Client Churn Prediction Model

Yuji Mori

18 May 2021

***Note: Report has been scrubbed of any identifying data/code.*

I. Introduction

Churn is the event where a customer/client ceases their relationship with a business. Losing clients in this manner has a direct impact on overall revenue, and it is imperative that the business monitors their customers' behavior in order to maximize their earnings. [COMPANY] partners with small to medium-sized businesses to offer assistance with marketing efforts and advertising campaigns. The company is interested in analyzing their client portfolio, specifically by developing a churn prediction model. The predictions from this model can help identify clients that are most likely to churn, which will enable [COMPANY] to focus their marketing and retention efforts.

II. Data

The data is in tabular (.csv) format with exactly 10,000 rows, each row representing a unique client. The data includes key metrics relating to the performance of the client campaign(s), such as client CPL (cost-per-lead), calls/clicks received, and their overall budget. There is also demographic information such as the client's industry and age (months) in the portfolio. Finally, there is a flag column indicating whether the client has churned or not.

III. Methodology

A. Data Pre-processing

A successful model requires clean data, and the first step is to handle missing values. I found that one CPL-related column had nearly 2,000 missing values. I decided it was appropriate to impute (fill) these missing values with a 0, as it would indicate 'no change' in CPL (the variable represents change in CPL compared to previous month). Afterwards, I also checked the dataset for duplicates and unexpected values.

B. Exploratory Analysis

It is crucial to understand the underlying data before beginning a modeling effort. In this step, I analyze the distributions of all the variables available, and take notes on any findings. These include skew of distributions/frequencies, outlier identification, and correlation plots.

C. Model Training

Model training first requires that the dataset is split into training and testing partitions. I reserve 80% of the observations for model training (n=8,000), and leave the remaining 20% as a hold-out (test) set in order to evaluate the model (n=2,000).

I initially trained three different models on the data using the following methods: Logistic Regression, Random Forest Classification, and Gradient Boosting.

D. Model Evaluation

The most important indicator of model performance is how often it can correctly classify the test data. Some commonly used metrics to evaluate this ability are:

1. Accuracy: the fraction of correct predictions over total sample
2. F-score: similar to accuracy, but with more weight on false predictions
3. AUC: measures how successfully the model labels a client as 'churned'

Based on these metrics, as well as their granular mathematical components, I determined that the **Gradient-Boosting Model** performed the best.

IV. Results

The final model is able to predict the event of churn about 37% of the time. While this percentage seems low, it can translate into significant revenue boosts. And impressively, the model is also able to detect event of non-churn at a success rate of 85%. The full technical overview of the metrics I considered are shown below (left). The percentages I referenced are True Positive Rate (TPR) and True Negative Rate (TNR).

I also identified the top contributing variables (right). The two CPL metrics exhibit the largest impact on the model, followed by other expected variables like duration, budget size, and number of clicks/calls.

	Random Forest	Gradient-Boosted
F Score	0.364326	0.380228
Accuracy	0.832500	0.755500
AUC	0.610403	0.611457
False Postive Rate (FPR)	0.015066	0.145637
True Positive Rate (TPR)	0.235872	0.368550
True Negative Rate (TNR)	0.984934	0.854363

CPL_wrt_BC	0.190763
CPL_wrt_self	0.144195
duration	0.121634
avg_budget	0.119779
clicks	0.103555
calls	0.051647
num_products	0.017325
client_state_FL	0.010348
BC_Home & Home Improvement	0.009048
client_state_CA	0.008933

V. Business Implications and Recommendations

The goal of this prediction model is to allow the business to focus marketing and retention efforts on the clients that are most likely to churn.

Suppose that, through these efforts, we are able to retain each client detected by the model (37% or 150/400 churned clients). The sum of their average budget comes out to be nearly $132,000, which is monthly revenue that [COMPANY] would have otherwise lost.

With these savings in mind, it is clear that a churn prediction model should be used on a regular basis to continuously monitor the portfolio. Because many data elements (such as budget, time in-file, change in CPL) are measured in monthly increments, I recommend running the model on monthly snapshots of the client population. Retention efforts should consist of active engagement with the clients to adjust campaigns (costs, targeting strategies, etc.) and make sure all their needs and expectations are met.

High CPL is an indicator that:

- Your ad's targeting settings are not defined
- The landing page is not optimized for desired outcome
- The product/service being offered is inadequate
- Price point too high

Top Marketing Performance Metrics:

- Social Media Engagement. ...
- Click-Through Rate. ...
- Landing Page Conversion Rate. ...
- Cost Per Lead . ...
- Customer Lifetime Value. ...