December 10th, 2017

MIS-637-A (Fall 2017)

PROJECT PROPOSAL – KKBOX STREAMING SERVICE CUSTOMER CHURN

PROBLEM STATEMENT

KKBox (herein referred to as "the client") has been at the forefront of the Asian music streaming market from when it was founded in 2004 up until 2014 when YouTube and Apple entered their market. The client is experiencing increasingly disturbing customer attrition and needs to put a handle on it before irreparable damage is done to their bottom line.

The client has published a sample of anonymized customer information, and one month's worth of transactions and listening behavior. They want valuable information on customer churn mined from this data to aid their decision-making going forward.

SAMPLE DATA - Link

Customer information: 6 features (user-id, city code, age, gender, registration channel, registration time) and 795,091 rows of observations.

members.csv

```
# As the name implies, this csv holds subscriber information

members = pd.read_csv('members_v2.csv')
members.head()
```

	msno	city	bd	gender	registered_via	registration_init_time
0	+tJonkh+O1CA796Fm5X60UMOtB6POHAwPjbTRVl/EuU=	1	0	NaN	7	20110914
1	WFLY3s7z4EZsieHCt63XrsdtfTEmJ+2PnnKLH5GY4Tk=	6	32	female	9	20110915
2	l0yFvqMoNkM8ZNHb617e1RBzlS/YRKemHO7Wj13EtA0=	13	63	male	9	20110918
3	OoDwiKZM+ZGr9P3fRivavgOtglTEaNfWJO4KaJcTTts=	1	0	NaN	7	20110918
4	4De1jAxNRABoyRBDZ82U0yEmzYkqeOugRGVNlf92Xb8=	4	28	female	9	20110920

Transactions: 9 features (user-id, payment method, payment plan, plan price, amount paid, auto renew or not, transaction date, membership expiration date, is a cancel or not) and 1,431,010 rows.

transactions.csv

```
trans = pd.read_csv('data_files/transactions_v2.csv')
trans.head()
```

	msno	payment_method_id	payment_plan_days	plan_list_price	actual_amount_paid	is_auto_renew	transacti
0	++6eU4LsQ3UQ20ILS7d99XK8WbiVgbyYL4FUgzZR134=	32	90	298	298	0	2
1	++lvGPJOinuin/8esghpnqdljm6NXS8m8Zwchc7gOeA=	41	30	149	149	1	2
2	+/GXNtXWQVfKrEDqYAzcSw2xSPYMKWNj22m+5XkVQZc=	36	30	180	180	1	2
3	+/w1UrZwyka4C9oNH3+Q8fUf3fD8R3EwWrx57ODIsqk=	36	30	180	180	1	2
4	+00PGzKTYqtnb65mPKPyeHXcZEwqiEzktpQksaaSC3c=	41	30	99	99	1	2

User Logs: Logs of daily listening behavior for one month 9 features (user-id, date, number of songs listened quarter of the way, 50% of the way, 75% of the way, 98.5% of the way, 100% of the way, number of unique songs listened to that day, total seconds of songs consumed that day) and 18,396,363 rows.

user_logs.csv

```
logs = pd.read_csv('data_files/user_logs_v2.csv')
logs.head()
```

	msno	date	num_25	num_50	num_75	num_985	num_100	num_unq	total_secs
0	u9E91QDTvHLq6NXjEaWv8u4QIqhrHk72kE+w31Gnhdg=	20170331	8	4	0	1	21	18	6309.273
1	nTeWW/eOZA/UHKdD5L7DEqKKFTjaAj3ALLPoAWsU8n0=	20170330	2	2	1	0	9	11	2390.699
2	2UqkWXwZbIjs03dHLU9KHJNNEvEkZVzm69f3jCS+uLI=	20170331	52	3	5	3	84	110	23203.337
3	ycwLc+m2O0a85jSLALtr941AaZt9ai8Qwlg9n0Nql5U=	20170331	176	4	2	2	19	191	7100.454
4	EGcbTofOSOkMmQyN1NMLxHEXJ1yV3t/JdhGwQ9wXjnI=	20170331	2	1	0	1	112	93	28401.558

METHODOLOGY

Following the CRISP-DM, with descriptive discovery run on the individual datasets and combined, null value imputation methods applied accordingly, I plan to run logistic regression as the benchmark algorithm and compare results with Random Forest and, if required, Stochastic Gradient Descent on the combined dataset to produce the probability of churn. I will then measure the log loss of my results against the client's prediction benchmark.

SOFTWARE

I will explore the data with a jupyter notebook running Python 2.7 and Tableau, combine the data with Alteryx, and train the respective algorithms using the scikit-learn data science package.