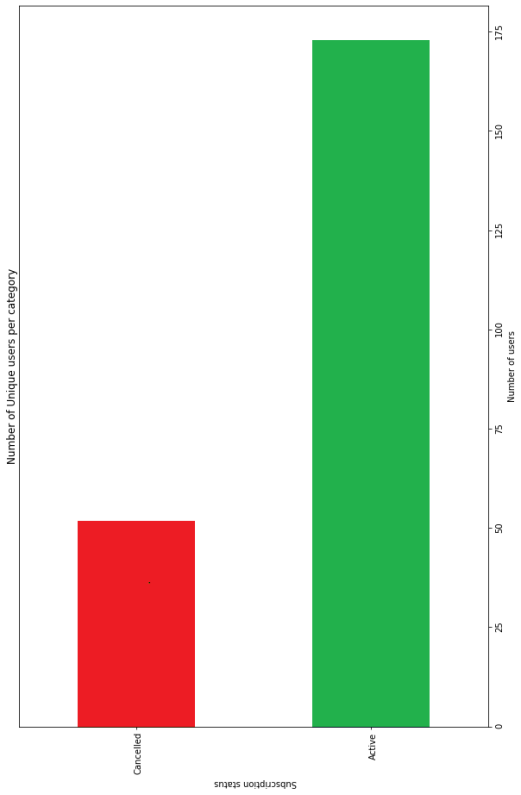
Hold to them dearly! A blog post on preventing customer churn by discovering patterns of data!

A blog post on preventing customer churn by discovering patterns of data!

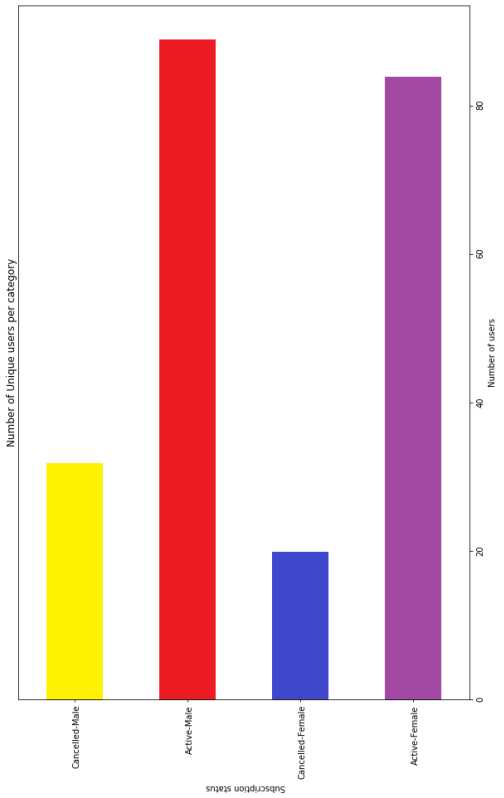
February 26, 2020

INTRODUCTION:
Now, the trend is companies adopting monthly subscription rather than one time billing. Since, in such cases, the customer may cancel, it is better to predict this with certain signatures based on Customer activity and nature, in order to maximize the membership and prevent Churn(i.e. Loss of existing Customer).
With this Aim, The Sparkify Capstone project of DSND is designed to help tackle such issue for a fictitious company 'Sparkify'. Since the given customers dataset is huge (12GB), we worked with only a 128MB slice of the original dataset.

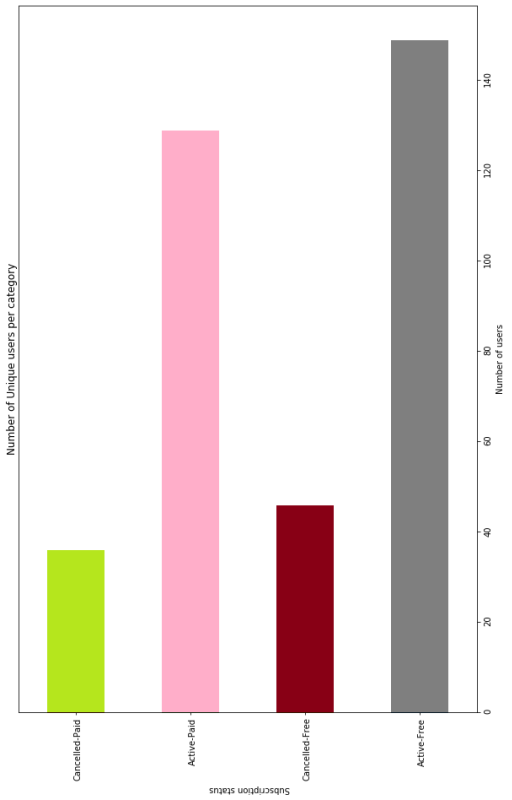


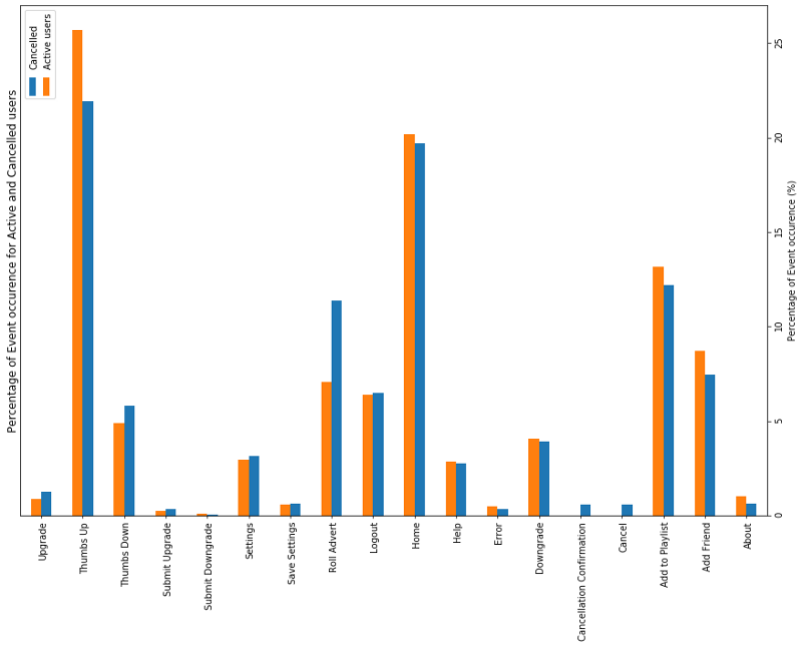
Some records contain empty userId or empty sessionID, these are the logged-out users, so we should first drop them all.
The dataset lacks an indicator of whether the user churns or not, we should add a churn field for that.

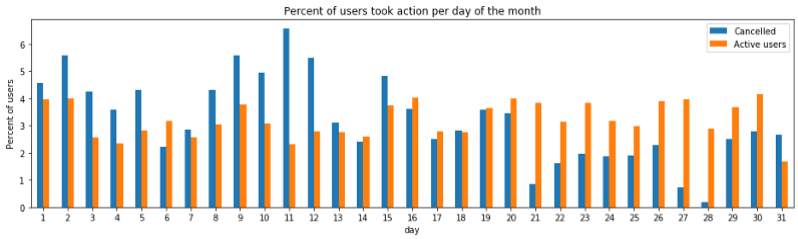
We see a 25% of cancellation, very large. We now explore the churn users by gender or by subscription.



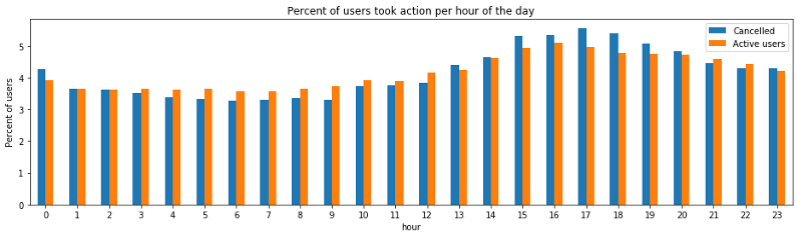
It appears that the paid/free status is not influencing termination of the account. Gender seems to affect churn decision.

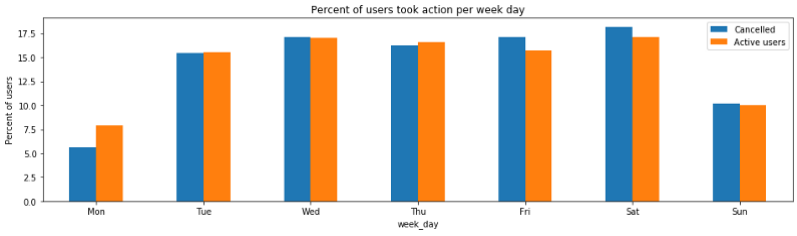




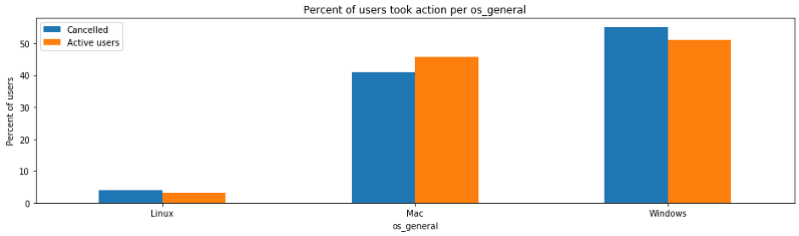


Cancellation high at end of month to avoid renewal fees. The usage is high at beginning of month.

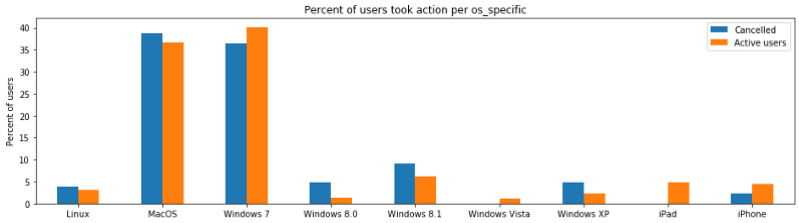




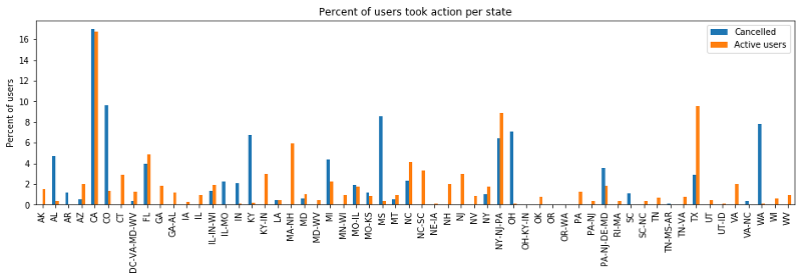
Both Active and Cancelled user is high on the weekdays.



Mac users have lower relative cancellation rate than others.



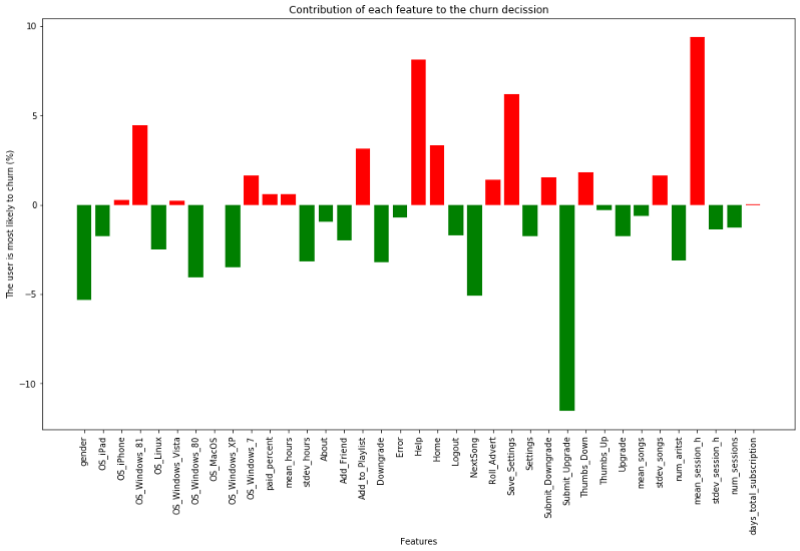
Ipad users are more satisfied due to no cancellation followed by Vista and Iphone. Maybe due to better product satisfaction of Ipad and Iphone.



DIFFERENT STATES AND ACTION TOOK BY USERS

MODELING:

We have normalized all the input features and combined them into one vector. Then we divided the dataset into 80% for training the model, and 20% for testing.Now, lets compare the efficacy of different ML models:

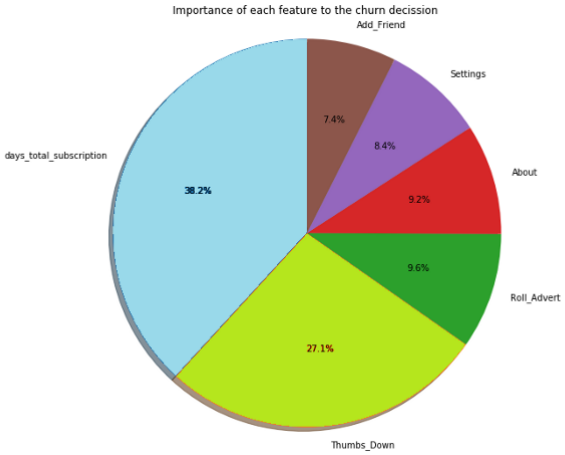


Logistic Regression

Logistic Regression Model:

The accuracy of the logistic regression model is relatively good, 82% and 75% for the training and the testing datasets. the other measures like precision, recall, and F-score are slightly fewer than the accuracy values. This shows good performance of the model to detect churn customers.

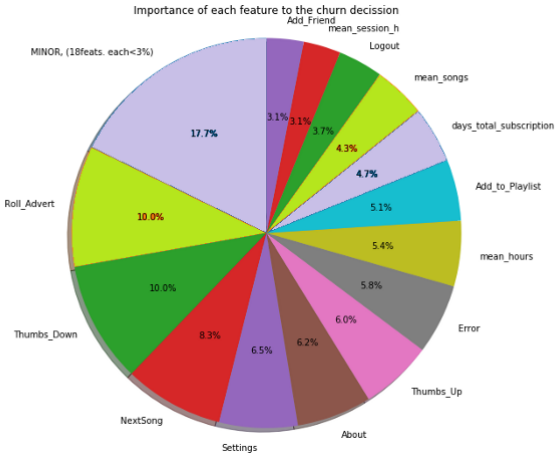
Happy Songs are feature indicating no churn. But, save settings,Help, mean_Session hours indicate chance of high churn.



Decision Tree Classifier

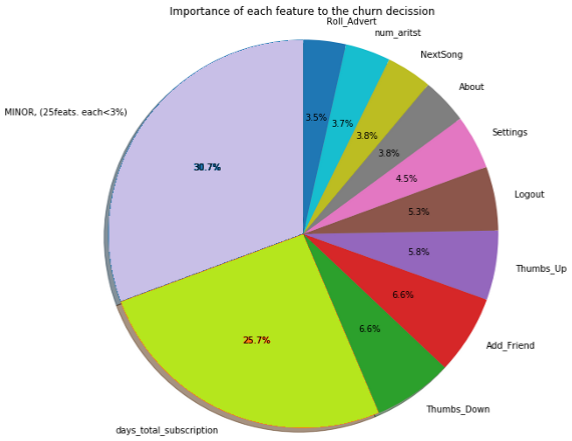
Decision Tree Classifier Model:

This model has a 'Features importance' output, which indicates that this feature influences the results more, regardless of whether it has a positive or negative effect. The features importance show that the most influencing feature is the days_total_subscription, which indicate an effect of the subscription length to the churn possibility. The second feature is the amount of thumbs_down , the Roll_advert, and the other shown features.



Gradient Boosted Trees

Gradient-Boosted Trees Model:
This model, has higher accuracy and performance measures on the training data set than the previous two, but the results on the test data set are worse, which means that the model over fits the data. The features' importance shows that the most important features are the NextSong visits (number of songs played), which appear to be an indicator of satisfied customer, as well as the Thumbs_UP indicator. The Error page is the second runner up here, which appears to indicate that the user is almost bored of errors, and will leave soon.



Random Forest

Random Forest Model:
This model, like the GBT before, has over-fitting, with very high training accuracy, and low testing accuracy. The random forest model agrees with the Decision Tree Classifier in the features' importance, as both show the most important indicators are the days_total_subscription and the Thumbs_Down while it agrees with the GBT on including all the features as important somehow.
Note: All the features that have less than 3% importance are collected in the MINOR category.

Conclusion:
The ML models sucessfully helped in finding the customer patterns leading to the cancellation. ‘Decision-Tree Classifier’ is the best among all! So, solution is to avoid the user cancellation by providing discounts a month end if the user is shown using features such as ‘Help’, ‘Save Settings’, is a male and is using Linux or Windows OS.

Enter your comment...