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:

$$z = \frac{x - \mu}{\sigma}$$

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

```
In [379]: std_pd.head(5)
```

Out[379]:

Hprofileview	Hcoachimport	Hmessage	Hsearchhit	Hcoacheval	Hemailopen	EAthlete newsletter	Eathlete_new	Eathlete_new_info_request	...
-0.334910	-0.082896	-0.118345	-0.863017	-0.034355	-0.132922	-1.570523	-0.050501	-0.012864	...
-0.334910	-0.082896	-0.118345	-0.063949	-0.034355	-0.132922	1.506564	-0.050501	-0.012864	...
0.663747	0.843619	0.773367	1.001475	-0.034355	-0.132922	0.480869	-0.050501	-0.012864	...
-0.334910	-0.082896	2.556791	-0.197127	-0.034355	4.445944	0.480869	-0.050501	-0.012864	...
-0.334910	-0.082896	-0.118345	-0.330305	-0.034355	-0.132922	0.480869	-0.050501	-0.012864	...

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)

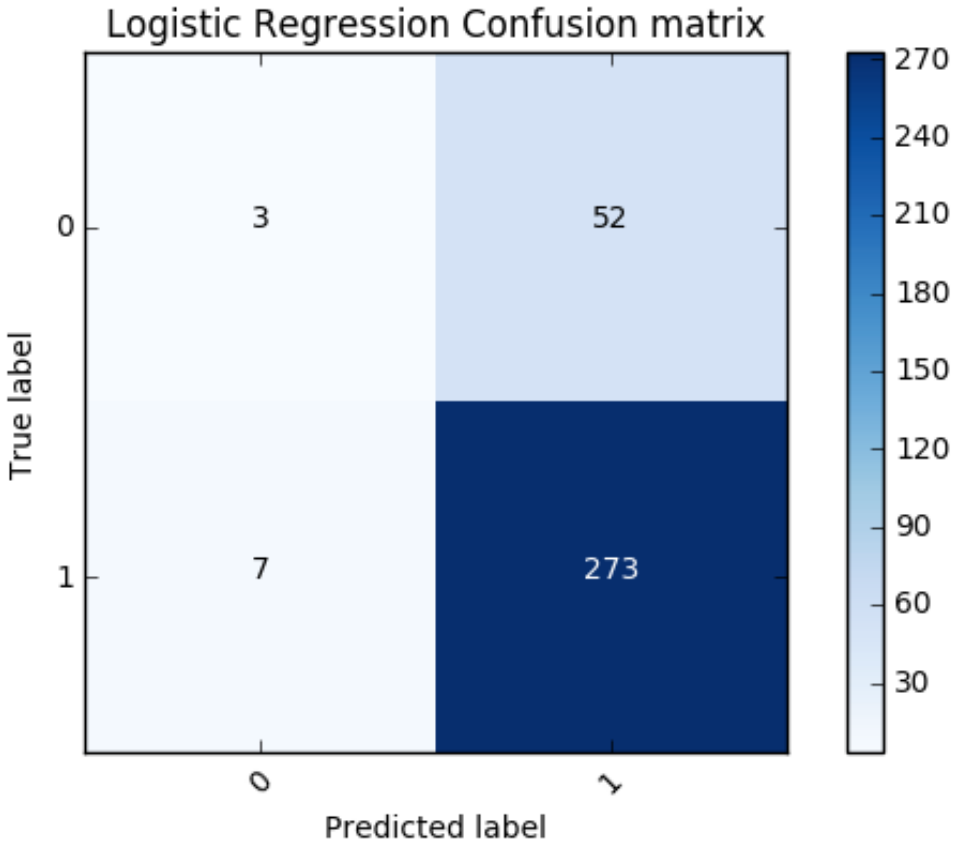
When it came to building models, we used 3 months worth of data (Jan – March 2014) and used it to predict April 2014 churn.

Training data: **827 rows 51 columns**
Test data: **: 335 rows 51 columns**

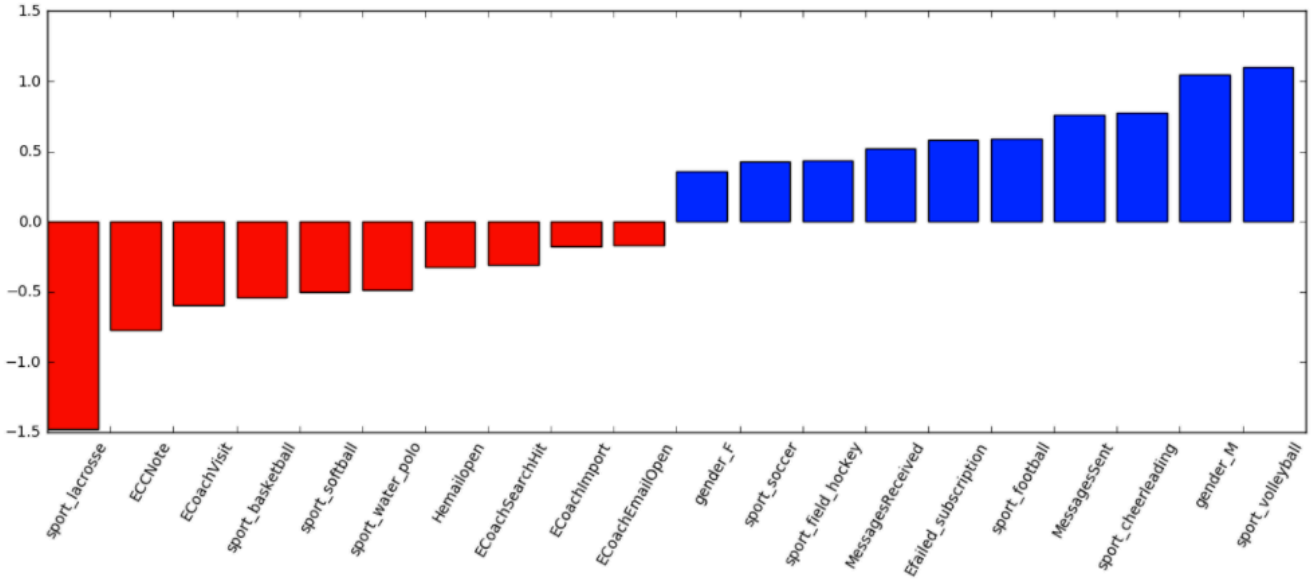
Summary of Precision Recall Values

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

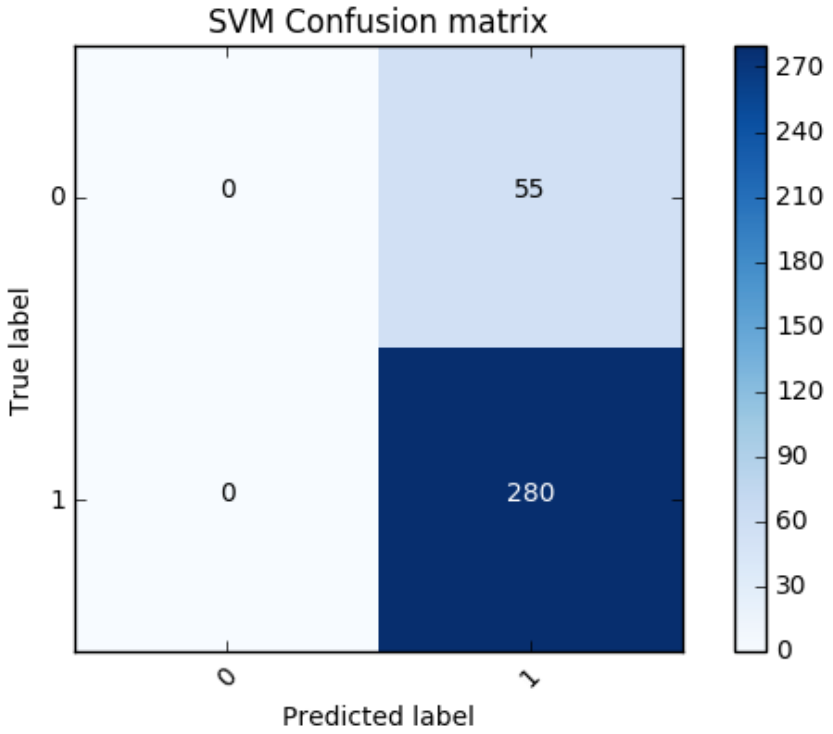
Logistic Regression Visualization



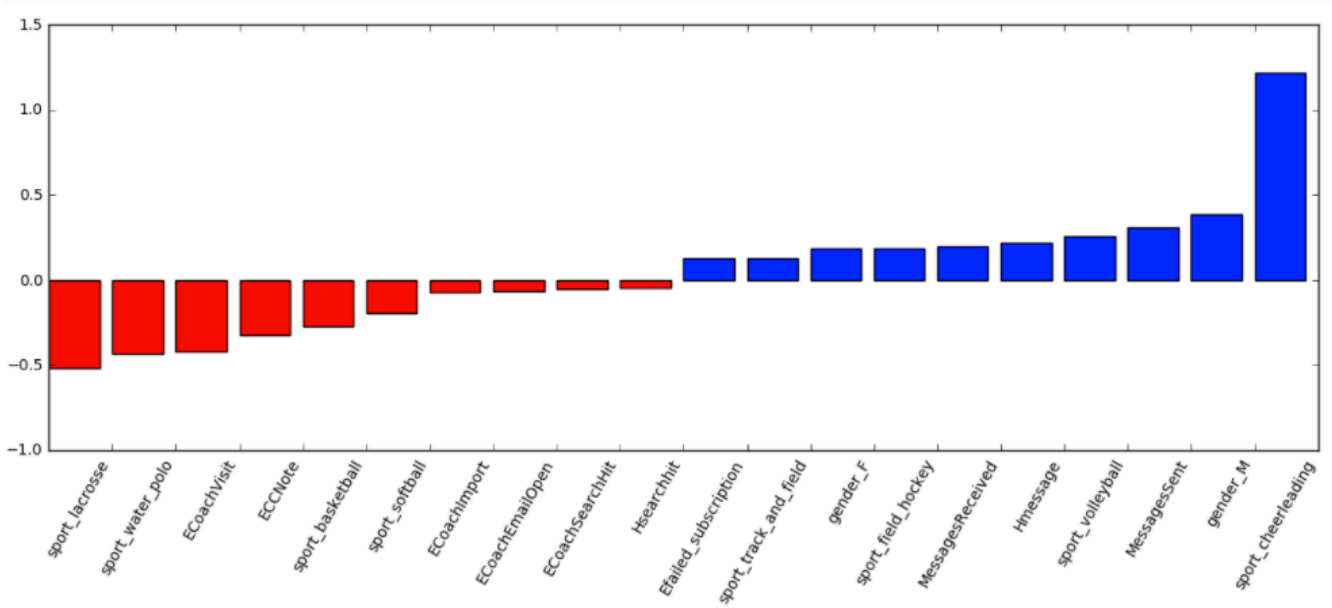
Important Features According to Logistic Regression



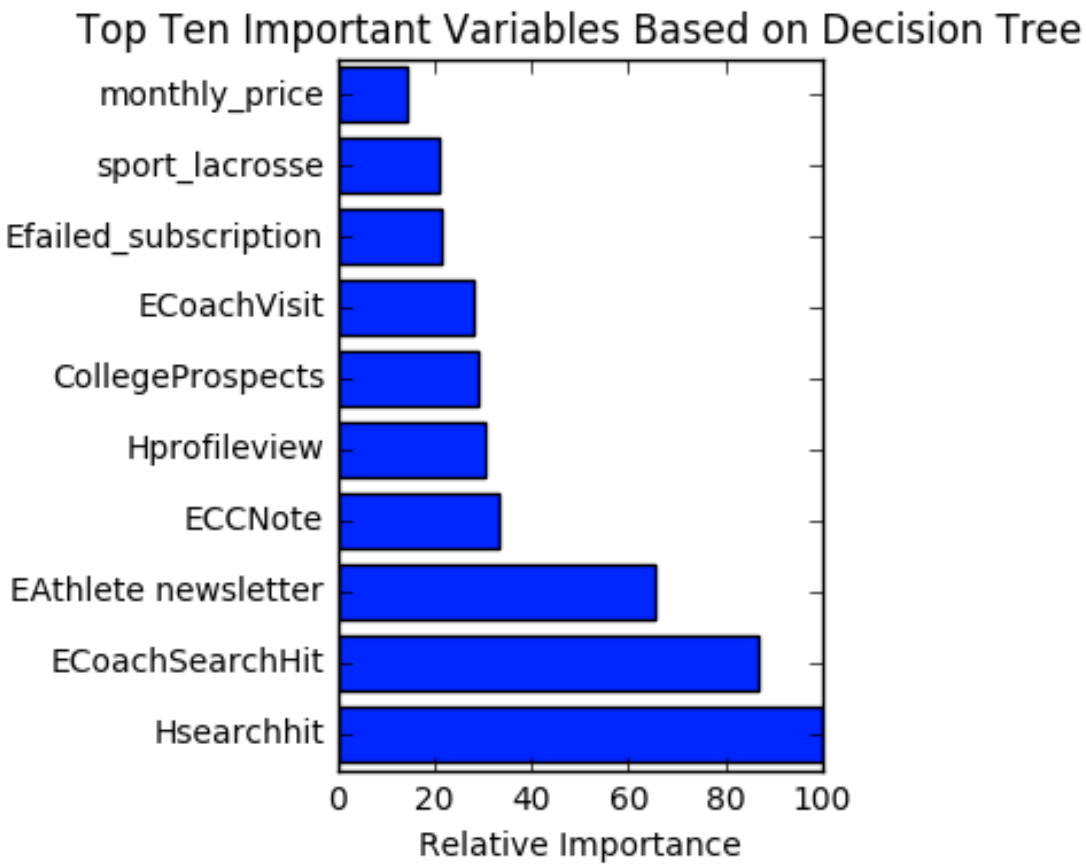
Support Vector Machine Visualization



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