DATA description

A mini subset of size 125 MB of the original 12 GB customer log json data file will be used for creating the prediction model. The small dataset has 286’500 log entries with 18 unique columns.

The schema and info of dataset is given below:

```
root
 |-- artist: string (nullable = true)
 |-- auth: string (nullable = true)
 |-- firstName: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- itemInSession: long (nullable = true)
 |-- lastName: string (nullable = true)
 |-- length: double (nullable = true)
 |-- level: string (nullable = true)
 |-- location: string (nullable = true)
 |-- method: string (nullable = true)
 |-- page: string (nullable = true)
 |-- registration: long (nullable = true)
 |-- sessionId: long (nullable = true)
 |-- song: string (nullable = true)
 |-- status: long (nullable = true)
 |-- ts: long (nullable = true)
 |-- userAgent: string (nullable = true)
 |-- userId: string (nullable = true)
```

Column’s Name	**Description**
artist	The artist being listened to
auth	Whether or not the user is logged in
firstName/lastName	Name of the user
gender	Gender of the user
itemInSession	Item number in session
length	Length of time for current row of specific log
level	Free or Paid user
location	Physical location of user, including City and State
method	Get or Put requests
page	Which page are user on in current row
registration	Users registration number

Column's Name	Description
sessionId	Session ID
song	Song currently being played
status	Web status
ts	Timestamp of current row
userAgent	Useragent of post or get in browser of users
userId	User ID

We use the `Cancellation Confirmation` events of `page` column to define the customer churn, and perform some exploratory data analysis to observe the behavior for users who stayed vs users who churned.

- churn

```
df_cleaned_cancel.dropDuplicates(['userId']).select('Churn').groupby('Churn').count().collect()
```

```
[Row(Churn=1, count=52), Row(Churn=0, count=173)]
```

So, there are 52 users have churned events in the dataset, it's about 23.1% churned rate. The rate of churn and not churn is roughly 1:3, so this is an unbalanced dataset.

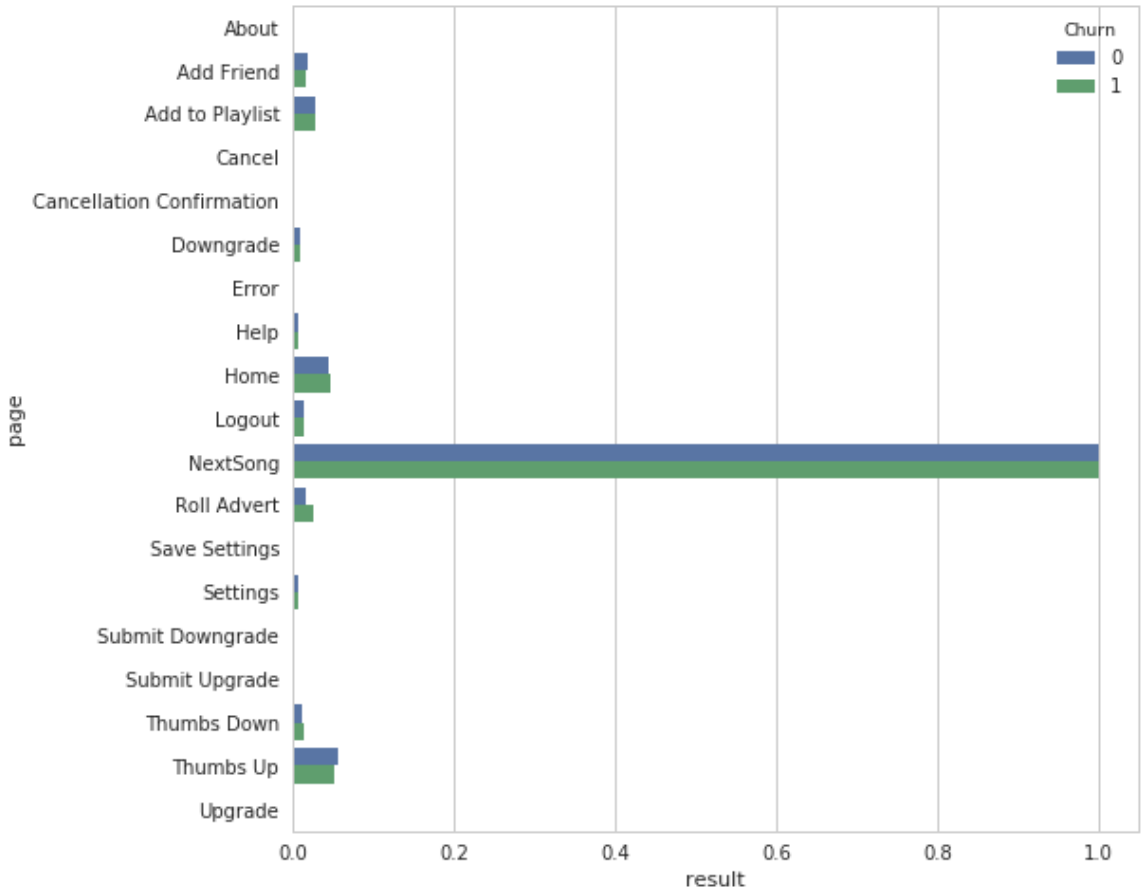
- gender

```
+-----+------+-----+
|Churn|gender|count|
+-----+------+-----+
|    0|     M|   89|
|    0|     F|   84|
|    1|     F|   20|
|    1|     M|   32|
+-----+------+-----+
```

Can we say the gender has effect on Churn or not ? We calculate the p-value and result is 0.20 over 0.05, so, we can't say like that.

- page

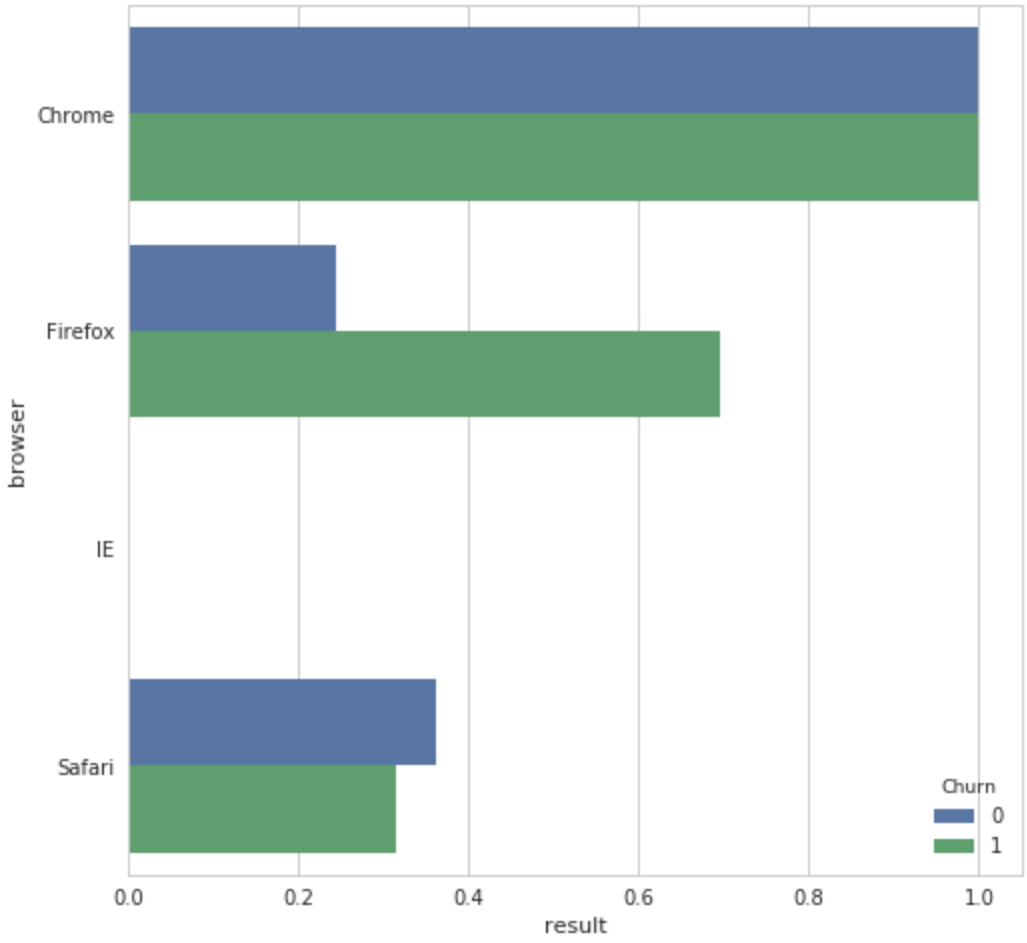
We count each item in `page` column of different group and normalized data.

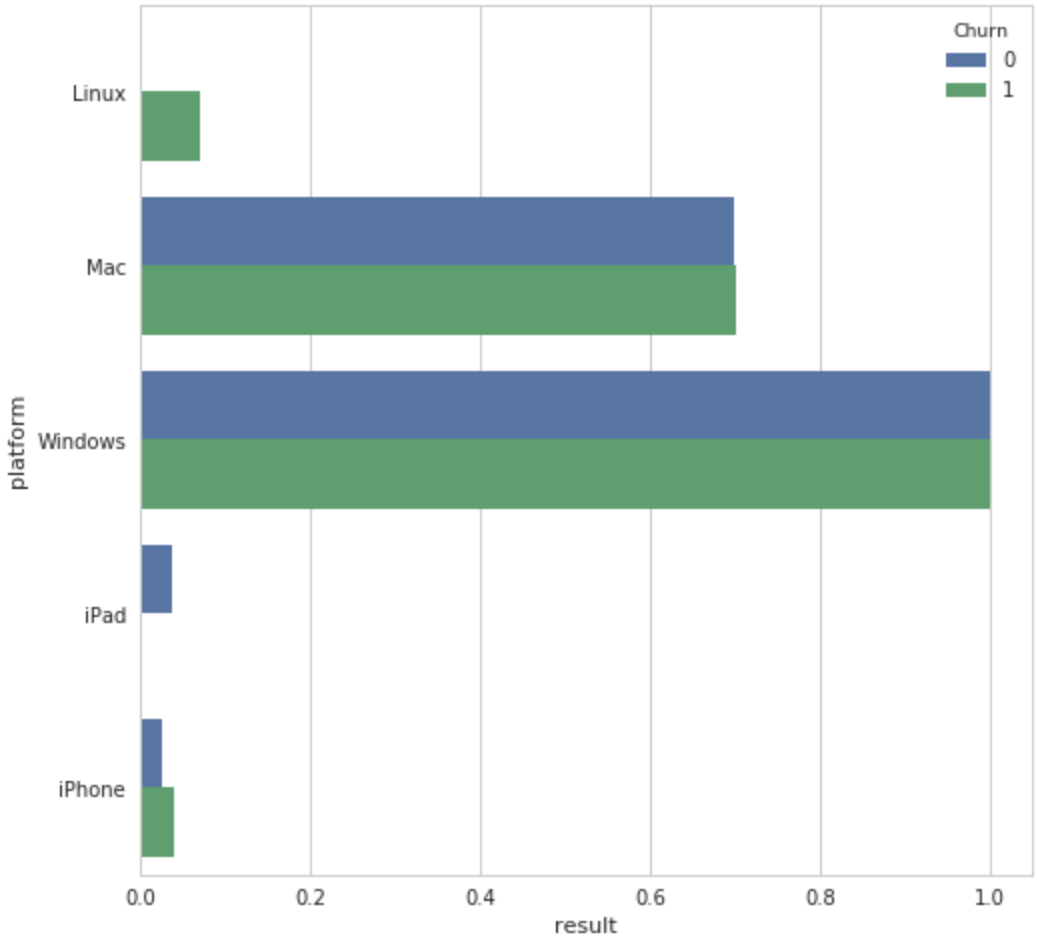


Obviously, NextSong has accounted for most of customers' events. Thumbs Up ,Thumbs Down , Home and Add to Playlist have effect on churn too.

- userAgent

We extract the browser and platform of customers from userAgent column.

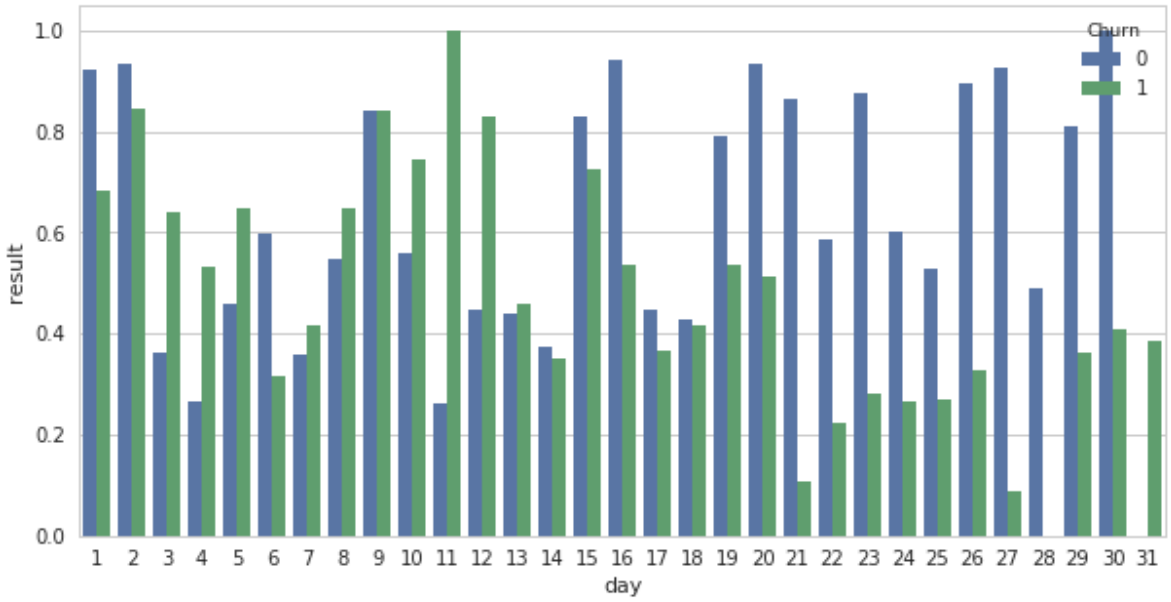

Chrome
Firefox
IE
Safari
browser
0.0
0.2
0.4
0.6
0.8
1.0
result
Churn
0
1

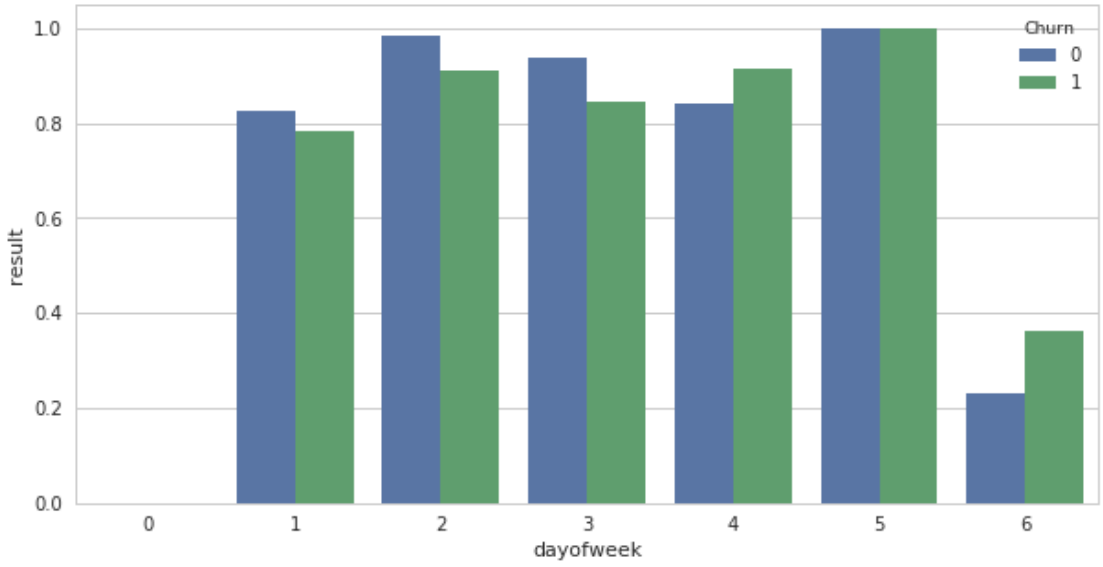


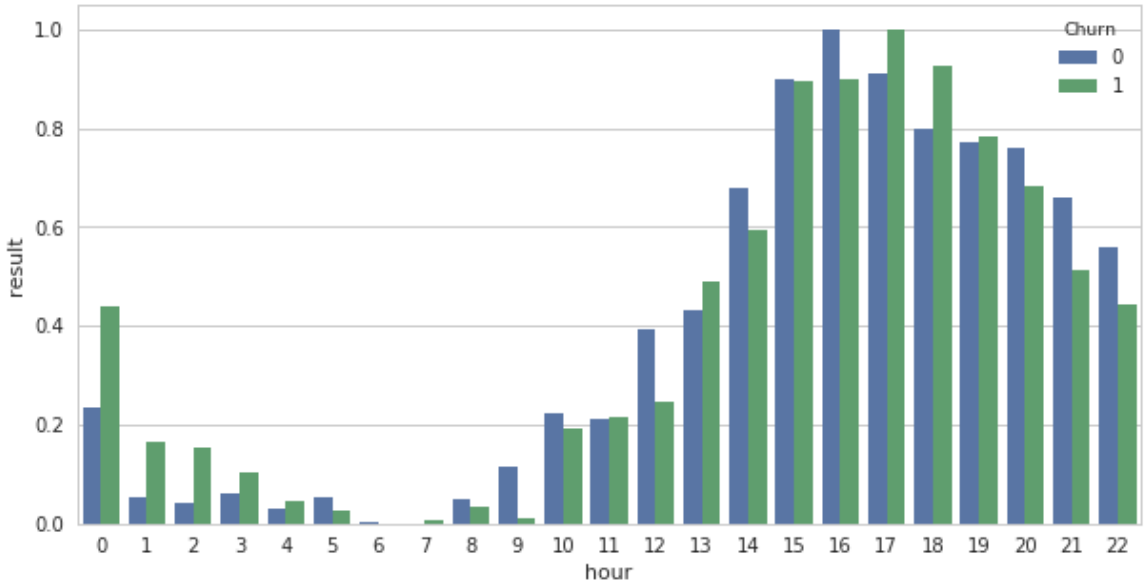
Customers using safari and iPad have more proportion in churn.

- time

We extract day-of-month, day-of-week and hour from `ts` column.


Churn
0
1
result
day
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31


Churn
0
1
result
dayofweek
0.0
0.2
0.4
0.6
0.8
1.0
0 1 2 3 4 5 6



Customers from churn group have more events after 15th in one month, and have less events in weekend.

After the preliminary analysis on the data. We split the full dataset into train and test sets. Test out the baseline of four machine learning methods: Logistic Regression, Linear SVC, Decision Tree Classifier and Random Forest Classifier. Conclusion will be made based on the result of the modeling.