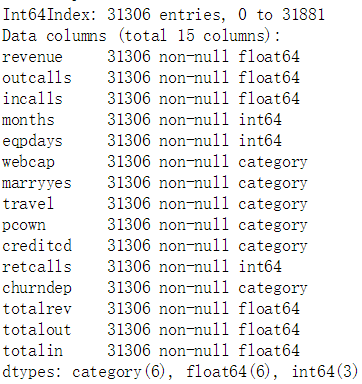
* 1. **Load data, explore data and clean data:**

I loaded data into Python and used several methods to explore data. The dataset has 31891 rows and 12 columns, and there's no null value in the dataset.

I excluded some abnormal values in 'eqpdays', 'revenue', 'outcalls' etc. and generate some new variables.

Finally, I got 31306 observations and 15 columns.



*(Basic information about the dataset)*

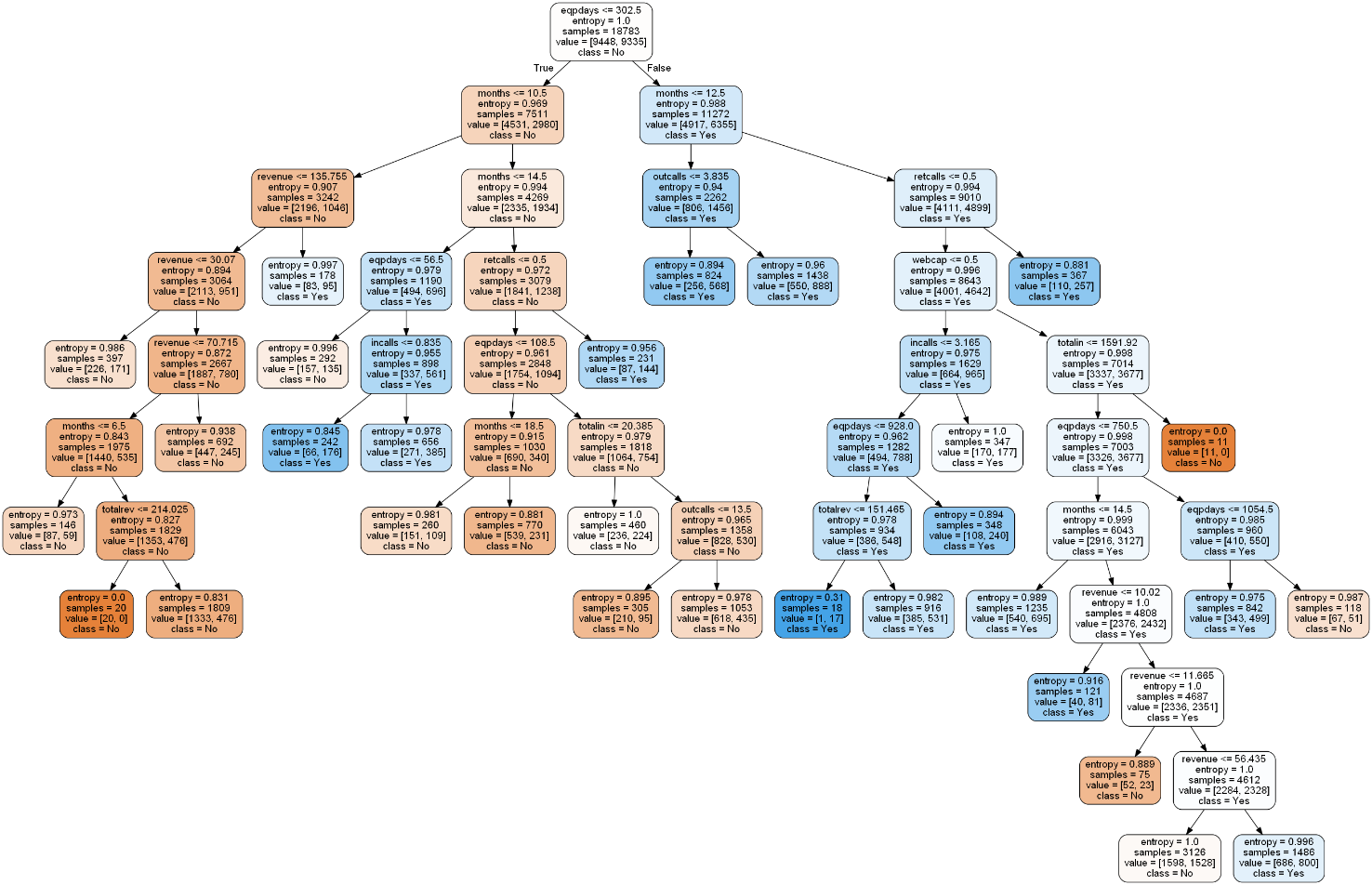
* 1. **Build model:**

I set 'churndep' as target variable (y), and other variables as predictor variables (x). To evaluate and test different models, I split the whole dataset into 3 datasets, which were training data set, validation data set and test data set, the proportion was 60%, 20%, 20%.

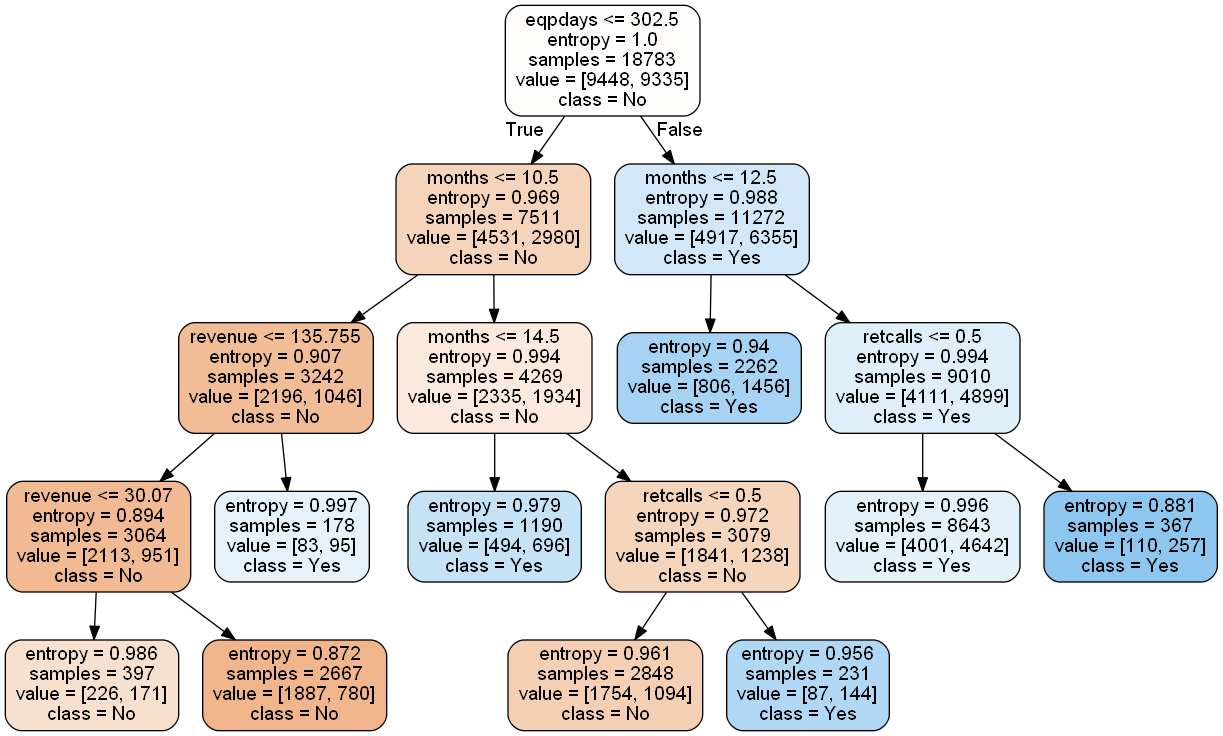
Since the purpose is to predict whether a consumer will terminate his/her contract, it’s a supervised learning and classification problem, and decision tree will be a good method to build the model and make it easier to visualize and explain the outcome, thus making it easier for the senior management to understand.

I used several different criterions to build the decision tree model, then I found 2 relatively optimal models.

For the first model, I controlled the max leaf nodes in the whole tree and the minimum number of samples required to split a node. The tree looks like this:



For the second model, I only controlled the criteria of impurity decrease. The tree looks like this:

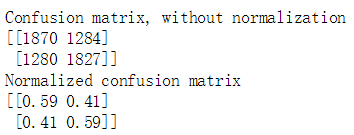
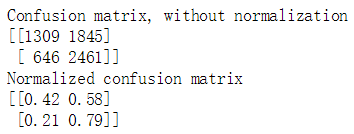


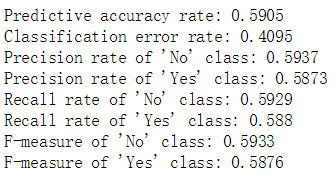
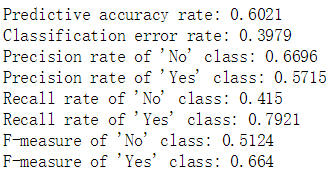
* 1. **Evaluate and compare model:**

Then I used the validation dataset to compare the outcome of the two models. I used the predictive accuracy rate, classification error rate, precision rate, recall rate, and f-measure to evaluate the accuracy of the model.

From the outcome, I found that the accuracy rate of the two models is very close, the first one is 59.05% and the second one is 60.21%. However, in terms of the confusion matrix and other indicators, I found that the first model is a more 'balanced' model, since the precision rate of "No" and "Yes" classes are similar. While for the second model, the prediction accuracy of class "Yes" and “True Positives” is higher than class "No" and “True Negatives”. Which means if using the second model, for class "Yes" we will have a higher chance that the prediction is accurate.

Since the purpose of this prediction is to reflect the customer's decision of terminating the contract or not. Knowing that a customer will terminate the contract is more valuable for the company. Plus, the tree map of the second tree only has four layers and is much easier for the senior management to understand. Thus, I chose the second model as the final model.

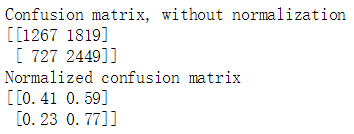
 

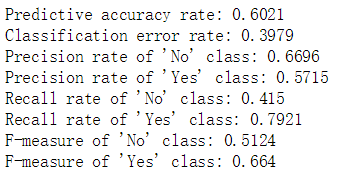
 

*(Goodness criterions of model 1) (Goodness criterions of model 2)*

* 1. **Test model & model interpretation:**

Finally, I used the test dataset to test the model I chose. The Predictive accuracy rate is 0.6021, the classification error rate is 0.3979, the precision rate of ‘No’ class is 0.6696, the precision rate of ‘Yes’ class is 0.5715, the recall rate of ‘No’ class is 0.415, the recall rate of ‘Yes’ class is 0.7921, the f-measure of ‘No’ class is 0.5124, and the f-measure of ‘Yes’ class is 0.664.





*(Goodness criterions of model 2 using test dataset)*

The model tells us that the most influential attribute to determine whether a consumer will terminate his/her contract is the Number of days the customer has had his/her current equipment, and the second most influential attribute is the Months in Service. These two attributes both indicate how long a customer has used our services. However, the model indicates that the longer a customer stays, the higher the possibility that he/she churns. This definitely shows some problems with the company’s ability to keep customers.

Probably, the company needs to improve the service it provides, do some marketing campaign targeting old customers, and let the customer service team make more phone calls to those customers.

* 1. **Thoughts of how to improve model accuracy:**

Though I chose a relatively high-accuracy model, the overall accuracy rate is still very low, only around 60%.

In my opinion, there are several ways to improve the accuracy. First is adding more related variables and try to see if the accuracy rate could be improved. Second is using another classification method to do prediction, ie. Logistic regression. Although the decision tree is a great way to show the result to the senior management, probably it not the fit classification method for this particular dataset.