DATA SCIENCE PRACTICUM: PREDICTING CHURN

JAMES MWAKICHAKO & MANOJ KUMAR

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Churn rate, according to the dictionary is the annual percentage rate at which customers stop subscribing to a service or employees leave a job. In the context of CaptainU, and specifically from the perspective of high school athletes, an athlete is considered to have churned if they cancel their subscription before making a college team or canceling before the spring semester of their senior year.

Therefore, the following scenarios are not considered churn:

1. When an athlete makes a team and then cancels his/her subscription
2. When an athlete cancels his/her subscription in the spring of their senior year

* *Department of Data Science, Illinois Institute of Technology, Chicago, United States*
[1] *Department of Data Science, Illinois Institute of Technology, Chicago, United States*

1 INTRODUCTION

The primary goal of the practicum was to predict athletes who are most likely to churn early enough so that steps could be taken to mitigate churn. To aid in answering this question, a two pronged approach was taken.

1. Predicting lifetime churn - This method sought to predict the likelihood of a athlete churning at some point in their high school career. This was an easier approach to take and it helped us understand important features and what machine learning models to implement. The main drawback to this approach is that it doesn't have a strong business usecase. Saying ' Athlete A will churn at some point is not as actionable as the same athlete churning in the next month or two. '
2. Predicting one month churn - In this approach we sought to answer, given the monthly following transaction of athlete A, what is his/ her probability of churning in the next month ? Modeling this problem is slightly more challenging than the first but more beneficial

1.1 Dataset

To train and test our machine learning models, we used data provided by CaptainU. Specifically we used MSG_RFM table. We also focused on active subscriptions. Active subscriptions refer to athletes who are paying a monthly fee to be on the system.

2 LIST OF FEATURES

In this section we shall briefly discuss the features we used for the machine learning models. For all the models we shall present, we used the same list of features. This list is a subset of the features in MSG_RFM table. The histograms that accompany gives the mean and standard deviation of each feature given active subscriptions.

Below is a summary of the features used.

Table 1: Final List of Features

gender_F	gender_M
EventsAttended	Hprofileview
Hcoachimport	Hmessage
Hsearchhit	Hcoacheval
Hemailopen	EAthlete newsletter
Eathlete_new	Eathlete_new_info_request
ECCNote	ECCNote_camp
Ecoach_list_known_updated	ECoachEmailOpen
ECoachEval	ECoachImport
ECoachSearchHit	ECoachVisit
Ecolleges_going_to_the_event	Efailed_subscription
Eparent_new	Eparent_welcome
Epost_event_email	Esms_update
CollegeProspects	MessagesReceived
MessagesSent	monthly_price
CaptainU_CHURN	

2.1 Gender

The histogram below shows that female interacted with CaptainU more that males.

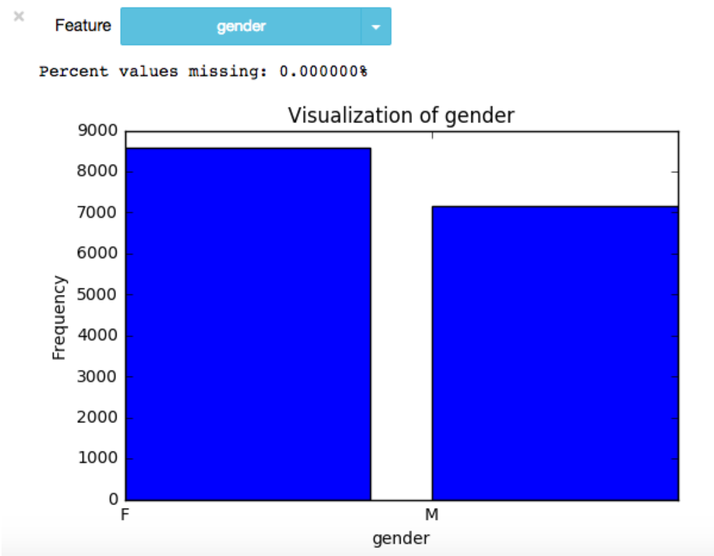

Figure 1: Gender Histogram

2.2 CaptainU_CHURN

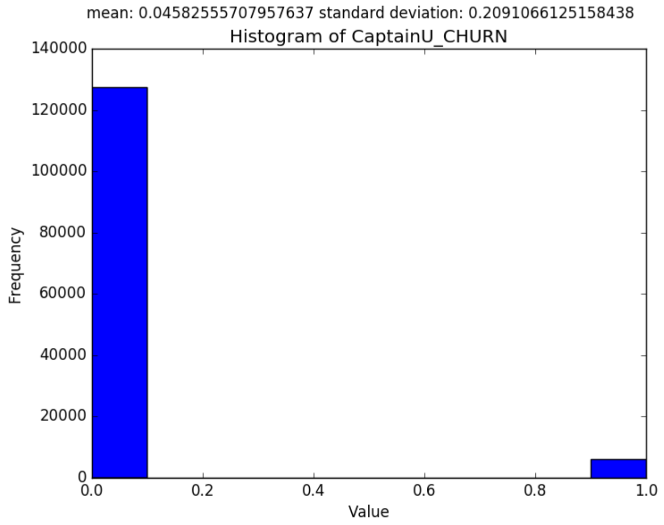

Figure 2: Gender Histogram

2.3 College Prospects

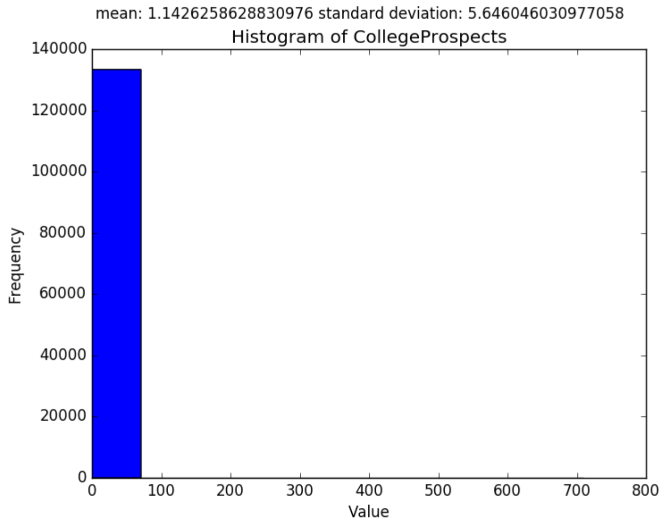

Figure 3: Gender Histogram

2.4 Duration

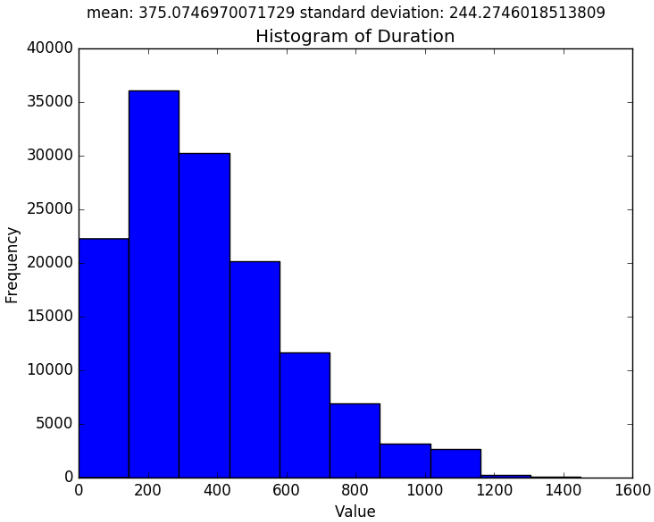

Figure 4: Gender Histogram

2.5 New Athlete Email

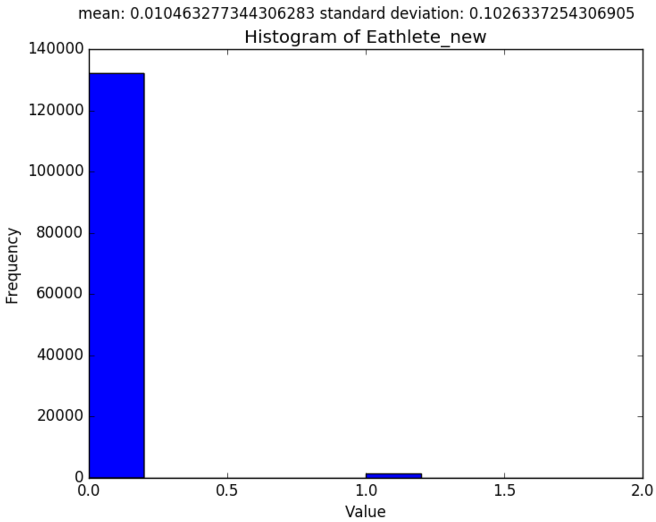

Figure 5: Gender Histogram

2.6 Athlete Newsletter

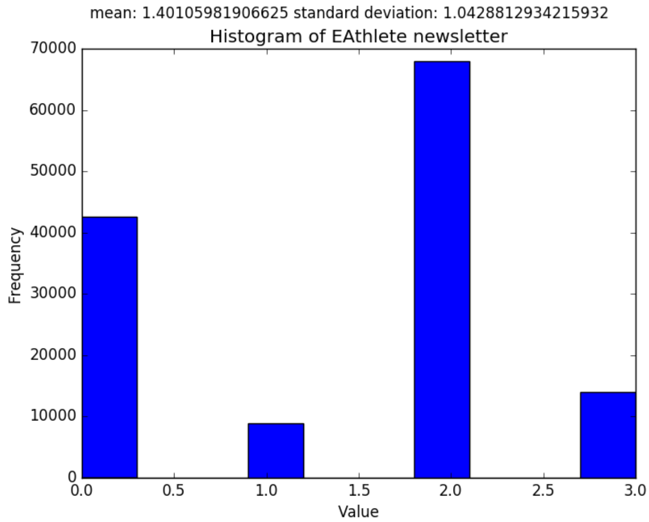

Figure 6: Gender Histogram

2.7 ECoachEmailOpen

This represents the number of times a coack opened an email an athlete sent to them.

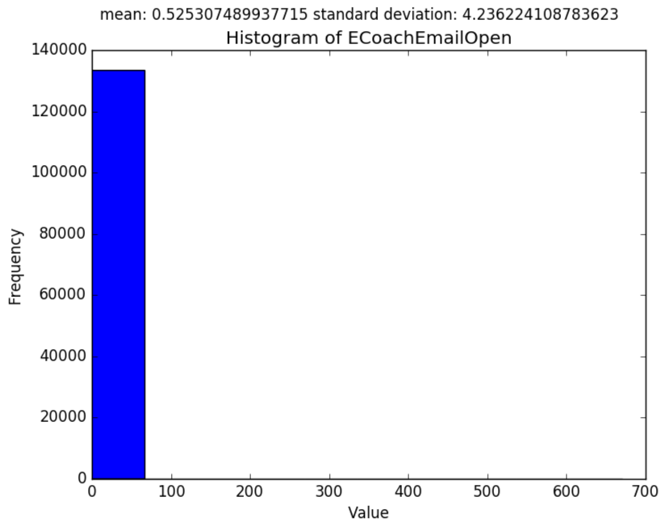

Figure 7: Gender Histogram

2.8 ECoachEval

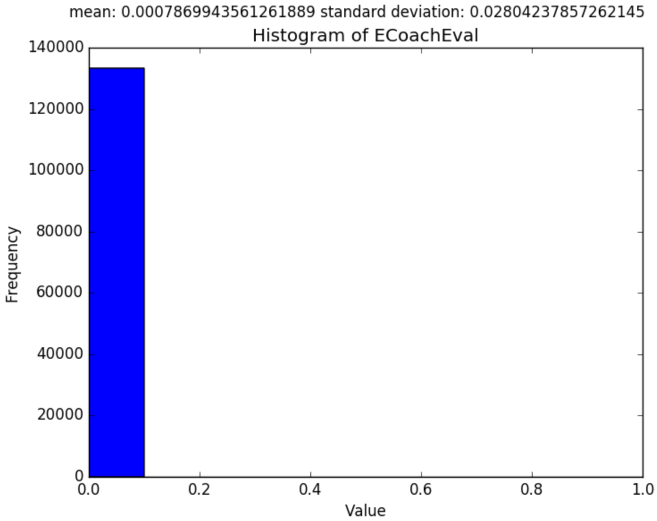

Figure 8: Gender Histogram

2.9 ECoachEval

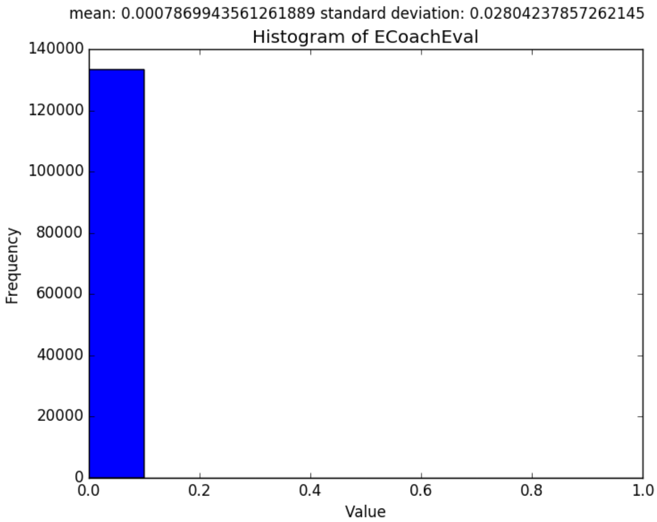

Figure 9: Gender Histogram

2.10 ECoachSearchHit

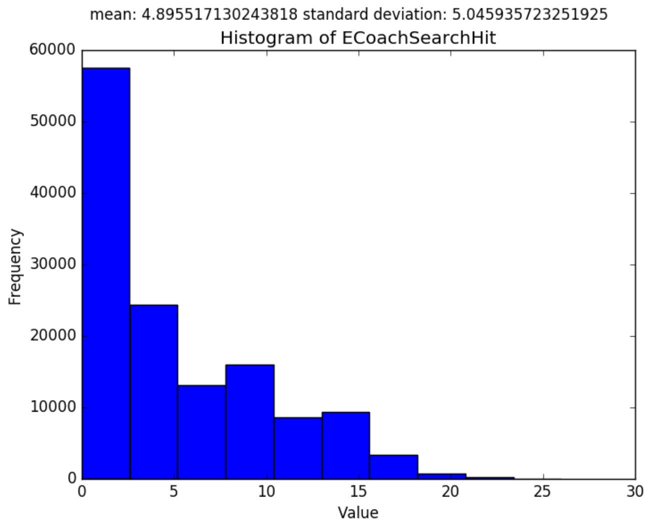

Figure 10: Gender Histogram

2.11 ECoachVisit

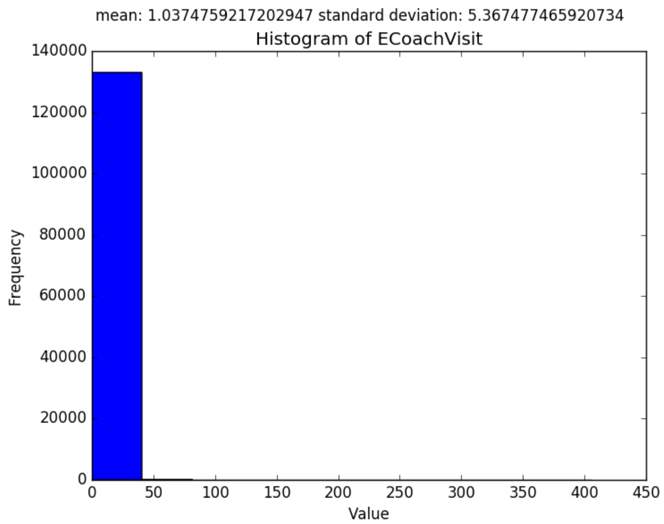

Figure 11: Gender Histogram

2.12 Ecolleges_going_to_the_event

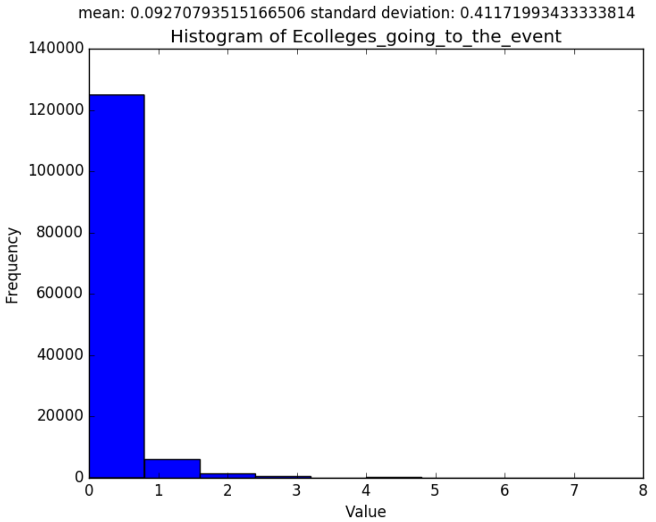

Figure 12: Gender Histogram

2.13 Efailed_subscription

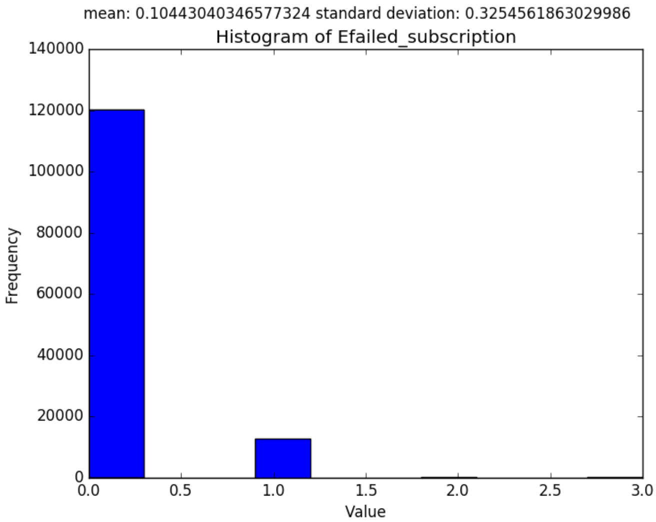

Figure 13: Gender Histogram

2.14 Eparent_new

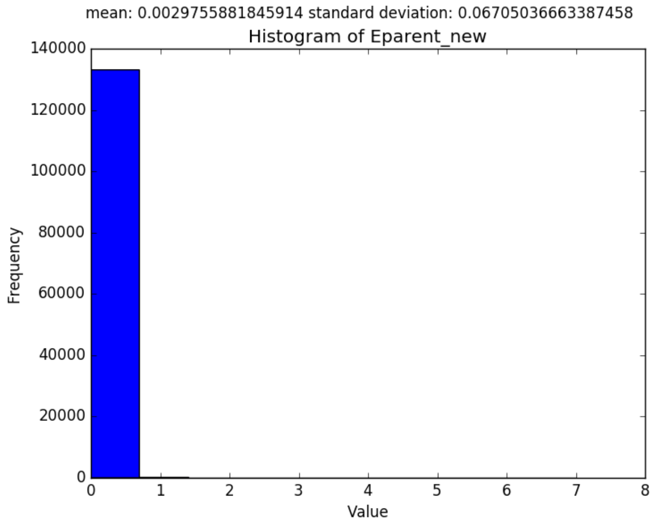

Figure 14: Gender Histogram

2.15 Eparent_welcome

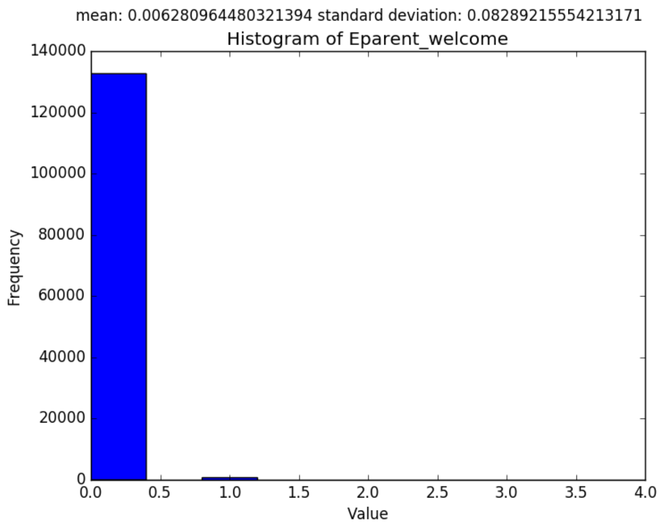

Figure 15: Gender Histogram

2.16 Epost_event_email

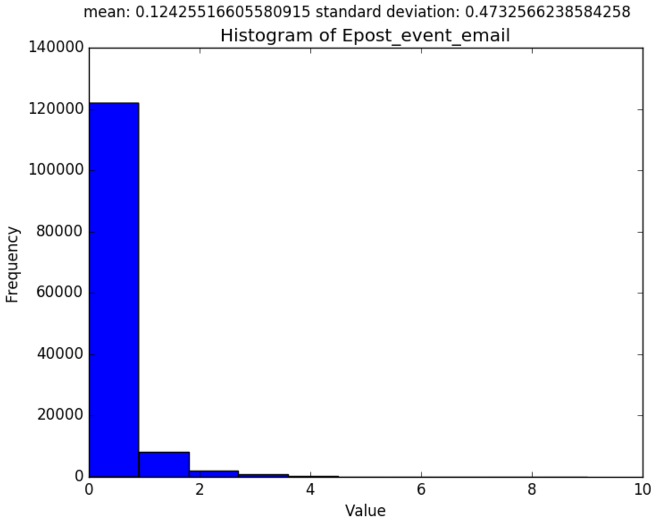

Figure 16: Gender Histogram

2.17 EventsAttended

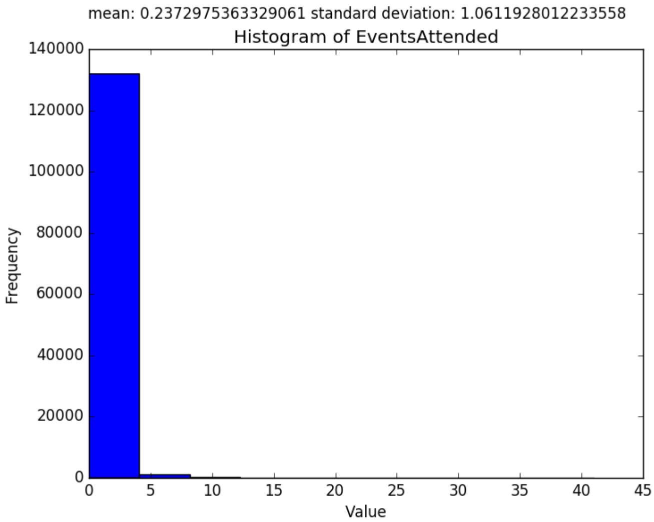

Figure 17: Gender Histogram

2.18 Hcoacheval

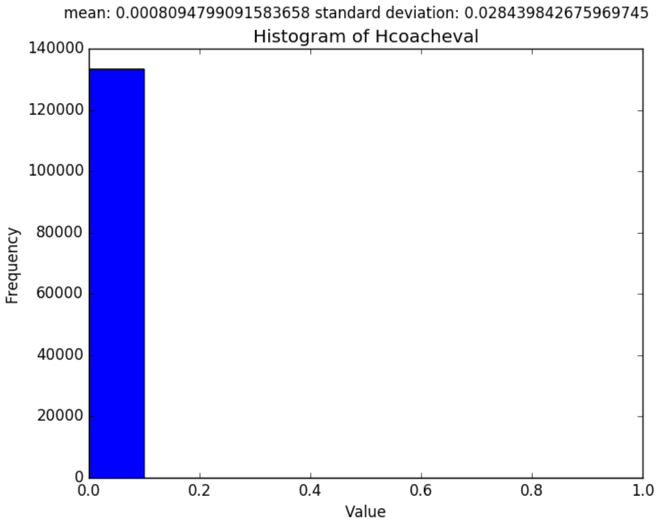

Figure 18: Gender Histogram

2.19 Hcoachimport

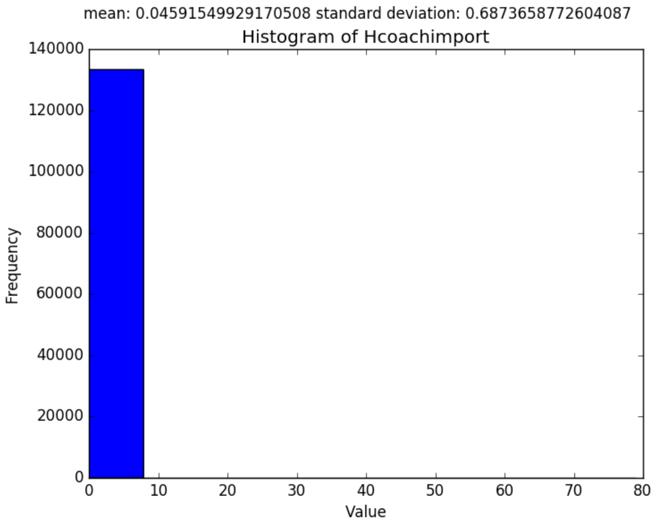

Figure 19: Gender Histogram

2.20 Hemailopen

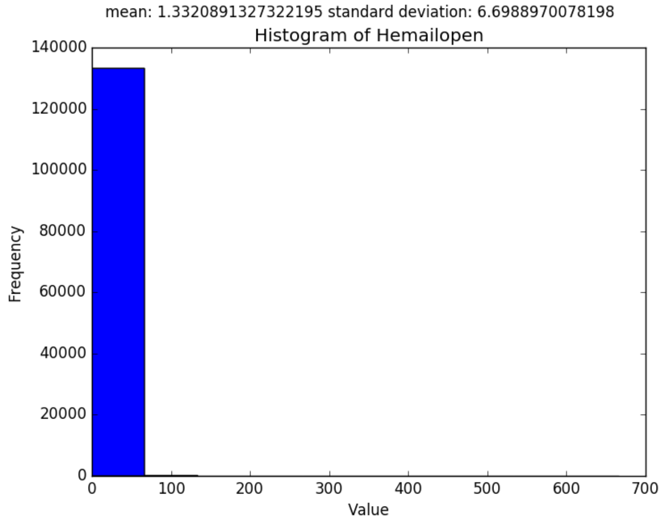

Figure 20: Gender Histogram

2.21 Hmessage

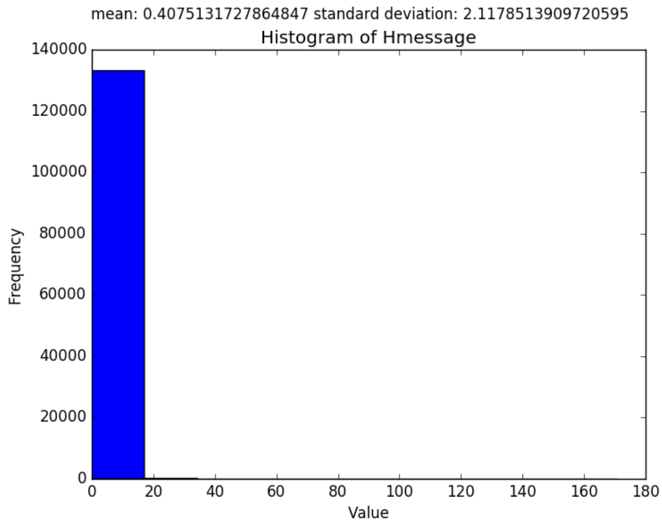

Figure 21: Gender Histogram

3 EVALUATION METRICS

These are measures we used to determine how good our machine learning models were. We used the testing data to evaluate the model fitted using the training data. We shall briefly explain the measures we used and present the formula. Before then, we shall present two variables we shall use in the proceeding equations.

- True Positive(TP) - These are the athletes who are predicted by the model to be most likely to churn and actually churned.

- True Negative(TN) - These are the athletes who are predicted by the model to be most likely to be retained and actually are retained.
- False Positive(TP) - These are the athletes who are predicted by the model to be most likely to churn and actually are retained.
- False Negative(TN) - These are the athletes who are predicted by the model to be most likely to be retained and actually churned.

3.1 Precision

In the churn context, precision refers to how pure our predicted set is. Given a set of prediction of churners, how many of them are actually churners(True Positive (TP))?

$$\text{Precision} = \frac{TP}{TP + FP}$$

3.2 Recall

Given a set of churners, recall refers to the fraction of churners that the model correctly returns.

$$\text{Recall} = \frac{TP}{TP + FN}$$

3.3 F1 Score

This is the harmonic mean of precision and recall.

$$\text{Recall} = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 2: Dataset Attributes

Dataset	Rows	columns	Churners	Non-Churners
Full	4360	26	601	3759
Training	3488	26	484	3004
Testing	872	26	117	755

4 ONE MONTH CHURN PREDICTION APPROACHES

We used March 2014 to predict churn in the next month. Predicting churn in the next month entailed setting the output variable (CaptainU_Churn) of the training and testing datasets to the output variable of the next month.

We implemented the following models in predicting churn:

1. Logistic Regression
2. Logistic Regression with (class_weight = 'balanced') - This mode uses the values of y(CaptainU_CHURN) to automatically adjusts weights inversely proportional to class frequencies in the test data. Effectively, more attention was paid to churners as they make up the minority class. You can read more about class_weight here
3. Gradient Boosting
4. Random Forest Classifier

We then calculated the precision and recall values for churning.

4.1 Using a Specific Month's Transactions Data

This is entailed tracking an athlete's monthly transactions and using that information to predict whether the athlete would churn in the next month. In our case for instance, we used March 2014's transaction data to predict if an athlete will churn in April 2014.

4.1.1 *Precision Recall Table*

Table 3: Precision Recall Values

Model	Precision	Recall	F1-Score
Logistic Regression	0.65	0.26	0.38
Logistic Regression(with class_weight = 'balanced')	0.43	0.56	0.49
Random Forest	0.53	0.35	0.42
Gradient Boosting	0.59	0.33	0.43

4.1.2 *Confusion Matrix: Logistic Regression*

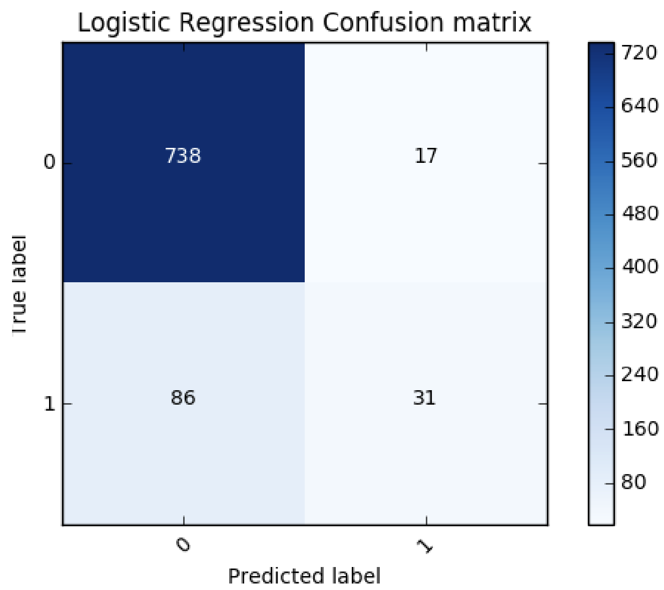

Figure 22: Lostic Regression Confusion Matrix

The above confusion matrix illustrates the following:
Given 117 athletes who are churners the model will:

- correctly predict 31 of the 117 athletes.
- incorrectly predict 17 athlete as having churned
- incorrectly predict 86 athletes as having been retained while in reality they are churners

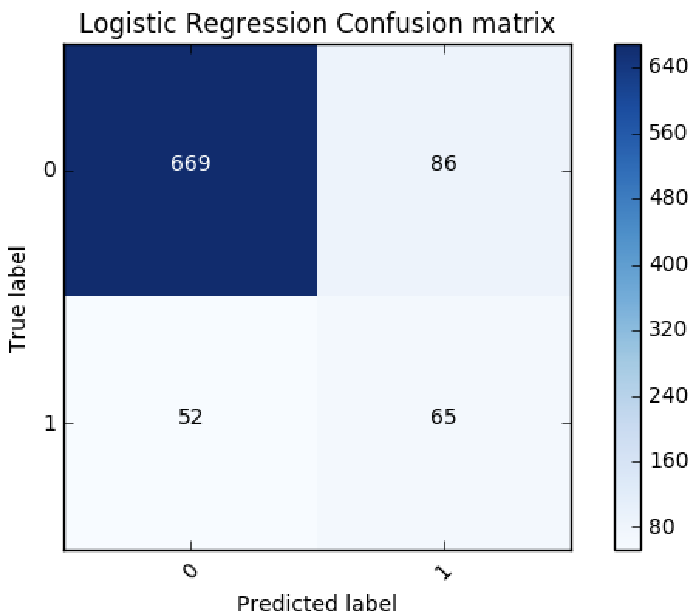

Figure 23: Lostic Regression(class_weight = 'balanced')

The above confusion matrix illustrates the following: Given 117 athletes who are churners the model will:

- correctly predict 65 of the 117 athletes.
- incorrectly predict 86 athlete as having churned
- incorrectly predict 52 athletes as having been retained while in reality they are churners

4.1.3 *Confusion Matrix: Random Forest*

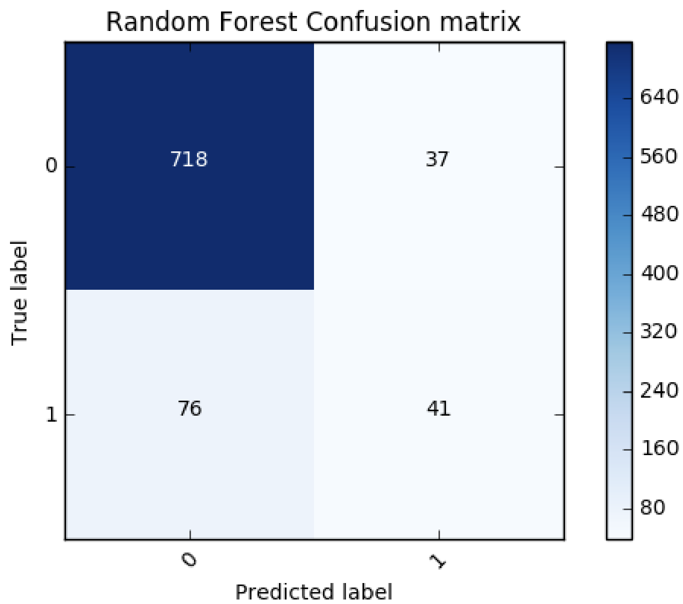

Figure 24: Random Forest Confusion Matrix

The above confusion matrix illustrates the following:
Given 117 athletes who are churners the model will:

- correctly predict 41 of the 117 athletes.
- incorrectly predict 37 athlete as having churned
- incorrectly predict 76 athletes as having been retained while in reality they are churners

4.1.4 *Confusion Matrix: Gradient Boosting*

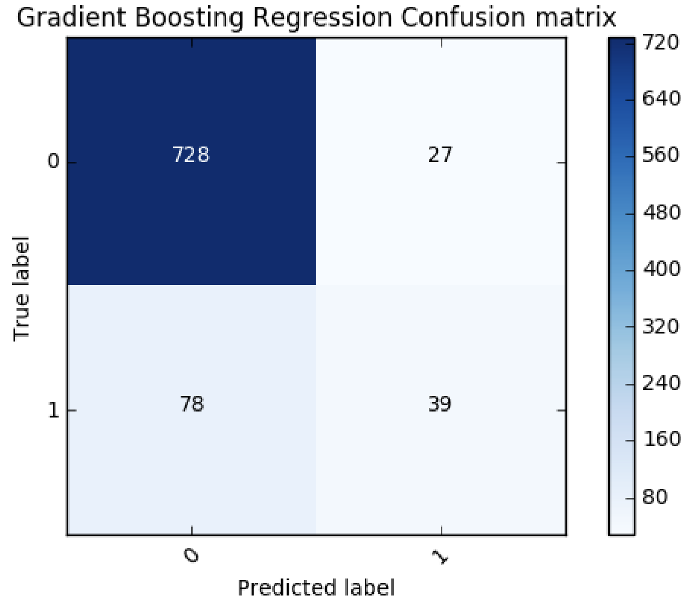

Figure 25: Gradient Boosting Confusion Matrix

The above confusion matrix illustrates the following:
Given 117 athletes who are churners the model will:

- correctly predict 39 of the 117 athletes.
- incorrectly predict 27 athlete as having churned
- incorrectly predict 78 athletes as having been retained while in reality they are churners

4.2 Using Aggregate Transactions Data

This is entailed tracking an athlete's aggregate transactions data up to a certain month and using that information to predict whether the athlete would churn in the next month. In our case, we used the cumulative sum of each feature up to March 2014 to predict churn in April 2014.

4.2.1 *Precision Recall Table*

Table 4: Precision Recall Values

Model	Precision	Recall	F1-Score
Logistic Regression	0.00	0.00	0.00
Logistic Regression(with class_weight = 'balanced')	0.18	0.61	0.28
Random Forest	0.00	0.00	0.00
Gradient Boosting	0.50	0.02	0.03

4.2.2 *Confusion Matrix: Logistic Regression*

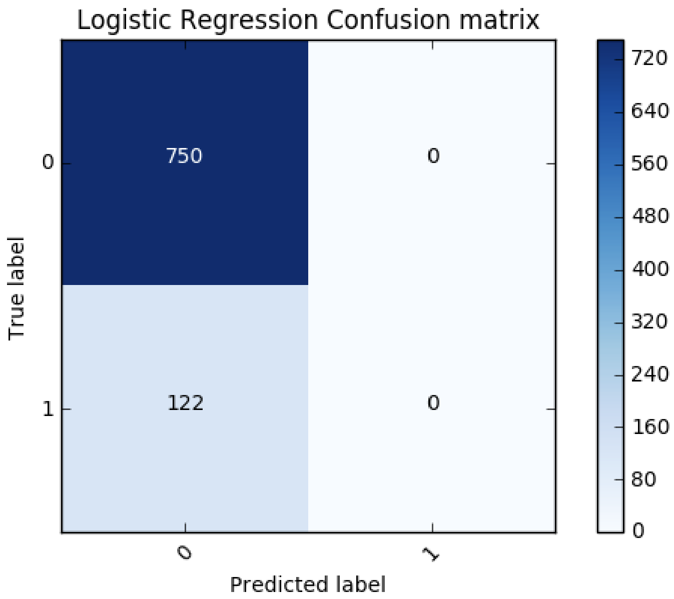

Figure 26: Lostic Regression Confusion Matrix

The above confusion matrix illustrates the following:
Given 117 athletes who are churners the model will:

- correctly predict none of the 117 athletes.
- incorrectly predict 0 athletes as having churned
- incorrectly predict 122 athletes as having been retained while in reality 117 of them are churners

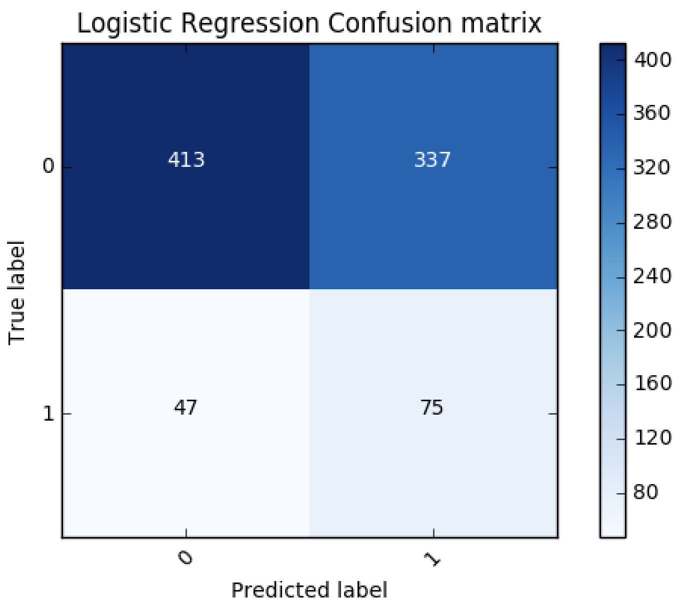

Figure 27: Lostic Regression with class_weight='balanced'

The above confusion matrix illustrates the following:
Given 117 athletes who are churners the model will:

- correctly predict 75 of the 117 athletes.
- incorrectly predict 337 athletes as having churned
- incorrectly predict 47 athletes as having been retained while in reality 117 of them are churners

4.2.3 *Confusion Matrix: Random Forest*

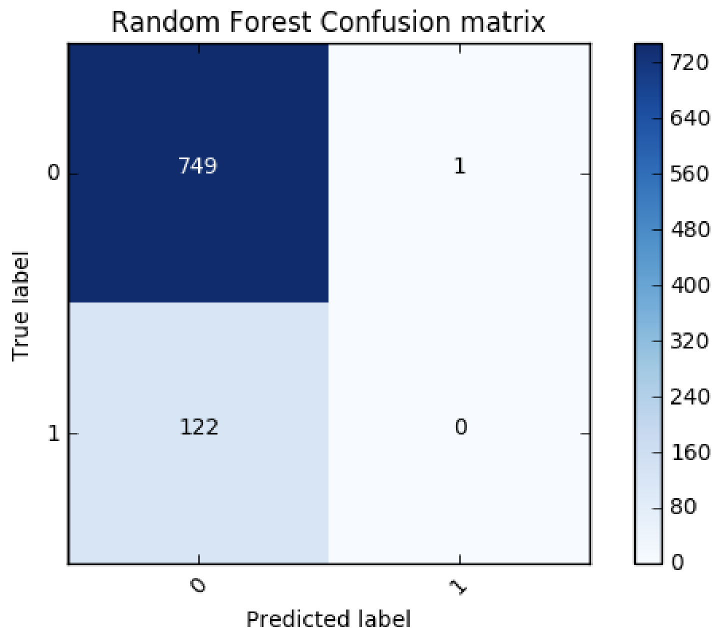

Figure 28: Random Forest Confusion Matrix.

The above confusion matrix illustrates the following:
Given 117 athletes who are churners the model will:

- correctly predict none of the 117 athletes.
- incorrectly predict 1 athlete as having churned
- incorrectly predict 122 athletes as having been retained while in reality 117 of them are churners

4.2.4 *Confusion Matrix: Gradient Boosting*

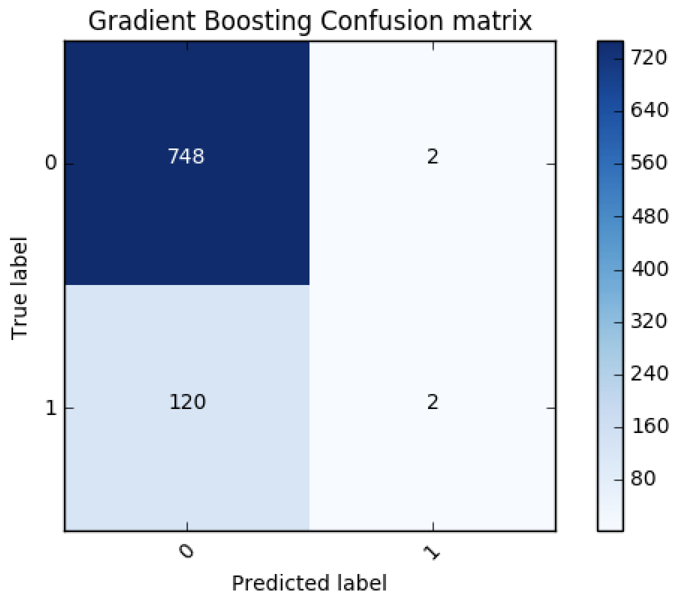

Figure 29: Gradient Boosting Confusion Matrix

The above confusion matrix illustrates the following:
Given 117 athletes who are churners the model will:

- correctly predict 2 of the 117 athletes.
- incorrectly predict 2 athletes as having churned
- incorrectly predict 120 athletes as having been retained while in reality 115 of them are churners

4.3 Using Monthly Difference in Transactions Data

This is entailed tracking an athlete's change in transactions data from one month to the next and using that information to predict whether the athlete would churn in the next month. For instance, we would track the change in the number of hits between February 2014 and March 2014 and Creating a Hits_diff column. For creating our testing and training data, we tracked the difference in transactions(interactions) between February 2014 and March 2014 to predict churn in April 2014.

4.3.1 *Precision Recall Table*

Table 5: Precision Recall Values

Model	Precision	Recall	F1-Score
Logistic Regression	0.00	0.00	0.00
Logistic Regression(with class_weight = 'balanced')	0.07	0.55	0.13
Random Forest	0.00	0.00	0.00
Gradient Boosting	0.10	0.02	0.03

4.3.2 ***Confusion Matrix**: Logistic Regression*

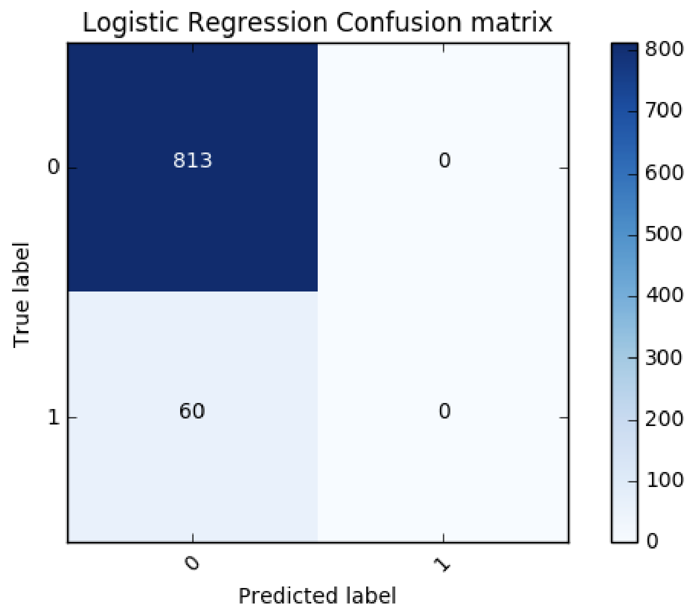

Figure 30: Lostic Regression Confusion Matrix

The above confusion matrix illustrates the following:
Given 117 athletes who are churners the model will:

- correctly predict none of the 117 athletes.
- incorrectly predict 0 athlete as having churned
- incorrectly predict 60 athletes as having been retained

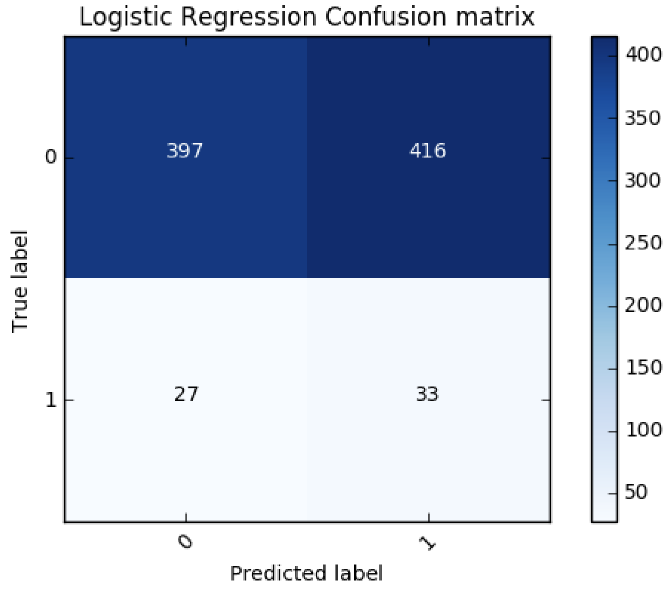

Figure 31: Lostic Regression with class_weight='balanced'

The above confusion matrix illustrates the following:
Given 117 athletes who are churners the model will:

- correctly predict 33 of the 117 athletes.
- incorrectly predict 416 athlete as having churned

4.3.3 *Confusion Matrix: Random Forest*

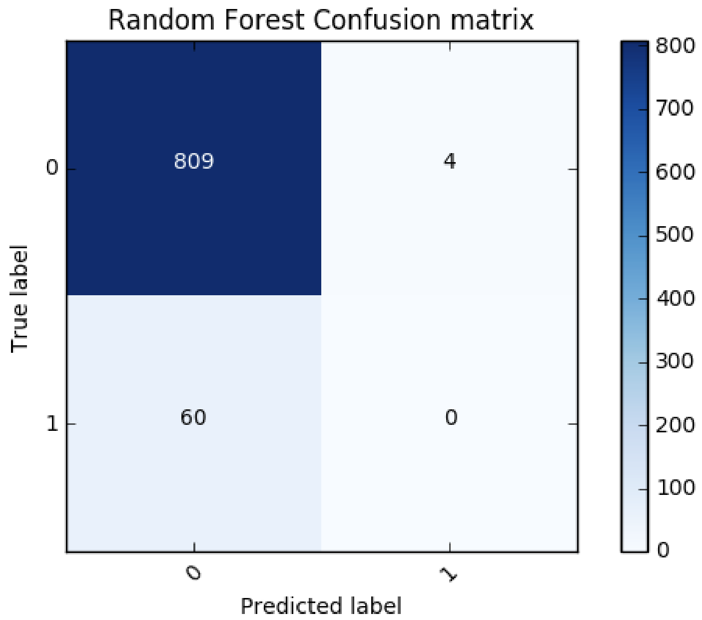

Figure 32: Random Forest Confusion Matrix. We Suspect Overfitting here

4.3.4 *Confusion Matrix: Gradient Boosting*

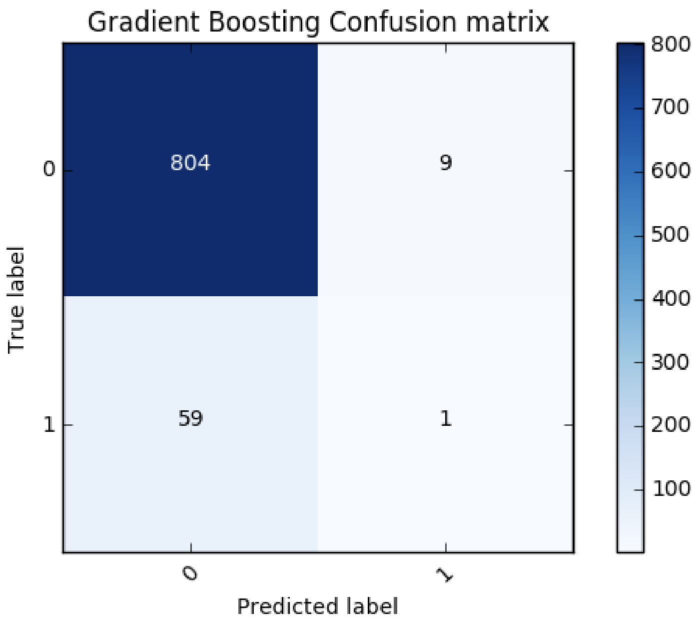

Figure 33: Gradient Boosting Confusion Matrix

The above confusion matrix illustrates the following:

Given 10 athletes who are predicted to have churned, only one actually churns.