Rizky Luthfianto

Dana Customer Churn Prediction

Given:

- training set of data of three months (Nov-Jan)
- Test set of data contains user_id and their last_trx, some transacted on february
- Attrition cutoff is 20 days

We want to:

- predict which users are going to churn on 1st of march

Dataset shape:

nov			dec			jan			feb			mar
1-10	11-20	21-30	1-10	11-20	21-31	1-11	12-20	21-31	1-10	11-20	21-28	1

The challenges:

- Dataset is order_id granularity
  - -> need to aggregate to user_id granularity
- Need to construct the label is_churn
  - -> will use the 12-31 January as a label because our attrition cutoff is 20 days. If user transacted on this period, we will label them as not_churn, otherwise we label them as churned
- There is a gap between in february where we don't know the user activity
- Each user has different known last transaction date, while we have to predict their probability on March 1st
  - -> Those two points are why I think we cannot just fit the dataset with the traditional binary classification approach, and have to look for a better approach. If we use plain logistic regression, we can only predict the proba for the next 20 days (12-31 jan) which does not answer our question.

Hence, I reformulate the problem into:

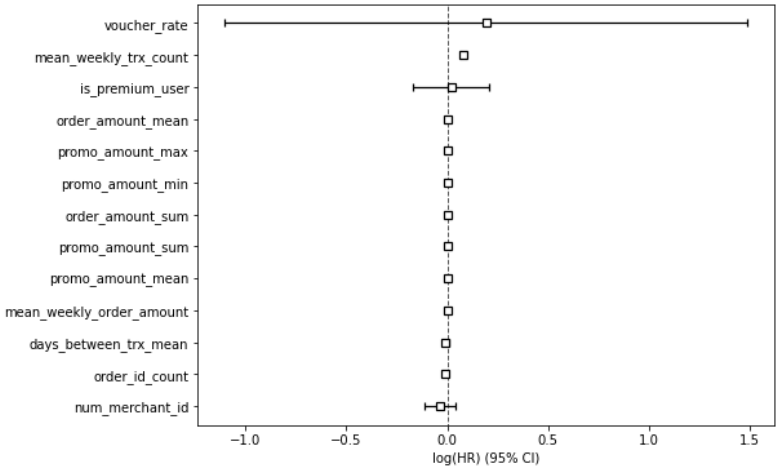
- Predicting the survival rate of customers, because it will give us the probability of their survival up to 90 days.
  We will look at **Day 48**, which is the difference between March 1st and January 11, because the survival model is trained with the label between 11 Jan and 31 Jan,
  we will ignore the last_trx_date on the test set

Feature engineering:

- is_premium user
- num_merchant_transacted
- promo_rate (num of trx with promo / num of all trx)
- promo_amount_{max, min, mean}
- Tenure (customer length of relationship)

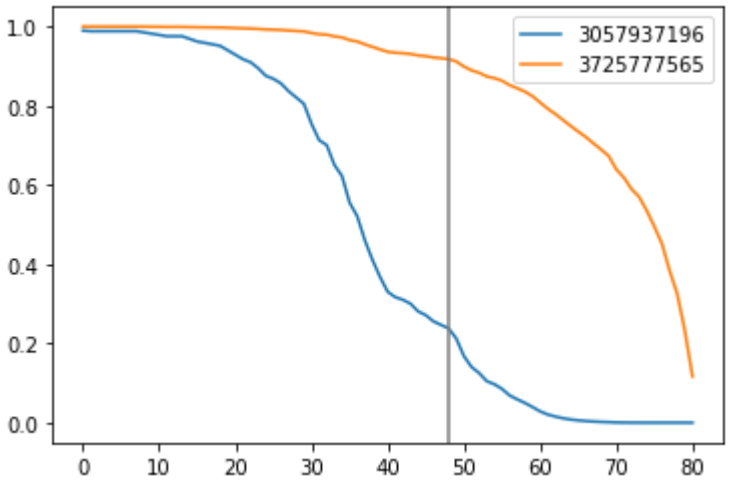
- features based on RFM:
    - Frequency:
        - mean days between order
        - mean weekly order amount
    - Monetary:
        - Mean order amount

Model Fitting Result



Here we see the coefficients. Voucher_rate & is premium_user has a large interval. Meanwhile num_weekly_trx_count is our strongest predictor, because of its small interval

Survival plot of a churned user and not churned user:



We see that on Day 48 (March 1st) the user 3057937196 that is not churned has a better survival rate than the user 3725777565 who is churn.
I conclude with Survival Analysis, we could predict who will churn on March 1st.

If there is more time to do this research, it would be nice to dig more on these:

- Better problem reformulation
- Better feature engineering: explore more features based on RFM, CLV, trends
- Feature selection. Remove features who are not strong predictors.
- More rigorous model evaluation and testing