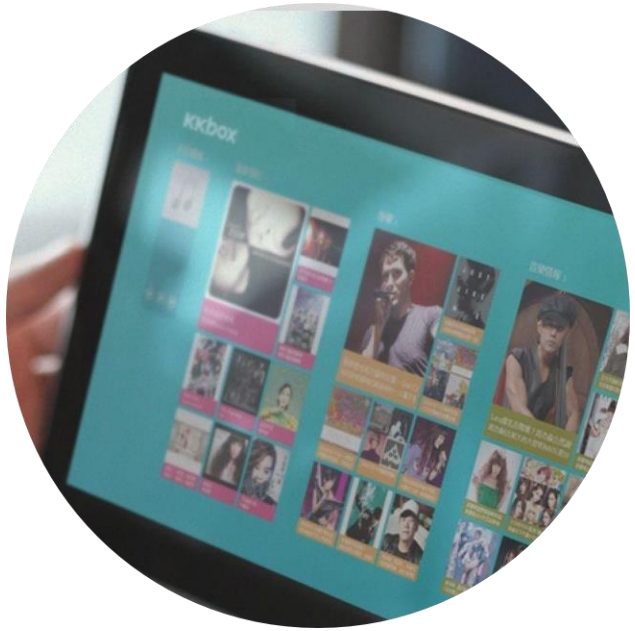
User Churn Prediction

KKBOX Music Streaming Subscription



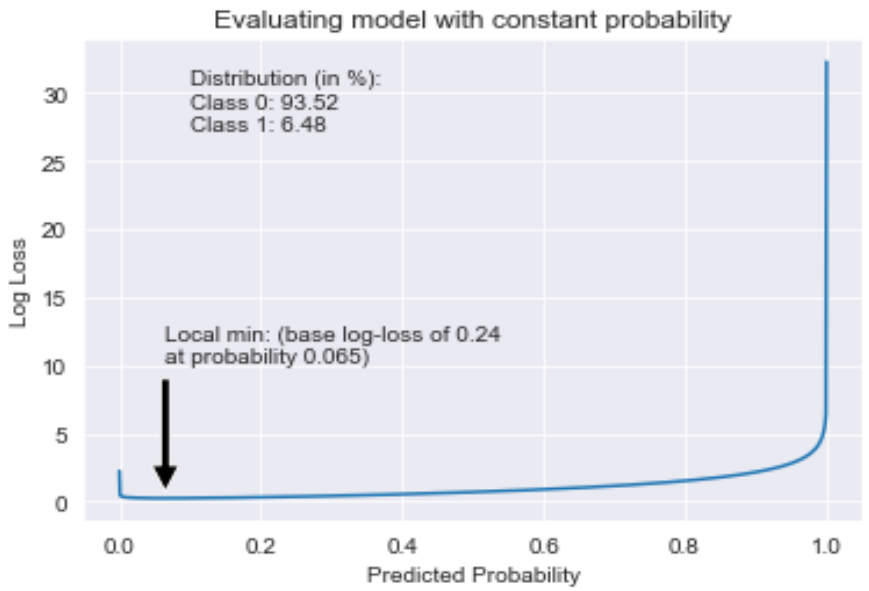
Gaurav Dembla

Feb 28, 2020

Problem Statement

Predict whether a user will churn after his/her subscription expires

- KKBOX is Taiwan's leading music streaming service, offering unlimited version to millions of people, supported by advertising and paid subscriptions
- Analysis to be done on user registration, sales transaction and daily usage data
- Binary classification problem with success criteria as log-loss score on unseen data (hold-out/test set) – should be better than baseline score of 0.24



Source Data

Four distinct datasets:

a) **Labeled** data – churn behavior of users during Mar' 17. Size/shape: 1mn x 2.

b) **Registration** data – basic service membership and demographic information about users, such as gender, age and registration method/source. Size/shape: 7mn x 6.

c) **Sales transaction** data – information about subscription transactions, whether a transaction was renewal or cancellation of service, amount paid if renewal, whether auto-renewal option was chosen by the user, and the length of membership plan selected. Size/shape: 22mn x 9.

d) **Usage logs** – daily listening behavior of a user, such as number of songs played once, replayed, skipped or completed, and the amount of time songs played during a day. Size/shape: 400mn x 9.

Source: Kaggle

Data Extraction

Issue

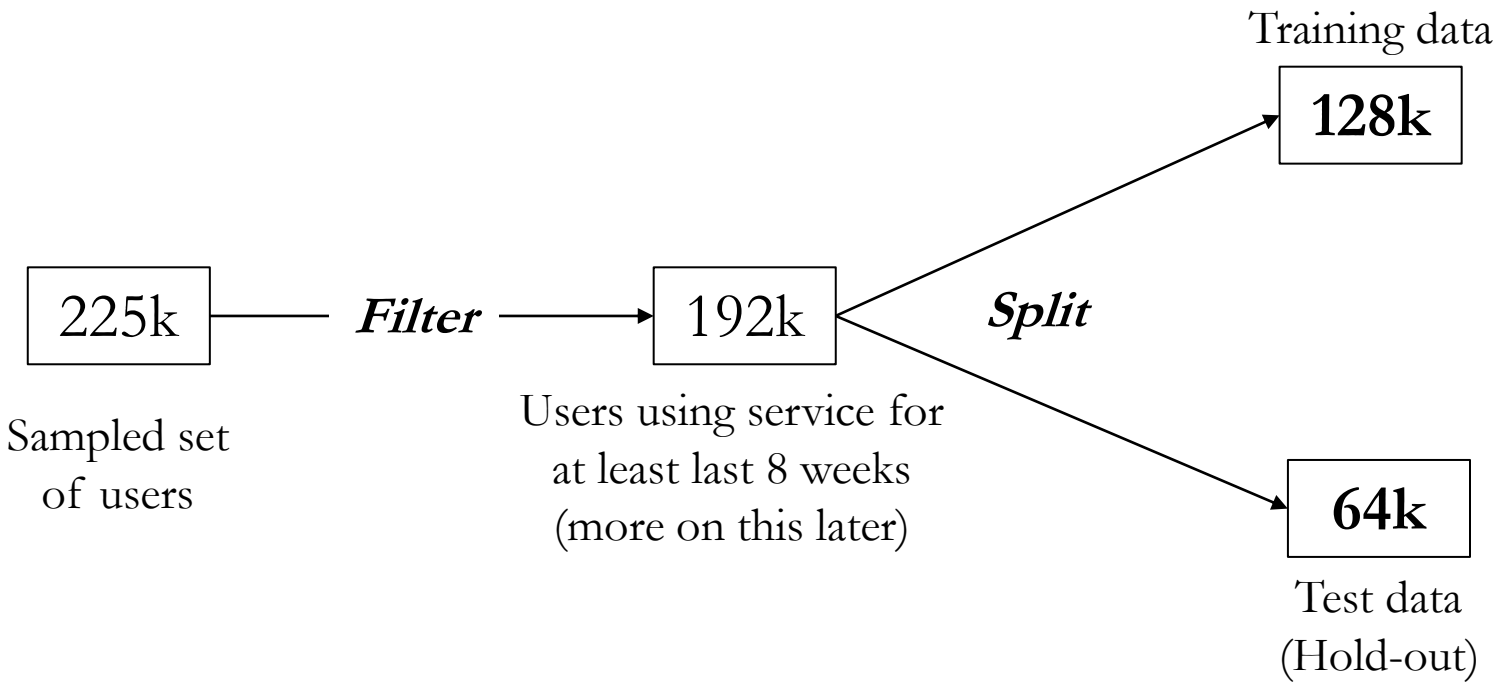
- 1mn users with 400mn records of raw usage records
- Limited computing power and memory resources on the available local machine

Solution

- Extract random sample **225k** users with manageable set of 55mn usage records

	Labeled	Registration	Sales	Usage
# records	225,000	198,761	3,601,833	55,668,485
# users	225,000	198,761	225,000	197,142

Data Manipulation



Data Manipulation

Raw Data

Aggregated Data (user level) with Metrics

Denormalized Data

Sales

user	churn	payMethod	autoRenew	cancel	trxDate
JYfXE	0	41	1	0	2/14/2016
JYfXE	0	41	1	0	3/14/2016
AnU3d	1	33	1	0	1/17/2016
AnU3d	1	33	1	0	2/17/2016
AnU3d	1	33	1	1	3/17/2016

user	churn	payMethod	autoRenew	cancel	daysLastTrx
JYfXE	0	41	1	0	351
AnU3d	1	33	1	1	348

(last transaction)

Usage

user	churn	date	songs
JYfXE	0	2/19/2017	3
JYfXE	0	2/24/2017	5
JYfXE	0	2/27/2017	11
JYfXE	0	2/28/2017	11
AnU3d	1	1/10/2017	5
AnU3d	1	1/11/2017	3

user	churn	songsPerActiveDay	daysLastLogin
JYfXE	0	3	1
AnU3d	1	4	38

(lifetime usage)

Registration

user	churn	gender	regSrc	city	regDate
JYfXE	0	female	7	6	11/5/2016
AnU3d	1	male	4	13	9/17/2016

user	churn	gender	regSrc	city	daysSinceReg
JYfXE	0	0	7	6	115
AnU3d	1	1	4	13	164

user	churn	daysLastLogin	autoRenew	cancel	daysSinceReg
JYfXE	0	1	1	0	115
AnU3d	1	38	1	1	164

Data Cleaning

Missing/null values –

1) **Gender** values missing for more than 50% of records. For remaining 50%, churn behavior of either of the genders was same (~8%) -> no potential importance in predicting churn behavior -> attribute dropped.

Unreasonable values –

2) **List price** less than paid amount for 100k out of 2.4mn sample sales transactions. For rest, meagre 300 odd users (out of 128k) availed non-zero discount in their last transaction -> attribute dropped.

3) **Age** of users outside reasonable range (18-60) for more than 50%. For remaining 50%, no potential importance in predicting churn behavior -> attribute dropped.

4) **Playtime** beyond conceivable range (0-86,400 seconds/day) for negligible amount of usage records, while having significant effect on the mean value (from 8,037 to 306,050 seconds). Playtime highly correlated with total number of songs completed -> drop the former, use the latter in its stead.

Exploratory Data Analysis

Auto-renew and Cancellation

Do renew users tend to stay on subscription auto-renewal feature more than churn users?

		Churn		
cancel	**autoRenew**	No	Yes	**churnRate**
No	No	11,229	5,549	33.1%
No	Yes	128,559	475	0.40%
Yes	Yes	565	3,623	86.5%
			150,000	

Insights

a) If the user is not on an auto-renewal plan, there is roughly 33% chance that the user will churn.

b) If the user is on an auto-renewal plan that has not been cancelled yet, he is most likely to continue with the service (~0.4% chance of churning) and let the service plan renew automatically next month (Mar' 2017).

c) If the user cancelled an auto-renewal plan, he is highly likely to leave the service and not come back to renew it the next month (~86% chance of churning).

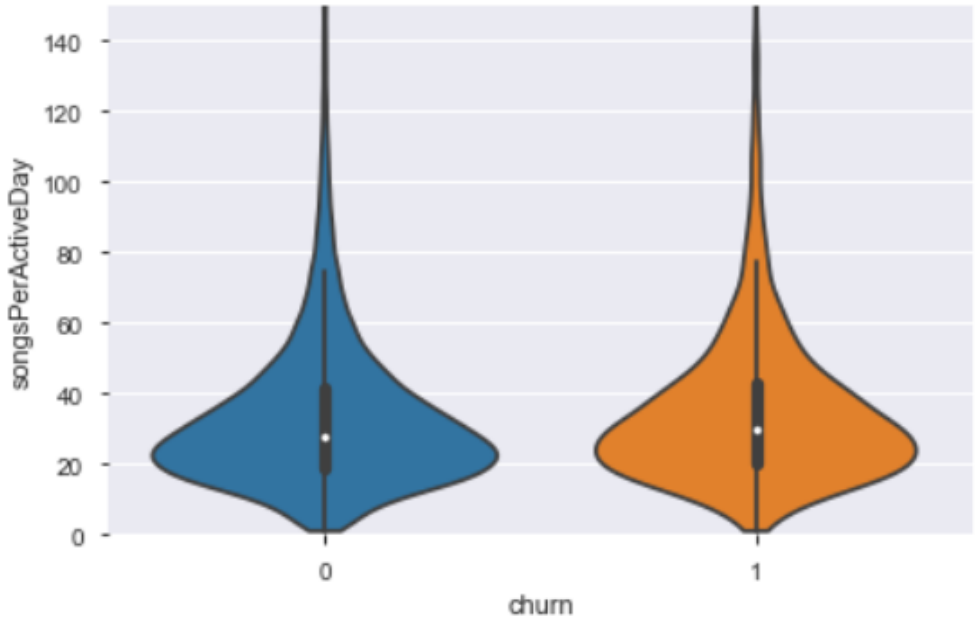
Lifetime Usage Trends

How did an average churn user behave in his lifetime compared to an average renew user?

Exploratory Data Analysis

Songs per active day

Does a renew user, on average, listen to more songs on an active day than a churn user?

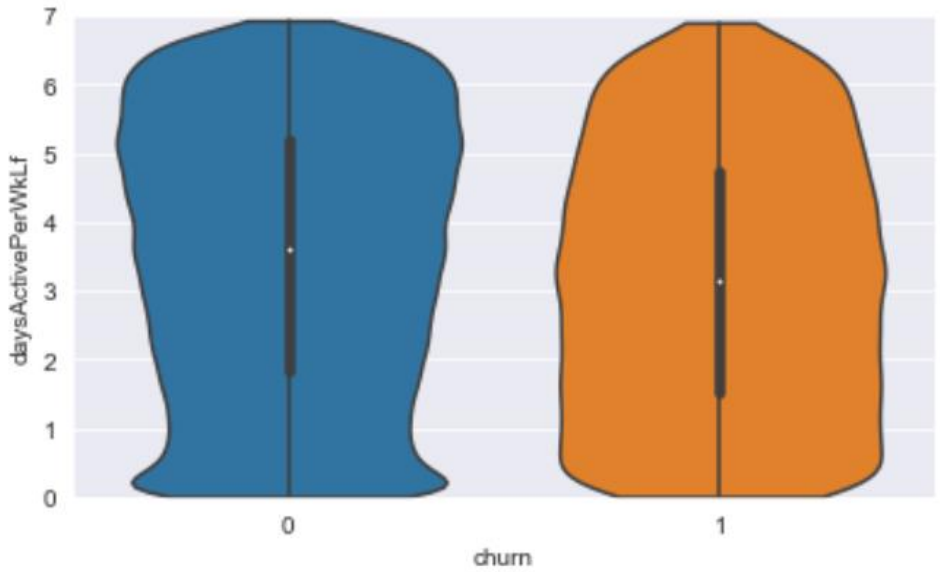


Insight: On an active day, both renew and churn users listen to same number of songs, on an average. Same for average number of songs played-once, replayed, completed or skipped on an active day.

Exploratory Data Analysis

Days active per week

Does a renew user, on average, use the service more often during his lifetime than a churn user?



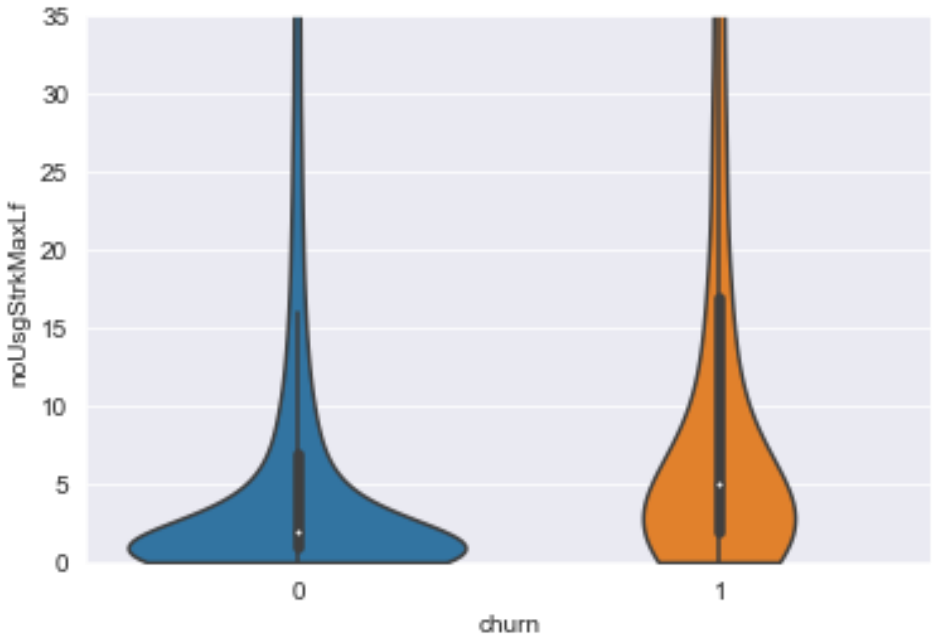
Insight: Churn users *might* be logging in the service (app) slightly less over their lifetime than renew users

Exploratory Data Analysis

Maximum no-usage streak (in weeks)

Do churn users stay away from the service for longer periods (weeks in succession), on average, than renew users?

Longest successive number of weeks user did not use the service for, during his lifetime



Insight: Maximum streak of no usage (of service) by churn users is more than that by renew users

Exploratory Data Analysis

Imagine you have a Netflix, Spotify or Amazon Prime (Video/Music) subscription account

- Are you planning to stop the subscription to any service?
- If yes, do you see a relation between your departure decision and your recent usage behavior?
- Have you already stopped using the service completely?
- If not, has your usage declined these days, compared to what it used to be earlier?

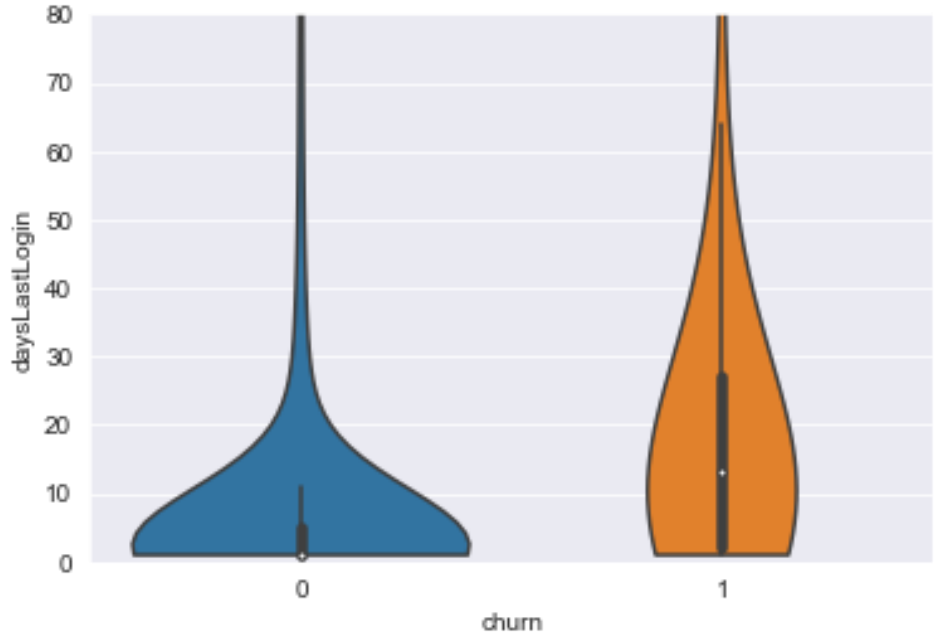
Recent Usage Trends

Did churn users behave differently from renew users in recent times?

Exploratory Data Analysis

Days since last login

Did churn users last log in and use the service lot further ago than renew users?

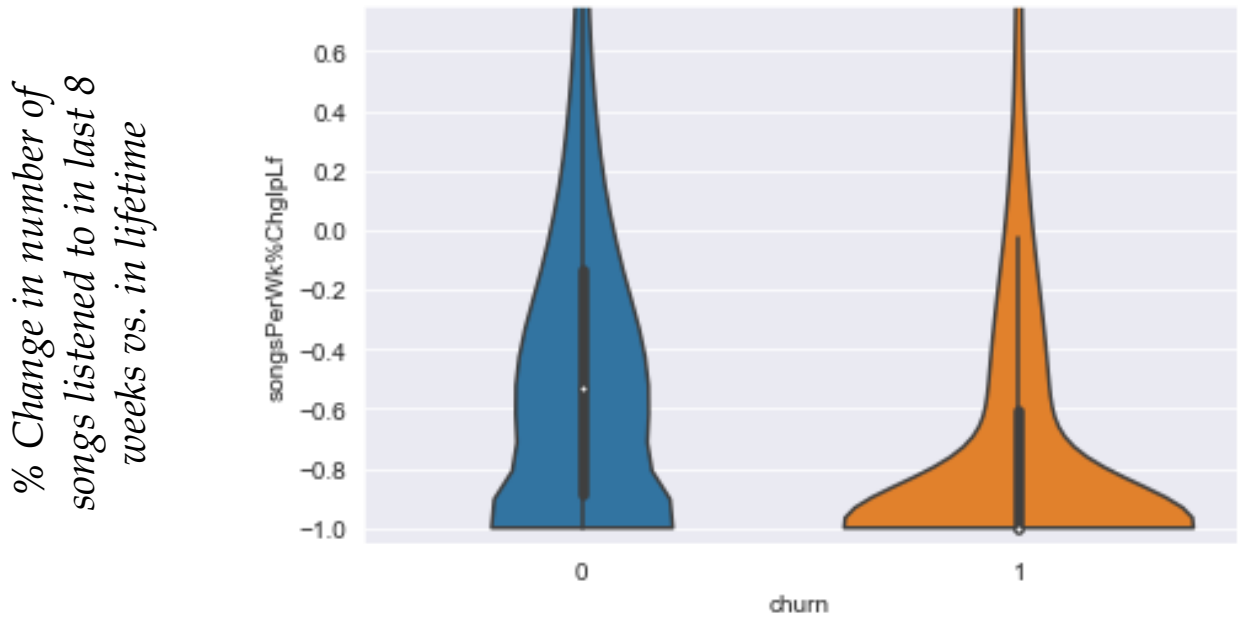


Insight: Churn users have not logged into the service for longer time than renew users, as of present day (Mar' 17). The median number of days since last login for churn and renew users is 13 and 1 days respectively.

Exploratory Data Analysis

Recent usage vs. lifetime usage

Has usage of churn users declined more than renew users in recent past compared to lifetime?



Insight: Usage of churn users in last 2 weeks has declined by a much higher degree than that of renew users

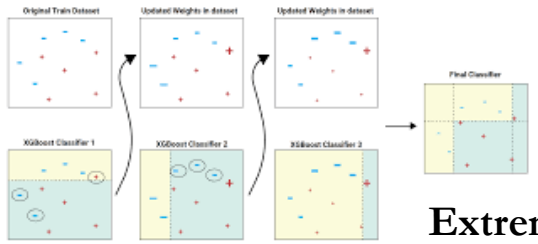
! Hence, the need to limit the scope to only the users who started using the streaming service at least eight weeks back, as shown in Data Manipulation slide (# 5).

Statistical Modeling

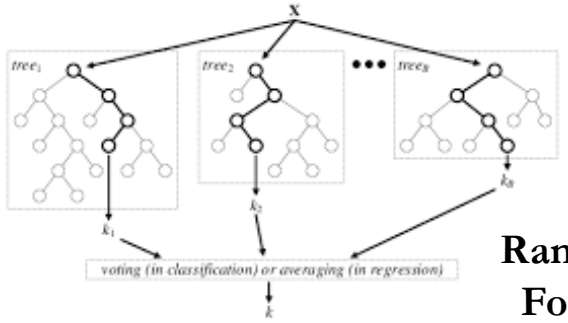
Feature Engineering

26 features normalized within a range of 0 and 1

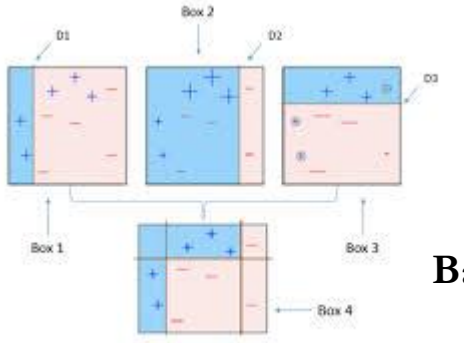
Statistical Models



Extreme Gradient Boosting (XGB)



Random Forest



Bagging

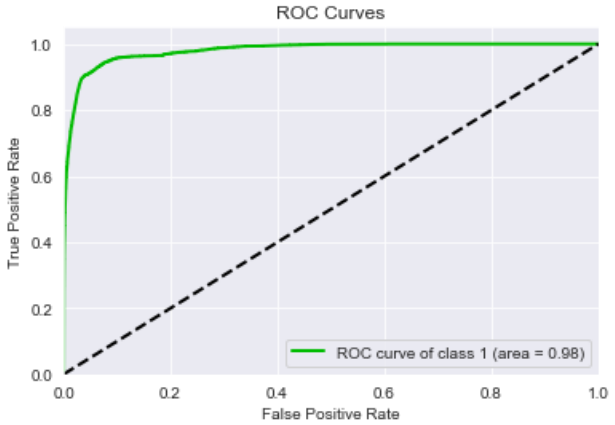
Hyperparameter tuning using Stratified 5-Fold cross validation

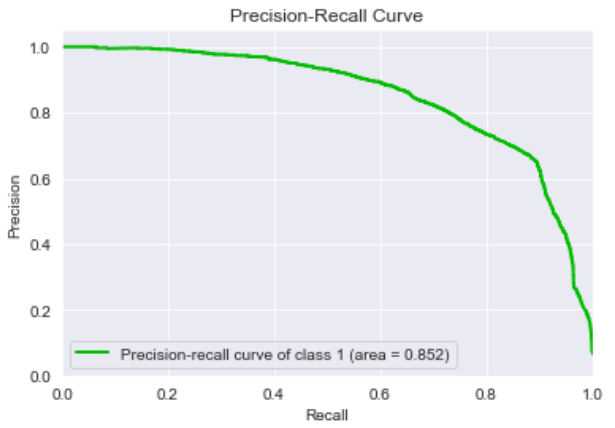
Model Evaluation

Base Log-Loss	0.241		
	XG Boosting	Random Forest	Bagging
Log-Loss	0.0752	0.0806	0.1014
Precision	83.1%	83.4%	83.5%
Sensitivity	68.7%	67.8%	64.8%
ROC AUC	0.98	0.976	0.972

✓ **XG Boosting**

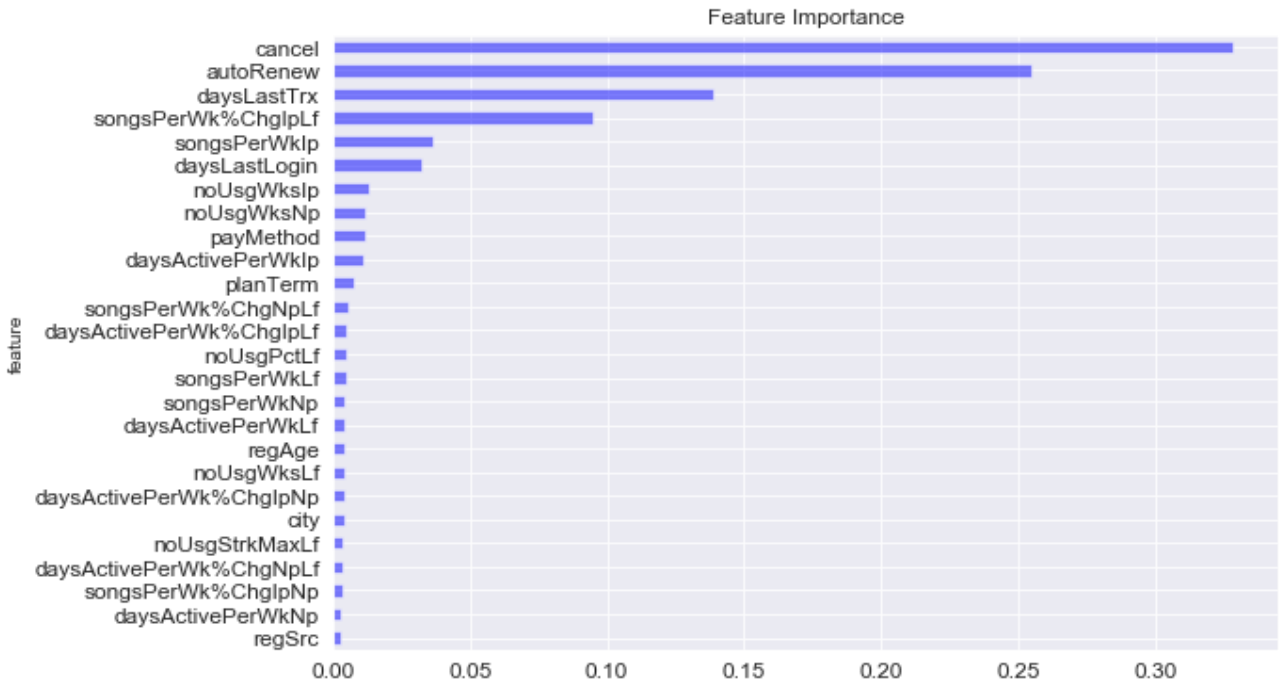
- Outperformed Random Forest and Bagging
- Beat baseline log-loss score of 0.24
  - Relatively complicated to comprehend





ROC-AUC = ~0.98

Feature Importance



S. No.	Data Category	Feature	Importance	Rank
1	Last Sales Transaction	cancel	0.3286	1
2		autoRenew	0.2547	2
3		payMethod	0.0115	9
4		planTerm	0.0076	11
5		daysLastTrx	0.1384	3
6	User Registration	regAge	0.0039	18
7		city	0.0037	21
8		regSrc	0.0025	26
9	Lifetime Usage	noUsgWksLf	0.0039	19
10		noUsgPctLf	0.0048	14
11		noUsgStrkMaxLf	0.0037	22
12		daysActivePerWkLf	0.0040	17
13		songsPerWkLf	0.0047	15
14	Recent Usage	daysLastLogin	0.0325	6
15		noUsgWksIp	0.0128	7
16		daysActivePerWkIp	0.0110	10
17		songsPerWkIp	0.0367	5
18		noUsgWksNp	0.0115	8
19		daysActivePerWkNp	0.0029	25
20		songsPerWkNp	0.0041	16
21	Recent vs. Longer-term Usage	daysActivePerWk%ChgIpNp	0.0038	20
22		songsPerWk%ChgIpNp	0.0033	24
23		daysActivePerWk%ChgIpLf	0.0049	13
24		songsPerWk%ChgIpLf	0.0950	4
25		daysActivePerWk%ChgNpLf	0.0035	23
26		songsPerWk%ChgNpLf	0.0058	12

Top 10 features highlighted in green

Conclusion

Key Takeaways

1) It's paramount that a user stays on an auto-renewal plan, as the probability of continuing with the service on auto-renewal plan is very high. If the user cancels an auto-renewal plan, then he is likely to leave.
2) It's important that a user logs into the application and uses it every now and then – couple of unused weeks in succession poses risk of churning.

Future Scope

1) Use Spark and AWS to process the entire dataset available (for 1mn users) and not limit to a sample of 198k users.
2) If KKBOX supplied information about user's profile such as number of playlists stored, number of songs favorited or saved in playlists, etc., prediction capability could be further improved, as they seem to be important factors of customer's stickiness to a music streaming service.

Useful Links

Project Repository

Detailed Report

Codebase (Jupyter Notebooks)

Intuition behind Log-loss score