Sparkify Music Service: Churn Prediction using Spark

Introduction

In this project, user transaction data of Sparkify music streaming service are analyzed to predict churn, which is cancellation of service or downgrading service level. Key factors associated with churn have been identified which could help Sparkify to optimize its service to reduce churn rates.

Because the dataset is large (12GB for full dataset), Spark is used for data processing and machine learning. Specifically, Spark Python API is used in this project. Spark SQL and Dataframe are used for data loading and manipulation, Spark ML module is used for classification and hyper parameter tuning.

The full dataset is analyzed on Amazon AWS EMR (Elastic MapReduce) service. Before the code was deployed to Amazon EMR, A medium sized (23 MB) subset of the dataset was used on a local machine for exploratory data analysis, data cleaning, feature engineering, and machine learning model selection.

Data Exploration and Feature Engineering

A medium sized subset of the full data is used for data exploration and feature engineering. The data is stored in a json file (medium-sparkify-event-data.json). There are 543,705 rows in the dataset. The schema of the dataframe is

```
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)
```

The users can be identified by columns `firstName`, `lastName`, and `userId`. The missing users can be removed by dropping the missing elements in `firstName` and `lastName`:

```
data = data.dropna(subset=['firstName', 'lastName'])
```

After data cleaning, there are 448 unique users in the medium sized subset of the data.

User transactions are logged in the column `Page`. It contains the following 19 transactions:

```
    +--------------------------+
    |page                      |
    +--------------------------+
    |Cancel                    |
    |Submit Downgrade          |
    |Thumbs Down               |
    |Home                      |
    |Downgrade                 |
    |Roll Advert               |
    |Logout                    |
    |Save Settings             |
    |Cancellation Confirmation|
    |About                     |
    |Settings                  |
    |Add to Playlist           |
    |Add Friend                |
    |NextSong                  |
    |Thumbs Up                 |
    |Help                      |
    |Upgrade                   |
    |Error                     |
    |Submit Upgrade            |
    +--------------------------+
```

These transactions are used to define churn and as predictors in the machine learning model.

Definition of Churn

There are two types of churns: 1) Cancellation of membership and 2) Downgrade from paid level to free level. The first type of churn is indicated by "Cancellation Confirmation" and the second type is by "Submit Downgrade" in `Page`. We will consider both types and focus more on the first type. In the medium sized dataset, the portion of users that have churned is 22.1% (99/448) and 21.7% (97/448) for cancellation and downgrade, respectively. As the crosstab table below shows, about 4.7% of users in the whole population have both downgraded membership level and canceled membership.

```
+-----------------------------+---+---+
|churn_Cancel_churn_Downgrade|  0|  1|
+-----------------------------+---+---+
|                            1| 78| 21|
|                            0|273| 76|
+-----------------------------+---+---+
```

In this project, churn due to cancellation is denoted by `churn_Cancel` and churn due to downgrade by `churn_Downgrade`. These two types of churns are treated separately.

Feature Engineering

Both continuous and categorical features are created from the dataset to be used as predictors in the classification model.

Continuous Features

The continuous features are related to the session properties and users' transactions.

Multiple sessions for each user are logged in the dataset. Mean and standard deviation of the temporal length of each session(`length`), number of items in each session (`itemInSession`) and number of songs (`song`) a user listened to in each session are calculated.

For features related to user transaction in the column `Page`, the ratio of the number of transactions for a type and the total number of interactions for each user is calculated for `Thumb Up`, `Thumb Down`, `Add Friend`, `Roll Advert`, `Add to Playlist`, `Submit Upgrade`, and `Error`.

In addition, the time difference between the last transaction (`max('ts')`) and the registration (`registration`) for each user is calculated.

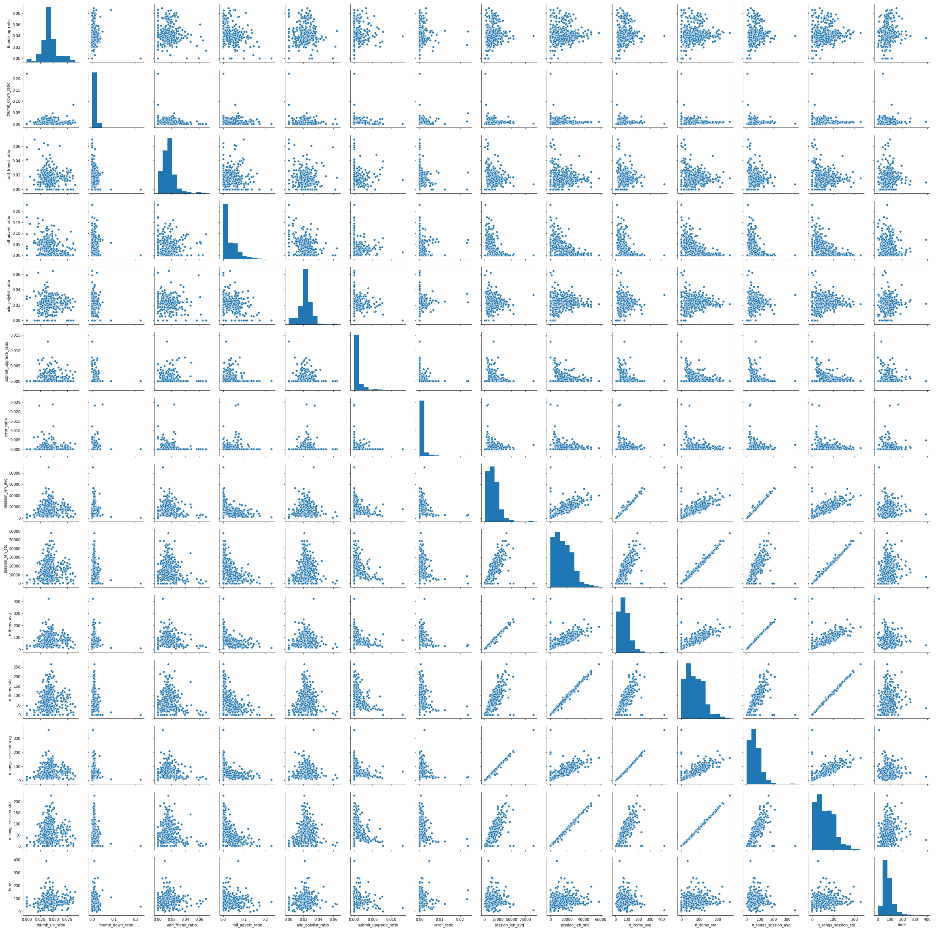


Fig. 1 Correlation between continuous features.

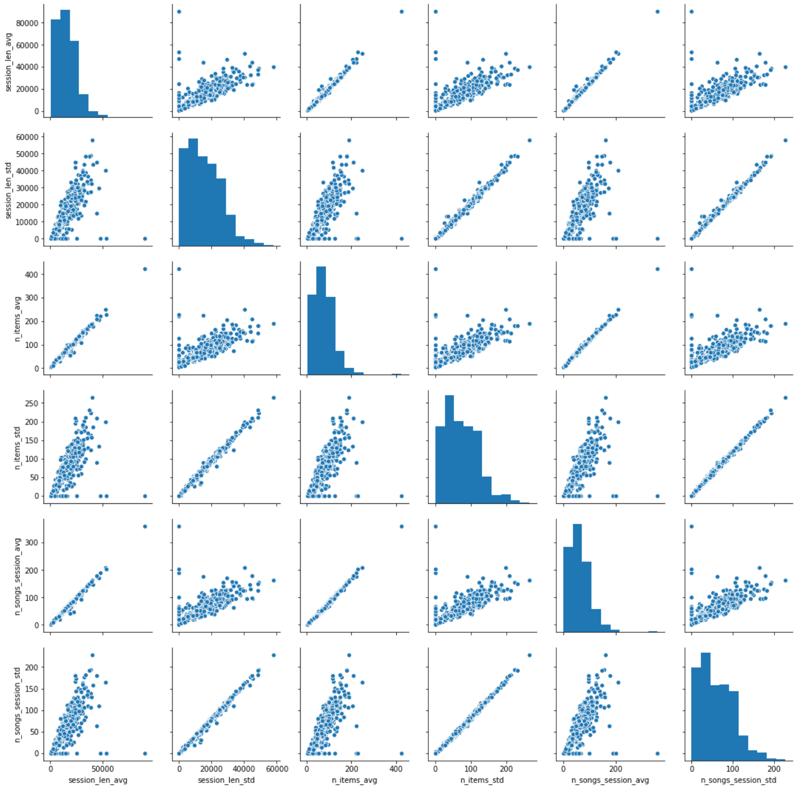
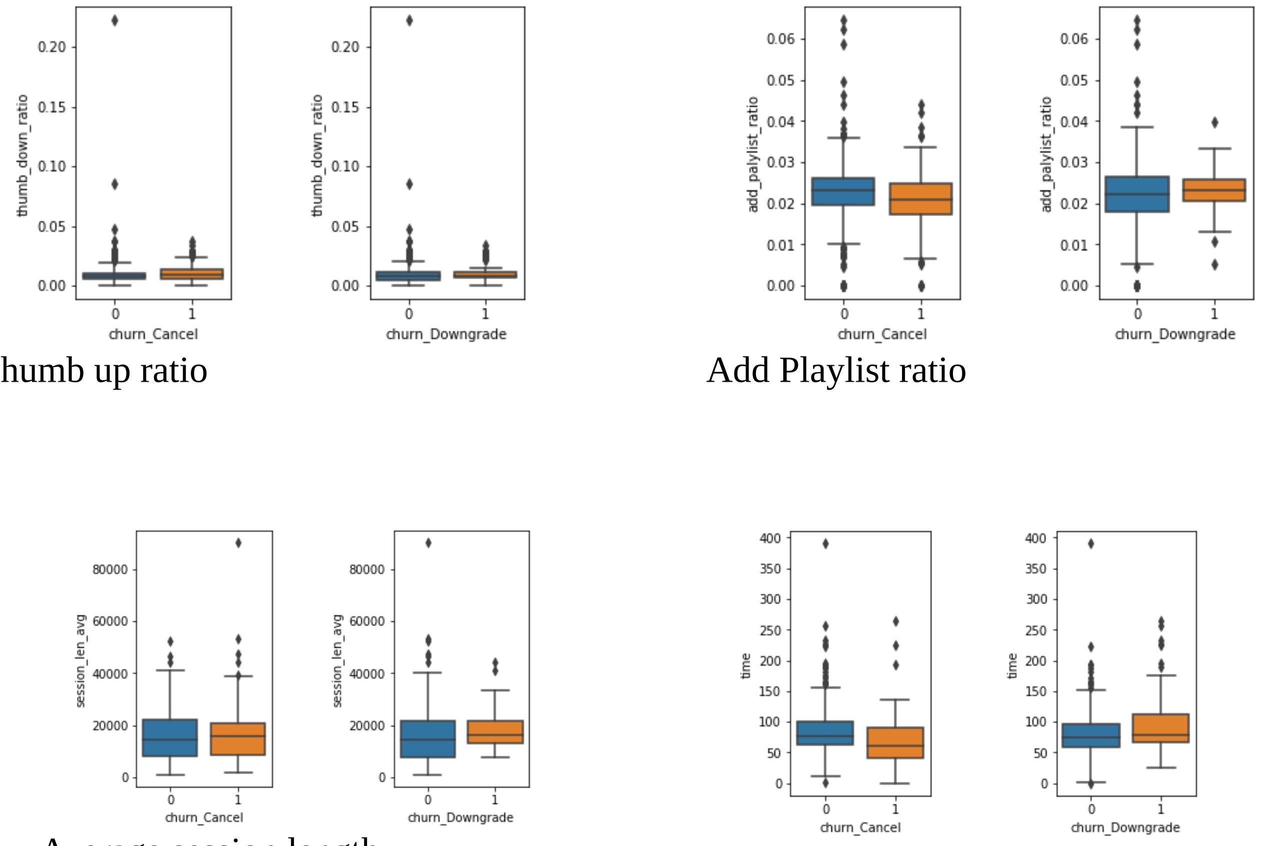


Fig. 2 Strong correlation between session length, number of items per session and number of songs per session.

Pairwise scatter plots in Fig.1 and Fig. 2 show that the correlation among most of the continuous features are weak except for session length, number of items and number of songs per sessions, which are strongly correlated. As a result, number of items per session and number of songs per session are removed from the feature list for the classification model.



Thumb up ratio

Add Playlist ratio

Average session length

Time between last transaction and registration

Fig. 3 Distribution of some continuous features for churned and non-churned users.

In Fig. 3 the distributions of some continuous features are plotted for users that have churned and users that have not. Both churns due to cancellation (`churn_Cancel`) and downgrade (`churn_Downgrade`) are included. The distribution is different for some features. For example, users that have churned have higher thumb up rates and lower add playlist rates.

Categorical Features

Three categorical features are included in the data analysis: Gender (`gender`) of the users, whether the transaction status (`status`) is 404 or not, and the device (`userAgent`) the users used to access the Sparkify service. `userAgent` is parsed to obtain the device or computer used.

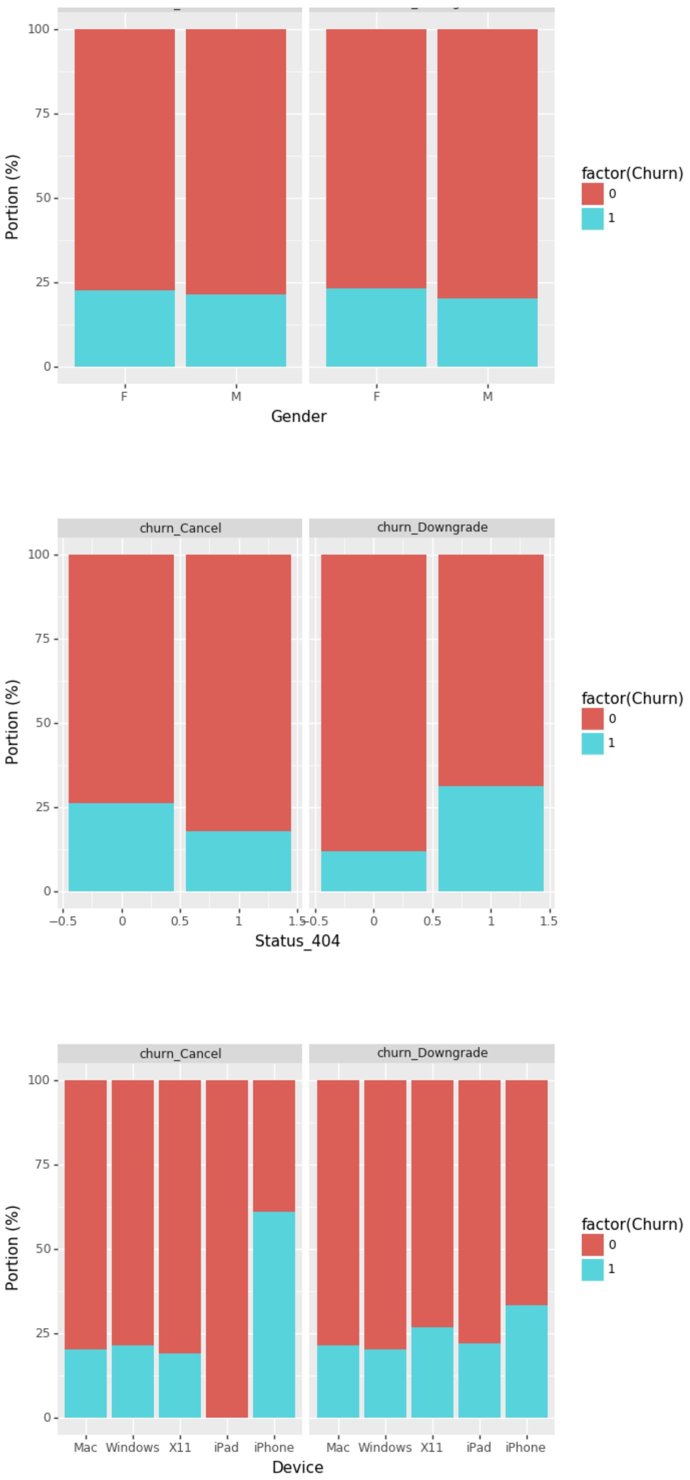


Fig. 4 Distributions of categorical features for users that have churned and users that have not.

The categorical features shows that there is very little difference in churn rates between diffferent `gender`. There is some but very small difference in churn rate between different `status_404`. It appears that churn rate for cancellation is high when iPad is used. But the significance is still unclear because of the smaller sample size for each device.

The resultant dataframe is transformed using Spark data pipeline which converts categorical features from string value to numerical value and then encoded with one-hot encoding.

Classification Model

After data cleaning and feature engineering, the resultant data is split into train and test datasets with 75% and 25% portions, respectively. Three classification models (logistic regression, random forest and gradient boosted trees) are trained with their default settings with the train dataset, and the prediction performance with the test dataset is obtained. We pick the classifier (gradient boosted trees) that has the overall best performance in predicting churns due to both cancellation and downgrade for further hypreparamter tuning using cross validation. The performance of the tuned model is evaluated and the feature importance is obtained with the fine tuned model so that the most important features affecting churns rates are identified.

Finally the random forest model is applied to the full dataset using Amazon EMR service.

Classifier Performance Metric

Because the dataset is imbalanced--churn rate is about 23%, we use F1 score as the classifier performance metric:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

Unfortunately in Spark ML library, `BinaryClassificationEvaluator` does not output F1 score, it reports area under the ROC curve or the precision-recall curve. `MulticlassClassificationEvaluator` reports F1, precision, recall and accuracy for multiclass classification, but the values are ***incorrect*** for binary classification. Therefore, an evaluator class with F1 score as metric is created:

```
class Fevaluator:
"""Evalautor classs with F1 score as metric"""
def __init__(self, labelCol='label', predictionCol='prediction'):
    self.label = labelCol
    self.prediction = predictionCol

def evaluate(self, dataset):
    _, score = performance(dataset, self.label, self.prediction)
    return score[0]  #f1 score
def isLargerBetter(self):
    return True
```

where function `performance` calculates F1 score, precision, recall and accuracy of a binary classification prediction. This evaluator can be used for model tuning with cross validation.

Result with Medium Sized Dataset

First the results with the medium sized subset are presented.

Classification Models: Logistic Regression, Random Forest and Gradient Boosted Trees

Logistic regression, random forest and gradient boosted tree classifiers with their default setting are trained with the train dataset. Then they are applied to the test dataset. The performance of the prediction for churn due to cancellation and for churn due to downgrade is summarized in the tables below.

Churn Due to Cancellation

Model	F1	Precision	Recall	Accuracy
Logistic Regression	0.45	0.82	0.31	0.80
Random Forest	0.32	0.75	0.21	0.77
GBT	0.53	0.75	0.41	0.81

Churn Due to Downgrade

Model	F1	Precision	Recall	Accuracy
Logistic Regression	0.071	0.20	0.043	0.76
Random Forest	0.22	0.75	0.13	0.81
GBT	0.48	0.48	0.48	0.78

Among the three models, gradient boosted trees (GBT) classifier is much slower than logistic regression and random forest. But GBT has best F1 score for both types of churn. Therefore we will further tune the GBT model and apply it to the full dataset on Amazon EMR. Since GBT is tree based model, the numerical features need not be scaled or normalized.

Model Tuning

3-fold Cross validation is used to fine tune the GBT model. `MaxIter` and `subsamplingRate` are scanned. For churn due to cancellation the default setting happens to be optimal. For churn due to downgrade, `subsamplingaRate=0.7` improves the GBT performance from the default value of `1.0`.

Churn Due to Cancellation

Model Setting	MaxIter	subsampingRate	F1 score
Default	20	1.0	0.53
Tuned	20	1.0	0.53

Churn Due to Downgrade

Model Setting	MaxIter	subsampingRate	F1 score
Default	20	1.0	0.48
Tuned	20	0.7	0.51

The model tuning improves the classification performance for churn due to downgrade. But the improvement is moderate. There is potential for further improvement if more parameters in the setting are searched.

Feature Importance

The trained gradient boosted trees model provides the importance score for each feature in the model. Importance of a feature is determined by how much and how effective the feature is used in the decision trees.

Fig. 5 shows feature importance retrieved from the models for churns due to cancellation and downgrade. We will look at the importance more closely when we summarize the full dataset results.

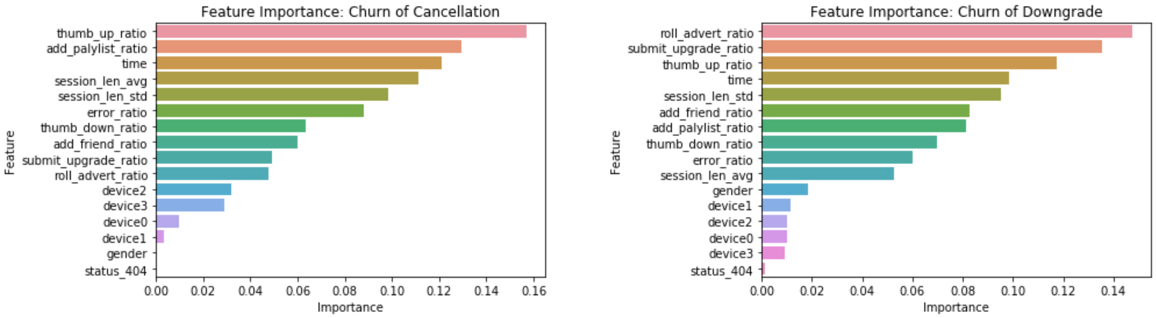


Fig. 5 Feature importance.

Result of Full Dataset on Amazon EMR

The full dataset results are summarized in this section. The computation is done on Amazon EMR with a cluster of 4 cores (1 master, 3 workers).

The full dataset has 26,259,199 rows. After removal of missing values in `firstName` and `lastName`, there are 25,480,720 rows and 22,277 unique users. The churn rate is 22.5% and 22.9% for cancellation and downgrade, respectively.

GBT classification Performance

Gradient boosted trees classifier is used for prediction. The settings obtained using the medium sized dataset are used. No further model tuning is conducted to avoid long computation time. As shown in the table below, the prediction performance is very similar to that with medium sized dataset.

GBT Classification Performance

Churn Type	F1	Precision	Recall	Accuracy
Cancellation	0.53	0.72	0.42	0.83

Downgrade	0.53	0.59	0.48	0.81

Feature Importance

Feature importance from the GBT model is plotted in Fig. 5. There are common features that are the most important predictors for both churn due to cancellation and churn due to downgrade, such as `roll_advert_ratio`, `thumb_up_ratio` `thumb down_ratio`, `add_friend_ratio`, `session_len_avg`, and `session_len_std`.

The feature with the highest importance, however. is different for cancellation and downgrade. For churn due to cancellation, it is `time`, which is the time span from registration to the latest transaction; for churn due to downgrade, it is `roll_advert_ratio`, which is likely related to advertisements that users have encountered.

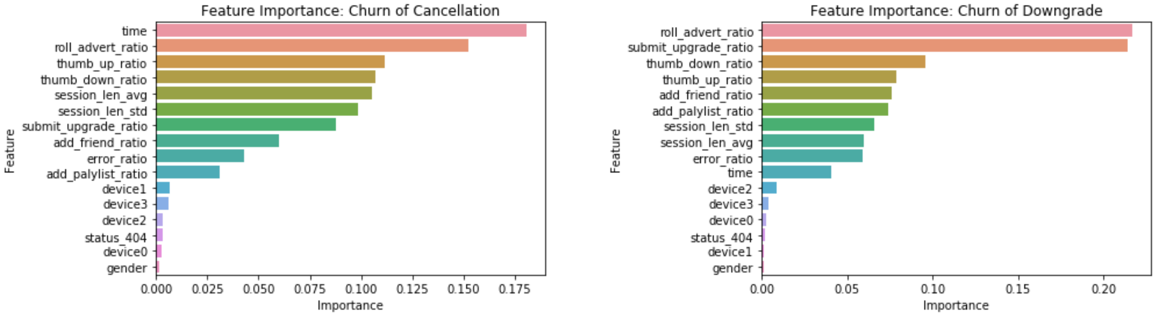


Fig. 5 Feature importance.

In order to see how the features affect churn, the mean values of the continuous features for the churn group and non-churn group are calculated. Then the relative difference in percentage between the churn group and non-churn mean values is calculated:

$$\text{Delta}(\%) = \frac{M_{\text{churn}} - M_{non-churn}}{M_{\text{non-churn}}} \times 100.$$

Fig. 6 shows Delta (%) for churn due to cancellation and churn due to downgrade. Blue color means that the feature value is higher for the churn group and red means the opposite.

By examining the features with the highest importance, we can see that the users who have churned tend to have the following characteristics:

1. received more advertisements (higher `roll_advert_ratio`),
2. gave more negative reviews (higher `thumb_down_ratio` and lower `thumb_up_ratio`),
3. stayed longer in session (higher `session_len_avg`),
4. added fewer friends (lower `add_friend_ratio`),
5. more likely to have upgraded membership (higher `submit_upgrade_ratio`),
6. have shorter time of being members for churn due to cancellation (lower `time`), have longer time of being members for churn due to downgrade (higher `time`).

Many of the above findings are consistent with common sense. However it may be hard to determine the causal relationship for some features. For example, in the case of churn due to cancellation, the users who churned have short membership length. This may be the consequence of churn by definition and does not necessarily mean that users tend to cancel their membership during the early phase of their membership. Additional information or controlled study may be needed.

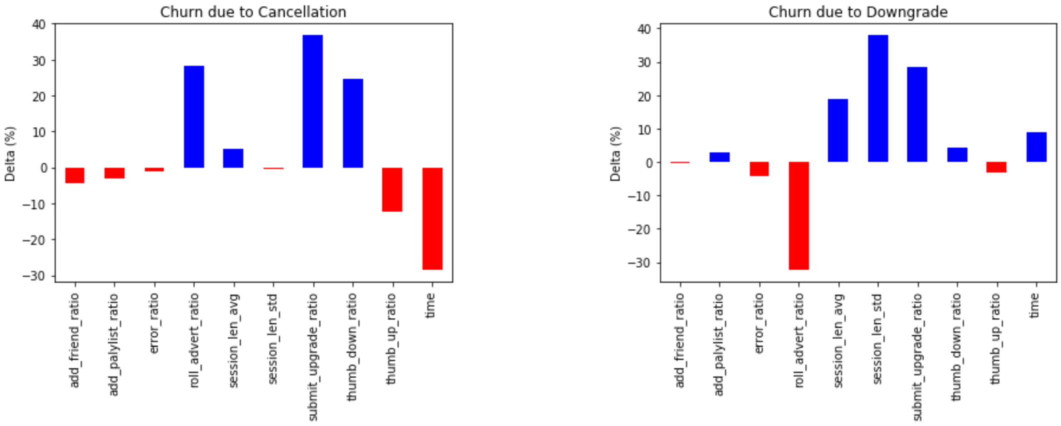


Fig. 6 Difference in continuous feature mean value between churn group and non-churn group. Blue means the feature value is higher for the churn group and red means lower.

Future Work

Overall performance of churn prediction in this project is not great: F1 score is about 0.53. Below are a few areas that potentially can improve the machine learning performance:

1. **Data imbalance:** The data is skewed. The churn rate is about 23%. It might be helpful to try oversampling of the churn population or under-sampling of the non-churn population to improve classifier prediction performance.
2. **Feature engineering:** Some of the user data are not included in the feature. For example, the month, day, and time of the day the users used the services. These features can be added and evaluated for their significance.
3. **Model tuning:** More model parameters and larger ranges of values can be tried to improve model performance.