Exponea internship assignment - report

Jakub Bahyl

March 5, 2018

1 Introduction and basic analysis

In this work I'm summarising my analysis with data of 5000 customers located in California and their inclination to churn. I worked overall with 707 churned customers described by 12 reasonable parameters (*account length*, *total day minutes spent*,..) using three predictive models (logit, decision tree and random forest). Below I'm picking the main results from each model and in conclusion, I have prepared two strong recommendations.

2 Logit model

This model has proved that the *number of service calls* customer made is reasonably affecting his/her probability to churn. For instance, the customer who called more than 4 times to service, has **25-times higher odds** to churn than a customer, who didn't call at all, whereas the customer with less than 5 calls to service has **only 7-times higher odds**. The best accuracy of the model was **83.8%** and has suffered from not independent customers data and its over-fitting property. Definitely not the model worth of make predictions.

3 Decision tree (DT)

DT provided a lot of info when the churn exactly happens. If customer's *total day mins* exceeds 250 mins, there emerges **50% chance of churn** (!). Moreover, if no *voice mail plan* was made and the most of the day mins was made in night (*total night mins*), the churn **will definitely occur**. Also, the model showed that churn happens for customers with less than 250 day mins spent too - if they made more than 4 service calls (this is a consistent information with logit model). Accuracy of decision tree model was **94%** and the main disadvantage of was it's high results sensitivity on - with new customers assing, the main decisions would stay same, but the ordering can easily interchange.

4 Random forest (RF)

RF was the best model for churn prediction (**95%** accuracy) and provided us also with the importance table, where all features are sorted based on their significance. Consistently with the DT model, the most important feature is *total day mins* (**25%** of churn information is hidden here). The second and third most important feature is whether a *international plan* was made and how many *service calls* was done (as expected). Only these three features cover around **60% of churn information**. The least important seems *eve mins, night mins*, but not negligibly. Even RF model is very powerfull, it didn't bring any description.

5 Conclusion

If I was asked to prepare any recommendations based on the models results, they would be definitely these two:

- Try to keep down the customer's spent minutes and number of service calls he/she has to make. This should ensure **97%** chance of churn **to not happen**.
- If customer's spent time is simply high and you cannot do anything with it, definitely make the *voice mail plan* and try to make his/her spent time **during day** (not night).