用戶流失指標與解決方案

Customer Churn Analysis and Prediction

涂瀞匀 Jean Tu

2022.01

資料分析流程

定義商業問題
Define Business Issue

資料準備
Data Preparation

資料前處理
Data Preprocessing

數據分析與預測
Data Analyze & Prediction

結果應用
Conclusion

Business Issue

Why Customers Churn ?

Internal Factors	External Factors
Plan Price	Spotify
Rental days	YouTube Premium
Use rage	Apple Music
Service Quality	Line Music

User Data

Machine Learning Algorithms

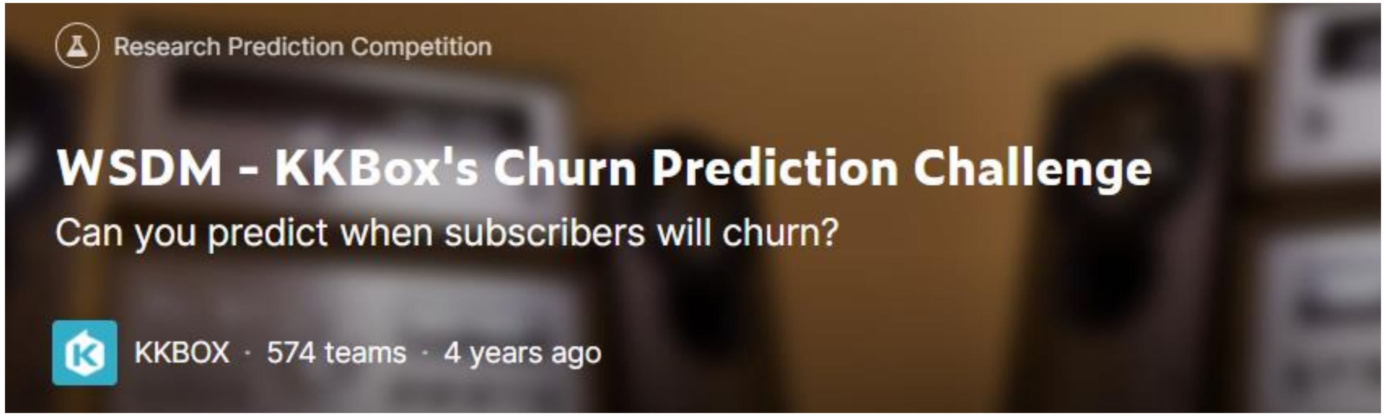
是否能用過去用戶的**行為數據**

找出**關鍵特徵**以**預測**用戶流失機率

針對**退租高風險**用戶進行續留

Data Preparation

kaggle



Data Explorer

8.95 GB

- WSDMChurnLabeller.scala
- members_v3.csv.7z
- sample_submission_v2.csv.7z
- sample_submission_zero.cs...
- train.csv.7z
- train_v2.csv.7z
- transactions.csv.7z
- transactions_v2.csv.7z
- user_logs.csv.7z
- user_logs_v2.csv.7z

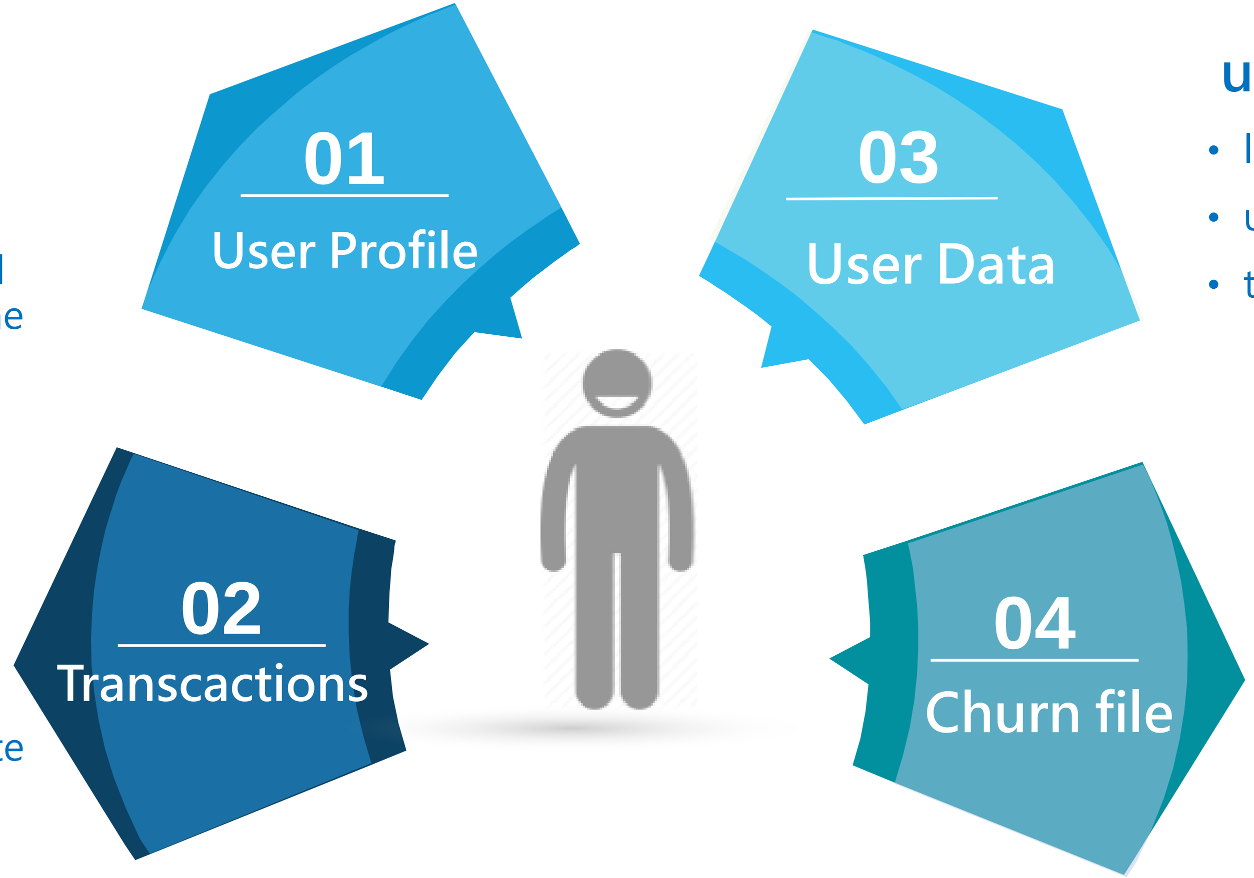
Glance at all tables

member.csv

- gender
- age
- city
- registration method
- registration_init_time

transcactions.csv

- payment method
- membership plan
- plan_list_price
- actual_amount_paid
- auto_renew
- membership_expire_date
- is_cancel



user_logs.csv

- listening behaviors
- unique songs played
- total seconds played

train.csv

- churn

User did not continue the subscription within 30 days

Data Preprocessing

Data Cleaning → **Data Exploration** → **Data Analysis**

Data Cleaning

- Missing Data
- Noisy
- Duplicate Data

Data Exploration

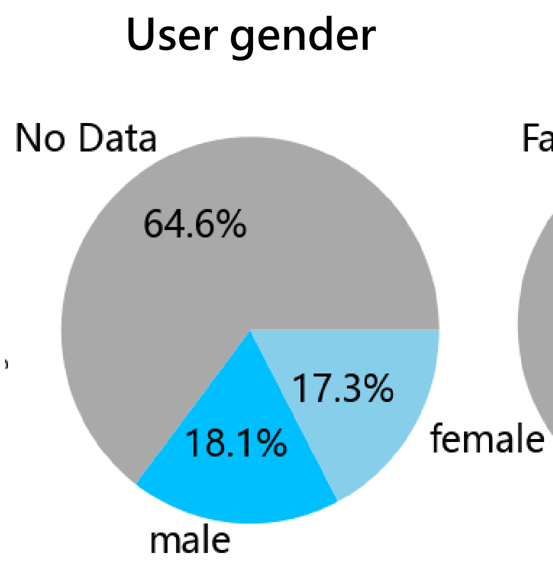
- Augmenting Data
- Descriptive analytics
- Visualization

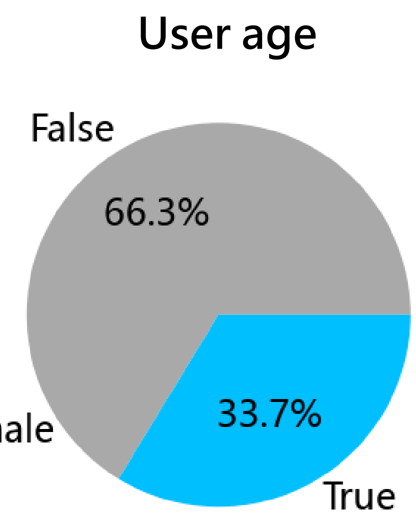
Data Analysis

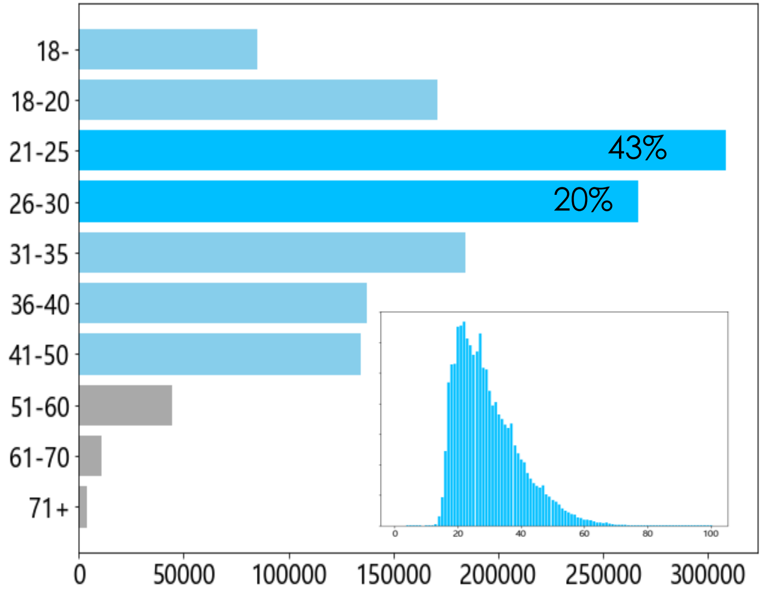
- Multivariate Analyses

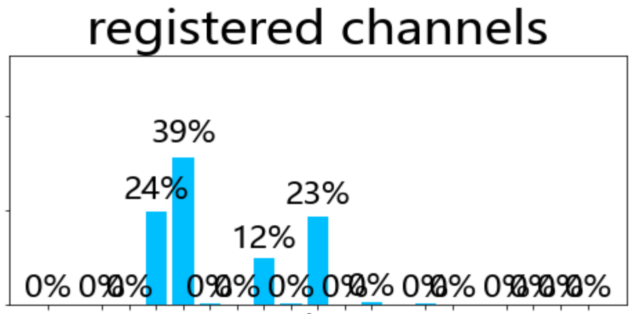
User Profile

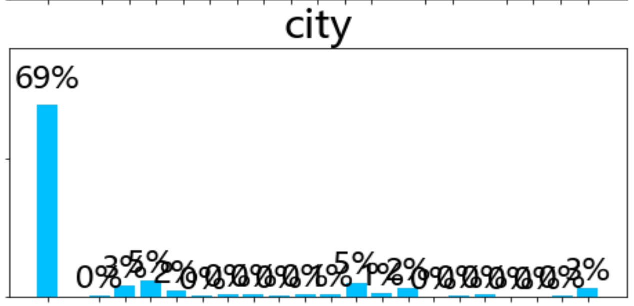
- **資料：** 部分缺失，僅**34%**用戶具生理資料
- **性別：** 有性別資料用戶，用戶男女比占比相當
- **年齡*:** 將用戶年齡分為10組，以20~25歲占43%為主；次之為26-30歲占23%
- **城市：** **69%**的用戶集中於單一縣市
- **註冊：** 兩主要管道，分別占**47%**、**37%**











Transactions

Data Cleaning | Data Exploration | Data Analysis

transcations.csv: (21547746, 9)

- **Remove duplicate user id by the latest transactions record**

```
df_T.nunique()
```

```
msno                      2363626
payment_method_id              40
payment_plan_days              37
plan_list_price                51
actual_amount_paid             57
is_auto_renew                   2
transaction_date              790
membership_expire_date       1559
is_cancel                       2
```

```
#drop  duplicates user id  & keep the lastest transactions record
df_t=df_T.sort_values('transaction_date',ascending=(False)).drop_duplicates('msno')
```

- **Count total transaction records as user loyalty and split it into 6 groups**

```
#calculate the total transaction records per user
df_T['t_times']=1
transactionstime=df_T.groupby(['msno','DATE'])['t_times'].count()
transactionstime=transactionstime.count('msno')
df_t=pd.merge(df_t,transactionstime,on='msno',how='left')
```

```
#grouping the transaction time as user loyalty
tt=df_t.t_times.value_counts()
bins = [0,1,4,12,18,24,max(df_t.t_times)]
df_t['user_level'] = pd.cut(df_t.t_times, bins, labels= np.arange(1, len( bins) ) )
df_t['user_level'] = df_t['user_level'].astype('int')
user_level=df_t['user_level'].value_counts().sort_index(ascending=True)
df_t.user_level.value_counts().sort_index()
```

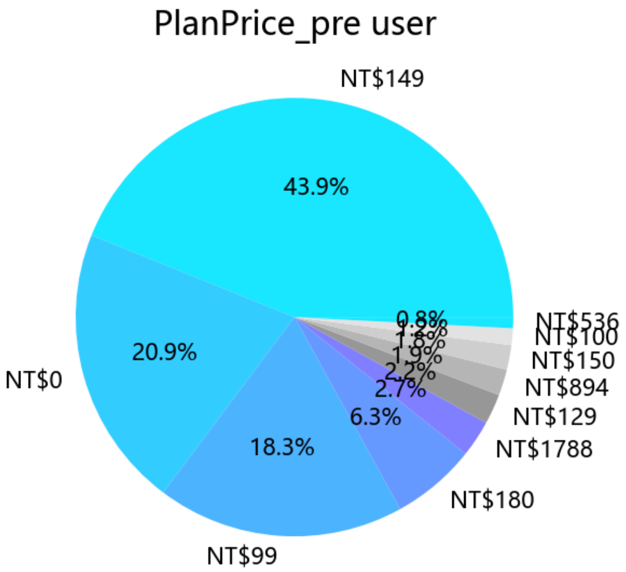
User Loyalty

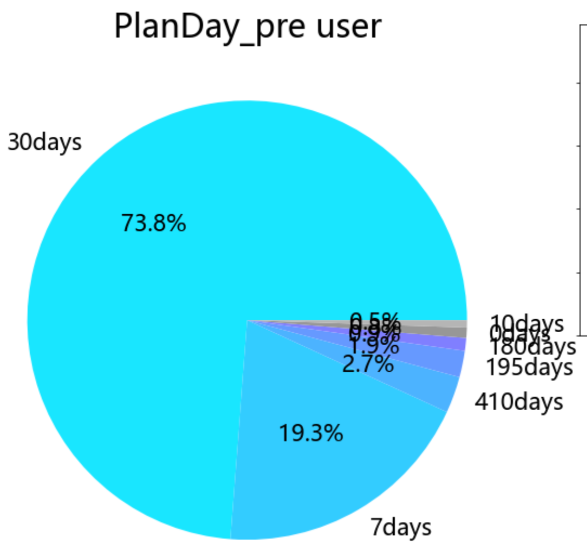
1	787689	New
2	386877	2-4 times
3	437879	5-12 times
4	305617	12-18 times
5	240935	18-24 times
6	204629	25+ times

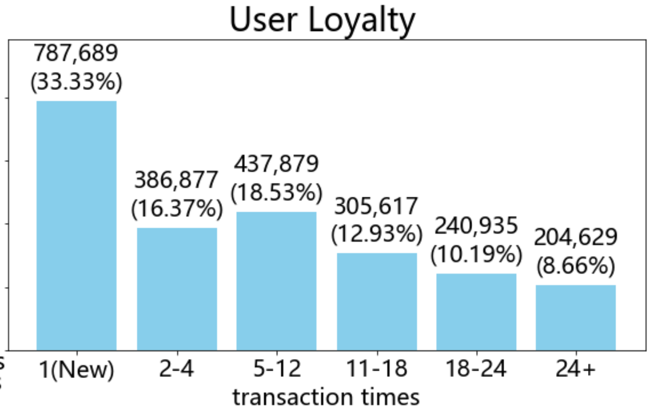
Transactions

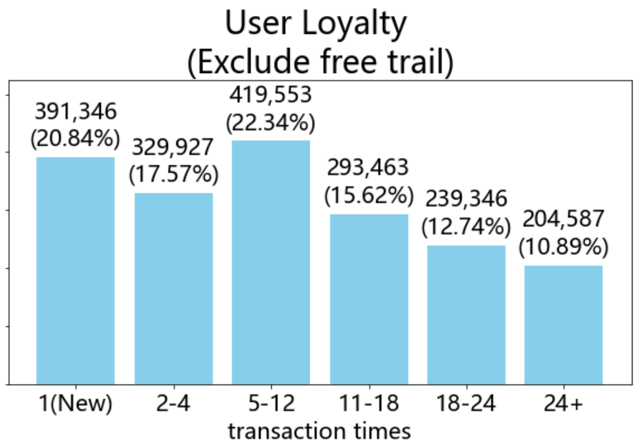
Data Cleaning | Data Exploration | Data Analysis

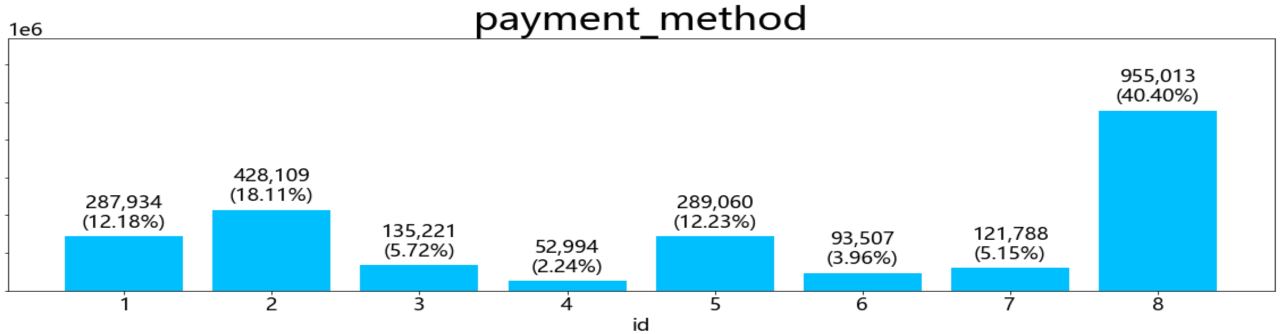
- **資料：** 每用戶具一至多筆資料，依方案到期日排序，取得最近資料
- **方案 ：** $149 個人方案占**43%**，$0 試用方案占**21%**， $99 享樂方案占**18%**
- **訂閱 ：** 30天月租占**74%**， 7天試用占**19%**
- **交易：** 依交易次數將用戶分為新客戶、老客戶共六個層級(0-5)
- **付款：** 兩主要管道，分別占**45%**、**20%**











User logs

Data Cleaning | Data Exploration | Data Analysis

`user_logs.csv` : (18396362, 9)

- **Calculate the average of unique songs, total seconds, and ratio of songs played per users**

```
df_userlog.nunique()
```

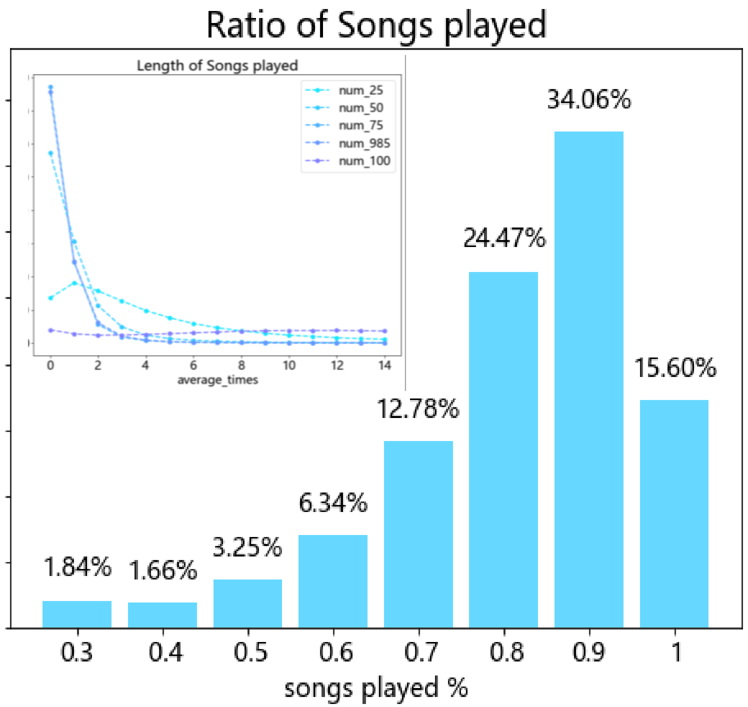
```
msno         1103894
date              31
num_25           743
num_50           356
num_75           193
num_985          340
num_100         1115
num_unq          776
total_secs  10701475
```

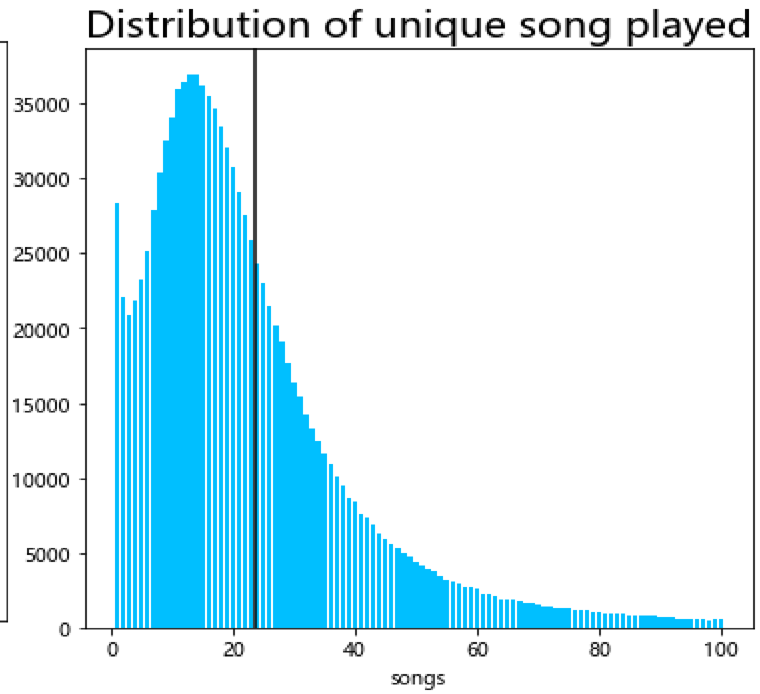
```
#last 1 month user data
#find avg seconds played per user/ ratio songs played of song length
user_engagement='''Select msno,COUNT(msno) as log_counts, round(avg(num_unq),0) avg_unisongs_listened,
round(avg(total_secs),0) 'avg_totalseconds' ,
round((sum(num_25)*0.25+sum(num_50)*0.5+sum(num_75)*0.75+sum(num_985)*0.985+sum(num_100))/
sum(num_25+num_50+num_75+num_985+num_100),1)  listening_rate
FROM userlog
GROUP BY MSNO'''
cursor.execute(user_engagement)
conn.commit()
```

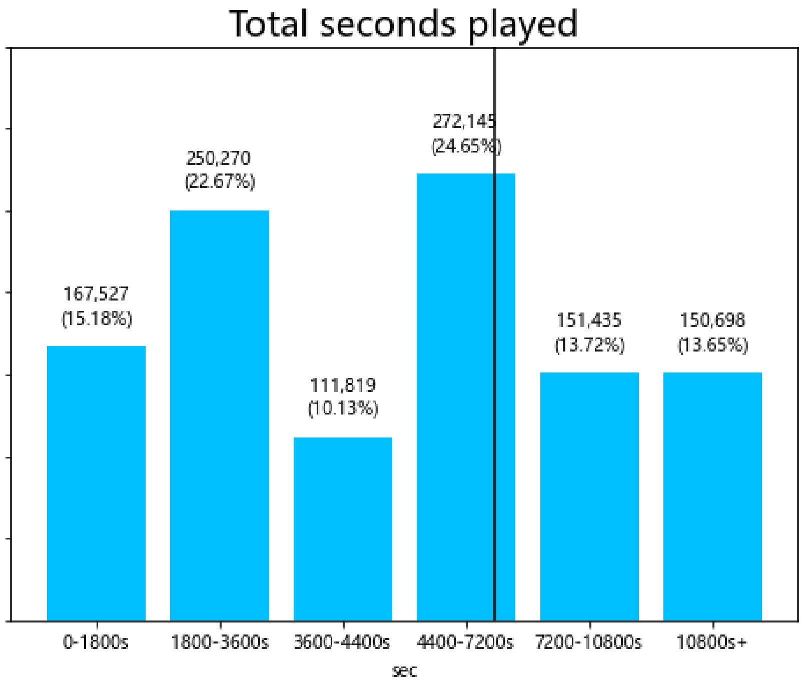
	count	mean	std	min	25%	50%	75%	max
log_counts	1103894.0	16.664971	10.303328	1.0	7.0	18.0	26.0	31.0
avg_unisongs_listened	1103894.0	23.981450	20.658422	1.0	11.0	19.0	30.0	1560.0
avg_totalseconds	1103894.0	6295.604971	6532.292049	0.0	2631.0	4573.0	7622.0	536354.0
listening_rate	1103894.0	0.814200	0.153109	0.3	0.7	0.8	0.9	1.0

User logs

- **資料：** 計算用戶每月平均聽歌長度、不重複歌曲數、和歌曲總播放秒數
- **聆聽率：** 以用戶平均聽歌完整率進行計算， 歌曲聆聽率90%達最高占34%，整體用戶歌曲完整聆聽率高
- **歌曲量：** 平均用戶每月聆聽不重複歌曲23首
- **聆聽量：** 平均用戶每月聆聽秒6,300s ，1.5- 2小時之間為最多(30min切割)



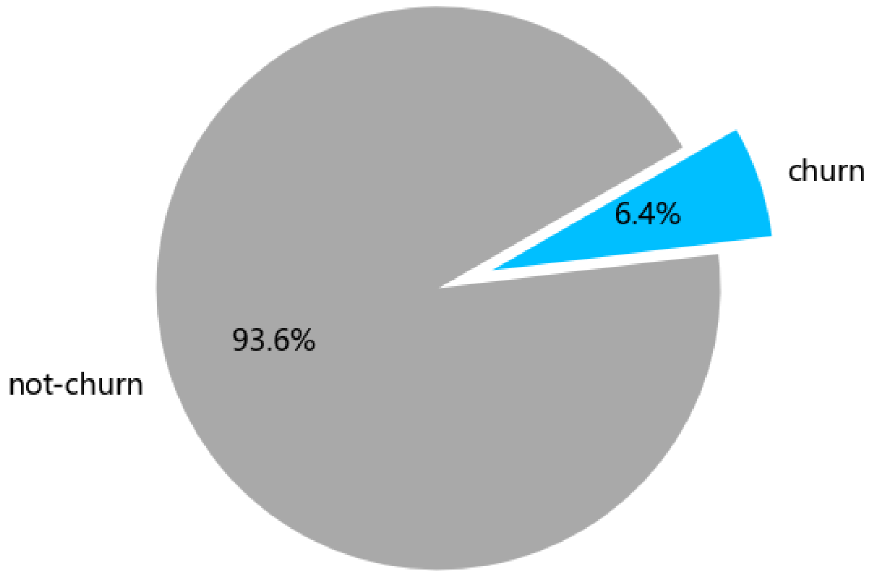




Variable Selection

Definition of Churn :
Not continue the subscription within 30 days

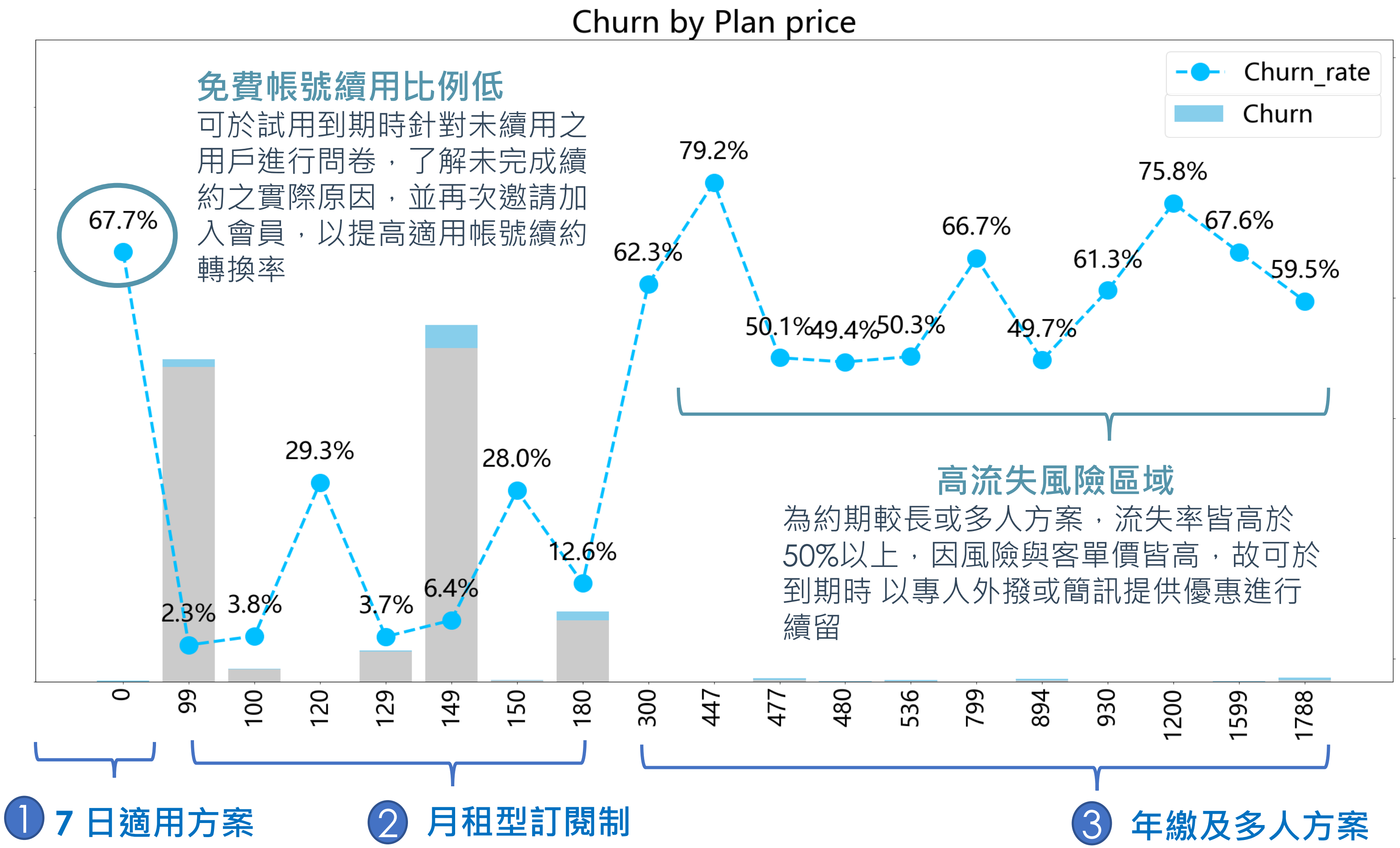
$$\text{Churn Rate} = \frac{\text{Number of Churn Customers}}{\text{Total Number of Customers}}$$

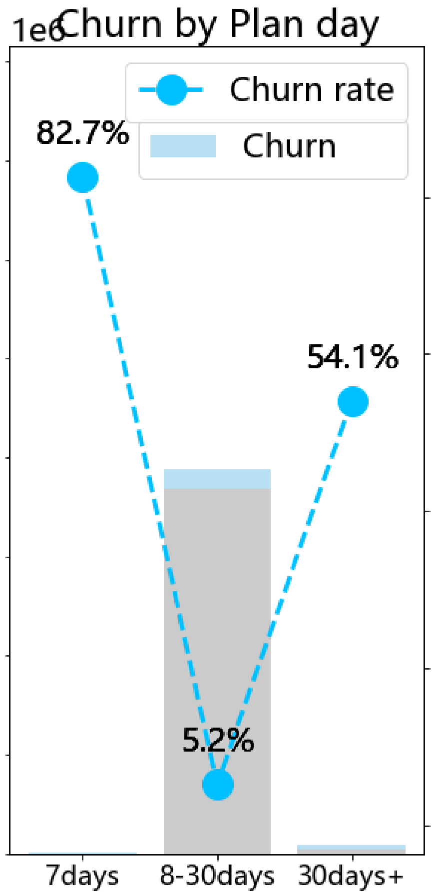


參考資料: 電信預測流失模型

Observations

資費：除$0適用外，用戶資費越高流失風險越大

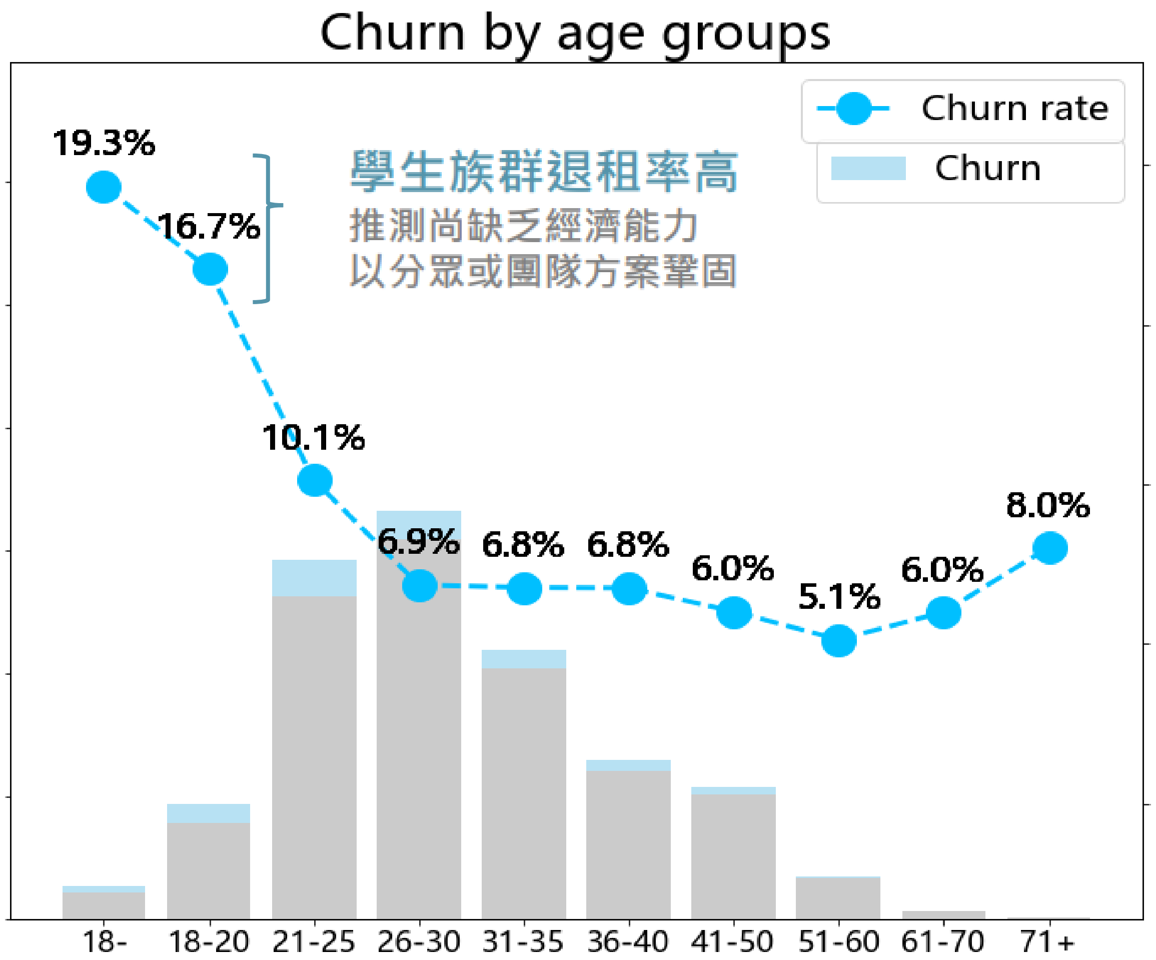


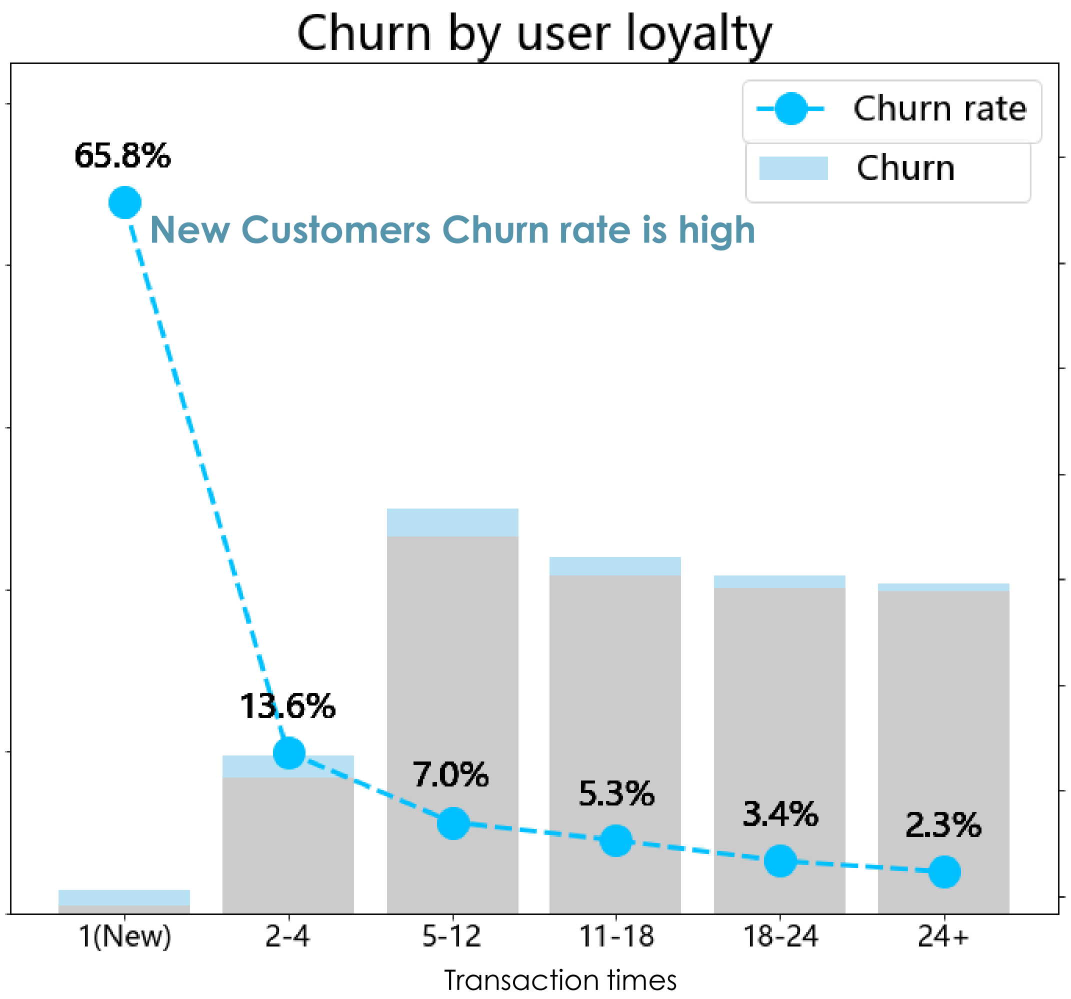


Observations

Age Groups :
不同年齡區間具特有退租行為，惟此類資料量不足

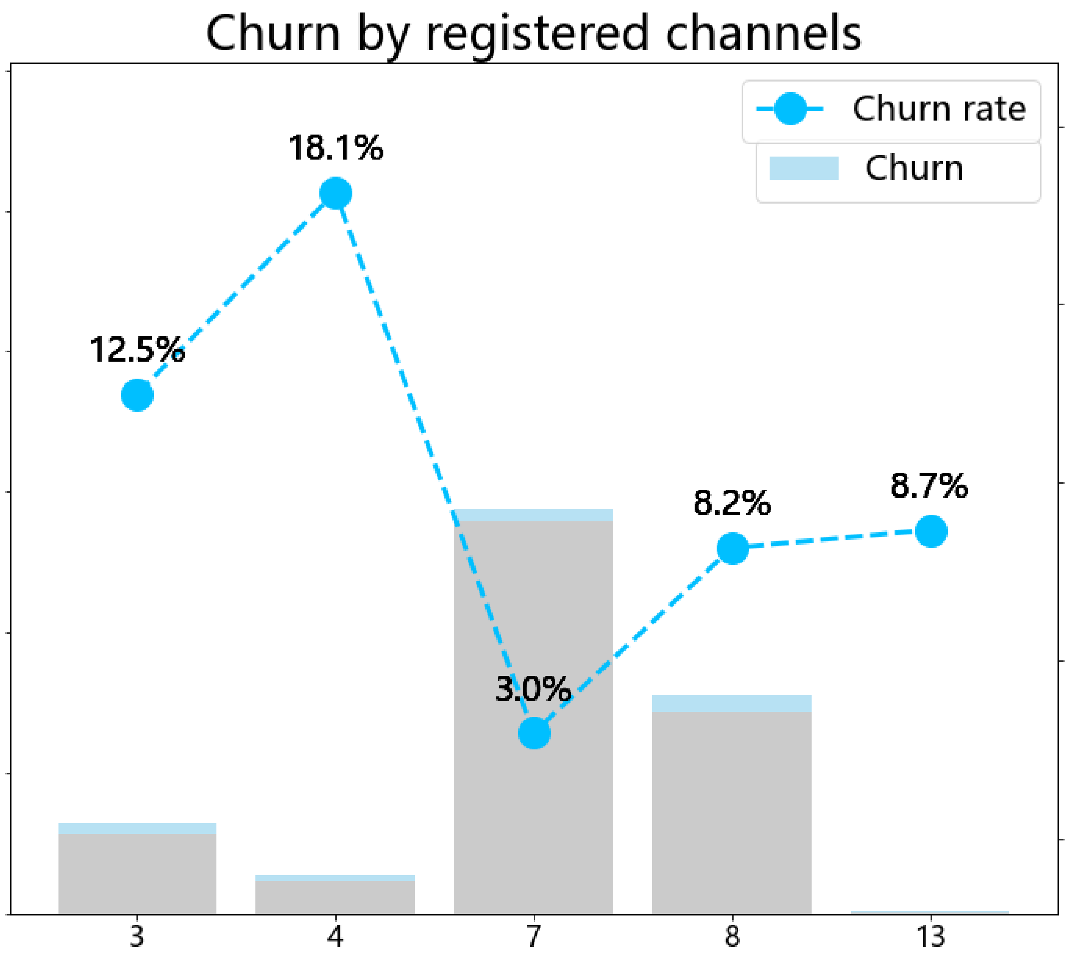
Loyalty :
新用戶/試用用戶之大幅高於其他



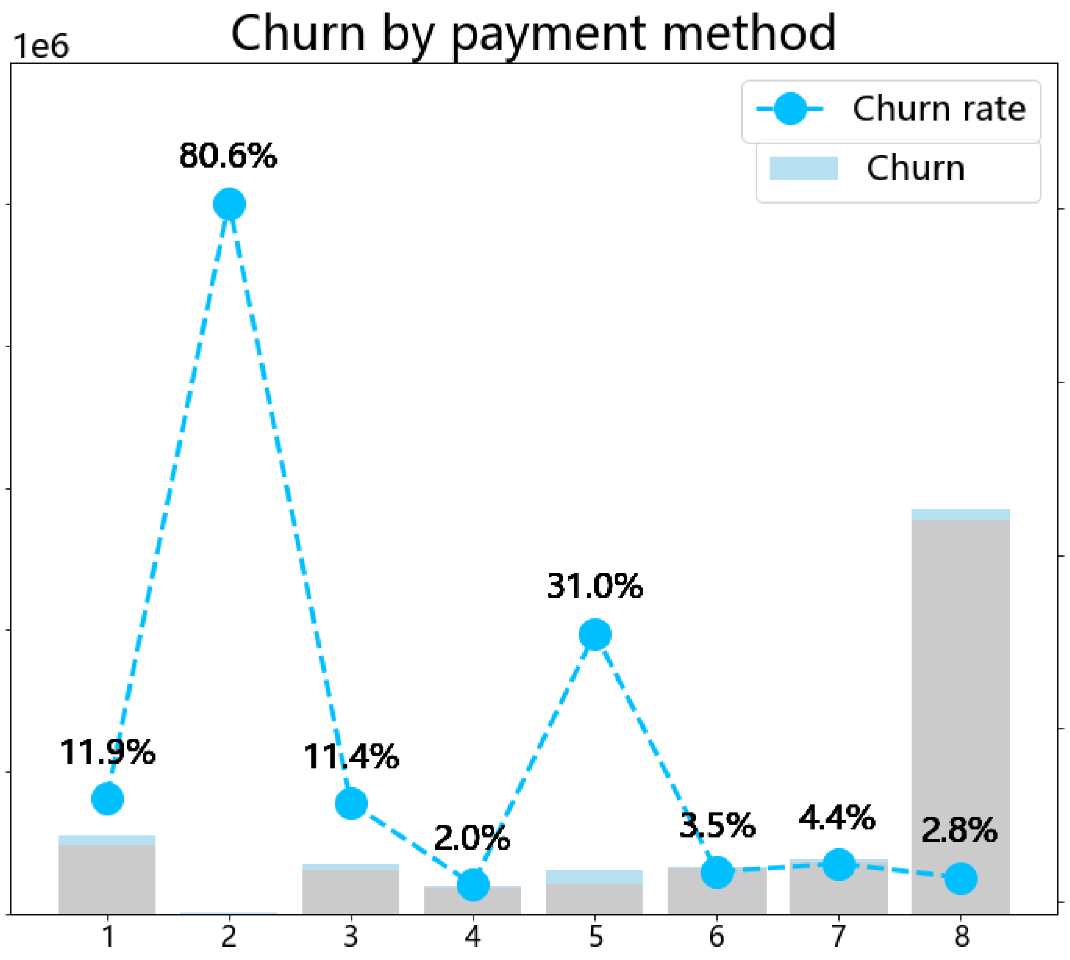


Observations

Register Channels :
不同的註冊方法具有不同的退租率高 (去識別化資料)



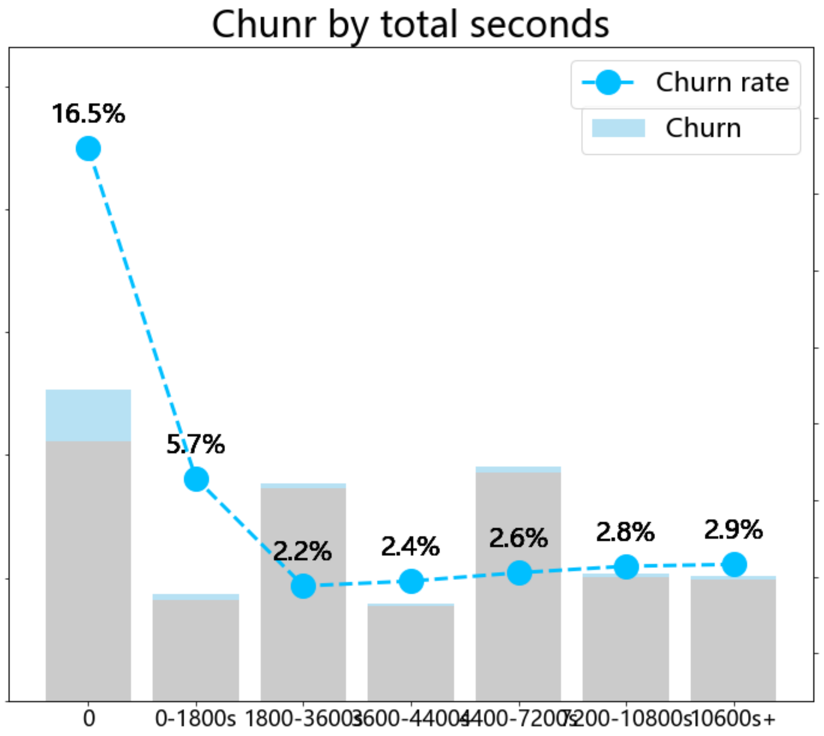
Payment methods:
特定付款方式之用戶退租機率高 (去識別化資料)

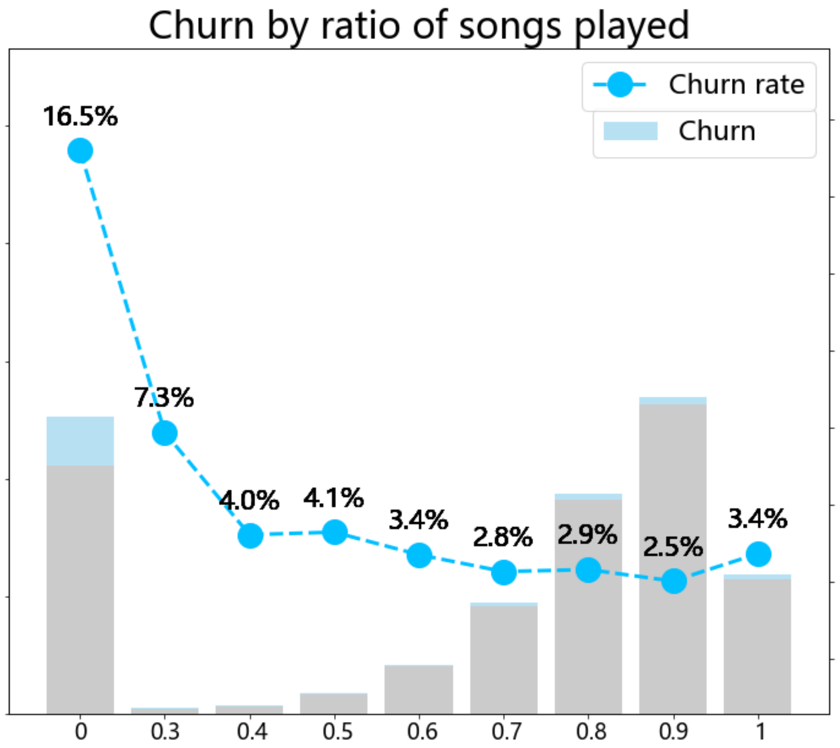


Observations

Total seconds played :
每月聆聽總長度超過每月30分鐘後即降低
可嘗試以超過30分鐘為觀察流失門檻

Songs played rate:
近一個月聆聽率越完整相對退租率越
低，不具有聆聽率之用戶，則風險高





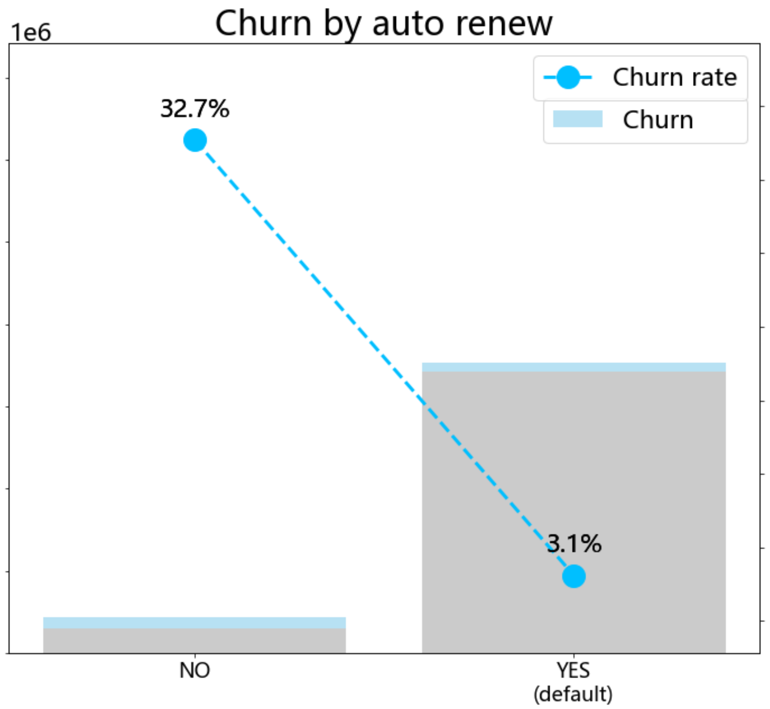
測試資料中25%的用戶
於近1個月內沒有聆聽紀錄

Churn rate
16.5%

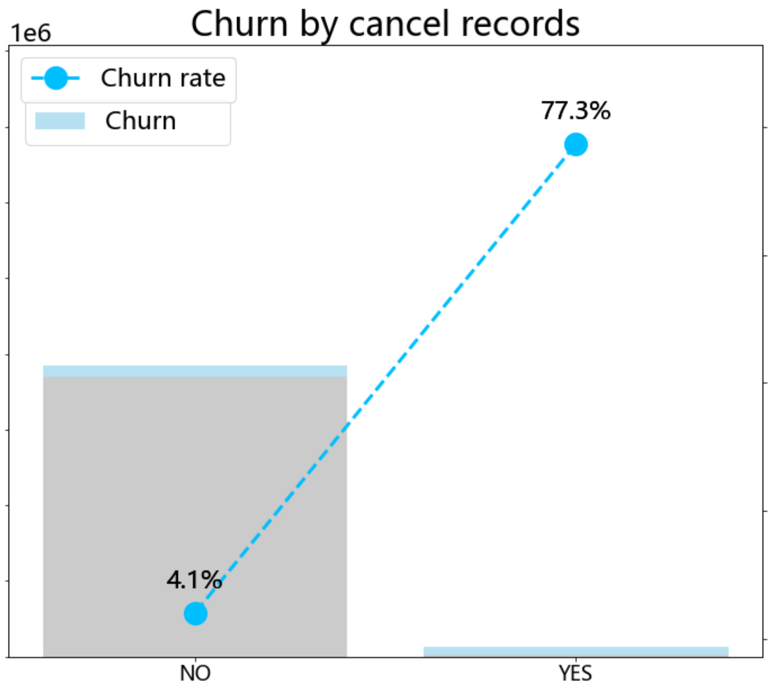
退租率為有聆聽紀錄的兩倍

Observations

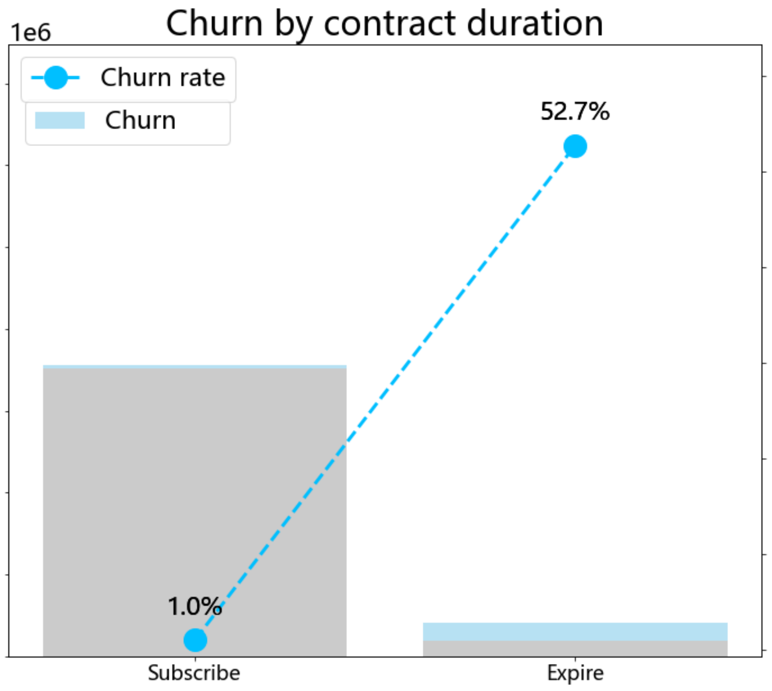
Auto Renew:
因此指標微系統預設
故此用戶行為可歸類為高流失指標



Cancel Records:
具取消行為用戶之流失率高
為未取消之19倍之高，達77%



Contract Duration:
合約已到期，但未自動續約
延長期到期日期者，超過52%流失機率



Prediction Model - Logistic Regression

	payment_method_id	plan_list_price	is_auto_renew	is_cancel	DATE	totalcharges	user_level	contract	listening_rate	secs_range	registered_via	Group
0	0.571429	0.0745	0.0	0.0	0.999787	0.018307	0.2	1.0	0.0	0.000000	0.692308	
1	0.714286	0.0745	1.0	1.0	1.000000	0.406561	0.8	0.0	1.0	0.166667	0.000000	
2	0.714286	0.0745	1.0	1.0	0.999787	0.183315	0.4	1.0	0.0	0.000000	0.692308	
3	0.571429	0.8940	0.0	0.0	0.999787	0.219683	0.2	1.0	0.7	0.666667	0.000000	
4	0.571429	0.0745	0.0	0.0	0.999787	0.439366	0.4	1.0	1.0	1.000000	0.692308	
5	0.000000	0.0745	1.0	1.0	0.999787	0.292911	0.8	1.0	0.0	0.000000	0.692308	
6	0.285714	0.0900	0.0	0.0	0.999787	0.217471	0.4	1.0	0.0	0.000000	0.000000	
7	0.571429	0.0745	0.0	0.0	0.999787	0.395872	0.8	1.0	1.0	0.166667	0.692308	
8	0.285714	0.0900	0.0	0.0	1.000000	0.452513	0.8	0.0	1.0	0.666667	0.692308	
9	0.571429	0.0745	0.0	0.0	0.999787	0.073228	0.4	1.0	1.0	0.333333	0.000000	
10	0.571429	0.0745	0.0	0.0	1.000000	0.457673	1.0	0.0	0.9	0.833333	0.000000	
11	0.285714	0.0900	0.0	0.0	0.999787	0.022116	0.0	1.0	0.0	0.000000	0.000000	
12	0.285714	0.0900	0.0	0.0	0.999787	0.264652	0.4	1.0	0.0	0.000000	0.000000	
13	0.285714	0.0900	0.0	0.0	0.999787	0.372282	0.6	1.0	0.0	0.000000	0.692308	
14	0.571429	0.2385	0.0	0.0	0.999787	0.186755	0.4	1.0	0.0	0.000000	0.692308	

```
coef of payment_method_id : -1.13876515
coef of plan_list_price : -0.50425810
coef of DATE : -0.02788801
coef of totalcharges : 5.63510357
coef of user_level : -3.14288920
coef of contract : 4.26235797
coef of listening_rate : -2.03088726
coef of secs_range : -1.05875721
coef of registered_via : -0.00091885
coef of Groupage : 0.54700893
coef of planday_group : -0.95127204
```

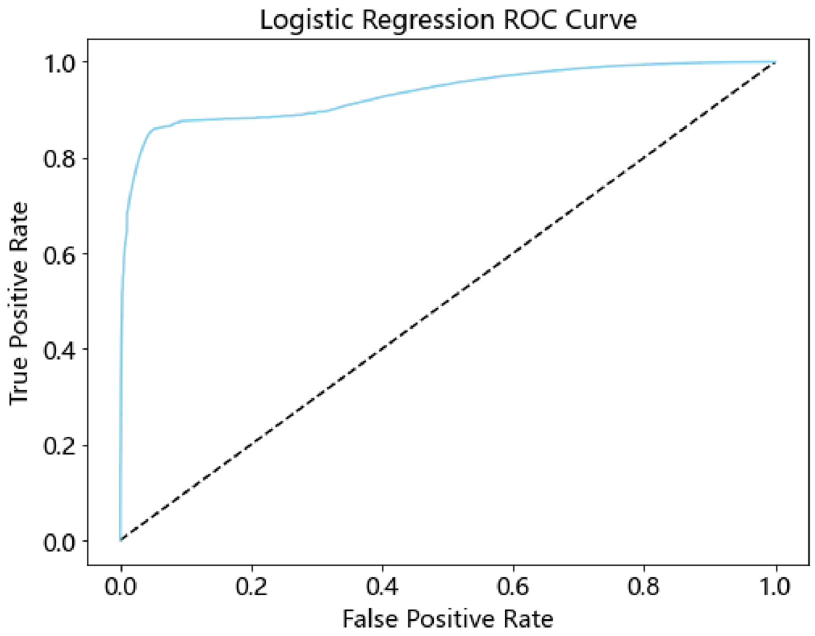
```
pred=logistic.predict(Xtest)
print('logistic Accuracy: ',logistic.score(Xtest,ytest))
```

```
logistic Accuracy:  0.9685754066867587

              precision    recall  f1-score   support

           0       0.98      0.99      0.98    371740
           1       0.78      0.70      0.74     25433

    accuracy                           0.97    397173
   macro avg       0.88      0.84      0.86    397173
weighted avg       0.97      0.97      0.97    397173
```



結果應用

1 年齡為用戶使用數據及退租行為的有效特徵，應強化此類數據蒐集
以完善分析準確性，以利優化使用者體驗及降低流失

如:學生族群流失率高，則可推學生組團方案，考慮經濟來源為父母則可往提供以親子/家族方案規劃
有資料用戶主要集中於21-25歲，可依不同年齡層調整平台歌單推薦內容及推薦順序

2 資費方案 種類及行為明確，可針對其制定對應分眾方案規劃

試用方案續約轉換率低，可於到期提供優惠，如延長免費2個月(需綁約一年)，再度強化免費誘因，以達締結
多人方案/長約方案因貢獻度高，且已具有此案行申辦意願，故可強化續約優惠折扣，透過人力外撥等

3 可透過使用行為數據 進行顧客分級貼標，用以行銷層面溝通，以強化顧客參與及黏著度

從用戶交易次數、聽歌時間、聆聽率等數據將分類，年資、使用率、累計金額等
並可藉由此標籤將用戶分及為不同等級，再續約或活動時提供分級的優惠內容，強化品牌黏著度

4 流失指標 取消自動續約、取消交易、約滿但未續約、近一個月內無聆聽紀錄
等行為皆為有效的流失指標，前三項可視為事件行指標，可依此進行用戶溝通

例如當用戶首次取消自動續約時，可觸發簡訊提供開啟自動續約，可換取$xx元等機制