Churn prediction

with highly imbalanced data

Yevheniia Mitriakhina

Problem

Churn - rate at which customers ***stop*** doing business with an entity.

Early detection of churn allows to take a proactive approach to retaining the existing customers or at the very least - forecast the cash flows which will be lost. This is especially important for businesses relying on subscription model.

Data - WSDM - KKBox's Churn Prediction Challenge

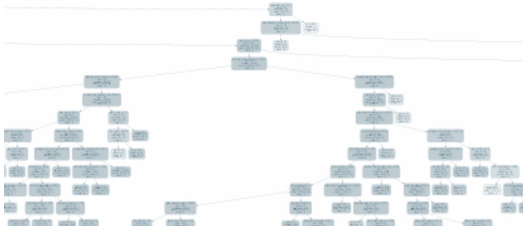
KKBOX is Asia’s leading music streaming service, holding the world’s most comprehensive Asia-Pop music library with over 30 million tracks. They offer an unlimited version of their service to millions of people, supported by advertising and paid subscriptions.

The key fields to determine churn/renewal are transaction date, membership expiration date, and is_cancel. is_cancel field indicates whether a user actively cancels a subscription. Subscription cancellation does not imply the user has churned. A user may cancel service subscription due to change of service plans or other reasons. The criteria of "churn" is no new valid service subscription within 30 days after the current membership expires.

Available Data

- payment_method_id: payment method
- payment_plan_days: length of membership plan in days
- plan_list_price: in New Taiwan Dollar (NTD)
- actual_amount_paid: in New Taiwan Dollar (NTD)
- is_auto_renew
- transaction_date: format `%Y%m%d`
- membership_expire_date: format `%Y%m%d`
- is_cancel: whether or not the user canceled the membership in this transaction.
- ~~num_25: # of songs played less than 25% of the song length~~
- ~~num_50: # of songs played between 25% to 50% of the song length~~
- ~~num_75: # of songs played between 50% to 75% of of the song length~~
- ~~num_985: # of songs played between 75% to 98.5% of the song length~~
- ~~num_100: # of songs played over 98.5% of the song length~~
- ~~num_unq: # of unique songs played~~
- ~~total_secs: total seconds played~~
- city
- bd: age.
- registered_via: registration method
- registration_init_time: format `%Y%m%d`
- expiration_date: format `%Y%m%d`

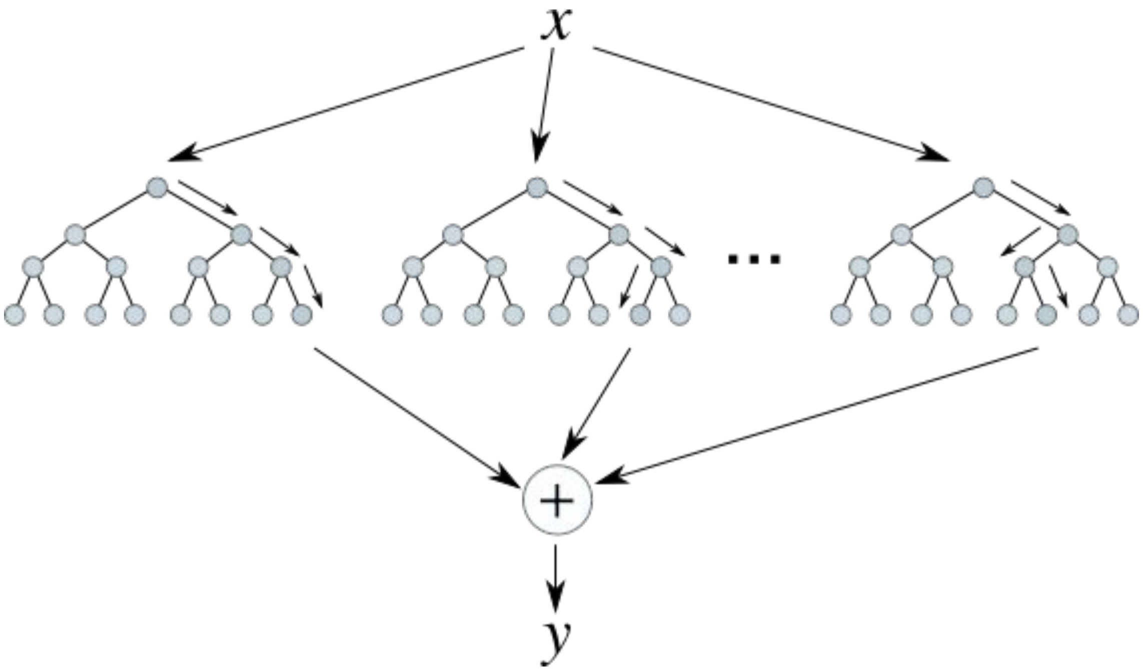
Data from the time domain → Features for the model



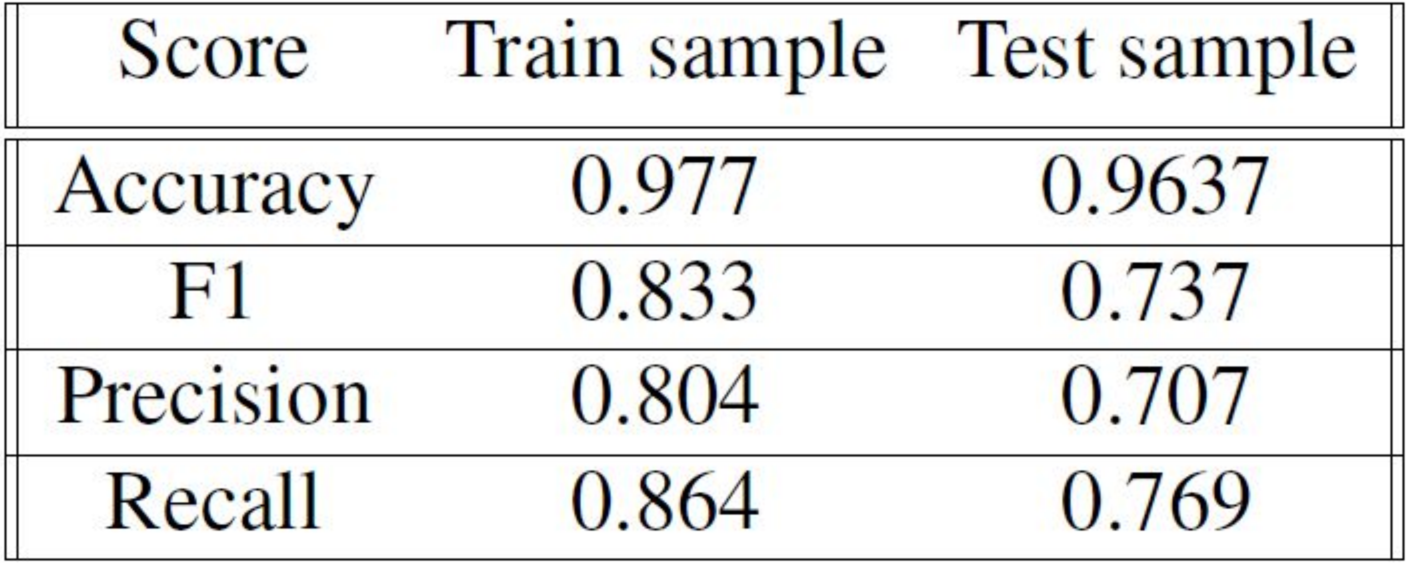
Added Features

Last transaction features + aggregated transaction metrics

Model

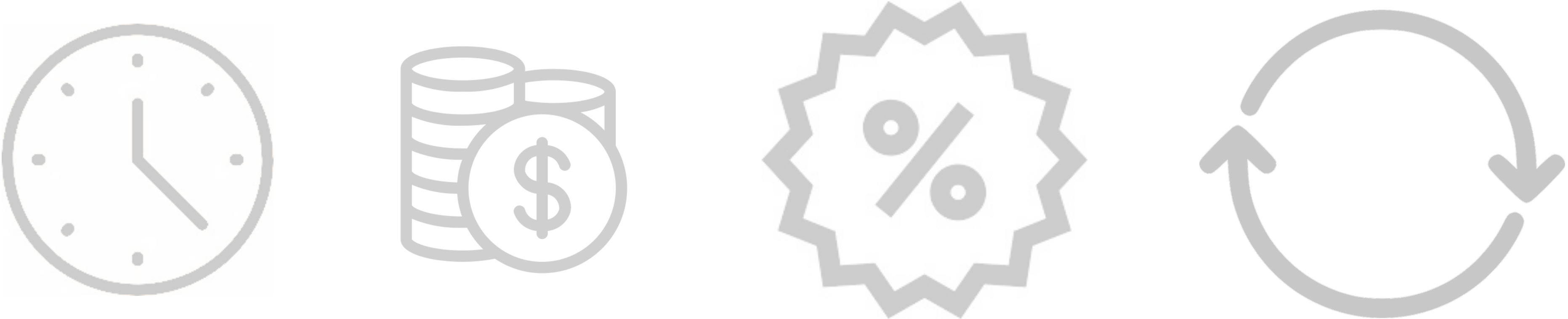


Results



Score	Train sample	Test sample
Accuracy	0.977	0.9637
F1	0.833	0.737
Precision	0.804	0.707
Recall	0.864	0.769

Most important variables



Questions?