

CHURN PREDICTION

EXECUTIVE SUMMARY

Data Scientist and author: Leomar Fonseca

1. Definitions

Without a system that predicts which users are likely to churn (or are more likely to), the company would either risk losing clients or employ resources (time, people and money) on all users, not optimizing the investment in client retention. The goal of the challenge was to develop a churn prediction model that identifies users at risk of churning within the next 30 days, filling this gap and allowing the company to better leverage resources.

At the beginning of this project, the focus was on creating the target variable. In order to do this, the information from the timestamp column was used. It is also worth mentioning that the column “previous_session_gap_hours” was incorrect, as it contained mostly nulls or values between 1 and 10 (which were not accurate).

The steps taken were as follows:

- Ordering the dataset by “user_id” and “timestamp”;
- Calculating the difference in hours between each user session and their next one
- Removing the last session rows per user, since there is no information available as to whether the user actually churned;
- Flagging churn users where the difference was higher than 720 (equivalent of 30 days in hours).

The picture below illustrates the final result for user_000383.

	user_id	session_id	timestamp	previous_session_gap_hours	next_session_gap_hours	churn
3569	user_000383	sess_00004675	2024-04-03 19:09:11.500651+00:00	408.686141	521.640204	0
4034	user_000383	sess_00006760	2024-04-25 12:47:36.233928+00:00	521.640204	496.688817	0
4440	user_000383	sess_00006919	2024-05-16 05:28:55.976022+00:00	496.688817	665.225008	0
5011	user_000383	sess_00006921	2024-06-12 22:42:26.004615+00:00	665.225008	180.656865	0
5173	user_000383	sess_00006934	2024-06-20 11:21:50.720364+00:00	180.656865	295.357550	0
5427	user_000383	sess_00007197	2024-07-02 18:43:17.900300+00:00	295.357550	157.060952	0
5558	user_000383	sess_00008166	2024-07-09 07:46:57.328446+00:00	157.060952	386.799021	0
5898	user_000383	sess_00008359	2024-07-25 10:34:53.804581+00:00	386.799021	595.627664	0
6400	user_000383	sess_00009916	2024-08-19 06:12:33.395563+00:00	595.627664	812.820274	1
7102	user_000383	sess_00009942	2024-09-22 03:01:46.380224+00:00	812.820274	NaN	0

The next step was to define how the model would be evaluated. The defined structure had two components:

- Use recall as main performance metric; and
- Compare the final model to a simpler alternative, which was called baseline.


In other words, two models were to be created, where the performance of the baseline model had to be an improvement on the hypothesis of not having a model, as well as having the final model be a considerable improvement on the baseline.

2. Modeling

All the traditional machine learning steps were taken to ensure a high-quality predictive model at the end, such as: EDA, feature engineering, feature selection and hyperparameter tuning, not to mention all steps taken to assure reproducibility. Some findings during EDA are worth mentioning:

- The churn sessions represented 21.82% of the dataset, which supports the idea of choosing recall as main metric;
- The vast majority of the data points was from the CLV bronze tier;
- Churn was identified only in the CLV bronze tier; and
- The CLV diamond tier is the one which brings the most revenue to the company considering the average per user.

At the end of the modeling phase, the final predictive model had a 85% recall versus 74% of the baseline model. Besides, an application was created to make predictions possible - either a single data point or a batch.



Churn Prediction

Instructions

Single prediction

Batch prediction

This app predicts customer churn probability and recommends retention campaigns based on user data.

Single Prediction

Should be used to predict a single data point. Fill in the fields and press "Run prediction".

Batch Prediction

Use this tab to process multiple users at once. Upload a CSV file with the same format as the provided training file and press "Run prediction". The output will be a file with predictions, campaign recommendations, and the features created.

Output Interpretation

Churn Prediction

- 0: User is not likely to churn
- 1: User is likely to churn

Campaign Recommendations

Based on churn risk and CLV, the app recommends:

3. Results

The first deliverable was the aforementioned application, which has overall great performance and ease-of-use. The second product was a file generated at the end of the notebook with all the recommended campaigns per user in the dataset. This was created by calculating the “campaign budget” considering the user “CLV-to-date”, their churn likelihood and the information passed on the provided PDF file (cost by campaign). In summary, out of 1419 users, the app recommends:

- No action to 1082 users (either due to low likelihood of churn or because their churn probability with “campaign budget” relation would return in loss of revenue;
- Sending an automated email to 322 users;
- Delivering a bonus offer to 11 users; and
- Having a phone call to the last four users.

As per next steps, more time employed on feature engineering might increase accuracy, while maintaining recall high (or above a company defined threshold) further optimizing revenue.