Music Box Project

BITTIGER

The Lifelong Learning Platform of Silicon Valley

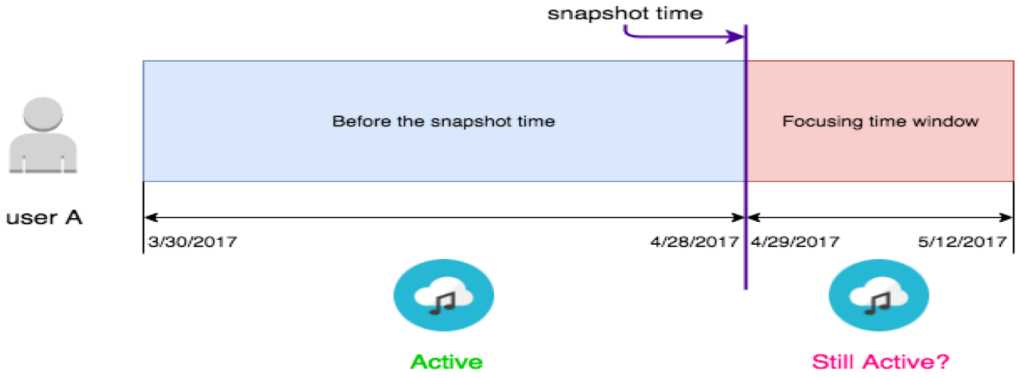
Outline

- Churn Prediction Overall Description
- Data Preprocessing
- Feature Engineering
- Modeling Design
- Performance Evaluation
- Expected Commercial Value

Churn Prediction

Target: Build an churn prediction model based on user's behavior from using our music box product

We want to know whether user A is still active during our focused time period who once has activities before the snapshot time?



Data Preprocessing

BITTIGER

- Raw Data Description
  - Only play log data included in my project
  - Data size: 14.1 G
  - Attributes: uid, device, song id, play time, song length, song name, singers

- Data Exploration
  - 871702 Users, 164,667,143 play records
  - Time series: 2018-03-01 to 2018-03-09; 2018-03-29 to 2018-05-12

- Platform: Local Computer (Macbook pro with 16G RAM)
  - Programming Language: python
  - Package: Spark(python based), pandas, scikit-learn, keras

Feature Engineering

BITTIGER

- Useful attributes: uid, date, song id, play time, song length

- Data Cleaning
  - Remove records satisfying any of the following:
    - Any attributes Including null values
    - Uid, song id, play time, song length including characters
    - Play time is larger than song length

- Feature Design
  - A total of 11 features from 3 categories
    - frequency on play log(last 1,3,7,14,30 days)
    - Recency
    - Play time percentage per song(last 1,3,7,14,30 days)

Modeling Design - (I)

BITTIGER

- Design the target(label)
  - Snapshot date: 2017-04-29
  - All time window: 2017-03-29 to 2017-05-12
  - Focusing time window: 2017-04-29 to 2017-05-12
  - potential churners: the users who have play activities before the snapshot date but no activity during the focusing time window(otherwise, it could be seen as the potential loyaltees)

- Balance the data
  - Original data
    - loyaltees : churners = ~200k: ~330k (38%: 62%) [churn rate: 39%]
  - Balanced data
    - loyaltees : churners = ~200k: ~200k (50%: 50%)

Modeling Design - (II)

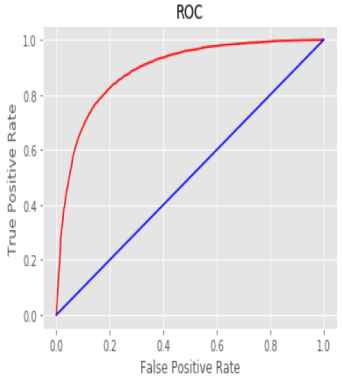
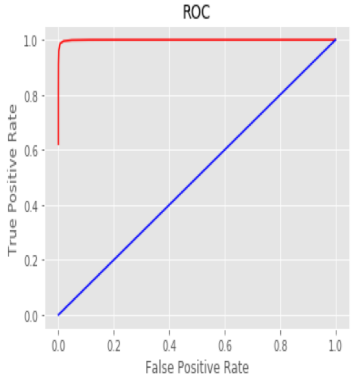
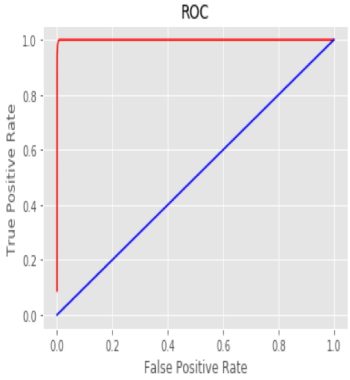
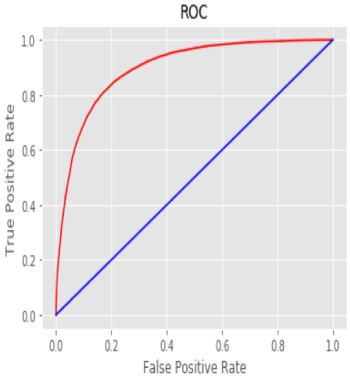
BITTIGER

- Model Selection

Model	Logistic Regression	Random Forest	RF with Tuning	Neural Network
Hyperparameter Setting	C = 0.1 L2 penalty	N_estimator = 10	N_estimator = 300, max_depth = 30	11->8->4->1

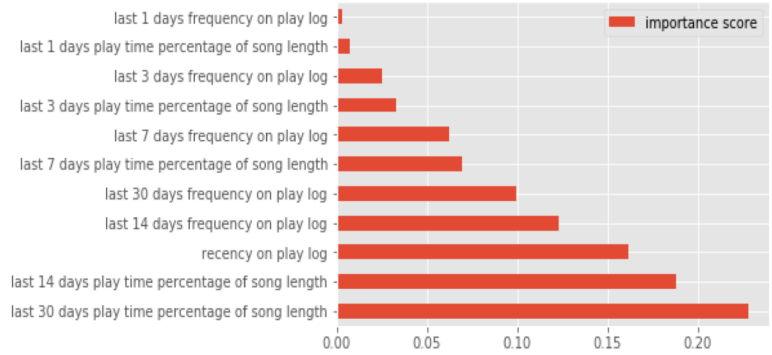
- Hyperparameter Tuning
  - Training data sets : Testing data sets = 80% : 20%
  - 5-fold cross validation on training sets for hyperparameter tuning.

Performance Evaluation

			
Logistic Regression	Random Forest	RF with Tuning	Neural Network
AUC: 0.7159	AUC: 0.8422	AUC:0.8564	AUC: 0.8198

Expected Commercial Value

- Top 3 features influencing the churn
  - last 30 days play time percentage of song length
  - last 14 days play time percentage of song length
  - recency



- Suggestions on retaining users
  - Send push notifications to users with high churn possibility, i.e. users who don't have any play activity for 14 days.
  - Recommend potential favorite songs to users, especially for those whose play time percentage per song has decreased significantly for last 14 or 30 days.

Thank you for your watching!

BITTIGER

The Lifelong Learning Platform of Silicon Valley