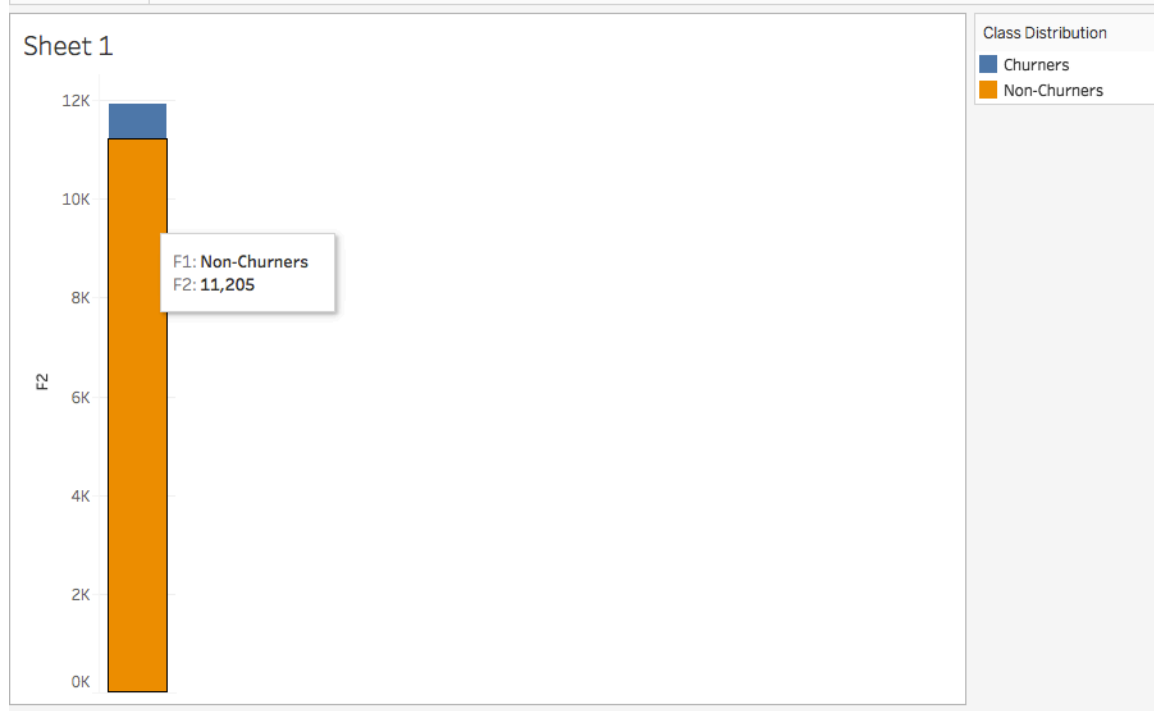# Preliminary Report

James Mwakichako Manoj Kumar

Predicting 4th Month Churn Rate

Number of Rows = 11925

Class Distribution:

The skewness towards Non-Churners adversely affects precision

Predicting 4th Month:

We used Jan-March 2014 for our analysis as this contained the most number of subscribers. We used decision tree and logistic regression models.
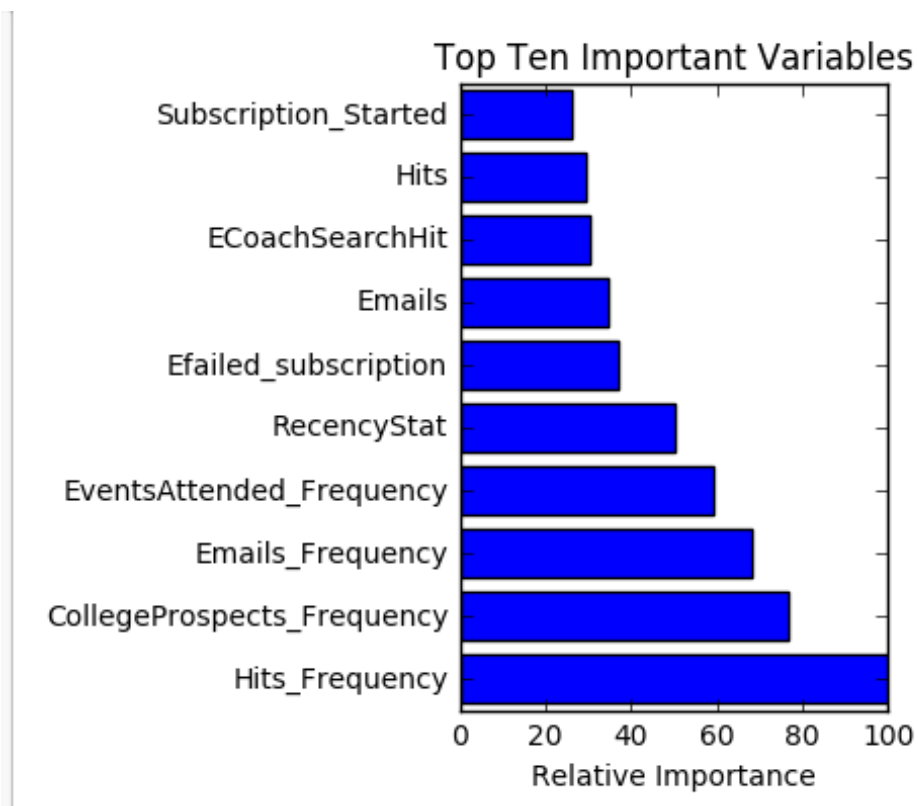
	Actual Churners	**Actual – Non Churners**
Predicted Churners	226	219

Predicted Non-Churners	55	4153

Precision = 0.51
Recall = 0.8

<u>Top 10 Most Important Features</u>

Using the decision Tree model, below are the top 10 most important features.

Top Ten Important Variables

| Variable	Relative Importance
Subscription_Started	~26
Hits	~29
ECoachSearchHit	~31
Emails	~35
Efailed_subscription	~37
RecencyStat	~50
EventsAttended_Frequency	~59
Emails_Frequency	~68
CollegeProspects_Frequency	~77
Hits_Frequency	100

Predicting Churn at Anytime

We cleaned the whole database and used 75% for testing and the remaining 25% for testing to predicting churn.

Number of Rows = 133419
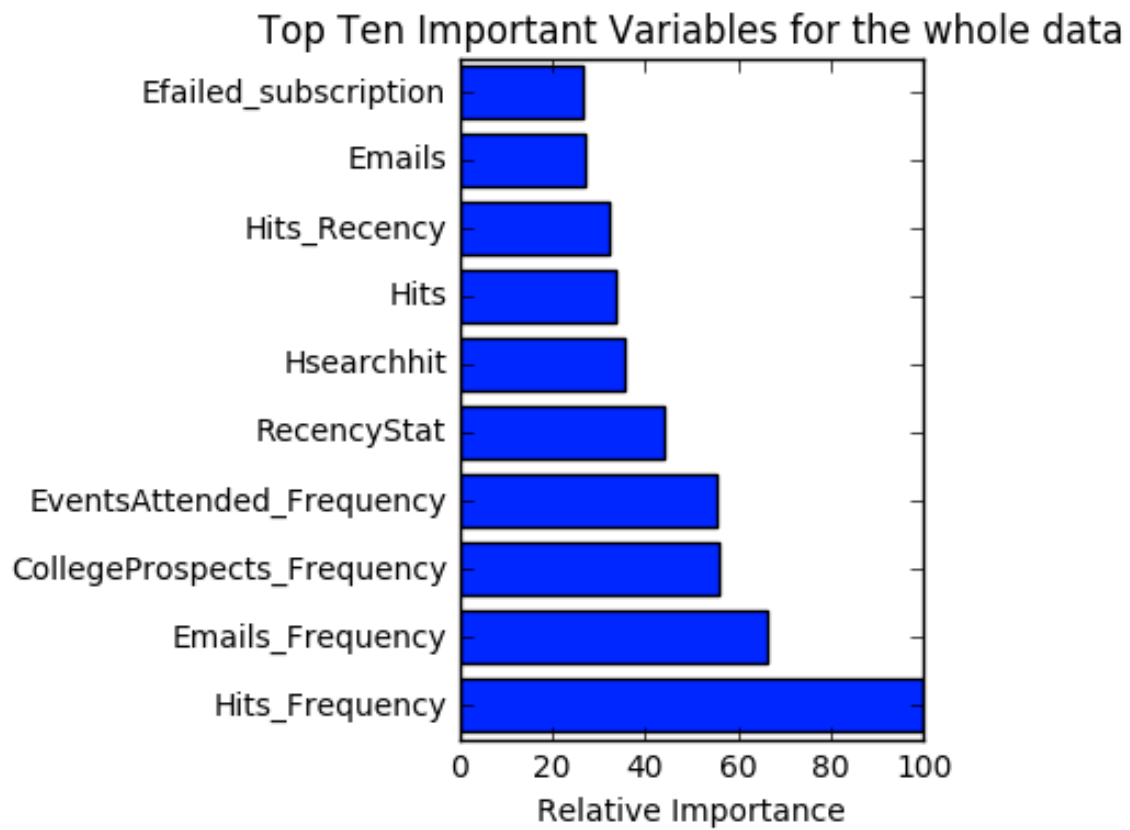
Class Distribution
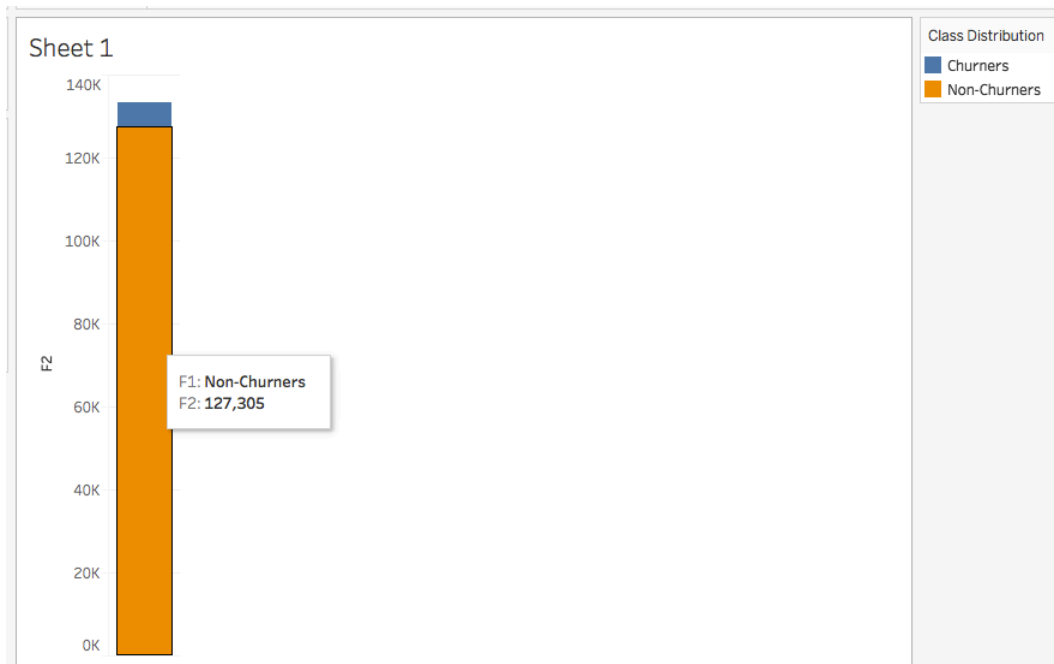
Confusion Matrix

	Actual Churners	Actual – Non Churners
Predicted Churners	1323	1332

Predicted Non-Churners	168	30532

Precision = 0.5
Recall = 0.89

Feature Importance

Top Ten Important Variables for the whole data

Variable	Relative Importance
Efailed_subscription	
Emails	
Hits_Recency	
Hits	
Hsearchhit	
RecencyStat	
EventsAttended_Frequency	
CollegeProspects_Frequency	
Emails_Frequency	
Hits_Frequency	

Further Work

- Explore Cohort Analysis and Time Series models
- Talk to the team more on significance of some columns eg Hits_Recency as this would impact how we impute missing values