in data mining whenever the huge amount of data is there, some data sets should be used for training the model the remaining data set should be used for testing the model. Now we are not going to do that way we are going to take all the dataset for build the model we are not going to test the model from sklearn.tree import DecisionTreeClassifier clf = DecisionTreeClassifier dt = clf.fit(xy) this was the dt this was the output.

(Refer Slide Time: 05:10)



```
In [13]: 1 from sklearn.tree import export_graphviz
2 from sklearn.externals.six import StringIO
3 from IPython.display import Image
4 import pydotplus

In [14]: 1 dot_data = StringIO()
2 export_graphviz(dt, out_file=dot_data,
3               filled=True, rounded=True,
4               special_characters=True, feature_names = feature_cols, class_names=['0', '1'])
5 graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
6 graph.write_png('buys_computer.png')

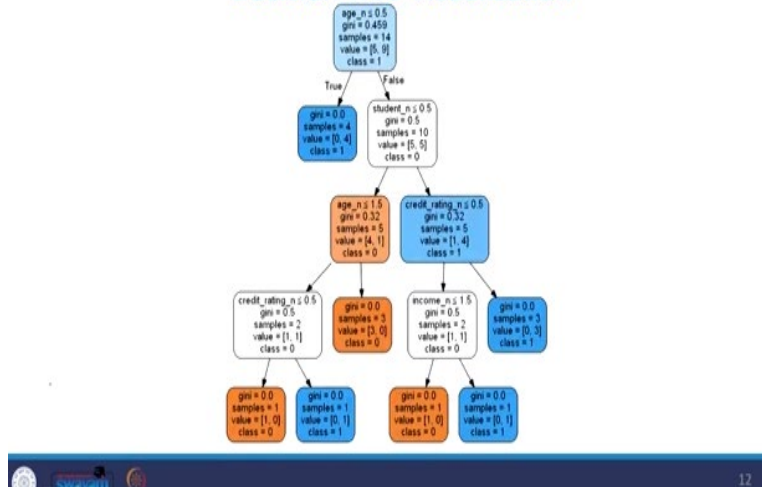
Out[14]: True

In [15]: 1 Image(graph.create_png())
```

Now we are going to visualize the decision tree from sklearn dot tree import export graphviz. From sklearn dot externals dot six import String IO. From IPython dot display import Image then import pydotplus. So dot underscore data = StringIO export underscore graphviz this was the commands for getting the graphical output of our CART model. Then I have specified what is the dependent variable.

(Refer Slide Time: 05:48)

Decision Tree Visualization



This is the output of our CART model that is the Decision Tree this stage, let us understand how to interpret this. When age underscore $n \leq 0.5$ there are two possibility it is true and false. I will explain the meaning of this 0.5 in the next slide. So whenever is a true, whenever this blue box represents it is a favorable decision for us. For example, orange colored box represents unfavorable.

What is unfavorable? That person will not buy the computer that is a class 0. If it is a class 1 means that person will buys the computer. Suppose if you want to interpret this blue box how to interpret is first look for the age, age we have seen, there are different levels in age. If it is true that person surely will buy the computer. If it is a false, then look for the student and the student also there is a two possibility Yes or No, this is true, this is false.

When you go for a student then this condition is failed it will go for a false then it will look for another attributes credit rating then it is true this is false. The credit rating when the false condition appears then this is favorable decision. If it is true, then look for another attribute income in that income if the false is applicable then it is favorable decision for us. So I will explain each values in the box and the meaning of the different color coding in coming slides.

(Refer Slide Time: 07:31)

Interpretation of the CART Output

Let us interpret the CART output.

(Refer Slide Time: 07:35)

Calculation of Gini(D)

- We first use the following Equation for Gini index to compute the impurity of D:

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2$$

$$= Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459.$$

RID	age	income	student	credit_rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

This was the data which we have taken we first used the following equation for the Gini index to compute the impurity of D this also I have explained. We know that formula $Gini D = 1 - \sum_{i=1}^m p_i^2$, m is number of levels in our dependent variable here the m = 2 because Yes or No so Gini index = 1 minus the 9 represents how many years? 1, 2, 3, 4, 5, 6, 7, 8, 9; 9 yes out of 14 - this 5 represents number of Nos so 5 divided by 14 the whole square this is 0.459. This is Gini index for our dependent variable, so D represents the dependent variable.

(Refer Slide Time: 08:26)

Income Attribute

- Low, Medium, High
- Option 1: {Low, Medium}, {High}
- Option 2 : {High, Medium}, {low}
- Option 3 : {High, Low}, {Medium}

RID	age	income	student	credit_rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Now let us take one attribute at this stage suppose I have taken income, income is over attribute we are going to use Gini index for this attribute we know Gini index always go for binary classification. There are three levels low, medium, high option 1 is we can group into two way because it is a binary classification low medium is one group high is another group. Next high medium is one group low is another group the last choice is high, low is one group medium is another group for all these three combinations, let us find out what is the Gini index.

(Refer Slide Time: 09:11)

Tuples in partition D1

- Low + Medium:

Low + Medium	Class: buys computer
Yes	3+4 = 7
No	1+2 = 3

RID	age	income	student	credit_rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

First, we will take low and medium in the low and medium we have to look at how many people answered Yes 1 No not this one 2, 3 that is how we got this 3 when it is medium, how many people answered Yes for our dependent variable, so this is 1 this is 2 this is 3 this is 4 so 3 + 4.

The same way when it is low, how many people answered No this case 1 only one options is there low and No.

Then the second case when it is medium, how many people answered No medium and No this is 1 then this, so 7 and 3 that is how we got this table this is one group part D1 actually what is happening here? so the income variable we are going to go for binary classification one is D1 another one is D2. D1 we are going to consider the two levels low and medium. In the another group we are going to consider only high so that is why D1 is low and medium so the D2 is only high.

(Refer Slide Time: 10:51)

Tuples in partition D2

• High :

High	Class: buys computer
Yes	2
No	2

RID	age	income	student	credit rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle age	medium	no	excellent	yes
13	middle age	high	yes	fair	yes
14	senior	medium	no	excellent	no

So D2 it is high how many people have answered when it is high, how many people answered Yes, this is 1 high Yes 2. So when it is a high, how many people answered No this 1 high No 2 yeah this is 2.

(Refer Slide Time: 11:13)

Gini index for income attribute

- The Gini index value computed based on this partitioning is

$Gini_{income \in \{high, medium\}}$

$$\begin{aligned}
 &= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2) \\
 &= \frac{10}{14} (1 - \frac{6}{10}^2 - \frac{4}{10}^2) + \\
 &\quad \frac{4}{14} (1 - \frac{3}{4}^2 - \frac{1}{4}^2) \\
 &= 0.45 = Gini_{income \in \{low\}}
 \end{aligned}$$

RID	age	income	student	credit rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	no	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

D_1 D_2

Now we are going to do the Gini index for income attributes the Gini index value computed based on this partitioning is, so Gini income belongs to low and medium so 10 to 12 / 14 how we got this 10 when you count low and medium, there will be 10 count it 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 divided by 14 Gini D1 the formula is $1 - (7/10)^2 - (3/10)^2$ whole square. The second option for D2 4 / 14.

How we got 4? When it is a high? We have to count it how many high is there in the income. 1, 2, 3, 4 so 4/14 Gini D2, D2 is 1 minus because number of Yes and number of Nos are same $1 - (2/4)^2 - (2/4)^2$ whole square. So this is the Gini index for low and medium that is 0.443 that is equivalent to Gini index for other level high. The Gini index value for the another option what is the another option.

So there are two options, option one D1 D2 this is D1 this is D2, D1 high and medium is there D2 low is there. So when you having this two splits the Gini index for income attributes, which belongs to that is high and medium we are getting the 0.45.

(Refer Slide Time: 13:02)

Gini index for income attribute

- The Gini index value computed based on this partitioning is

$$\begin{aligned} \text{Gini}_{\text{income} \in \{\text{high, low}\}} &= (8/14) (1 - (5/8)^2 - (3/8)^2) + \\ &\quad (6/14) (1 - (2/6)^2 - (4/6)^2) \\ &= 0.458 = \text{Gini}_{\text{income} \in \{\text{medium}\}} \end{aligned}$$

id	age	income	student	credit rating	class buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	no	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

h, L m

The third option is high and low is one group then medium is another group. The same way when you continue, we are getting the Gini index is 0.4.

(Refer Slide Time: 13:18)

Gini index for income attribute

- $\text{Gini}_{\text{income} \in \{\text{low, medium}\}} = 0.443 = \text{Gini}_{\text{income} \in \{\text{high}\}}$
- $\text{Gini}_{\text{income} \in \{\text{high, medium}\}} = 0.45 = \text{Gini}_{\text{income} \in \{\text{low}\}}$
- $\text{Gini}_{\text{income} \in \{\text{high, low}\}} = 0.458 = \text{Gini}_{\text{income} \in \{\text{medium}\}}$

Now by comparing all the combinations, the lowest Gini index value is this one 0.443 that is where when income is high.

(Refer Slide Time: 13:31)

Gini index for Age attribute

- The Gini index value computed based on this partitioning is

$$\text{Gini}_{\text{Age} \in \{\text{Youth, middle_aged}\}}$$

$$= 0.457 = \text{Gini}_{\text{Age} \in \{\text{senior}\}}$$

$$\text{Gini}_{\text{Age} \in \{\text{Youth, Senior}\}}$$

$$= 0.357 = \text{Gini}_{\text{Age} \in \{\text{middle_aged}\}}$$

$$\text{Gini}_{\text{Age} \in \{\text{senior, middle_aged}\}}$$

$$= 0.393 = \text{Gini}_{\text{Age} \in \{\text{Youth}\}}$$

AID	age	income	student	credit rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Now the Gini index for age attribute, so Gini index for senior it is 0.457 for middle age it is 0.357 for youth it is 0.393. The lowest Gini value is this one that is 0.357.

(Refer Slide Time: 13:51)

Gini index for student attribute

- The Gini index value computed based on this partitioning is

$$\text{Gini}_{\text{student} \in \{\text{Yes, No}\}}$$

$$= \frac{7}{14} (1 - (\frac{6}{7})^2 - (\frac{1}{7})^2) +$$

$$\frac{7}{14} (1 - (\frac{3}{7})^2 - (\frac{4}{7})^2)$$

$$= 0.367$$

AID	age	income	student	credit rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Similarly, we will take another attribute that attribute is student, so for student attribute also when you find the Gini index it is 0.367.

(Refer Slide Time: 14:01)

Gini index for credit_rating attribute

- The Gini index value computed based on this partitioning is

$$\begin{aligned} \text{Gini}_{\text{credit_rating}} E(\text{fair, Excellent}) \\ &= \frac{8}{14} (1 - (\frac{6}{8})^2 - (\frac{2}{8})^2) + \\ &\quad \frac{6}{14} (1 - (\frac{3}{6})^2 - (\frac{3}{6})^2) \\ &= 0.428 \end{aligned}$$

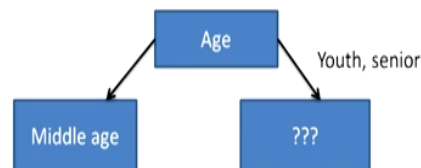
RID	age	income	student	credit_rating	Class buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

The next attribute is the credit_rating. For this, the Gini index is 0.428.

(Refer Slide Time: 14:09)

Choosing the root node

The attribute with minimum Gini score will be taken, i.e. Age (Gini_{Age} E(Youth, Senior) = 0.357 = Gini_{Age} E(middle_aged))



Attribute	Gini score
Age	0.357
Income	0.443
Student	0.367
Credit_rating	0.428

Now when you bring all the generic index for different attributes, the age one is having the least Gini score that is 0.357 this attribute should be chosen for the classification that is the preference should be given for age. So the attribute with the minimum Gini score will be taken that is the age because that Gini index is 0.357. When you go for age that are 3 levels did middle age, youth and senior when you look at middle age, they have answered Yes for buying all people have answered Yes for buying computers with that we can stop it. Now there are youth and senior then we have to continue which attribute has to be chosen for here.

(Refer Slide Time: 14:54)

Gini index for different attributes for sample of 10

- After separating 4 samples belonging middle age, total 10 are remaining:

RID	age	income	student	credit_rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

So since the all the middle aged people have answered Yes see that middle age is Yes, Yes, Yes. Now the next calculations we are going to drop these rows. After dropping these rows again here we are going to find out the Gini index, so after separating this 4 samples belonging to middle-age total 10 rows are remaining out of 14.

(Refer Slide Time: 15:22)

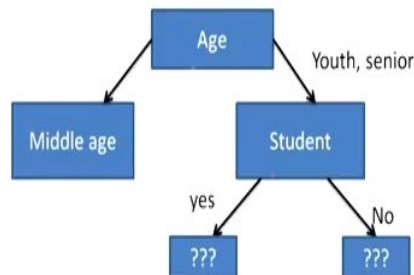
Gini index for different attributes for sample of 10

- Gini (D) = $(1 - (5/10)^2 - (5/10)^2) = 0.5$
- Gini_{Age} = 0.48
- Gini_{Credit Rating} = 0.41
- Gini_{Student} = 0.32
- Gini_{income} = 0.375
- Take student as node as it have mini. Gini Score

So for that 10 rows we are finding Gini index for our dependent variable 0.5 Gini index for age is 0.48 Gini index for our credit rating is 0.41 for a student it is 0.32 for income it is 0.375. Again, we have to look at which attribute is having the lowest Gini value. Take the student as node as it is having minimum Gini score.

(Refer Slide Time: 15:53)

Drawing cart



So what we have to do after first priority is age next, we have to take student as a classifier. In the student there was a 2 level was there Yes and No. If the student says Yes, then which attribute has to be chosen? If the student is No then which attribute has to be chosen, we will see that.

(Refer Slide Time: 16:10)

For branch Student = No

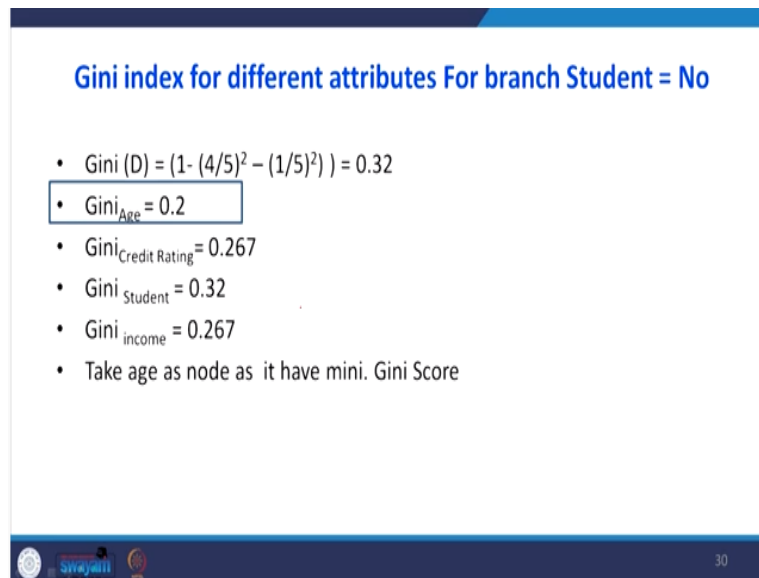
- Omit the marked rows (Data entry), either belonging Age = middle_aged or student = Yes
- Total 5 rows are remaining

RID	age	income	student	credit_rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Now what we have to do omit the marked rows either belonging age equal to middle-aged or student equal to Yes. So wherever we are talking about here in which variable which attribute has to be chosen. So what we are going to do, if it is a middle-age the dataset if the middle age is there that has to be dropped, the students are answering the student level is Yes that also is going to be dropped. So middle-aged is dropped wherever the student is Yes that also dropped. So how

many of you are going to drop 1, 2, 3, 4, 5, 6, 7, 8, 9 so out of 14, 9 rows we are going to drop it. So the remaining rows for further iteration is 5 rows.

(Refer Slide Time: 17:03)



Gini index for different attributes For branch Student = No

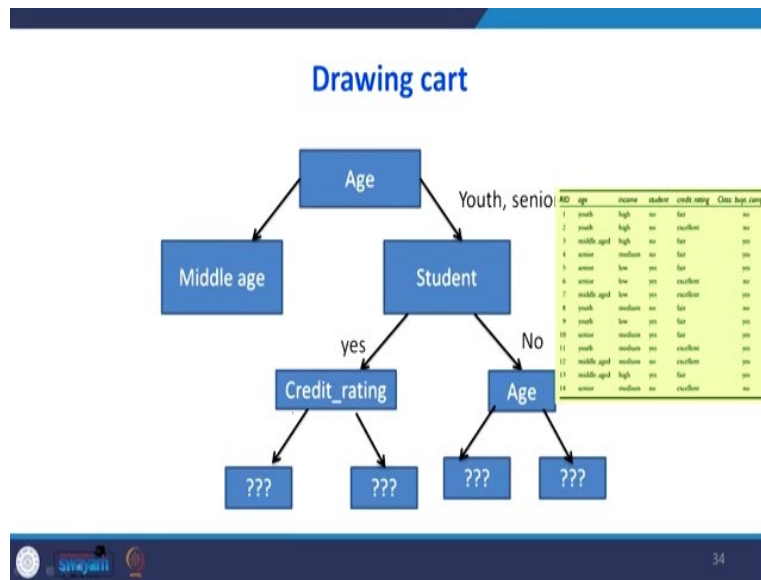
- $Gini(D) = (1 - (4/5)^2 - (1/5)^2) = 0.32$
- $Gini_{Age} = 0.2$
- $Gini_{Credit\ Rating} = 0.267$
- $Gini_{Student} = 0.32$
- $Gini_{income} = 0.267$
- Take age as node as it have mini. Gini Score

30

For that 5 rows again, we are going to find out the Gini index for the dependent variable 0.32 for the age 0.2 for credit rating, 0.267 for a student it is 0.32 for income it is 0.267. Again we have to look at which attribute is having the least Gini score. So the age is having the least Gini score. When the student level is No, we have identified what is next attribute. Similarly, when the student level is Yes, we have to find out what does the next year attribute for further classification for that omit the marked rows either belonging age equal to middle aged or student equal to No.

So when you do that one the remaining 5 rows are remaining using these 5 rows again, we are going to find out the Gini index for our dependent variable 0.32 for age 0.267 for credit rating it is 0.2 for student it is 0.32 for income it is 0.267. Again, you have to look at which attribute is having minimum Gini index though the credit rating is having the minimum Gini index.

(Refer Slide Time: 18:20)



What is to be done so the credit rating attribute has to be brought here for further classification.
Like this, we have to continue to satisfy all the conditions
(Refer Slide Time: 18:32)

Coding scheme

Age	Code
Youth	2
Middle Age	0
senior	1

Student	Code
Yes	1
No	0

Credit rating	Code
Fair	1
Excellent	0

Income	Code
High	0
Low	1
Medium	2

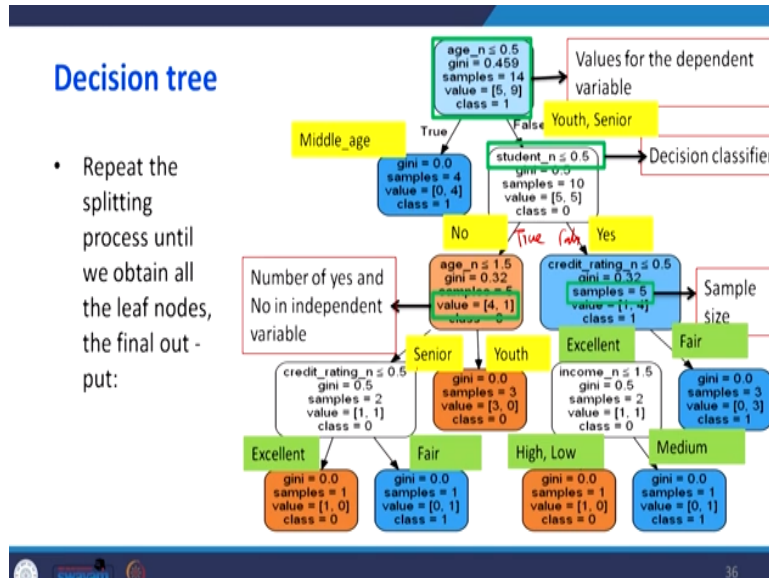
Buys computer	Class
Yes	1
No	0

Now I am going to explain the coding scheme because this coding scheme is important if you are going to interpret the python output for our CART model. There was a one attribute called age, student ,credit rating, income, Buys computer. So there are 4 independent variable, one dependent variable youth is coded as 2, middle-age is 0, senior is 1. So because this coding is important for interpreting the output.

(Refer Slide Time: 19:00)

Decision tree

- Repeat the splitting process until we obtain all the leaf nodes, the final output:



This was our python output Now if we look at the first variable is age underscore n when it is less than 0.5 this Gini index we got it 0.459 when you go back and see that this one, whatever the value which are appearing in the python output we have manually solved it so that you will feel more confidence when the age underscore n is the less than 0.5. Now you have to see what is the meaning of this 0.5 because we look at this coding the age is coded, youth is 2, middle-age is 0, senior is 1.

So if $n \leq 0.5$ that represents the middle age that is why we got this one when this condition is true this is middle age when the condition is false, they will belongs to youth and senior what is the meaning of this 5,9 the 5 represents No 9 represents Yes. Since in the middle aged group all are see that 0, 4 all are answered Yes. So the first represents No second represents Yes all are answered Yes so, we are not going for further classification.

Then, we are taking student underscore n as a another attribute for further classification when it is less than 0.5 we have to see what is less than 0.5 the student we have coded Yes = 1, No = 0, when I say $n < 0.5$ that represents No when the student is No that is why when the condition is true it is No this is true, this is false this is one branch when it is Yes there is another branch. Then next we choose age underscore n ≤ 1.5 .

Because now in the age only two group is there one is youth and senior when you go here youth and senior is there they when $n \leq 1.5$ that represents the senior when the condition is true that represents senior when the condition is false that represents youth we will go to the next one when the credit rating underscore $n < 0.5$. There are two options when the condition is true we got excellent. How are we got this excellent? Go to credit rating coding fair represents 1 excellent represents 0.

So when the condition is less than 0.5 it is excellent when the condition is true it is excellent if the condition is false it is fair. Then you go for income underscore $n \leq 1.5$ when the condition is true high and low how it is let us go for this coding income high = 0, low = 1, medium = 2 if it is 1.5 if $n < 1.5$ that represents low and medium when the income underscore $n < 1.5$ what is the meaning you look at this table.

So when it is less than 1.5 this group, this is less than 1.5 high and low, so high and low is one group medium is another group. So as a manager, how to interpret this first, the classifier is the age if it is true then go for middle age, the middle aged people will have the positive response for buying the computer that is why 1. If this condition fails then go for the student if they answer Yes then look for another attribute credit rating if it is fair, there is favorable response if it is excellent, we should go for further classification.

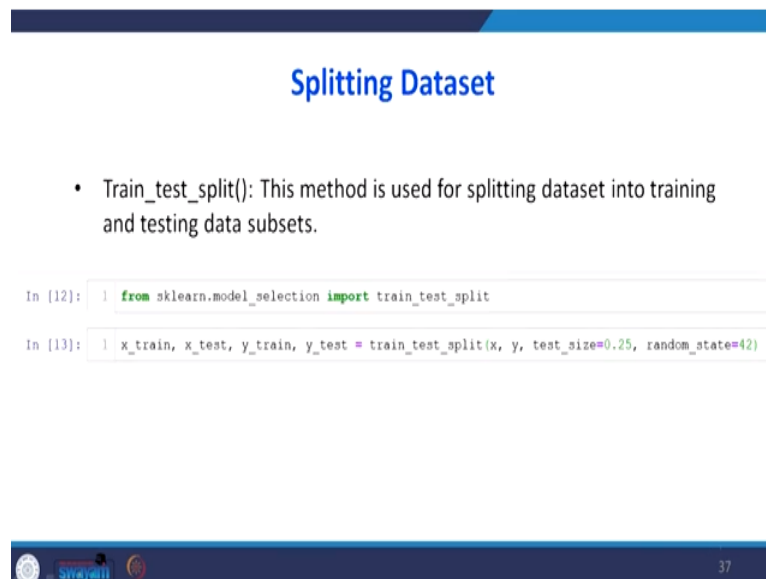
When it is excellent the next classifier is income when the income if it is true, they belongs to high and low then they will not buy the computer when it is medium then they will buy the computer. When we look at the left hand side, when the student $n \leq 5$ it is No then we will go for next attribute age because the age already we have dropped middle age the remaining is youth and senior if it is true it is senior, if it is false it is youth.

So if it is senior then you have to look for credit rating if it is youth there is a not favorable response. The 4 represents number of Nos 1 represents number of Yes in our dependent variable. So and another thing is look at the wherever the class is 1 there is only number of Yes is there see here that blue one 0, 4 only Yes is there 1, 4 only Yes is there 0,3 only Yes is there 0, 1 Yes is there here also 0, 1 Yes is there.

So the blue boxes are which will give you the favorite decision for us the orange boxes and look at this, see that the here there is no see 1, 0 only No is there here also 1, 0 only No is there here also 3, 0 only No is there. So the orange box represents it is not going to give a favorable decision. The white box represents that is intermediate in the sense we have to go for choosing some more attributes for further conclusion.

So what does 14 represents values for the dependent variable there are 14 then the sample represents the sample size and so on. So repeat the splitting process until we obtain all leaf nodes and the final output. The leaf node represents this one this is leaf nodes.

(Refer Slide Time: 25:07)



Splitting Dataset

- `train_test_split()`: This method is used for splitting dataset into training and testing data subsets.

```
In [12]: 1 from sklearn.model_selection import train_test_split
```

```
In [13]: 1 x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state=42)
```

37

Now we are done the data without splitting, but in the data mining generally we used to split the dataset. So we split the dataset so testing data set 25% data is going to be used for testing the remaining data set is for the training.

(Refer Slide Time: 25:22)

Evaluating the Model

```
In [16]: 1 from sklearn import metrics

In [17]: 1 y_pred = clf.predict(x_test)

In [18]: 1 print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

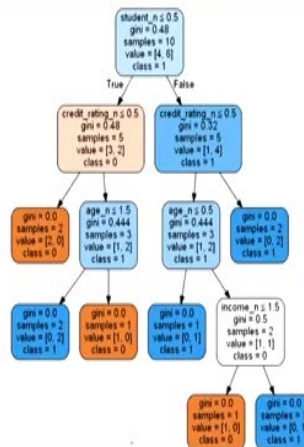
Accuracy: 0.75

True: [1 1 0 1]
pred: [1 0 0 1]
```

So you run the python code, what I told previously. So here we are going to get the accuracy is 0.75 what is the meaning of this 0.75 our classifier, this decision tree model able to classify whether they are going to buy the computer or not with the 75% of accuracy. Then visualize the decision tree.

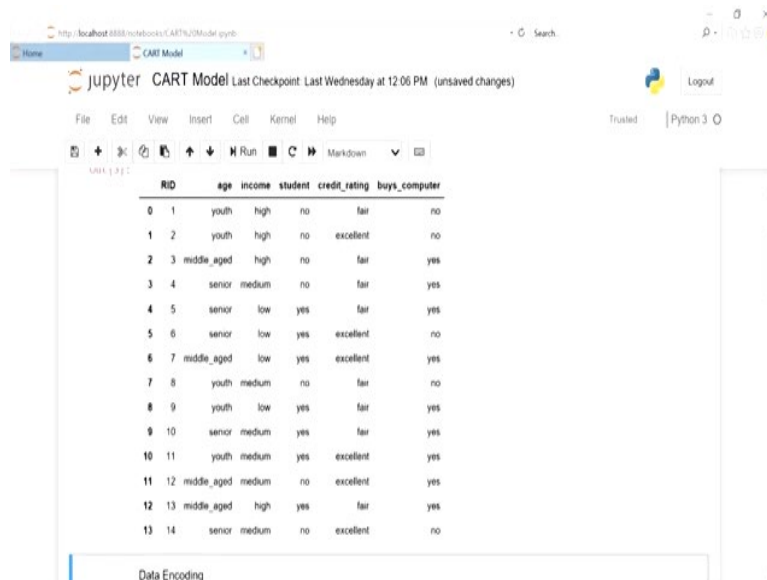
(Refer Slide Time: 25:46)

Decision Tree Visualization



So what is happening? you see when you are splitting the data set now the node variable is changed. The node is student previously it was the age, so what is happening since our data set is very only 14 dataset, we are getting this different result.

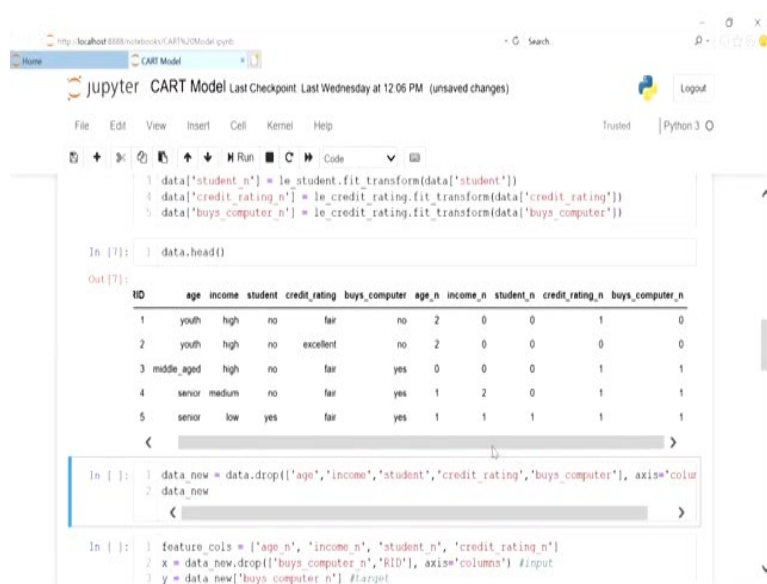
(Refer Slide Time: 26:03)



	RID	age	income	student	credit_rating	buys_computer
0	1	youth	high	no	fair	no
1	2	youth	high	no	excellent	no
2	3	middle_aged	high	no	fair	yes
3	4	senior	medium	no	fair	yes
4	5	senior	low	yes	fair	yes
5	6	senior	low	yes	excellent	no
6	7	middle_aged	low	yes	excellent	yes
7	8	youth	medium	no	fair	no
8	9	youth	low	yes	fair	yes
9	10	senior	medium	yes	fair	yes
10	11	youth	medium	yes	excellent	yes
11	12	middle_aged	medium	no	excellent	yes
12	13	middle_aged	high	yes	fair	yes
13	14	senior	medium	no	excellent	no

Now I am going to explain the python code for running the problem which I have explained. First import the pandas as pd than other libraries like numpy, matplotlib I have input. Then I am going to import the data. The data you know that that are age income, student ,credit rating, buys underscore computer. So this is in the text form, but I have to encode this data because they want it the numerical form I am encoding.

(Refer Slide Time: 26:46)



```

1 data['student_n'] = le_student.fit_transform(data['student'])
2 data['credit_rating_n'] = le_credit_rating.fit_transform(data['credit_rating'])
3 data['buys_computer_n'] = le_credit_rating.fit_transform(data['buys_computer'])

In [7]: data.head()

Out[7]:
RID    age income student credit_rating buys_computer age_n income_n student_n credit_rating_n buys_computer_n
1    youth  high     no      fair         no         2      0      0      1      0
2    youth  high     no  excellent         no         2      0      0      0      0
3  middle_aged  high     no      fair         yes         0      0      0      1      1
4    senior  medium     no      fair         yes         1      2      0      1      1
5    senior   low     yes      fair         yes         1      1      1      1      1

In [ ]: data_new = data.drop(['age', 'income', 'student', 'credit_rating', 'buys_computer'], axis='column')
1 data_new

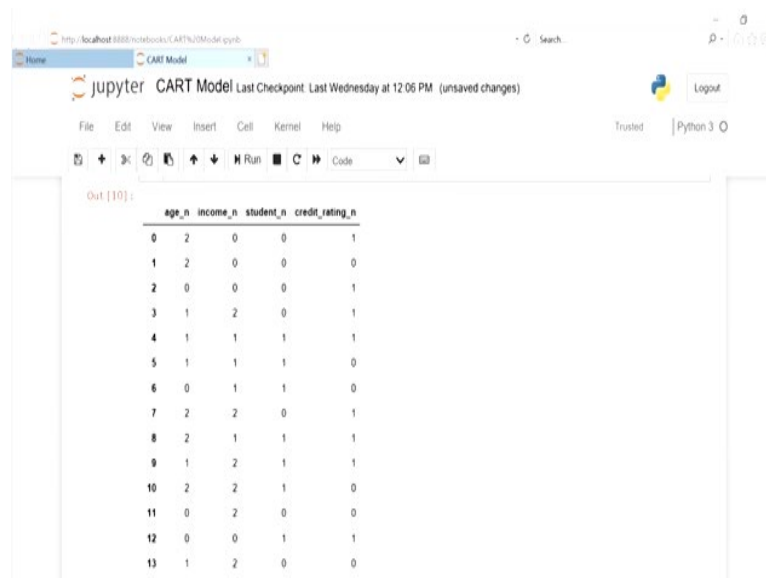
In [ ]: feature_cols = ['age_n', 'income_n', 'student_n', 'credit_rating_n']
2 x = data_new.drop(['RID'], axis='columns') #input
3 y = data_new['buys_computer_n'] #target

```

Now this shows that after encoding the right hand side we are able to see the equivalent numerical values. Now we are going to drop this text values because we are going to use only the numerical values for that building the CART model. So this was only the numerical values

because there is a 14 dataset. Now we are going to declare what our independent variable, what are the dependent variable.,

(Refer Slide Time: 27:14)

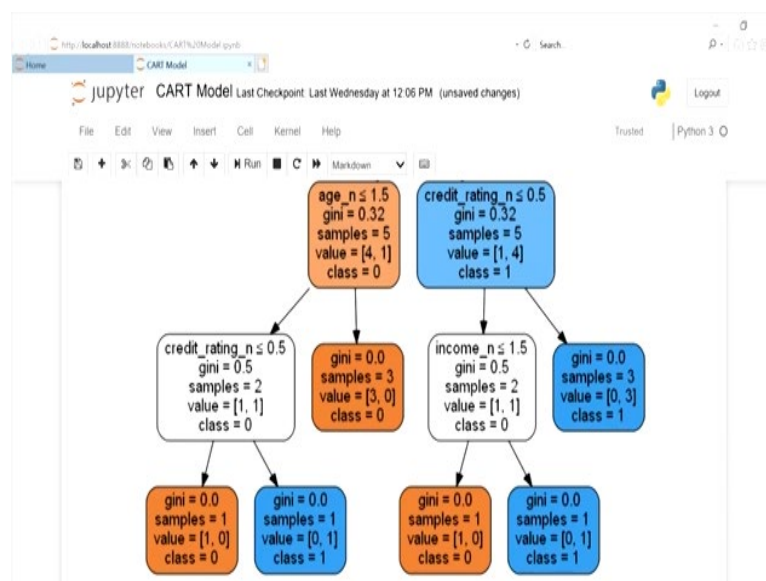


Out [10]:

	age_n	income_n	student_n	credit_rating_n	
0	2	0	0	0	1
1	2	0	0	0	0
2	0	0	0	0	1
3	1	2	0	0	1
4	1	1	1	1	1
5	1	1	1	1	0
6	0	1	1	1	0
7	2	2	0	0	1
8	2	1	1	1	1
9	1	2	1	1	1
10	2	2	1	0	0
11	0	2	0	0	0
12	0	0	1	1	1
13	1	2	0	0	0

So the x these are the independent variable what are they? independent variable is there age, income, student, credit rating. Let us see what is the dependent variable? Dependent variable is buys underscore computer that has two levels, 0 and 1 then I am building the decision tree model for getting this output you need to install these packages.

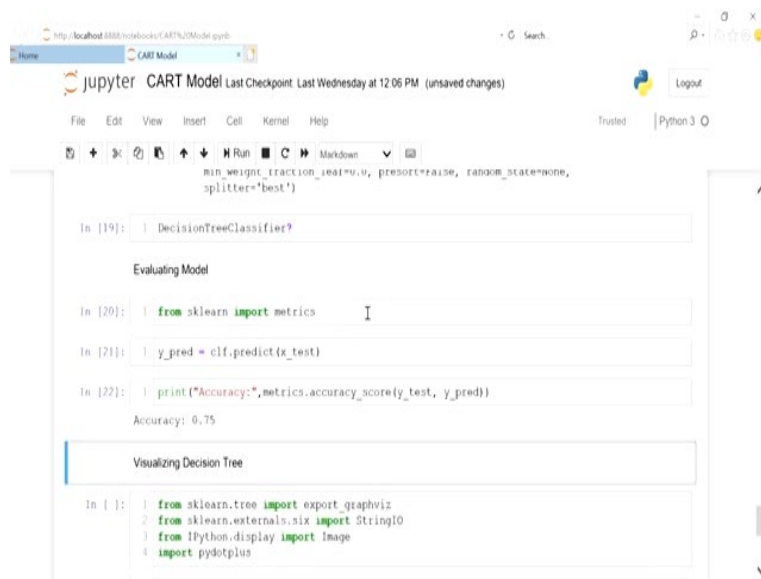
(Refer Slide Time: 27:54)



This shows our CART output, so age is the first attribute and the age is true then they we are getting the leaf node it is a false then we are choosing another two attributes if it is then age,

credit rating, further we say credit rating then we go for income if it is false, we are stopping. So here what do you need to understand the blue, the blue circle represents, the blue rectangle represents the favorable decision for us, the orange one represents the unfavorable, white one represents in between that means we need to do further analysis. Now we are going to split the dataset splitting the ratio of 75, 25 so 75% days of data set is for training remaining 25% is for testing.

(Refer Slide Time: 28:59)



```
min_weight_fraction=0.0, presort=False, random_state=None,
splitter='best')

In [19]: 1 DecisionTreeClassifier?

Evaluating Model

In [20]: 1 from sklearn import metrics

In [21]: 1 y_pred = clf.predict(x_test)

In [22]: 1 print("Accuracy:", metrics.accuracy_score(y_test, y_pred))

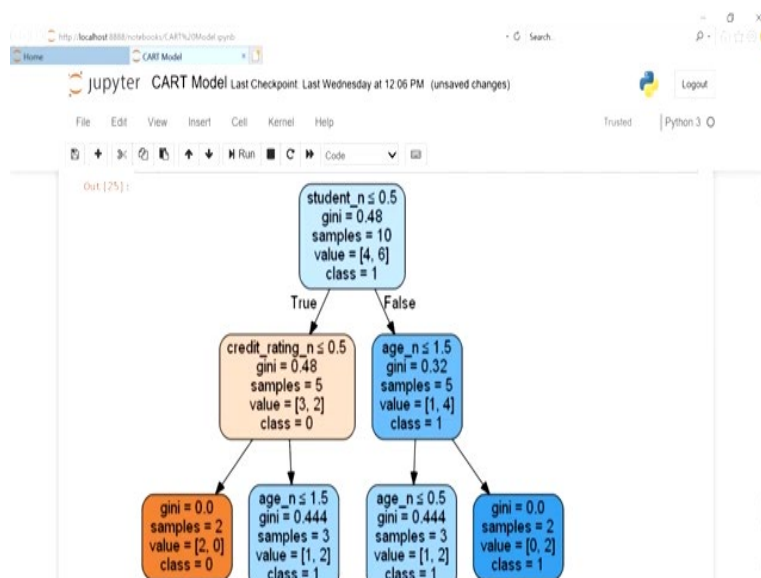
Accuracy: 0.75

Visualizing Decision Tree

In [ ]: 1 from sklearn.tree import export_graphviz
2 from sklearn.externals.six import StringIO
3 from IPython.display import Image
4 import pydotplus
```

So after splitting the dataset, we are running the CART model then we are going to evaluate there is accuracy of our model. So the accuracy is 0.75 now we will visualize the CART model.

(Refer Slide Time: 29:40)



This was the output of data set where we are doing the splitting. So now the student is taken as the primary node for splitting if it is true, we will go for credit rating if it is false then you go for age for the next classifier. Then age we will go for the condition and this age underscore $n \leq 1.5$ I explained what is the meaning of 1.5 everywhere there are more possibility of favorable decision because class = 1.

There is orange rectangles which represents 0, 0 which is not favorable decisions. So what it means that when we split the data set, our decision making become very simple because our tree is in the very simple form, easy to interpret it. In this lecture I have explained how to do the CART model with the help of python. I have taken an example problem with the help of sample problem first I have got the CART model without splitting the dataset.

After that, after splitting the dataset, then I got output then I have compared, and I have explained in detail the output of the CART model. With that we are concluding this course data analytics with the python. Thank you very much for attending this course. Thank you.