**Data Science Practicum Report**

**Feb 28th 2017**

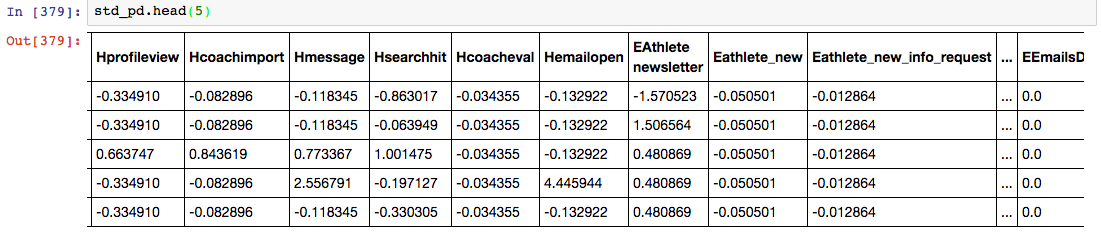
**Preprocessing**

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 33 features was as follows:

|  |  |  |
| --- | --- | --- |
| sport | 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 |  |  |

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 and sport .

Final table shape: **16117 rows 51 columns**

**Model Building**

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 3 months worth of data (Jan – March 2014) and used it to predict April 2014 churn.

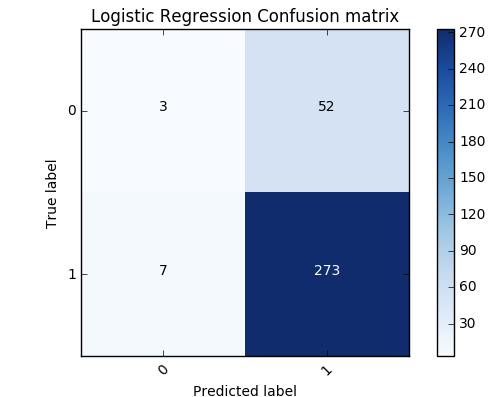
Training data: **827 rows 51 columns**

Test data: **: 335 rows 51 columns**

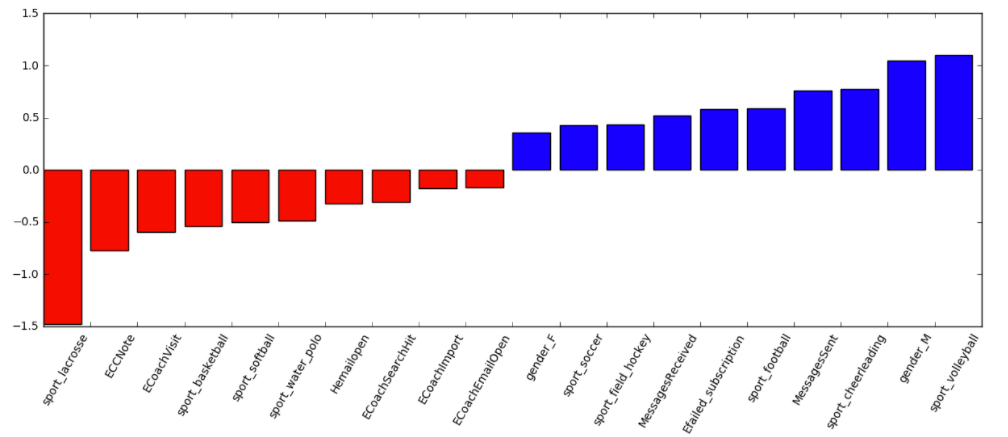
Summary of Precision Recall Values

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1 Score** |
| Decision Trees | 0.84 | 0.84 | 0.84 |
| Logistic Regression | 0.84 | 0.97 | 0.90 |
| SVM | 0.84 | 1.0 | 0.91 |

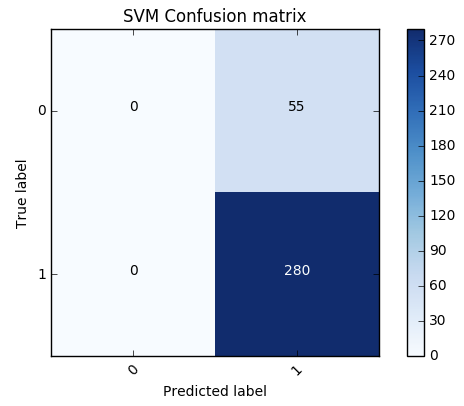
**Logistic Regression Visualization**



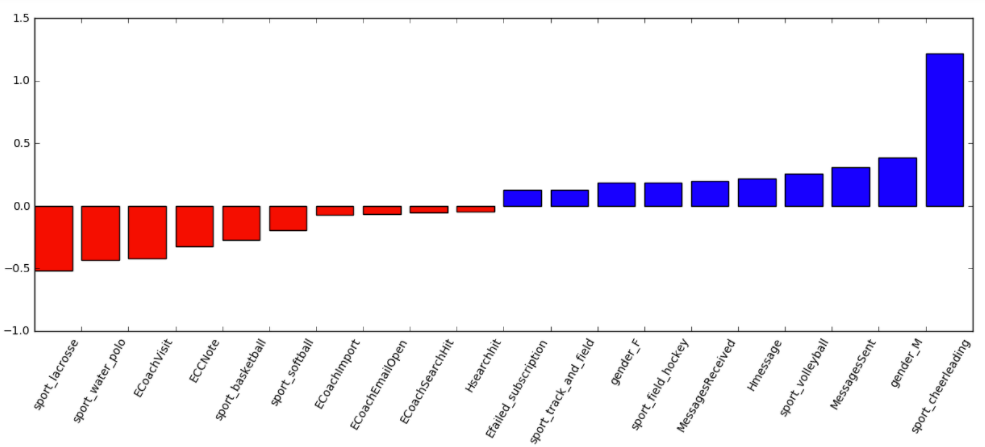
**Important Features According to Logistic Regression**



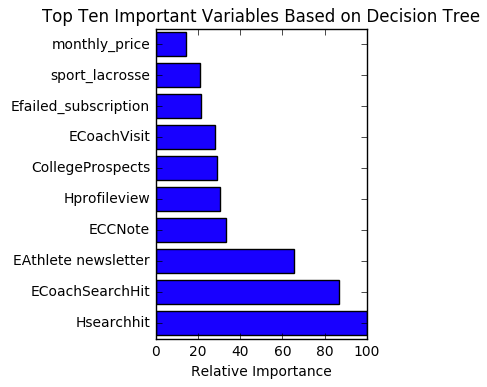
**Support Vector Machine Visualization**



**Important Features According to SVM**



**Feature Importance of Decision Trees**



Feature importance is based on the gini-index of each feature .

**Top 10 features that drive Churn**

|  |  |
| --- | --- |
| **Logistic Regression** | **SVM** |
| Sport\_volleyball | Sport\_cheerleading |
| Gender\_M | Gender\_M |
| Sport\_cheerleading | Message\_Sent |
| Message\_Sent | Sport\_volleyball |
| Sport\_football | Hmessage |
| Efailed\_subscription | MessagesReceived |
| Sport\_field\_hockey | Sport\_field\_hockey |
| MessagesReceived | Gender\_F |
| Sport\_soccer | Sport\_track\_and\_field |
| Gender\_F | Efailed\_subscription |

**Note:** Values are order based on descending order of significance**Top 10 features that drive Retention**

|  |  |
| --- | --- |
| **Logistic Regression** | **SVM** |
| Sport\_lacrosse | Sport\_lacrosse |
| ECCNote | Sport\_water\_polo |
| ECoachVisit | ECoachVisit |
| Sport\_basketball | ECCNote |
| Sport\_softball | Sport\_basketball |
| Sport\_water\_polo | Sport\_softball |
| Hemailopen | Ecoachimport |
| ECoachSearchHit | ECoachEmailOpen |
| Ecoachimport | ECoachSearchHit |
| ECoachEmailOpen | Hsearchhit |

**Note:** Values are order based on descending order of significance

**Meeting Notes**

* The SVM’s recall value is particularly too good to be true. We will need more data to confirm this. We will take a different slice of the data and see if this holds.
* We believe adding features with a cumulative values for all current features might help improve precision