**Sparkify-Capstone-Project**

**1 Project Introduction**

Imagine you are working for music streaming company like Spotify or Pandora called Sparkify. Millions of users stream their favorite songs every day. Each user uses either the Free-tier with advertisement between the songs or the premium Subscription Plan. Users can upgrade, downgrade or cancel their service at any time. Hence, it’s crucial to make sure that users love the service provided by Sparkify. Every time a user interacts with the Sparkify app data is generated. Events such as playing a song, logging out, like a song etc. are all recorded. All this data contains key insights that can help the business thrive. The goal of this project is then to analyze this data and predict which group of users are expected to churn — either downgrading from premium to free or cancel their subscriptions altogether. In this post I am going to walk you through the steps I have taken to build the model using Spark

**2 Strategy to solve the problem:**

First we load the dataset and then clean the dataset after than we Perform EDA to get insights from data and get to know the distribution of data we define label and explore the dataset. After data preprocessing we do feature engineering and prepare the data using vector assembler and string indexer after that apply random forest on this dataset and evaluate the performance and then do hyper-parameter tuning of random forest and improve the performance.

Steps are given below:

1. Load and clean the dataset
2. Exploratory Data analysis
   * Define Churn and label
   * Explore the Data
3. Feature Engineering: Create features for each user:

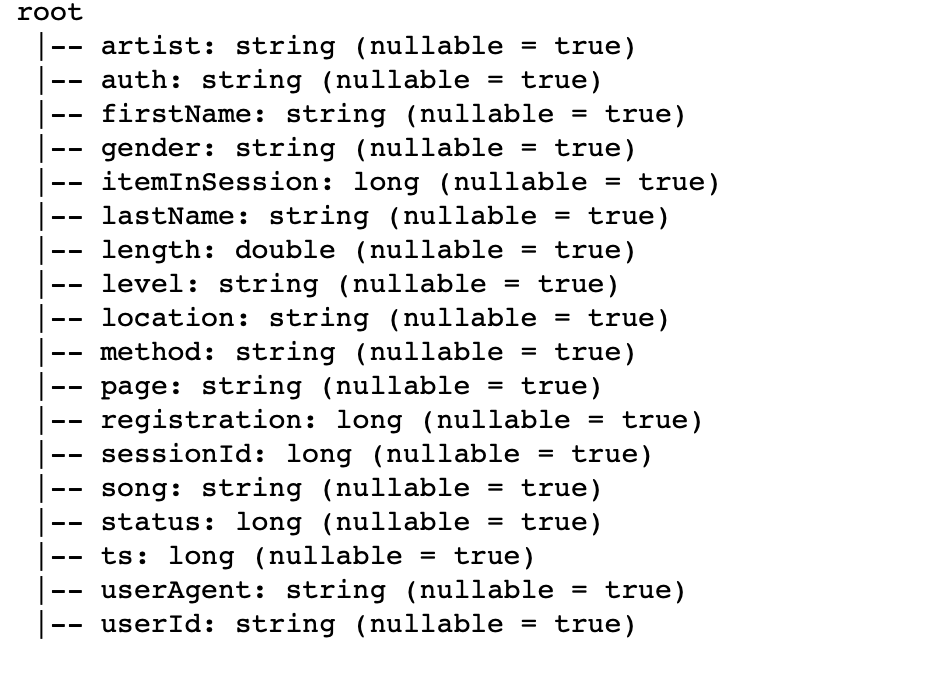
this data will be used as input to the model

1. Data transformation, data splitting and model training:  
   - Transform feature engineered data.  
   - Split data into training, validation and test data.  
   - Build a machine learning model to train using training data

Features for the dataset are as depicted below. The dataset provided has 18 features and 286500 total records.

* 1. **Load the dataset**



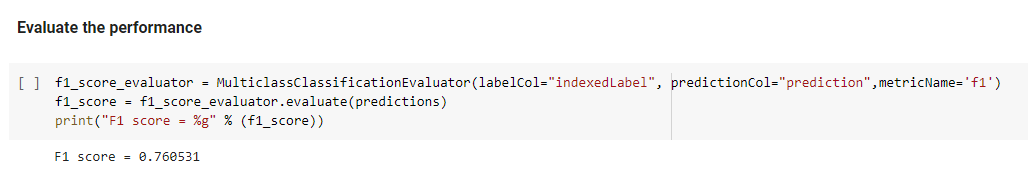


By looking at the schema we can make educated assumptions as to which of the features can be good indication of a customer which might walk away. Here are some of the features I thought are in such category.

* *song* : a song played by each user
* *registration*: user registration timestamp
* *page* : all the pages visited by a user
* *level* : free or paid

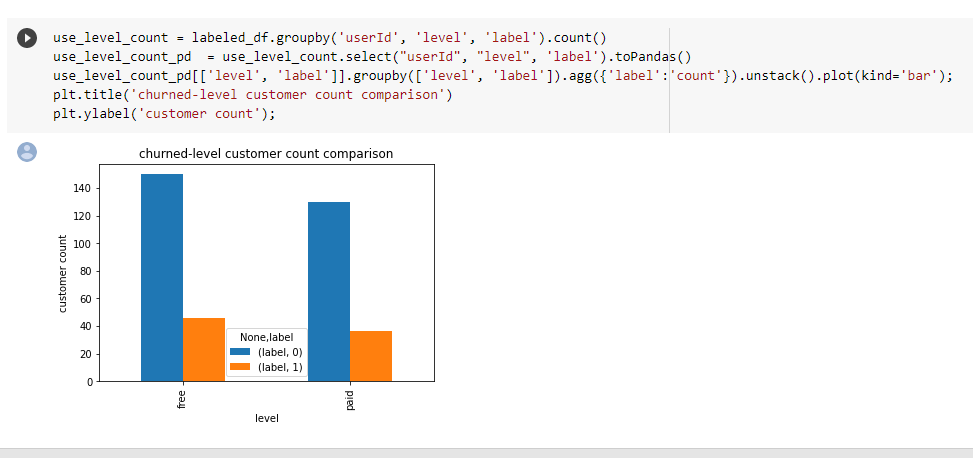
Further analysis indicates that **Cancel** page is followed by **cancellation confirmation page** and both indicate user is canceling thier subscreption unless they change thier mind once they are on the **cancel** page which seems unlikely event at least for this dataset. So we will use both of these events to define Churn. We have also seen that about 23% of the customers who downgraded have also cancelled thier subscription. However, we’ll exclude those who downgraded but didn’t cancel yet for this analysis even though they seem candidates who will eventually cancel their subscription.

**3 Metrics:**



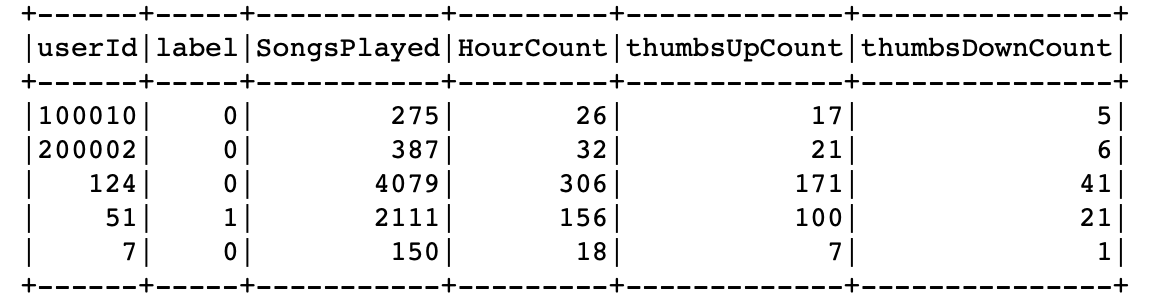
**4 EDA**

Below is the distribution after labeling our dataset





customer distribution for free vs paid {customer churned — 1 , active=0}



**5 Modeling:**

Steps in this section are:

* Split the full dataset into train, validation and test set.
* Tune selected machine learning alogirthm using the validation dataset.
* Score tuned machine learning model on the test dataset to verify it generalizes well.

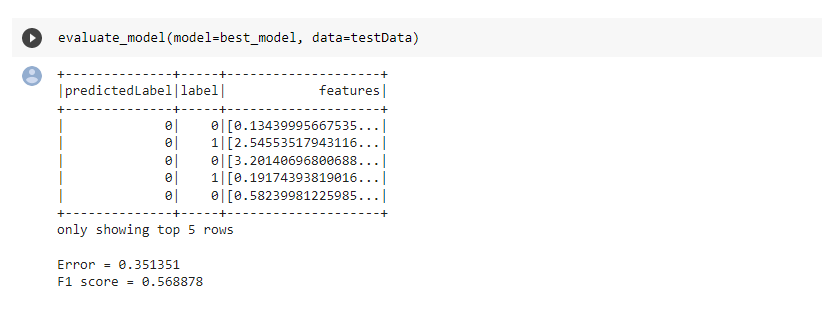
We can also select a couple of models and tune using the validation dataset. Score them on the test dataset and pick the one with the best performance on the test dataset. However, for this post I have picked only Random Forest Classifier as it has been proved to be the best off the shelf model.



**6 Hyper-parameter tuning:**

We can also select best parameter of models and tune using the validation dataset. Score them on the test dataset and pick the one with the best performance on the test dataset. However, for this post I have picked only Random Forest Classifier as it has been proved to be the best off the shelf model.

****

****

**7 Results**

We have analyzed the sparkify dataset and come up with new features to predict churn. We then created a machine learning model and tuned it to improve its performance. We achieved an accuracy score of 60% and F1 score of 56% on the test dataset.

**8 Conclusion**

The model performance is satisfactory with default parameter and results improve a little with hyper-parameter tuning so the main thing is to add the features of dataset so model need to train completely. After that model have many instances to boosting and decision tree is also good option instead of random forest

**9 Improvement:**

The model performance can be further improved by creating additional features and including some of the features that I have left out for this analysis. The model should also be tested using samples from the left out big dataset which hasn’t been used for this analysis. We can also pick other models to see if any of them can perform better. Once we are satisfied with the result, a large scale of the model can be implemented on the cloud.