The Problem:
The modern-day lifestyle has turned into one that revolves around the subscription business model. Nowadays, you can get fresh ingredients delivered to your door, clothes to update your wardrobe, movies and TV shows to watch across different online devices, and even access to a variety of music to listen to during long commutes; all for a reasonable recurring monthly fee. The critical business question is no longer just "How will I find customers for my product?" but will also include "How will I retain my customers?"

KKBOX, Taiwanese subscription based music streaming service, shared their customer data on Kaggle to learn more about predicting churning. When users signs up for our service, users can choose to either manually renew or auto-renew the service. Users can actively cancel their membership at any time. For this project, the criteria of "churn" is no new valid service subscription within 30 days after the current membership expires.

The Inquiry:
Can we predict if a user will make a new service subscription transaction within 30 days after the current membership expiration date based on their behaviors and interactions with the product?

The Client:
The client is KKBOX. KKBOX will be able to use this information to offer special deals to customers who are likely to churn or on the verge of churning.

The Data:
KKBOX has provided 4 datasets:

1. Train Set, contains the user ids and whether they have churned for March 2017
    a. msno: user id
    b. 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.
2. Transactions Data, contains the transactions data until 3/31/2017
    a. msno: user id
    b. payment_method_id: payment method
    c. payment_plan_days: length of membership plan in days
    d. plan_list_price: in New Taiwan Dollar (NTD)
    e. actual_amount_paid: in New Taiwan Dollar (NTD)
    f. is_auto_renew
    g. transaction_date: format %Y%m%d
    h. membership_expire_date: format %Y%m%d
    i. is_cancel: whether or not the user canceled the membership in this transaction.
3. User Logs Data, contains listening behaviors of a user until 3/31/2017
    a. msno: user id
    b. date: format %Y%m%d

    c. num_25: # of songs played less than 25% of the song length
    d. num_50: # of songs played between 25% to 50% of the song length
    e. num_75: # of songs played between 50% to 75% of of the song length
    f. num_985: # of songs played between 75% to 98.5% of the song length
    g. num_100: # of songs played over 98.5% of the song length
    h. num_unq: # of unique songs played
    i. total_secs: total seconds played

4. Members Data, contains user information
    a. msno
    b. city
    c. bd: age. Note: this column has outlier values ranging from -7000 to 2015, please use your judgement.
    d. gender
    e. registered_via: registration method
    f. registration_init_time: format %Y%m%d

The Method:

1. Explore the data provided to get an idea of the data (missing values, outliers, etc)
2. Combine different columns from different datasets to see which dependent variables affect the churn rate
3. Use machine learning to create an algorithm that predicts the probability of churning

The Deliverables:

This project will include the code on Github and a report outlining the business case