**Preliminary Report**

James Mwakichako Manoj Kumar

**Preprocessing**

The preprocessing step entailed three main steps:

1. Removing rows representing inactive subscription (Subscription = 0)
2. Removing columns whose data was completely missing (Status, )
3. Introducing dummy variables for categorical data eg (Sport and Gender)
4. Imputing missing values

This yielded 133419 rows and 3680 columns

**Predicting 4th Month Churn Rate**

Training data size(Jan-March 2014): 11925 rows

Testing data(April 2014): 4653 rows

Class Distribution:



The skewness towards Non-Churners adversely affects precision

Predicting 4th Month:

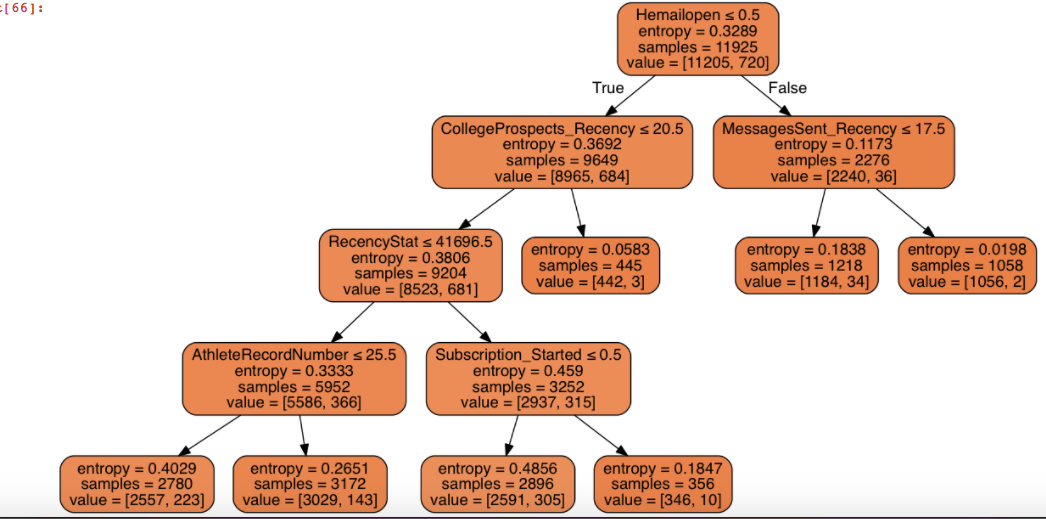
We used Jan-March 2014 for our analysis as this contained the most number of subscribers. We used decision tree and logistic regression models.

|  |  |  |
| --- | --- | --- |
|  | **Actual Churners** | **Actual – Non Churners** |
| **Predicted Churners** | 226 | 219 |
| **Predicted Non-Churners** | 55 | 4153 |

Precision = 0.51

Recall = 0.8

Tree Visualization



Above is the visualization of the Decision Tree based on entropy as the impurity criterion. Based on the preliminary results, in predicting churn in the following month, Hemailopen seems to be the most important feature. But this is very early to conclusively say that.

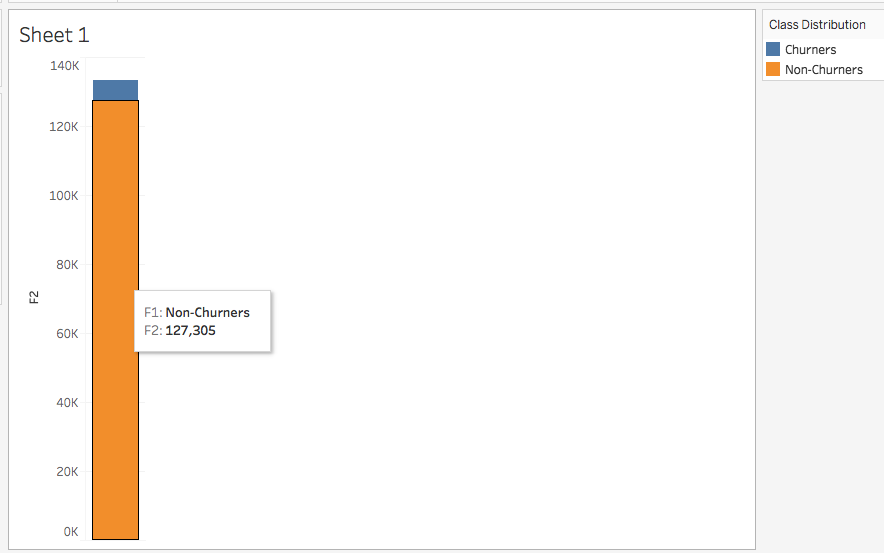
We played with a few DT parameters and finally settled on limiting maximum depth to 5 and number of leaf nodes to 7.

**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**



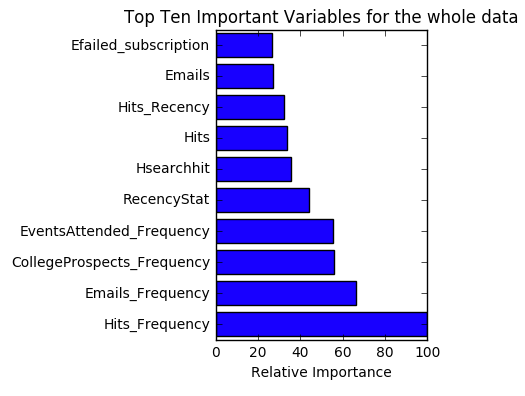
**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | **Actual Churners** | **Actual – Non Churners** |
| **Predicted Churners** | 1323 | 1332 |
| **Predicted Non-Churners** | 168 | 30532 |

**Precision = 0.5**

**Recall = 0.89**

**Feature Importance**

****

Feature importance here implied ‘Gini Importance’. The normalized reductions of gini index brought about by a particular feature. The more important a feature, the higher it’s feature importance. In the above example, Hits frequency is deemed most important.

**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