CHURN PREDICTION

LEOMAR FONSECA

CHURN PREDICTION
STAGE 1 - DEFINITIONS

Goal: develop a churn prediction model that identifies users at risk of churning within the next 30 days.

Evaluation metrics: recall (and accuracy).

Application example: user_000383

Column created using timestamp information to "look forward"

	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

Rows removed from the dataset, since there is no churn information available

CHURN PREDICTION
STAGE 2 - MODELING

1. **EXPLORATORY DATA ANALYSIS (EDA)**
2. **FEATURE ENGINEERING**
3. **FEATURE SELECTION**
4. **BASELINE MODEL**
5. **PREDICTIVE MODEL**
6. **APPLICATION/REPRODUCIBILITY**

- IMBALANCED CLASS: 21.28% OF CHURN
- THE VAST MAJORITY OF THE DATA POINTS WAS FROM THE BRONZE TIER (CLV)
- ONLY THE BRONZE TIER HAD CHURN
- THE DIAMOND TIER (CLV) BRINGS THE MOST REVENUE ON AVERAGE TO THE COMP

BY NOT USING A MODEL AS
PERFECT RECALL - INFLATIN
BASELINE MODEL HAD A G
OF 85% ON THE TESTING S
CAMPAIGNS.

CHURN PREDICTION
STAGE 3 - RESULTS

The **first** deliverable was an application backed by a machine learning model which predicts the likelihood of churn per user session with great performance metrics.

The **second** product was a file generated at the end of the notebook with all the recommended campaigns per users 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 user, the app recommends:

- No action to 1082 user (because they either have a 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 as recall is kept high (or above a company defined threshold) further optimizing revenue.