Predicting the Attendance and Running Times for the 2017 Miami Marathon

Nikolaus Yasui
nikolaus.yasui@mail.mcgill.ca
260558797

Robert Fratila
robert.fratila@mail.mcgill.ca
260615193

Srinivas Nadella
srinivas.nadella@mail.mcgill.ca
260531213

I. INTRODUCTION

This report outlines an application of three fundamental machine learning methods: logistic regression, naive Bayes, and linear regression. We apply these methods to data on roughly 30,000 past runners to predict attendance and running times in the 2017 Miami Marathon.

II. PROBLEM REPRESENTATION

A. Features

1) Id (Categorical)
2) We removed all the people with the id ”3327” because their name ”Private”, and they appeared in the data multiple times within the same year, with different age and sex values. We considered this to be noise and poor data collection.
3) Age (Categorical)
   - Instead of using Age in it’s raw form, we converted it into a set of decile categories (see Age Decile).
4) Sex (Categorical)
5) Time
   - We standardized all race times into seconds.
6) Year (Categorical)
   - We removed all data from 2003 before calculating the values of any features. It was removed because the data from 2003 only represented half-marathon times.
7) Age Decile (Categorical)
   - We created 10 age deciles spanning 0 to 100. In preparing this features, we dropped the smallest category [0-10), because all of the raw age values were 0, indicating noise in the data or poor data collection. This left us with 9 age categories for all runners.
8) Day of Year (Continuous)
   - This was found through an external data source that highlighted the specific day the marathon was run each year[1]. We represented it as the raw day in the year (ie. 33 indicates February 2.)
9) Temperature (Continuous)
   - This was found by referencing an external data source that highlighted the specific day the marathon was run each year[2] and then looking up these dates on Weather Underground[3]. We represented it in Fahrenheit.
10) Flu Incidence (Continuous)
    - We referred to the Florida Influenza Surveillance Reports[4] to find the the percentage of visits to percent of visits to Florida health care providers for ”Influenza Like Illnesses”[5]. We used this number as an index of the occurrences of the common cold/flu in Florida each year for the week the race was run. We represented this as a percentage.
11) Number of Marathons Run (Categorical)
    - This was calculated by observing the number of years we had in our data for a given ID.
12) Average Humidity (Continuous)
    - This was found by referencing an external data source that highlighted the specific day the marathon was run each year[6] and then looking up these dates on Weather Underground[7]. We represented as a relative humidity value[8].
13) Standard Deviation of Race Time (Continuous)
    - This value represents the standard deviation of the times for each ID, and where we only had one race time for an ID, we used the average race time across all years. This was calculated before the data was folded or split into training/testing (discussed further in Training Methods, section B).
14) Whether They Ran More Than Once (Categorical)
    - We represented this as a boolean 0,1 for whether the given runner in our dataset had run the Miami marathon in the range of years [2004,2015]
15) Number of Times Participated in the Last 3 years (Categorical)
    - This was represented as a value between 0 and 3, calculated by counting the number of times a

[1] http://www.marathonguide.com/races/racedetails.cfm?MIDD=1850170129; accessed January 25, 2017

[3] https://www.wunderground.com/history/airport/KMIA; accessed January 25, 2017

[4] http://www.floridahealth.gov/diseases-and-conditions/influenza/florida-influenza-surveillance-reports/index.html

[5] Influenza-like illness is a fever $>$ 100°F AND sore throat and/or cough in the absence of another known cause.

[7] https://www.wunderground.com/history/airport/KMIA; accessed January 25, 2017

[8] The ratio of water vapor contained in the air to the maximum amount of water vapor that can be contained in the air at the current temperature.

runner had participated in the Miami marathon in the range of years [2014,2016].

16) Number of Years Since Last Marathon (Categorical)
    - This was calculated per ID by finding their most recent marathon and subtracting it from 2016. It was represented as a value the range [0,12].
17) Age Category at Last Marathon (Categorical)
    - This was found using our previously mentioned feature "Age Category", represented the same way. It's value is the most recent age category for a given runner in our data.

B. TRAINING METHODS

We used bootstrap .632+ (Efron, Tibshirani 1997) and to evaluate the predictive performance of our models. With L as the loss function and f_b as the predictor trained on the bth bootstrap sample, the leave-one-out bootstrap estimator is defined as

$$\widehat{Err}_{(1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, f_b(x_i)),$$

where C^{-i} is the set of indices of bootstrap samples that do not include observation i. We combine $\widehat{Err}_{(1)}$ with the naive error estimate

$$\overline{err} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)),$$

where f is the predictor trained on the entire dataset $\boldsymbol{X}$, to create the .632 estimator proposed by Efron (1983):

$$\widehat{Err}_{.632} = 0.368\cdot\overline{err} + 0.632\cdot\widehat{Err}_{(1)}.$$

$\widehat{Err}_{.632}$ modifies the weights in the sum to reduce bias:

$$\widehat{Err}_{.632} = (1-\hat{w})\cdot\overline{err} + \hat{w}\cdot\widehat{Err}_{(1)},$$

$$\text{with}\quad \hat{w} = \frac{0.632}{1-0.368R} \quad\text{and}\quad R = \frac{\widehat{Err}_{(1)} - \overline{err}}{\hat{\gamma}-\overline{err}}$$

where γ is the error rate if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent,

$$\hat{\gamma} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L(y_i, f(x_j)).$$

The form of mean squared error, $MSE = Bias^2 + Variance$, shows that an estimator with a small bias but highly reduced variance could be preferable to an unbiased estimator with high variance. This trade-off is exactly what .632+ bootstrap claims to accomplish.

1) Linear Regression: The linear regression model was first fit using the closed-form equation:

$$\hat{\beta} = (X^\top X)^{-1}X^\top Y \tag{1}$$

and samples were redrawn during bootstrap if $X^\top X$ was singular. However, under (1) we observed a positive association between model complexity and estimated MSE, which is a symptom of overfitting. To combat this, L_2 regularization was introduced, adding the lambda term to the least squares solution:

$$\hat{\beta} = (X^\top X + \lambda I_p)^{-1}X^\top Y \tag{2}$$

where p is the number of features. We used a lambda value of 100, chosen using the R package 'glmnet'. Due to time constraints, feature-sets were chosen based on forward selection in R using adjusted R^2 as the selection criterion.

2) Naive Bayes: We fitted a naive Bayes model that modeled the conditional distribution of each feature differently depending on the data in the feature. Binary, integral, and continuous features were modeled as Bernoulli, multinomial, and normal random variables respectively. Laplace smoothing was used to avoid zero probabilities. Five models were considered, each using a single predictor. Since naive Bayes assumes conditionally independent features, we did not consider interaction terms. The features considered were day number in the year, flu incidence, temperature, sex, and age decile.

3) Logistic Regression: Since logistic regression does not have a closed form, gradient descent was used to minimize the cross-entropy. Several features are included in this model namely: age bracket and sex of the individual, number of marathons the person ran, the temperature and day of the year the marathon took place, and the prevalence of the flu virus for that year. These specific features where chosen to model the probability of an individual competing in the next marathon by analyzing the bootstrap error and training cycles to determine if the model was learning or not. A range of α values was used to narrow down the interval where the model can begin converging on the solution. From there, bootstrap was used to pinpoint the alpha value that returned the lowest error. When testing the model, the training set was composed of 70% of all the samples from the curated data set while the other 30% became the testing samples. To prevent the creation of not-a-number(nan) matrices, clipping of probabilities was put in place to prevent extreme values from occurring as a result of the sigmoid function (i.e 1.0 or 0.0). The data matrix was randomly shuffled to reduce the amount of bias that exists in the ordered list. From preliminary tests, it becomes evident that the order matters since the model was able to obtain 95% accuracy.

C. RESULTS

1) Linear Regression: Linear regression results were fairly consistent, other than some models that were especially prone to overfitting. The best performing model is Model B with $\lambda = 100$.

TABLE I: MSE Estimates for Simple Linear Regression
(Millions of seconds)

Model:	1	*day_no*	*temp*	*flu*	*sex*	*sd_time*
$\lambda = 0$	9.21	9.20	9.17	9.20	24085	9.03
$\lambda = 100$	9.21	9.33	9.29	9.21	8.71	9.04

Estimates were created using .632+ Bootstrap with 200 samples. The columns refer to the features included in the model, with '1' indicating that only the intercept was used.

TABLE II: MSE Estimates for Multiple Linear Regression (Millions of seconds)

Model:	A	B
$\lambda = 0$	30801919	24776875
$\lambda = 100$	8.48	8.18

Estimates were created using .632+ Bootstrap with 200 samples. Model A corresponds to the feature-set $day_{n}o * flu * temp + sd\_time + sex$, where $*$ denotes all combinations of interactions between the features. Model B corresponds to the feature-set $day\_no * flu * temp + sd\_time * no\_runs + sex + age\_decile$.

2) Logistic Regression: Following the procedure outlined in the previous section, the optimal α was determined to be 1. Keeping all features the same throughout model evaluation process, this value for α returned the lowest bootstrap error of 0.34708174 when it fit the model for 200 samples. The loss associated to other α values are displayed in Table 3. The training curve for this model is shown in Fig. 1.

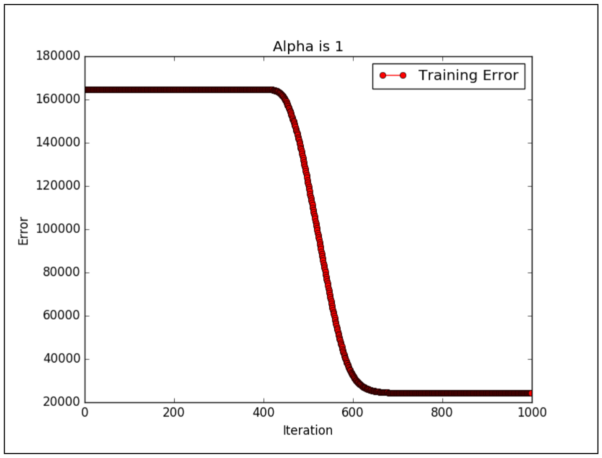

Fig. 1: Logistic regression SGD training error over the course of 1000 iterations.

TABLE III: Cross entropy loss for different values of alpha (Percent Error)

Model:	$1E(4)$	$1E(2)$	$1E(0)$	$1E(-2)$	$1E(-4)$
Error:	115014.26	164704.01	2301.64	34819.21	19833.84

Keeping all the features the same, the following learning rates yielded results varying greatly in error rates.

Using the new weights after gradient descent has completed all iterations in its training cycle, the model is able to accurately predict whether or not a runner will compete in the Miami marathon 67% of the time.

3) Naive Bayes: Each naive Bayes model fitted using a single feature produced the same predictions of all zeros. Due to the difficulty of posing the problem in a way that translates to the 2017 marathon and the computational burden of evaluating further models, we did not investigate combinations of the features.

TABLE IV: MSE Estimates for Naive Bayes (Percent Error)

Model:	day_no	$temp$	sex	flu	age_decile
Error:	32.53	32.70	32.67	32.67	32.69

Each column corresponds to a naive Bayes model using only that feature, with error estimated by 200 sample .632+ bootstrap.

III. DISCUSSION

Our best performing linear regression model achieved a bootstrap-estimated MSE of 8.18 million, which corresponds to an R^2 – the proportion of variance in the data that our model has explained – of 0.11. The poor performance can be explained by the highly noisy features, many of which do not affect the distribution of running times. However, it is a better model than predicting only the mean running time in the data, and we expect it to do better in prediction due to the accuracy of bootstrap's error estimation.

A. Summary of results

We successfully explained 11% of the variance in the data using our linear regression model. Our classification models were unable to find rules that predicted any 1s, ie everyone would not attend the 2017 marathon. This led to an overall accuracy of 67%.

B. 2017 Speculation

We expect to do well on the classification accuracy, as most people do not run again. Given that only the top 3000 race results will be available to compare against, it is very likely that we can achieve upwards of 80% or 90% accuracy in the 2017 data.

Regression accuracy is more difficult to predict, as the data is very noisy. We expect to have mean squared error of less than the total variance of the 2017 running times.

1) Misalignment of objectives vs. data used
   The data used is not all runners of the Miami Marathon - instead it is actually only the top 3000 participants. This skews us from our original objective of predicting attendance and race times of **all** participants and instead predict whether they would be in the top 3000 in 2017. To further highlight the scope of data missing from our methodology, according to Wikipedia, over 25,000 runners attend the Miami Marathon.
2) Missing key data related to our prediction task
   Upon reflection of our two objectives in this report: (1) predicting race time & (2) race attendance, we would posit that we were missing many of the true variables that will affect the outcomes that will be observed in the 2017 Miami marathon. If we were to produce this report again, we would want data points that included such factors as:
   a) Whether the given runner trained for the Miami Marathon that year
   b) Whether the given runner suffered any injuries in the twelve trailing months

c) The relative amount of training they did in the twelve trailing months (perhaps measured in hours they trained or miles ran)

d) Whether the given runner suffered any injuries in the twelve trailing months

e) Which, if any, other marathons they ran that year

f) Their times at any other marathons they ran

g) Whether they were from Miami

h) If they weren't from Miami, then the distance they traveled to participate

i) Their times at any other marathons run

j) If they are running the marathon with any friends

C. Limitations in Implementation

1) Our representation of standard deviation
   We used the standard deviation of the runner's times as a feature in our models. This presents an issue when folding our data for validation because the the standard deviation being used a feature may in fact be calculated from times in our testing data (for that fold). Therefore, this feature is biased by the outcome we are meant to predict.
2) Our choice of λ for regularization
   Our choice of λ was based on the R function cv.glmnet, which provides a λ that minimizes mean cross-validated error and a λ that gives the most regularized model with error within one standard deviation of the minimum. We picked a λ between the two.
   While this gives a good approximation for picking an optimal λ, there may be a more sophisticated way to pick a value for λ.
3) Consideration of Wind After writing this report, we reflected on extra features that we did not use but had potential value in predicting running times & attendance. One of these was wind speed & direction. As any runner can attest to, wind a huge factor in speed. Here is small sample of the past 5 marathons:

TABLE V: Wind speed and direction for past 5 marathons

Year:	Wind speed	Direction
2012	17 KM/H	NE
2013	14 KM/H	ENE
2014	14 KM/H	ESE
2015	10 KM/H	NW
2016	13 KM/H	NNW

Data captured from Weather Underground based on each day the marathon was run in the respective year. [9]

   We can see that wind conditions, and more specifically the direction of the wind, is quite varied year to year.
4) Other Age Features We also believe there was more potential in exploring more age related features. For example, a possible feature could have been whether a runner fell into the age bracket of 15-65, or were outside of this range. This feature was highly relevant to our attendance prediction task because the age range 15-65 contained practically all of the repeat runners.

D. Limitations in Analysis

Since the focus was on using .632+ bootstrap, we did not create any training-error/validation-error graphs discussed in class. We also failed to make any significant progress in Y1; methods like precision, recall, and F1 measure were not very useful in comparing models that only predicted zeros. Linear regression would have been improved by considering Id as a categorical feature, but we did not have sufficient computational resources to perform the matrix operations.

IV. STATEMENT OF CONTRIBUTIONS

Niko lead efforts to program the naive bayes and linear regression algorithms, as well our bootstrap and data cleansing. Robert lead development of our logistic regression algorithm and implemented k-fold cross validation. Vasu built features, contributed to cleaning data and writing the report, with help from Robert and Niko.

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

References

[1] Efron, Bradley, and Gail Gong. "A leisurely look at the bootstrap, the jackknife, and cross-validation." The American Statistician 37.1 (1983): 36-48.

[2] Efron, Bradley, and Robert Tibshirani. "Improvements on cross-validation: the 632+ bootstrap method." Journal of the American Statistical Association 92.438 (1997): 548-560.

[3] Weather Underground. Retrieved 04:40, January 27, 2017, from "Weather History for Miami, FL" https://www.wunderground.com/history/airport/KMIA/.

[4] Marathon Guide. "Miami Marathon Race Results" Retrieved 04:40, January 27, 2017, from http://www.marathonguide.com/results/browse.cfm?MIDD=1850160124

[5] Miami Marathon. (2017, January 8). In Wikipedia, The Free Encyclopedia. Retrieved 04:40, January 27, 2017, from https://en.wikipedia.org/w/index.php?title=Miami_Marathon&oldid=758982870