Predicting runners' performance 2017 Miami Marathon

Arun Rawlani- 260568533 –arun.rawlani@mail.mcgill.ca
Sorin Muchi – 260575810 - sorin.muchi@mail.mcgill.ca
Gandharv Patil – 260727335- gandharv.patil@mail.mcgill.ca

Abstract-The goal of this project was to accurately predict the participation and performance of participants in the upcoming 2017 Miami Marathon. We used two different classification methods to determine participation: Logistic Regression (accuracy: 85.9%) and Naïve Bayes Regression (accuracy: 87%). Performance results were calculated using Logistic Regression (accuracy: 99%).

I. Introduction

The Miami marathon consists of the 13.1-mile half marathon and the 26.2-mile full marathon and happens once every year. Our goal is to predict the following two objectives:

Y1: Will they participate in the 2017 Miami marathon?

Y2: What will be their finishing time?

Below we discuss on how we sanitized & normalized the data, trained our model and evaluated our predictions for objective.

II. Problem Representation

A. Data Sanitization

Initially, the data contained entries for half-marathon for one particular year. We graphed the data and realized that Year 2013 times were the one of a half marathon. After removing those entries, we were eventually left with 280176 entries, which is only 9% lesser than the original number. We considered using the *Reigel's formula* [1] to possibly make those entries useful for prediction. Considering the trade-offs between lost data and using manipulated data, we decided to avoid using manipulated data and simply eliminated entries for 2013. We also removed some outliers and entries with duplicate entries & with missing important data like Time & Gender.

We chose to normalize our data using feature scaling for all values, such that the parameter weights will represent significant relative activity between each variable. Essentially, each coefficient value will represent how the variable acts towards the objective function and magnitudes of these coefficients will comparatively show their degree of influence in the objective function.

The team also decided to completely eliminate the *Pace* data that was being presented to us, because we already had data for the *Finishing Time*. Eliminating *Pace* as a feature helped us avoid interdependency, because *Finishing Time* and *Pace* are strongly related features.

All time durations were converted form HH:MM:SS to seconds such that it is easier to plot the data and use it in our prediction algorithms.

TABLE I
FEATURE SELECTION FOR CLASSIFICATION

FEATURE	TYPE	DESCRIPTION
Sex	Binary	0 for women, 1 for men
Age Category	Continuous	Mean avg age of the athlete from all the participation entries in different years.
No_of_Participations	Discrete	Number of times an athlete has participated in the Miami Marathon
Participation_in_previous_year	Binary	If an athlete participated in previous year. 0 for No, 1 for Yes

B. Data Representation

Our objective in this project was to make predictions using *classification* and *regression*. As a result, different modified datasets were created for different predictions.

Feature Selection for Objective Y1:

In this case, we chose four important features: Age Category, Sex, Participated in Previous Year and Number of Times Participated as described in Table 1.

Age: After plotting the age ranges over years, we found that the average age of marathon athletes was 21-62 years old. (APPENDIX1) As a result, there was noticeable evidence that age played a major part in returning to the marathon next year with strong correlation in the 27-39 ranges.

Sex: It has been shown that women have lower resting metabolic rates than men. As a result, gender plays a role in returning for a marathon because averagely, women's metabolism starts decreasing around 34 - 43 years while for men it starts usually during 39-46. [2] Consequently, men have a higher chance of coming back. The graph to the right reinforces the claim.

Additionally, we observed strong interdependency between sex and participation numbers. For every year, the number of men participants was much higher than the number of women participants. As a result, it will be more likely for men to return to the marathon next year than for women. Hence, we decided to select this as a feature.

Number of Times Participated: We observed that more experienced athletes tend to come back to the marathon. People who have ran over three times before have a higher

probability of coming back compared to those who have ran less than three times. Participated in Previous Year: Runners who participated in the previous year have had a higher chance of running up till *2012*. However, lower attendance was shown in the years following 2012, which went against the first trend. This could have been due to various reasons: multiple occasions of bad weather, poor organization of the event, fewer young Miami residents taking part in marathons.

Feature Selection for Objective Y2:

TABLE II
FEATURE SELECTION FOR REGRESSION

FEATURE	TYPE	DESCRIPTION
Age	Continuous	Average age of the athlete from all the participation entries in different years.
Sex	Binary	0 for women, 1 for men
Pro or Amateur	Binary	If an athlete's average running time > 10800s its 0 (Amateur), else 1(Pro)
Rank	Continuous	Average rank of the athlete from all the participation entries in different years.

Age: After analyzing the data, we observed that the age of an individual plays a major part in determining the time taken to complete the marathon. That is as the participant gets older, the average running time increases to complete the marathon. As a result, for each participant we took an average of all the mean age entries for different years. For participants with missing age values, we assumed that their age was 39, which is the average estimated mean age of all the participants who have taken part in the marathon.

Sex: After looking at the plots of gender against time (Add graph here), we saw that for every year men had a lower finishing time than women. This suggests that there is a correlation between the gender and finishing time in the marathon. Consequently, we decided that it will be helpful to take this feature into account as domain knowledge

Pro_Or_Amateur: We decided that finishing time of an athlete will be determined if the person is a pro marathon runner or a first time marathon amateur. We set a threshold of 10800s, after doing some research on the running times of professional marathon runners across a number of events. For each participant, we used their average finishing time and then used the threshold to determine which category (i.e. Pro or Amateur) will they be chosen for.

Rank: This feature stored the average rank of the participant over multiple participations in different years. Our team decided that an individual's rank throughout multiple years will be a good indicator of participants performance in the next marathon.

Clarification: Why did we not choose average running time as a feature?

Since we chose to add the *Pro_Or_Amateur* and *Rank* features we did not end up choosing *Average Time* per participant as a feature. The reason for this was that *Pro_Or_Amateur* and *Rank* ended up encompassing a lot of the information that will overlap with the information covered in *Average Time*. As a result we did not want to introduce multiple features that are highly correlated with each other. So as to avoid a case of **multicollinearity**, we decided to eliminate *Average Time* per person and instead keep *Rank* & *Pro_Or_Amateur* features.

III. Training Method

A. Logistic Regression

Values for hyper parameters were selected after running the training with different values, while keeping everything else constant and then testing the classifier's accuracy. After running with several combinations of hyper parameter values, starting from α as 0.5, we decided to eventually choose *the step-size/learning rate* α as 0.001.

The specific value for α was chosen just as to find the optimal tradeoff between slower computation times for smaller step sizes and high oscillatory behaviors for larger step sizes. Given lesser constraints on computation resources, a much smaller step size can help in finding averagely more accurate solutions. Our team also focused on finding an accurate value for the *regularization constant*. However, after running several runs with varying values for the regularization constant, we did not note significant prediction improvements, so we chose to not use regularization as part of our training.

For cross validation, 187144 data entries were used for training while the remaining 93572 entries were saved up for the validation set. For the prediction, we decided to train our classifier with data from all years leading up to 2011 and then make prediction for participants in 2012. The reason for specifically choosing 2012 was that it had the highest number of participants during the 2003-2016 period. This indicated that external parameters that may have avoided people from returning to the marathon (*e.g. bad weather, disease breakouts, poorer lifestyles*) were probably not a major factor in the year 2012. Hence, we chose it as our prediction year to evaluate our classifier once it was trained on the given dataset. In order to evaluate how the above strategy worked for our classifier, we used our trained classifier to make further participation predictions for year 2014 and 2015 and received high accuracies for it. These results were a good indicator of the correctness of our training technique.

Finally, we used a standard gradient descent algorithm with the sigmoid function as part of the hypothesis function to train the classifier.

B. Naïve Bayes

We used the same features as above for Naive Bayes as well. Since our data sample is large enough, and chosen from an significantly larger population ~7bn, by Central Limit Theorem, we can safely assume that traits such as Sex, Age, Number of previous participations, and therefore Participation

in the last marathon, to be *normally distributed*. These assumptions were positively corroborated by our initial investigations, and visualizations of the dataset.

C. Linear Regression

The data set provided for this problem had a wide scope of modifications. The data primarily consisted multiple entries of contestants over a range of thirteen years. After drawing inferences from the statistical properties of data evaluated in the previous implementations we decided to average the values for Age, Rank and Speed of the contestant. This strategy provided a twofold advantage of reducing the data points without having to compensate for the loss of information.

Finally, the Linear Regression model was trained using two strategies:

1. Gradient Descent
2. Normal Equations (closed form implementation)

This choice was mainly made to monitor effect of hyper parameters on the convergence of the error function and to help choose their optimum values.

After running for various combinations of epochs it was empirically observed that the gradient descent function converged after approximately after 7000 iterations.

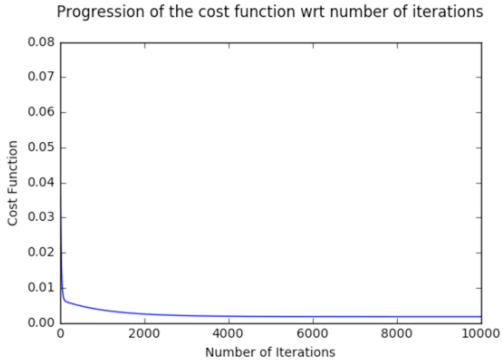


Figure 1: Cost function varying for gradient descent in linear regression

IV. RESULTS

The evaluation criteria for the classifiers correctness was to first produces predictions using the training data and then comparing the errors. We focused on the error not being very close to zero, which may indicate severe over fitting in the data. Once we had compared errors with training data, we used the other ⅓ of the data, which was left as the validation set.

We used this data to make predictions and then crosschecked with the real data values to calculate the validation error. This error, coupled with the training error, was a good indicator of how accurate our classifier was.

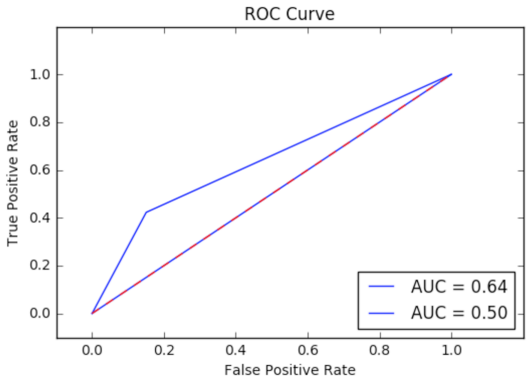
TABLE III
ACCURACY FOR LOGISTIC REGRESSION

SET TYPE	DESCRIPTION
Training Set	0.878966317241
Validation Set	0.859065284597

A. Logistic Regression

The classifier's accuracy started at 71% when we set the hyper parameter value of alpha to 0.5. After gradually dropping the values in minor steps, we eventually reached an accuracy of 85% in validation set at alpha = 0.001.

This change in accuracy behavior can be explained where large step sizes can lead to high oscillatory behavior in the classifier which prevents it from reaching the global minimum in the cost function. However, once the alpha value is small enough to find the global minimum, accuracy may be increased by a huge margin as can be seen in our case.



B. Naïve Bayes

With Naive Bayes we managed to reach a training accuracy of 0.798, when training on years 2003-2011, and comparing y_2012 with predicted yhat_2012, and a test accuracy of 0.872 when testing on 2015 data, and comparing y_2016 with predicted yhat_2016. Out of a total of 30442 participants we managed to correctly predict 26532, and misclassify only 3904 as either 1279 false positives or 379 false negatives.

ACCURACY FOR NAÏVE BAYES

SET TYPE	ACCURACY
Training Set	0.798965517241
Validation Set	0.871559030287

CONFUSION MATRIX FOR NAÏVE BAYES

TP = 26153	FP = 1279
FN= 2625	TN = 379

C. Linear Regression

The strategy of running two training models (Normal Equations & Gradient Descent) in parallel was a relatively high accuracy rate and an optimum R squared value.

On cross validating the model on the test data we have obtained metrics in Table IV the absolute mean squared error is approximately 33.66 minutes. With the accuracy of 99% the

model performs fairly well when modeled without higher order features.

TABLE IV
METRICS FOR LINEAR REGRESSION

CROSS VALIDATION RESULTS ON TRAINING SET	
METRICS WITH GRADIENT DESCENT	
MSE	33.6233764052 minutes
R-squared	81.2511515497 %
Accuracy	99.655231344
METRICS WITH NORMAL EQUATIONS	
MSE	33.6213625898 minutes
R-squared	81.3010369853 %
Accuracy	99.655231344
METRICS WITH SCIKIT LEARN	
MSE	33.6213625898 minutes
R-squared	81.2511515497 %
Accuracy	99.655231344

CROSS VALIDATION RESULTS ON VALIDATION SET	
METRICS WITH GRADIENT DESCENT	
MSE	33.6698178184 minutes
R-squared	80.5690488001 %
Accuracy	99.6340719181
METRICS WITH NORMAL EQUATIONS	
MSE	33.6660504179 minutes
R-squared	80.6602949888 %
Accuracy	99.6340719181
METRICS WITH SCIKIT LEARN	
MSE	33.6642184191 minutes
R-squared	80.5690488001 %
Accuracy	99.6340719181

IV. DISCUSSION

Naive Bayes predicted lesser attendance than Logistic Regression, perhaps due to the fact that NB classifies our participants based on generalizations of the participant's past performance, such as attendance count, average age, etc. Perhaps a more accurate prediction we could have made would be to have a meta-predictor combining the probabilistic output of both Logistic Regression and NB.

We can also improve our strategy of dealing with outliers, such as the half marathon times or undefined gender entries. Using the Riegel's formula is one of the possible techniques that can be utilized in this case in able to use the half marathon times as well. Similar, we can use the mean of the underlying distributions of the data to assume values for missing data entries.

We can also use external data such as weather, region's lifestyle habits, participants' personal health data as features in our models. These features can help us in computing much more accurate predictions, although on the other hand it'll definitely increase the complexity of models and the interdependencies amongst features.

In terms of hyper parameter optimization, in logistic regression, instead of trying to compute the value of each hyper parameter one-by-one, we could have used the *Grid Search [3] algorithm*. Grid Search can make it faster and optimal to determine the optimal values for the hyper parameters. Using such an algorithm can improve the performance of logistic regression on future datasets for marathon predictions.

Moreover, in logistic regression, instead of using the standard gradient descent algorithm to determine the weights of the parameter values, we can instead one of the more sophisticated algorithms like Conjugate Gradient[4] or BFGS. The Conjugate Gradient algorithm is similar to the gradient descent but instead uses the concept of conjugate vector. Implementation of these advanced algorithms is beyond the scope of this course, but if given enough time and resources, can help us make it much faster to find optimal values for the parameters.

IV. STATEMENT OF CONTRIBUTION

Arun Rawlani

I did the data sanitization using pandas module. Also, created the files containing the data that will be used for the Classification Objective Y1. Focused on completing the manual implementation of Logistic Regression algorithm using Python. Also wrote the Introduction, Data Representation, Problem Representation, Training/Evaluation of Logistic Regression Classifier, Discussion parts of this report.

Sorin Muchi

I was responsible for predicting attendance using a Naive Bayes classifier, and describing the methods and results. I also performed data sanitization and pre-processing for using the pandas python library, which I used to train, and test the Naive Bayes classifier, and I also participated in the group discussion about data sanitization, feature selection, and general research methodology.

Gandharv Patil

I did some work in the data sanitization and then switched to working on the implementation for Logistic Regression and Linear Regression. I did the training of the Linear Regression Model and the evaluation of the predictor. Also, assisted my teammate in the implementation of

We hereby state that all the work presented in this report is that of the authors.

REFERENCES

[1] Riegel, "Athletic Records and Human Endurance," American Scientist, pp. 285–290, Jun.1981.

[2] PAUL J. ARCIERO, *Resting metabolic rate is lower in women than in men*, http://goranlab.com/pdf/20.pdf

[3] James Bergstra, Yoshua Bengio Random Search for Hyper-Parameter Optimization http://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf.

[4] An Introduction to the Conjugate Gradient Method, http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf

APPENDIX:

1. **Showing the age ranges with the highest number of participants for each year.**