EMORY UNIVERSITY

Introduction to Business Data Analytics

Homework #1

______Jie Zhu_______

(put your name above)

Note: This is an individual homework. Discussing this homework with your classmates is a violation of the Honor Code. If you borrow code from somewhere else, please add a comment in your code to make it clear what the source of the code is (e.g., a URL would sufficient). If you borrow code and you don't provide the source, it is a violation of the Honor Code.

Total grade: ________ out of ___150___ points

This homework has ***<u>seven</u>*** ***questions. Please answer them all and submit your assignment using Canvas. In particular, you need to submit:***

- ***a) This file with your answers. Please transform the doc file to a PDF file. Then, submit the PDF File in the Assignment labeled as "Homework 1"***
- ***b) Your code and/or Rapidminer repositories for the hands-on exercise. Submit the (1) Python code in Jupyter format (2) the Rapidminer repositories and (3) the dataset(s) of the Assignment labeled as "Homework 1_Code & Repositories"***

1) (24 points) Choose the data technology (Q, H, U, or S) that is most appropriate for each of the following business questions/scenarios.

Q – SQL Querying
H – Statistical Hypothesis Testing
U – Unsupervised Data Mining/Pattern Finding
S – Supervised Data Mining

a) _H_ For my on-line advertising the decision tree model yields a response rate of 0.5% and the old, manual targeting model yields 0.3%. Is the decision tree model really better?

b) _Q_ I want to know which of my customers are the most profitable.

c) _Q_ I need to get data on all my on-line customers who were exposed to the special offer, including their registration data, their past purchases, and whether or not they purchased in the 15 days following the exposure.

d)_U_ I would like to segment my customers into groups based on their demographics and prior purchase activity. I am not focusing on improving a particular task, but would like to generate ideas.

e) _S_ I have a budget to target 10,000 existing customers with a special offer. I would like to identify those customers most likely to respond to the special offer.

f) _U_ I want to know what characteristics differentiate my most profitable customers.

2) (16 points) Label each case as describing either data mining (DM), or the use of the results of data mining (Use).

g) __Use__ Choose customers who are most likely to respond to an on-line ad.

h) __DM___ Discover rules that indicate when an account has been defrauded.

i) __DM___ Find patterns indicating what customer behavior is more likely to lead to response to an on-line ad.

j) __Use___ Estimate probability of default for a credit application.

3) (15 points) MTC (MegaTelCo) has decided to use supervised learning to address its problem of churn in its wireless phone business. As a consultant to MTC, you realize that a main task in the business understanding/data understanding phases of the data mining process is to define the target variable. In one or two sentences, please suggest a definition for the target variable. Be as precise as possible—someone else will be implementing your suggestion. ***(Remember: it should make sense from a business point of view, and it***

should be reasonable that MTC would have data available to know the value of the target variable for historical customers.)

The target variable should tell whether the customer stayed or left in a certain period of time (for example 6 months). MTC can then use other variables to identify the potential patterns of its churn problem.

4) (20 points) A predictive model has been applied to a test dataset and has classified 87 records as fraudulent (31 correctly so) and 953 as non-fraudulent (919 correctly so).

- **Present the confusion matrix for this scenario.**

Predicted	Actual	
	Fraudulent	Non-Fraudulent
Fraudulent	31	56
Non-Fraudulent	34	919

- **Calculate the error rate and accuracy rate.**

Accuracy rate = (31+919)/(31+56+34+919) = 950/1040 = 0.9135

Error rate = 1 – Accuracy rate = 1 – 0.9135 = 0.0865

- **Calculate the precision, recall, and f-measure values for each of the two outcome classes (i.e., fraudulent and non-fraudulent records);**

Precision(fraudulent) = 31/(31+56) = 31/87 = 0.3563

Precision(non-fraudulent) = 919/(919+34) = 919/953 = 0.9643

Recall(fraudulent) = 31/(31+34) = 32/65 = 0.4769

Recall(non-fraudulent) = 919/(919+56) = 919/975 = 0.9426

**f-measure(fraudulent) =
2*(Precision(fraudulent)*Recall(fraudulent))/(Precision(fraudulent)+Recall(fraudulent)) =
2*(0.3563*0.4769)/(0.3563+0.4769) = 0.4079**

f-measure(non-fraudulent) = 2*(Precision(non-fraudulent)*Recall(non-fraudulent))/(Precision(non-fraudulent)+Recall(non-fraudulent)) = 2*(0.9643*0.9426)/(0.9643+0.9426) = 0.9533

- **Also, calculate the accuracy rate that would be achieved by naïve (majority) rule on this data.**

Accuracy rate = (56+919)/(56+919+31+34) = 975/1040 = 0.9375

5) (50 points) [Implement this exercise with both RapidMiner (20 points) and Python (30 points)] Use the decision tree classification technique on the *HW1* dataset. This dataset is provided on the course website and contains data about consumers and their decisions to terminate a contract (i.e., consumer churn problem).

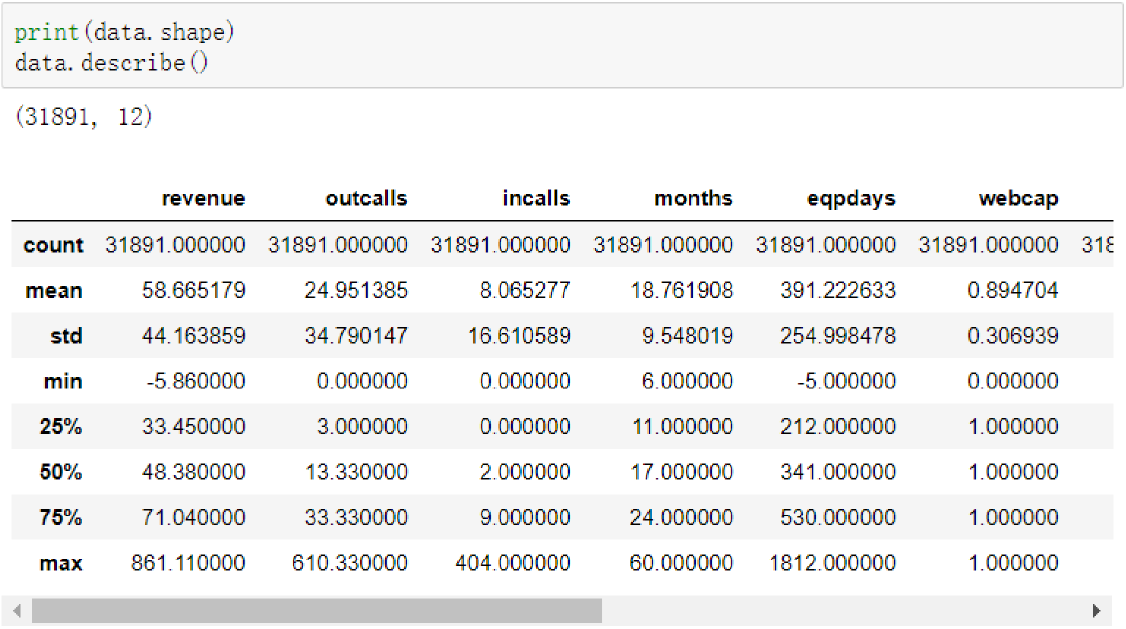
Data description:

Col.	Var. Name	Var. Description
1	revenue	Mean monthly revenue in dollars
2	outcalls	Mean number of outbound voice calls
3	incalls	Mean number of inbound voice calls
4	months	Months in Service
5	eqpdays	Number of days the customer has had his/her current equipment
6	webcap	Handset is web capable
7	marryyes	Married (1=Yes; 0=No)
8	travel	Has traveled to non-US country (1=Yes; 0=No)
9	pcown	Owns a personal computer (1=Yes; 0=No)
10	creditcd	Possesses a credit card (1=Yes; 0=No)
11	retcalls	Number of calls previously made to retention team
12	churndep	Did the customer churn (1=Yes; 0=No)

Build a decision tree model that predicts whether a consumer will terminate his/her contract. In particular, I would like for you to create a decision tree using entropy with no max depth. Explore how well the decision trees perform for several different parameter values (e.g., for different splitting criteria). Interpret the model (decision tree) that provides the best predictive performance.
Some possible issues / hints to think about: using training vs. test datasets.
Present a brief overview of your predictive modeling process, explorations, and discuss your results. Make sure you present information about the model "goodness" (please report the confusion matrix, predictive accuracy, classification error, precision, recall, f-measure).

Present a brief overview of your predictive modeling process. That is, you need to lay out the steps you have taken in order to build and evaluate the decision tree model. For instance, how did you explore the data set before you built the model? Write this report in a way that the upper level management of the team would understand what you are doing. Why is the decision tree an appropriate model for this problem? How can we evaluate the predictive ability of the decision tree? If you build decision trees with different splitting criteria, which decision tree would you prefer to use in practice?

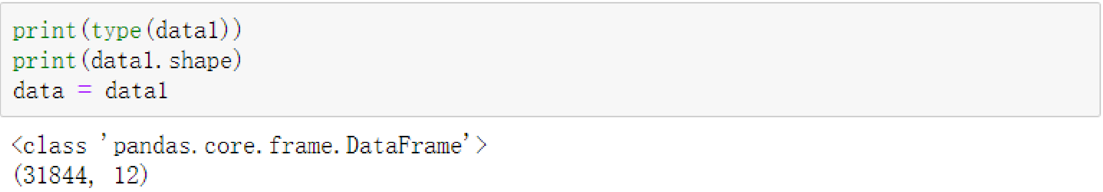
I. Import data and data cleansing
After import the .csv file, I find that there are some data points having negative revenue and negative eqpdays, which doesn't make sense. I remove these data points and the final data set has 31844 rows rather than the initial 31891 rows.



```
print(data.shape)
data.describe()
```

```
(31891, 12)
```

	revenue	outcalls	incalls	months	eqpdays	webcap	
count	31891.000000	31891.000000	31891.000000	31891.000000	31891.000000	31891.000000	318
mean	58.665179	24.951385	8.065277	18.761908	391.222633	0.894704	
std	44.163859	34.790147	16.610589	9.548019	254.998478	0.306939	
min	-5.860000	0.000000	0.000000	6.000000	-5.000000	0.000000	
25%	33.450000	3.000000	0.000000	11.000000	212.000000	1.000000	
50%	48.380000	13.330000	2.000000	17.000000	341.000000	1.000000	
75%	71.040000	33.330000	9.000000	24.000000	530.000000	1.000000	
max	861.110000	610.330000	404.000000	60.000000	1812.000000	1.000000	

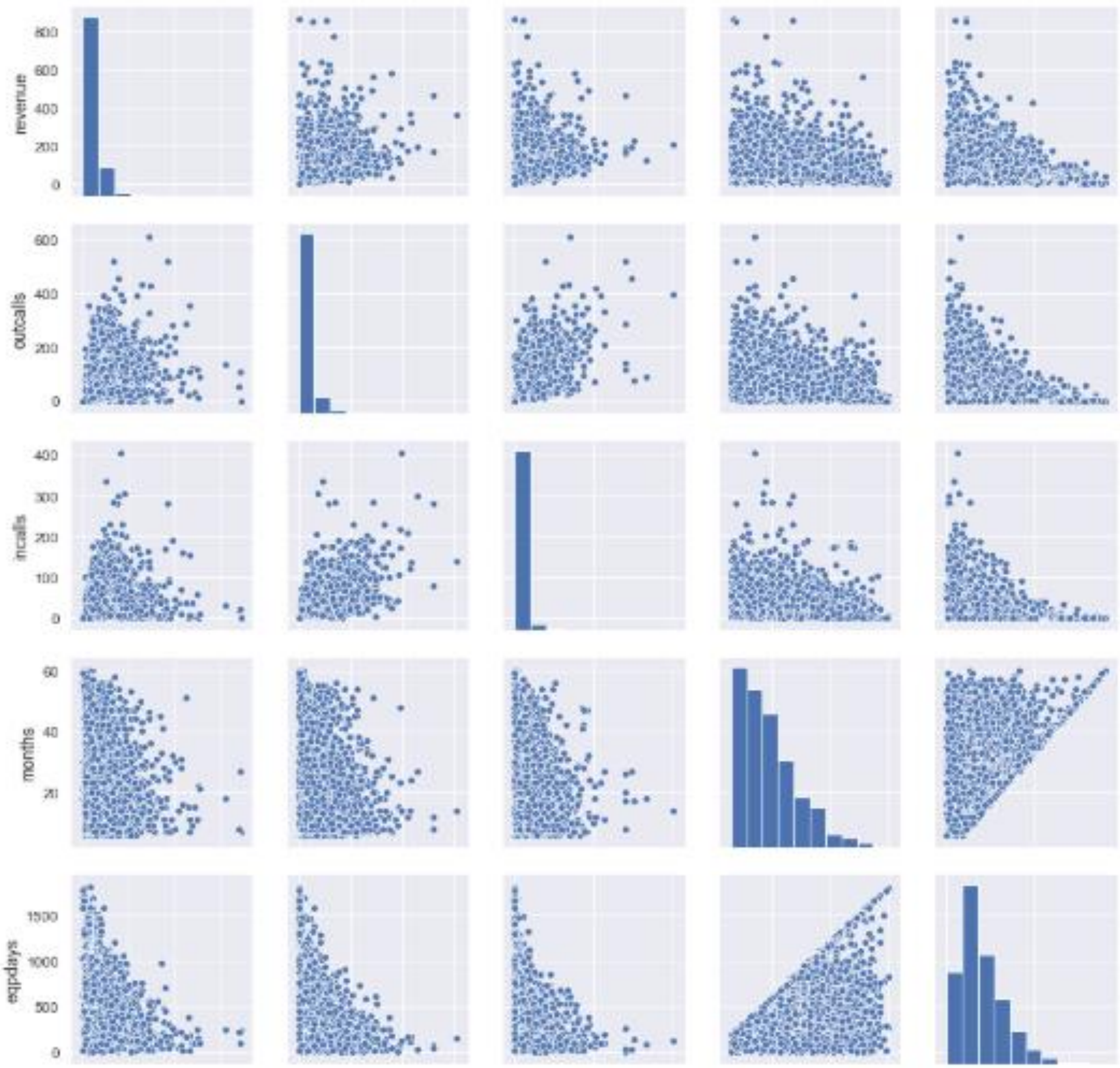


```
print(type(data1))
print(data1.shape)
data = data1
```

```
<class 'pandas.core.frame.DataFrame'>
(31844, 12)
```

II. Data Exploration

After making a pairplot using the seaborn package, I have following findings. REVENUE, OUTCALLS, INCALLS, MONTHS, and EQPDAYS are all highly right-skewed. This makes sense because most people have similar tele using habits, while only a few make more outcalls and incalls. Also, OUTCALLS and INCALLS have moderate positive correlation. This can be explained by the fact that people making more outcalls also receive more incalls.

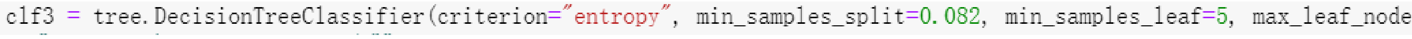


III. Why decision tree?

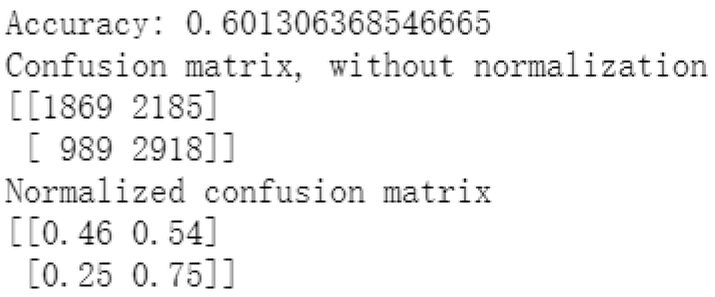
The decision tree is a useful tool in this problem, because (1) this problem has a specific purpose – analyzing and determining CHURNDEP; (2) the target variable CHURNDEP is categorical and binary (YES/NO) and finding the pattern of CHURNDEP is a class probability estimation.

IV. How to evaluate?

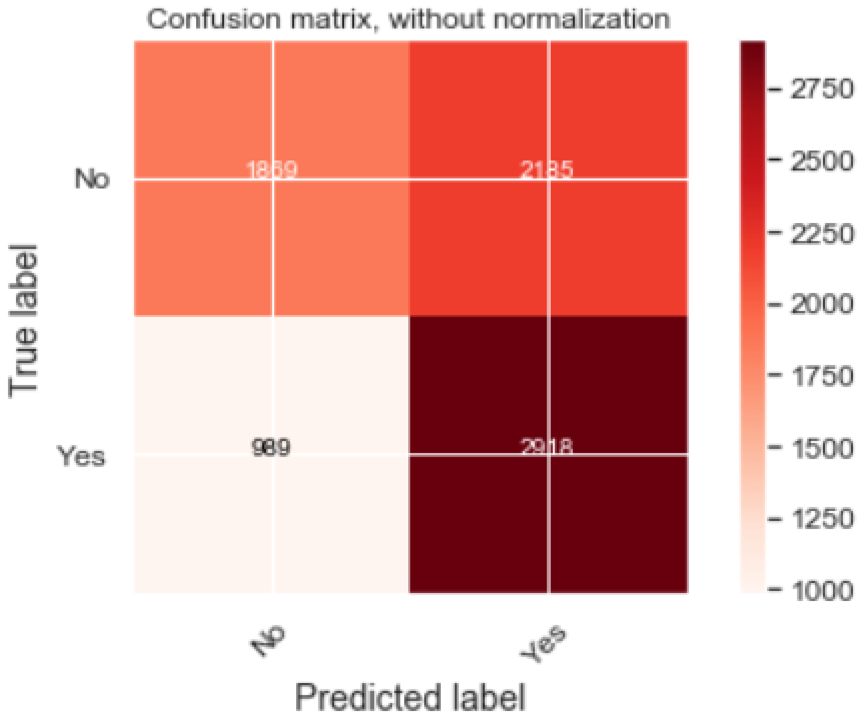
The performance of the decision tree model can be evaluated by the accuracy rate and the confusion matrix.



```
clf3 = tree.DecisionTreeClassifier(criterion="entropy", min_samples_split=0.082, min_samples_leaf=5, max_leaf_node
```



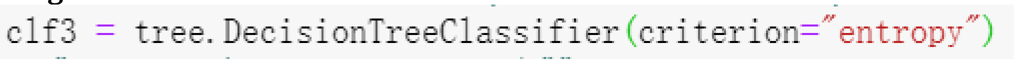
```
Accuracy: 0.601306368546665
Confusion matrix, without normalization
[[1869 2185]
 [ 989 2918]]
Normalized confusion matrix
[[0.46 0.54]
 [0.25 0.75]]
```



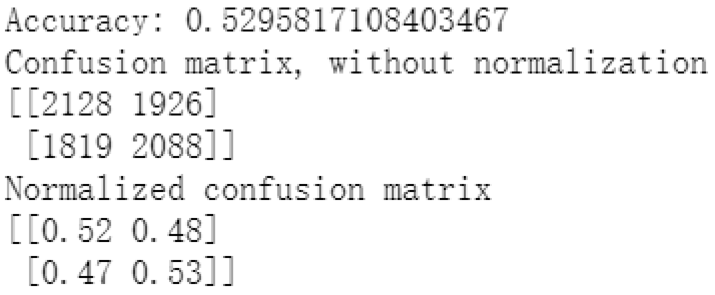
Previous image indicates the accuracy of this model is 60.13%. True(Yes) = 2918; True(No) = 1869; False(Yes) = 2185; False(No) = 989.

V. **Which decision tree criteria?**

a. **Original Tree**

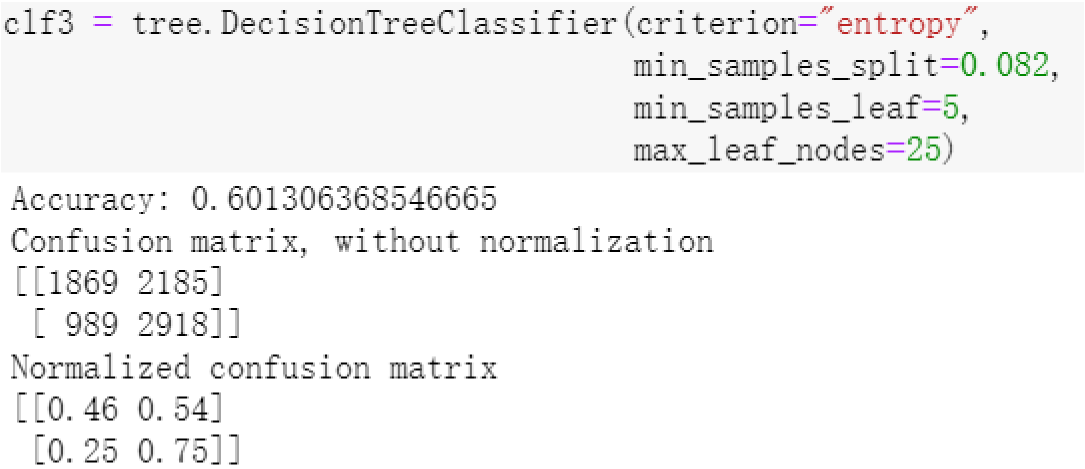


```
clf3 = tree.DecisionTreeClassifier(criterion="entropy")
```



```
Accuracy: 0.5295817108403467
Confusion matrix, without normalization
[[2128 1926]
 [1819 2088]]
Normalized confusion matrix
[[0.52 0.48]
 [0.47 0.53]]
```

b. **Final Tree**

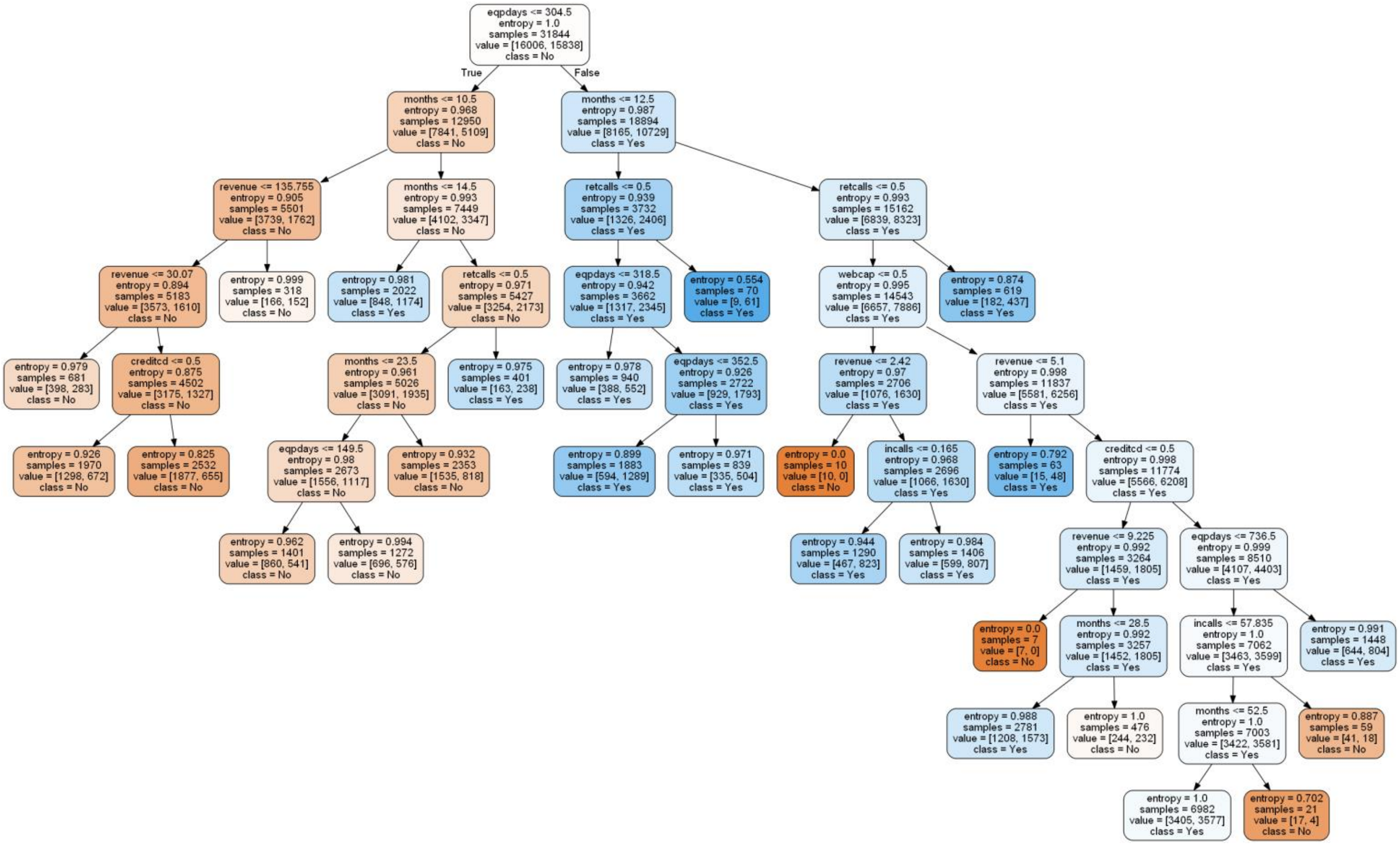


```
clf3 = tree.DecisionTreeClassifier(criterion="entropy",
                                   min_samples_split=0.082,
                                   min_samples_leaf=5,
                                   max_leaf_nodes=25)
```

```
Accuracy: 0.601306368546665
Confusion matrix, without normalization
[[1869 2185]
 [ 989 2918]]
Normalized confusion matrix
[[0.46 0.54]
 [0.25 0.75]]
```

Compared to the original tree with no splitting criteria, the final decision tree I use have specified minimum samples split (0.082), minimum samples leaf of 5, and maximum leaf nodes of 25.

The final tree looks like this:


eqpdays <= 304.5 entropy = 1.0 samples = 31844 value = [16006, 15838] class = No
True
False
months <= 10.5 entropy = 0.968 samples = 12950 value = [7841, 5109] class = No
months <= 12.5 entropy = 0.987 samples = 18894 value = [8165, 10729] class = Yes
revenue <= 135.755 entropy = 0.905 samples = 5501 value = [3739, 1762] class = No
months <= 14.5 entropy = 0.993 samples = 7449 value = [4102, 3347] class = No
retcalls <= 0.5 entropy = 0.939 samples = 3732 value = [1326, 2406] class = Yes
retcalls <= 0.5 entropy = 0.993 samples = 15162 value = [6839, 8323] class = Yes
revenue <= 30.07 entropy = 0.894 samples = 5183 value = [3573, 1610] class = No
entropy = 0.999 samples = 318 value = [166, 152] class = No
entropy = 0.981 samples = 2022 value = [848, 1174] class = Yes
retcalls <= 0.5 entropy = 0.971 samples = 5427 value = [3254, 2173] class = No
eqpdays <= 318.5 entropy = 0.942 samples = 3662 value = [1317, 2345] class = Yes
entropy = 0.554 samples = 70 value = [9, 61] class = Yes
webcap <= 0.5 entropy = 0.995 samples = 14543 value = [6657, 7886] class = Yes
entropy = 0.874 samples = 619 value = [182, 437] class = Yes
entropy = 0.979 samples = 681 value = [398, 283] class = No
creditcd <= 0.5 entropy = 0.875 samples = 4502 value = [3175, 1327] class = No
months <= 23.5 entropy = 0.961 samples = 5026 value = [3091, 1935] class = No
entropy = 0.975 samples = 401 value = [163, 238] class = Yes
entropy = 0.978 samples = 940 value = [388, 552] class = Yes
eqpdays <= 352.5 entropy = 0.926 samples = 2722 value = [929, 1793] class = Yes
revenue <= 2.42 entropy = 0.97 samples = 2706 value = [1076, 1630] class = Yes
revenue <= 5.1 entropy = 0.998 samples = 11837 value = [5581, 6256] class = Yes
entropy = 0.926 samples = 1970 value = [1298, 672] class = No
entropy = 0.825 samples = 2532 value = [1877, 655] class = No
eqpdays <= 149.5 entropy = 0.98 samples = 2673 value = [1556, 1117] class = No
entropy = 0.932 samples = 2353 value = [1535, 818] class = No
entropy = 0.899 samples = 1883 value = [594, 1289] class = Yes
entropy = 0.971 samples = 839 value = [335, 504] class = Yes
entropy = 0.0 samples = 10 value = [10, 0] class = No
incalls <= 0.165 entropy = 0.968 samples = 2696 value = [1066, 1630] class = Yes
entropy = 0.792 samples = 63 value = [15, 48] class = Yes
creditcd <= 0.5 entropy = 0.998 samples = 11774 value = [5566, 6208] class = Yes
entropy = 0.962 samples = 1401 value = [860, 541] class = No
entropy = 0.994 samples = 1272 value = [696, 576] class = No
entropy = 0.944 samples = 1290 value = [467, 823] class = Yes
entropy = 0.984 samples = 1406 value = [599, 807] class = Yes
revenue <= 9.225 entropy = 0.992 samples = 3264 value = [1459, 1805] class = Yes
eqpdays <= 736.5 entropy = 0.999 samples = 8510 value = [4107, 4403] class = Yes
entropy = 0.0 samples = 7 value = [7, 0] class = No
months <= 28.5 entropy = 0.992 samples = 3257 value = [1452, 1805] class = Yes
incalls <= 57.835 entropy = 1.0 samples = 7062 value = [3463, 3599] class = Yes
entropy = 0.991 samples = 1448 value = [644, 804] class = Yes
entropy = 0.988 samples = 2781 value = [1208, 1573] class = Yes
entropy = 1.0 samples = 476 value = [244, 232] class = No
months <= 52.5 entropy = 1.0 samples = 7003 value = [3422, 3581] class = Yes
entropy = 0.887 samples = 59 value = [41, 18] class = No
entropy = 1.0 samples = 6982 value = [3405, 3577] class = Yes
entropy = 0.702 samples = 21 value = [17, 4] class = No

6) (20 points) Is a node’s entropy generally lower or greater than its parent’s? Is it ever possible for a node’s entropy to be higher than its parent’s entropy? Please justify your answer. Be precise but concise.

A node’s entropy is generally lower than its parent’s. However, it is possible for a node’s entropy to be higher than its parent’s entropy. For example, the population could have 17 instances(16 dots, 1 star). If the nodes are divided into two sets, 15 dots, and 1 dot & 1 star. The second set will have entropy of 1, higher than the parent’s entropy.

7) (5 points) What are the differences between supervised and unsupervised methods in machine learning? Is the decision tree algorithm a supervised or an unsupervised method? Be precise but concise.

Supervised methods in machine learning are used when there’s a problem with a specific quantifiable target. Also, we have to have enough data prior to our decision (at least 500 for each type of classification). In comparison, unsupervised methods have no specific purpose or target specified, and there is no guarantee that the final results will be meaningful or useful. Since decision tree algorithm has a clearly defined categorical target variable, it is a supervised method.