**Data Science Practicum Report**

**Mar 20 th 2017**

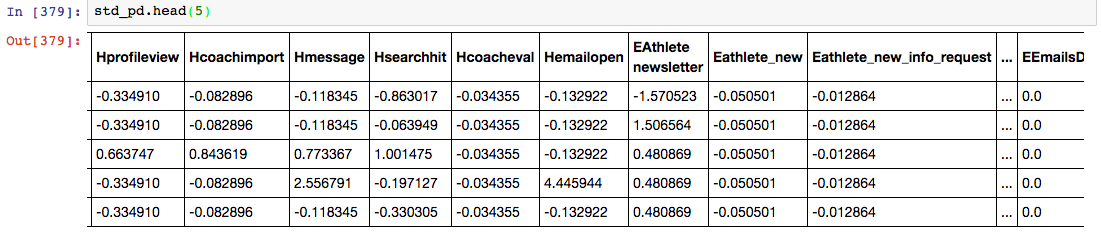
**Preprocessing**

As agreed in the meeting on Thursday 23rd, in addition to remove columns with >80% missing values, we kept columns suggested by Joel in the excel file sent on Monday Feb 27th. Our final list of 79 features was as follows:

|  |  |  |
| --- | --- | --- |
|  | ECCNote | Eparent\_welcome |
| EventsAttended | ECCNote\_camp | Epost\_event\_email |
| Hprofileview | Ecoach\_list\_known\_updated | Esms\_update |
| Hcoachimport | ECoachEmailOpen | CollegeProspects |
| Hmessage | ECoachEval | MessagesReceived |
| Hsearchhit | ECoachImport | MessagesSent |
| Hcoacheval | ECoachSearchHit | CaptainU\_CHURN |
| Hemailopen | ECoachVisit | NumYear |
| EAthlete newsletter | Ecolleges\_going\_to\_the\_event | NumMonth |
| Eathlete\_new | Efailed\_subscription | monthly\_price |
| Eathlete\_new\_info\_request | EEmailsDigest | Eparent\_new |
| Gender | Hprofileview\_Freq | Hcoachimport\_Freq |
| Hmessage\_Freq | Hsearchhit\_Freq | Hcoacheval\_Freq |
| Ecolleges\_going\_to\_the\_event\_Freq | Efailed\_subscription\_Freq | Esms\_update\_Freq |
| Hits\_Frequency | College\_Prospects\_Frequency |  |
| Hemailopen\_Freq | Ecoach\_list\_known\_updated\_Freq | ECoachVisit\_Freq |

We then took columns with continuous values and normalized the values by the mean and the standard deviation:

Hence each feature had a mean of 0 and a standard deviation of 1. Below is a snapshot of the data:



We then created dummy variables from features with categorical features: gender .

We built models ignoring the sport a student plays.

Final table shape: **16117 rows 57 columns**

**Predicting Churn in Girls**

This week, we focused on predicting lifetime churn. We implemented three models machine-learning models recorded and visualized Precision and Recall Values.

Below are the three models we chose:

* Decision Trees
* Logistic Regression
* Support Vector Machines (SVM)

When it came to building models, we used 80% of the data for training and the remaining 20% for testing.

Training data: **7020 rows 57 columns**

Test data: **: 1755 rows 57 columns**

**Class distribution of Churners and Non-Churners**

**Training set**

|  |  |
| --- | --- |
| **Non-Churners** | **4060** |
| **Churners** | **2960** |

**Testing Set**

|  |  |
| --- | --- |
| **Non-Churners** | **1006** |
| **Churners** | **749** |

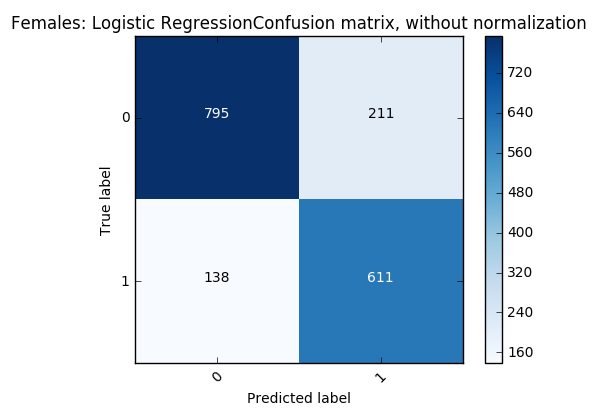
**Summary of Precision Recall Values of Churning**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1 Score** |
| Decision Trees | 0.71 | 0.73 | 0.72 |
| Logistic Regression | 0.74 | 0.81 | 0.78 |
| SVM | 0.73 | 0.83 | 0.78 |

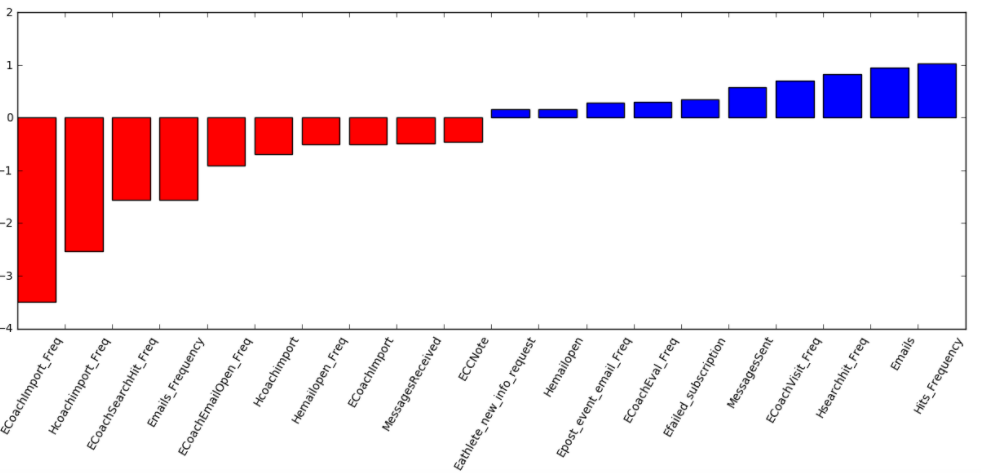
**Summary of Precision Recall Values of Retention**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1 Score** |
| Decision Trees | 0.79 | 0.78 | 0.78 |
| Logistic Regression | 0.85 | 0.79 | 0.82 |
| SVM | 0.86 | 0.79 | 0.82 |

**Logistic Regression Visualization**



**Important Features According to Logistic Regression**



**Predicting Churn in Boys**

**Summary of Precision Recall Values of Churn**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1 Score** |
| Decision Trees | 0.67 | 0.65 | 0.66 |
| Logistic Regression | 0.74 | 0.63 | 0.68 |
| SVM | 0.72 | 0.62 | 0.67 |

**Summary of Precision Recall Values of Retention**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1 Score** |
| Decision Trees | 0.84 | 0.83 | 0.83 |
| Logistic Regression | 0.83 | 0.89 | 0.85 |
| SVM | 0.83 | 0.89 | 0.85 |

**Class distribution of Churners and Non-Churners**

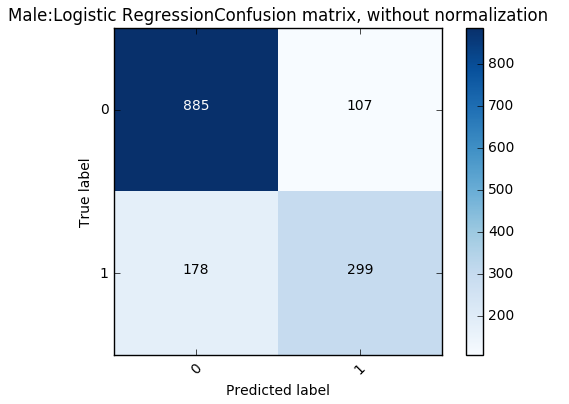
**Training set**

|  |  |
| --- | --- |
| **Non-Churners** | **4016** |
| **Churners** | **1857** |

**Testing Set**

|  |  |
| --- | --- |
| **Non-Churners** | **992** |
| **Churners** | **477** |

**Logistic Regression Visualization**



**Important Features According to Logistic Regression**

