Sparkify Churn Prediction

Prakash Dhimal
Manav Garkel
George Mason University
CS 657 Mining Massive Datasets

Problem Statement

- Acquiring new customer is more costly than retaining current customer
- Goal: Identify current customers who are likely to churn/cancel subscription

Step 1

Data Cleaning

Remove Null Values from various columns

Step 2

Feature Engineering

Extract features from user log data

Step 3

Cross Validation

User grid search cross validation for hyper-parameter tuning

Step 4

Prediction

Test the model using 20% of the dataset

Data

- Sparkify is an imaginary digital music service similar to Spotify.
- The dataset contains 12GB of user interactions with this service.

```
In [4]: data.printSchema()

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)
```

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

Data preprocessing

Data selection

Columns that were not significant to the modelling process were dropped

- Firstname
- Lastname
- Id_copy

userID was retained as it was used for feature engineering step

Unit Conversion

Registration and TS were given in milliseconds

These fields were converted to seconds by dividing the values by 1000

Create Churn Label

Dataset only contains user log data

Used Page column to identify churners:

- Visiting *Cancellation Confirmation* page indicated a churned user
- Created a label column where 1 indicates a churned user and 0 indicated otherwise

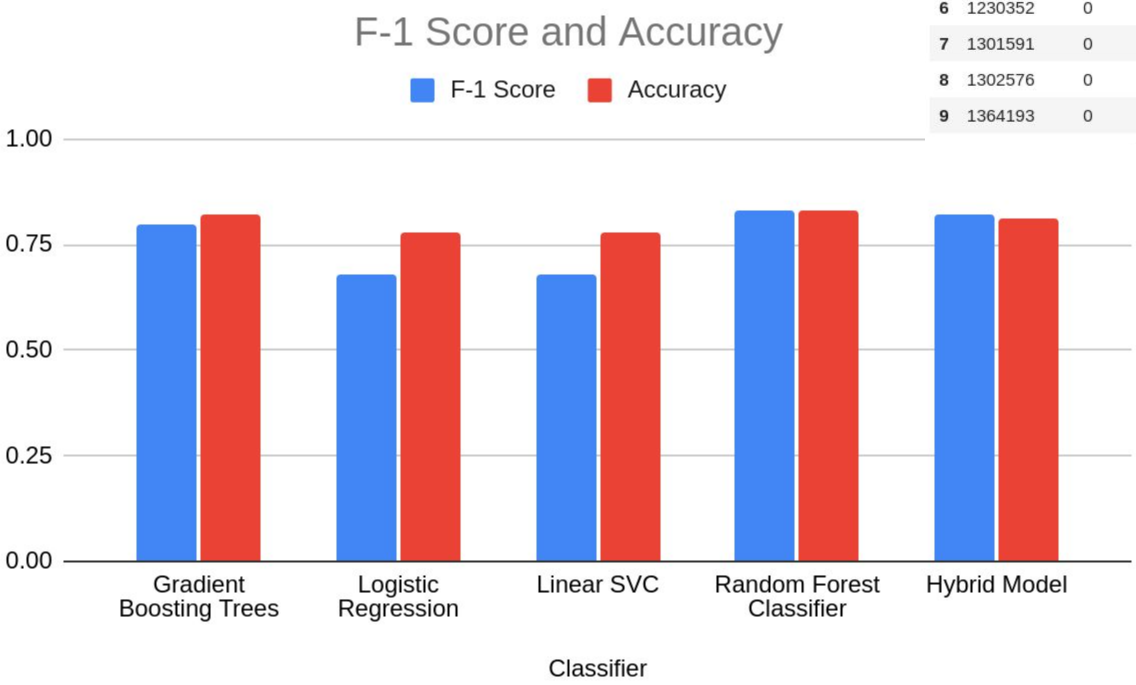
Feature Engineering

- Meaningful data has to be created from the user log data that could be used by the prediction models
- The following features were used
  - Time since registration
  - Number of friends referred
  - Total songs listened to
  - Total songs liked
  - Total songs disliked
  - Number of songs in user playlist
  - Average songs played
  - Number of artists listened to
  - Number of user sessions logged
- More features were used initially but discarded after observing less than 1% feature importance during training of models

Modelling

- Dataset was split into 80-20 train test split
- Grid Search Cross Validation with three folds was used to built the following models
  - Gradient Boosting Trees
  - Random Forests
  - Logistic Regression
  - Support Vector Machine
  - Hybrid Model
- Goal was to maximize F-1 score since the dataset is highly imbalanced

Results



	userId	label	prediction_GBT	prediction_LGR	prediction_SVC	prediction_RF	prediction
0	1064059	0	0.0	0.0	0.0	0.0	0.0
1	1100362	0	0.0	0.0	0.0	0.0	0.0
2	1116168	0	0.0	0.0	0.0	0.0	0.0
3	1122323	0	0.0	0.0	0.0	0.0	0.0
4	1127870	0	0.0	0.0	0.0	0.0	0.0
5	1144509	0	0.0	0.0	0.0	0.0	0.0
6	1230352	0	1.0	0.0	0.0	0.0	1.0
7	1301591	0	0.0	0.0	0.0	0.0	0.0
8	1302576	0	0.0	0.0	0.0	0.0	0.0
9	1364193	0	1.0	0.0	0.0	0.0	1.0

Fig: Predictions from all models

Conclusion

- Random Forest and Gradient Boosting Trees outperformed Logistic Regression and Support Vector Machines
  - Random Forest Classifier performed the best
- Support Vector Machines did not predict any churners
- Areas for improvement
  - Try over sampling, under sampling techniques to balance the dataset and run the same classifiers to analyze results
- Other resources:
  - Project report
  - Sparkify tutorial notebook
  - Project github page: https://github.com/pdhimal1/Sparkify