Capstone Project-1 Proposal

Project Name: WSDM - KKBox's Churn Prediction Challenge

Can you predict when subscribers will churn?

Proposal Details:

1. What is the problem you want to solve?

Predict whether a user will churn after his/her subscription expires. Specifically, we want to forecast if a user make a new service subscription transaction within 30 days after the current membership expiration date.
Looking at the relationship between cancelled subscriptions and churn rate.

2. Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

The client is Wisdom KKBox who is Asia's leading music streaming service, holding the world's most comprehensive Asia-Pop music library with over 30 million tracks. They offer a generous, unlimited version of their service to millions of people, supported by advertising and paid subscriptions. This research is to predict whether a user will continue with subscription or not, From this research they will know:-

- Create a model for accurately predicting churn of paid users.
- Discover new insights to why users leave and retain them for longer periods.

3. What data are you using? How will you acquire the data?

The key fields to determine churn/renewal are transaction date, membership expiration date, and is_cancel. Note that the 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. The Data is available in this Kaggle competition.

The Overview of Data Features is as follows:-

Train.csv

the train set, containing the user ids and whether they have churned.

- msno: user id
- is_churn: This is the target variable. Churn is defined as whether the user did not continue the subscription within 30 days of expiration. `is_churn = 1` means churn,`is_churn = 0` means renewal.

Transactions.csv

- msno: user id
- 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.

Members.csv

user information. Note that not every user in the dataset is available.

- msno
- city
- bd: age. Note: this column has outlier values ranging from -7000 to 2015, please use your judgement.
- gender
- registered_via: registration method
- registration_init_time: format `%Y%m%d`
- expiration_date: format `%Y%m%d`, taken as a snapshot at which the member.csv is extracted. Not representing the actual churn behavior.

4. Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

The first goal is to exhaustively determine the relationship between different variables

1. Data Wrangling will occur to clean the data as per requirements and check for any inconsistencies like missing data, balanced /Imbalanced data and fix it.
    - Join of three(train.csv, members.csv, transactions.csv) tables based on msno:user id
    - Transaction_date, Membership_expire_date,Registration_init_time formatting needs to be done.
    - Needs to investigate the feature "payment_plan_days" with 0 value, if plan_days are zero it means no more days left in plan so why there is some value for corresponding actual_amount_paid.
    - Feature"members age" needs to deal with outliers.
2. Exploratory Data Analysis(EDA) will occur to check for possible trends and/or correlations between different characteristics/Features
    - Trends/correlations between features :-

      is_churn and is_cancel, member's age and is churn/is_cancel, payment_method and age/auto_renewal.

3. Statistical Inferences
    - Hypothesis test - we will be using T-test to test the hypothesis.
4. Final model building and testing
    - Determine Accuracy using ROC curves/Confusion Matrix.

5. What are your deliverables? Typically, this includes code, a paper, or a slide deck.

The Final Products will be:-

- Code Ipython Notebook
- Final Report
- Slide Deck