McGill University
COMP551 - Fall 2016

Mini-project #1: Predicting the winner!
Project Report

Robert Belfer - 260634641 - robert.belfer@mail.mcgill.ca
Uong Bao Lam - 260567988 -lam.uong2@mail.mcgill.ca
Wei-Di Chang - 260524917 - wei-di.chang@mail.mcgill.ca

I. Introduction

Using data scraped from www.sportstats.ca for about 8700 runners from Canada and their performance in various races from 2012 to 2016, two variables were predicted for each participant:

- *Y1*: Will they participate in the 2016 Montreal Marathon ?
- *Y2*: Assuming they participate, what will be their finishing time ?

To answer Y1, we used two methods: the first being logistic regression and the second being Naive Bayes. To answer Y2, linear regression was used.

II. Problem representation

A. Y1 Classification Data

1. Full Feature Set

To predict *Y1*, our approach was to use domain knowledge to find features that would help us obtain a good prediction score while exploiting the data we had on hand as much as we could. In order to achieve all of the above, all of the information fields except for running time were used to develop a feature set, which came out to contain 386 features. The first four features are : sex, age, "marathon ratio", and the number of previous Oasis events attended.

In addition to these features, we use 382 binary values, which we'll call the "Event attendance" features, representing the participant's attendance to each of the 382 distinct events found in the data.

Participants' sex and age:

The sex variable is encoded as a binary feature, equal to True/1 if the participant is male or no sex data found and False/0 if female.

The age is a floating point number computed by taking the average of the age bracket of the last event the participant ran in. If no age bracket was given, the age was set to 40 as it is the mean of the age data.

Participant "Marathon Ratio"

This feature is a floating point number between 0 and 1, computed by parsing the number of events as well as the count of "Marathon" event types a participant is listed in. It is the ratio between the number of full marathons ran and the number of events the person participated in.

Number of previous Oasis events attended

This feature is an integer value computed from the event name data. Similarly to the marathon ratio, it comes from the idea that if a participant has run in an Oasis event in the past, he is more likely to compete in an Oasis event again.

Event Attendance

These 382 features are binary features, each of them corresponding to a race event, taking the value True/1 if the participant ran at that event and False/0 otherwise.

This allows our prediction algorithms to potentially derive correlations between some of these events which might not be apparent initially

Event and Marathon Attendance

This feature was only used for the response variable Y and corresponds to attendance at a specific event. It is a binary variable taking value 1 if the participant attended this event and ran a full marathon, and 0 if either of these conditions are not fulfilled.

This response variable attempts to get as close to Y1 (2016 Oasis full marathon participation) as possible.

We decided to approximate it using the 2015 Oasis full marathon data for training. Hence for this value to be 1, the participant needs to have the date “2015-09-20” listed and “Marathon” listed as the corresponding event type.

2. Logistic Regression

For the Logistic Regression classifier, the full feature set described above was used to predict Y1 for each participant.

3. Naive Bayes

For the Naive Bayes classifier the following subset of the full feature set were chosen to predict Y1 for each participant:

- Sex
- Age
- Number of Oasis Events attended previously
- Marathon ratio
- Participation in 2014-09-28 event
- Participation in 2013-09-22 event
- Participation in 2013-02-17 event

Our insights on the importance of past participation event date on Y1 were validated by feature selection. Each of the 386 dates as a single feature were tested on the Naive Bayes classifier and the above three dates give the most accurate prediction on Y1, shown in table 4. Thus, these 3 features were chosen for the model.

The differences in number of dates features used by Logistic Regression and Naive Bayes would be elaborated in the discussion section.

B. Y2 Prediction Data

To predict Y2, we use four features. The first feature used is sex. Domain knowledge suggests that, on average, males run faster than females, which justifies the inclusion of sex as a feature. We assume that the runners who did not participate in a category that lists their sex are male.

The second feature is the estimated age of the contestant. Its inclusion accounts for the fact that as people age, they start running slower. To get an estimate of someone’s age, we first assume that all runners are aged between 13 and 70. We then look through each runner’s event history, and examine their category to find the tightest bounds for their age. We then average the lower and upper bounds to get an age estimate. For people whose age is completely unknown (i.e. they did not participate in a category that gave an age estimate), we assume that their age is 40. Similarly, for the classification feature set, we chose 40 because it is the average estimated age of the runners for whom we could estimate it.

Before discussing the next two features, we first need to talk about comparing the performance of two individuals. Indeed, the data contains information about many events, and not all of them involved running. For example, some events involved biking, some were obstacle courses, and some were triathlons. We considered events that only involved running. Comparing times between non-running events and running ones would involve a lot of assumptions and could very easily lead to a bad model. In addition, we only considered running events that had a distance between 20km and 65km, which we shall refer to as ‘relevant races’. We restricted ourselves to studying these events because they are a format very similar to marathons. In addition, we do not expect a person who has never participated in at least a half-marathon to participate in the upcoming marathon, considering that the half-marathon is offered on the same day as well. In other words, ignoring the non-relevant races will reduce the number of assumptions we have to make, and have minimal impact on the MSE for the test set.

With that being said, we are ready to introduce our last two features. The third feature is the number of relevant events a candidate has participated in. The more events a candidate participates in, the more experience they get, and the better they are expected to perform.

The final feature we considered was each runner’s time in their previous relevant race. Of course, since we are considering races of various length, we have to scale the time. In other words, given a runner’s time T_1 and distance D_1, we try to predict T_2, the time it would take the runner to complete a marathon. To do so, we use

Riegel's formula, $T_2 = T_1(D_2 / D_1)^{1.06}$. Riegel's formula is frequently used because of its impressive predictive accuracy [2].

III. Training Method

Each of the implemented prediction algorithms were trained and tested using 5-fold cross-validation, which the team agreed on as a good balance between model selection/evaluation method vs the increase in computation time required. With approximately 9000 training examples in the dataset, 5-fold cross-validation leaves around 7000 examples for training and 2000 examples for validation testing. For linear regression a significant number of examples had to be left out, and we end up with about 1600 training examples and 400 examples for validation testing.

A. Logistic Regression

Training for Logistic Regression is done using a standard gradient descent algorithm. However a lot of optimization can be done using the selection of hyperparameters such as the λ value for regularization, the convergence/termination criteria and the learning rate/step size α for gradient descent.

In order to select each of these hyperparameters, a variation of each parameter while keeping the others constant was done. Instead of evaluating each individually, an exhaustive search of all possible value combination using an approach such as Grid Search would have been more optimal. However, time did not permit implementing a Grid Search algorithm.

1. Convergence criteria

By varying the convergence criteria from 1 to 0.0005 and keeping the learning rate constant at 0.005, we selected the convergence criteria value that would maximize accuracy. The relatively low constant value for the learning rate was chosen for this experiment to avoid any oscillatory behavior which can happen for higher values of the step size, while keeping computation times low relative to the time available to complete the project.

2. Learning rate α

Similarly to the convergence criteria, the learning rate was lowered incrementally from 1 to 0.0005. A constant value of 0.01 was chosen for the convergence criteria as we have found out in the previous section that it gives good accuracy while keeping computation time low. The results of this experiment are presented in table 1 and Figure 1.

3. Regularization

By maintaining fixed values of both the learning rate α and the convergence criteria to 0.01 and 0.005 respectively, we evaluated different values for the regularization parameter λ for both L1 and L2 regularization.
However based on initial testing with a broad range of values for λ, results were consistently worse when compared to no regularization, λ = 0.
We therefore chose to leave out regularization.

B. Naive Bayes

For the Naive Bayes classifier, there was no difference in 5 fold cross validation accuracy in the likelihood calculation, P(x|y), of gender as a Gaussian or by numeration. Gender ratio of all the participants can be accurately represented by a gaussian distribution.

There is a higher cross validation accuracy of 1 - 2 % where the likelihood of participation in the three dates is assumed to follow a Gaussian distribution . As a result, we choose to represent the participations of candidates as Gaussians.

We tested k-fold cross validation for k=1 to 10, and decided to pick k = 2, since it gave the lowest validation error, as shown in Table 3.

C. Linear Regression

Linear regression was optimized in closed form. For training, participants were split into three groups. Those who did not participate in a relevant race were not used for any training. Those who participated in a single relevant race were used to train the predictor P_1. P_1 uses features that are only a function of a participant's sex and age, and is meant to predict the time of a person who has never participated in a relevant race before. As discussed previously, it is unlikely that such people will participate in the marathon. Those who participated in two or more relevant races were used to train the predictor P_2. P_2 uses all of the features described in the problem representation section. To increase the number of available data points, we consider all but the first races ran by these runners(i.e. if a participant ran 4 races, we consider his three most recent ones to be independent data points, given the features we discussed). We also removed outliers by excluding any races for which a runner had an adjusted time of over 6 hours.

We applied linear regression on functions of features of order 1, 2, and 3, with no cross terms. Since our feature count was much lower than the size of the training set, regularization was not used.

IV. Results

For logistic regression, the results of various hyperparameter tests are presented in Table 1, Figure 1, Table 2, and Figure 2.

For the convergence criteria,, the prediction accuracy starts to plateau off starting from a convergence criteria value of 0.01, only giving marginal accuracy gains up to the last tested value of 0.0005. While small, these gains should not be ignored as we wish to maximize accuracy.

Hence, ideally, we would pick a convergence criteria as small as time and computation resources permit. In our case that is 0.0005.

The learning rate's accuracy stays constant at about 67% from $\alpha = 1$ down to $\alpha = 0.005$, at which point it abruptly increases to approximately 95%. From $\alpha = 0.005$ to the last tested value of $\alpha = 0.0005$, the accuracy then gradually drops down to around 91%. We know that a learning rate that is too high prevents access to the global minimum, while one that is too low can get the algorithm stuck in a local minimum, and these results reflect exactly that. Therefore, for our purposes we selected a learning rate $\alpha = 0.005$.

Using all of the 382 Event attendance features, the logistic regression method gave the best prediction performance of approximately 95% cross validation accuracy while with the same feature set, Naive Bayes only gave a near random result of 53%. Instead, Naive Bayes, obtains its best prediction of 86% validation accuracy using only three important dates.

For linear regression, the performance of the predictors was measured by calculating the MSE on the validation set. It is worth noting that the validation set also contains the adjusted times of people who did not necessarily participate in the marathon. The results are summarized in table 5 of the appendix. We observe that the order 2 fit produces the best results. Thus, to get the weights for P_1 and P_2 , we apply the order 2 fit regression on the entire training set. The obtained formulas for P_1 and P_2 are shown in the appendix, under equations 1 and 2 respectively.

V. Discussion

For Y1 prediction using Naive Bayes, feature selection played an important role in validation accuracy.

As shown in Table 4, age and sex only played a small role in Y1 prediction compared to the Marathon ratio and the three dates.

This suggests Naive Bayes weighs every feature equally for its prediction since the P(x|y) incorporates the product of all of the features x_i. Hence, if a lot of unimportant date features were used, the method would obtain a lower accuracy since the assumption of independence in features is held.

Logistic regression on the other hand can assign different weights to its features. If the features are highly correlated or repetitive, logistic regression can compensate by varying the weights. This results in a better prediction accuracy for Y1 since many more features can be used and their correlations can be accounted for by variation in weights.

By exhaustively testing every event date as a single feature using Naive Bayesian, we determine that the three dates 28/09/2014, 22/09/2013 and 17/02/2013 are the three most important features on determining Y1 in table 3. It is no coincidence that two of these events are also the dates in which the previous Oasis Montreal Marathon took place.

For Naive Bayes, there is little difference between training error and validation error, as seen in table 3. This suggests the model is relatively simple and no overfitting occurs. This simplicity comes from the model assumption of independence between features.

One of the limitations of our prediction is that we have to assume that the trends displayed for the 2015 Oasis Marathon will also happen for the 2016 one.

For the linear regression model, all of our initial assumptions are confirmed. Both predictors predict that men should, on average, run faster than women, and that people run slower as they age. In addition, P_2 assigns a weight of large magnitude to a runner's previous time, which suggests that it is an important value. This would explain why P_2 performs better than P_1 on the validation set. Finally, participating in multiple races will reduce the time predicted by P_2.

A possible concern would be the fact that we used the dataset containing the adjusted times of runners. While this increases the size of our dataset, it also means we use fabricated values to test the accuracy of the regression model. However, since Riegel's formula has been shown to be empirically accurate, this should have little effect of the model's performance on the test set. Finally, we could have tested predictors that include cross terms, but that would have been too time consuming, as we would have to test over 40 cases for the degree 3 fit.

Given the data we have available, we feel quite satisfied with our model. Since our training and validation errors barely differ, we believe that we will have around 95% prediction accuracy and a MSE of 0.5 hours^2. A possible way of improving it would have been to scrape the weather for each of the events. Indeed, we believe that the weather should have an effect on both attendance and performance.

VI. Statements of Contribution

All team members discussed and had equal contributions in deciding which features to use for each of the three problems. Wei-Di answered Y1 using logistic regression, Lam answered Y1 using Naive Bayes, and Robert answered Y2 using linear regression. Everyone wrote about their findings separately, and then everyone worked together to make the final report. Wei-Di unified our code so that it could easily be ran. We hereby state that all the work presented in this report is that of the authors.

VII. References

[1] http://www.runningusa.org/marathon-report-2016?returnTo=main , "2015 Running USA Annual Marathon Report", *Running USA Organization*- Accessed on 23-09-2016

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

VIII. Appendix

		Accuracy (%)										
Convergence Criteria	**Alpha**	2345_1		1345_2		1245_3		1235_4		1234_5		**Test Avg**
		Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	
1	**0.005**	33.47	32.24	31.63	32.70	32.32	32.53	32.89	32.39	32.13	32.58	**32.49**
0.5	**0.005**	66.53	67.76	68.37	67.31	67.68	67.48	67.11	67.63	67.87	67.44	**67.51**
0.1	**0.005**	90.70	91.29	91.96	91.00	90.47	91.43	91.50	91.10	91.28	91.26	**91.18**
0.05	**0.005**	92.71	93.69	93.69	93.36	93.69	93.37	93.46	93.49	93.57	93.51	**93.42**
0.01	**0.005**	94.20	95.21	96.21	94.96	95.12	95.37	94.72	95.29	95.35	95.26	**95.12**
0.005	**0.005**	94.20	95.68	96.38	95.51	95.35	95.84	95.41	95.80	95.70	95.80	**95.41**
0.001	**0.005**	95.69	96.79	96.50	96.58	95.81	96.71	96.21	96.48	96.16	96.60	**96.07**
0.0005	**0.005**	95.52	97.06	96.67	96.81	96.04	96.90	96.21	96.81	96.21	96.76	**96.13**

Table 1: Accuracy results for Y1 predictions using logistic regression with no regularization or normalization while varying the convergence criteria for gradient descent

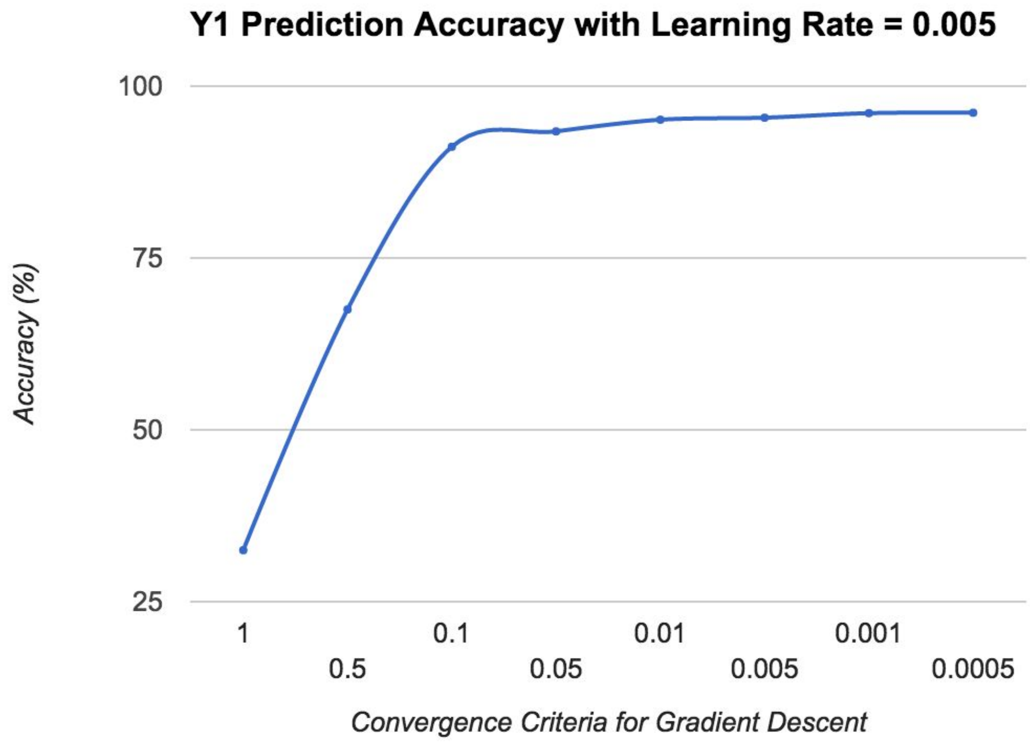

Figure 1: Y1 Prediction Accuracy as a function of the convergence criteria

		Accuracy (%)										
Convergence Criteria	**Alpha**	2345_1		1345_2		1245_3		1235_4		1234_5		**Test Avg**
		Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	
0.01	**1**	66.53	67.76	68.37	67.30	67.68	67.47	67.11	67.61	67.87	67.42	**67.51**
0.01	**0.5**	66.53	67.76	68.37	67.30	67.68	67.47	67.11	67.61	67.87	67.42	**67.51**
0.01	**0.1**	66.53	67.76	68.37	67.30	67.68	67.47	67.11	67.61	67.87	67.42	**67.51**
0.01	**0.05**	66.53	67.76	68.37	67.30	67.68	67.47	67.11	67.61	67.87	67.42	**67.51**
0.01	**0.01**	66.53	67.76	68.37	67.30	67.68	67.47	67.11	67.61	67.87	67.42	**67.51**
0.01	**0.005**	94.20	95.21	96.21	94.96	95.12	95.37	94.72	95.29	95.35	95.26	**95.12**
0.01	**0.001**	92.71	93.69	93.69	93.36	93.69	93.37	93.46	93.49	93.57	93.51	**93.42**
0.01	**0.0005**	90.70	91.29	91.96	91.00	90.47	91.43	91.50	91.10	91.28	91.25	**91.18**

Table 2: Accuracy results for Y1 predictions using logistic regression with no regularization or normalization while varying the convergence criteria for gradient descent

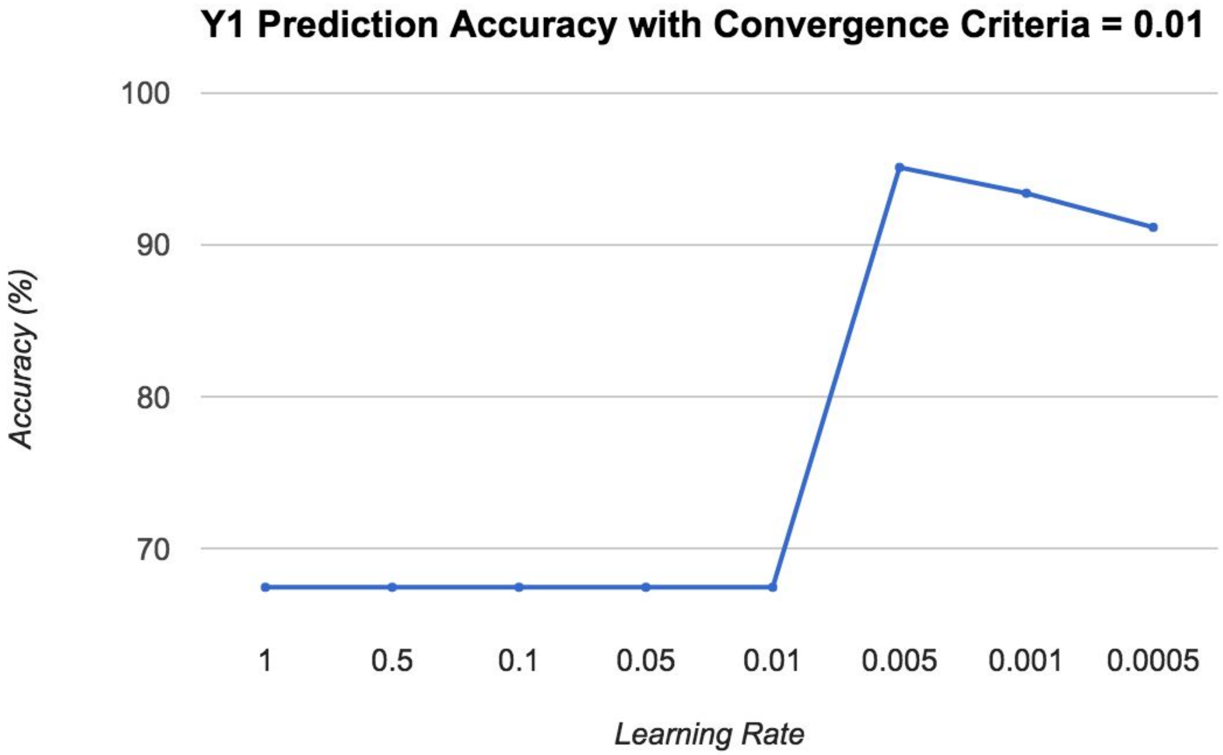

Figure 2: Y1 Prediction Accuracy as a function of the learning rate α

K	2	4	6	8	10
Cross validation accuracy on Test set	86.11	86.03	86.06	86.01	86.09
Cross validation accuracy on Training set	86.05	86.09	86.10	86.09	86.10

Table 3: Cross Validation Accuracy results for Y1 predictions using different K values with Naive Bayes.

Feature	Maratho n ratio to all events	Sex	Age	Number of Oasis events	28/09/2014	22/09/2013	17/02/2013
Training Accuracy	0.61	0.53	0.54	0.53	0.68	0.64	0.63

Table 4: Training Accuracy results for Y1 predictions using different single feature with Naive Bayes.

Predictor	Order of fit	Average Training Error (hours^2)	Average Validation Error (hours^2)
P_1	1	0.500	0.500
	2	0.493	0.494
	3	0.493	0.495
P_2	1	0.311	0.312
	2	0.301	0.305
	3	0.299	0.309

Table 5: Average training and validation errors order for P_1 and P_2

$$P_1(x) = 5.279 - 0.4\, I(x\ is\ male) - 3.63 * 10^{-2} * A(x) + 5.33 * 10^{-4} * A(x)^2 \qquad \textbf{Equation 1}$$

$$\begin{aligned} P_2(x) = &-1.096 - 7.26 * 10^{-2}\, I(x\ is\ male) - 1.48 * 10^{-2} * A(x) + 1.97 * 10^{-4} * A(x)^2 + \\ &+ 2.01 * T(x) - 0.162 * T(x)^2 - 2.84 * 10^{-2} * N(x) + 2.36 * 10^{-3} * N(x)^2 \qquad \textbf{Equation 2} \end{aligned}$$

Where x is the runner, I is the indicator function, A(x) is the runner’s age, T(x0) is the runner’s previous time (in hours), and N(x) is the runner’s number of previous events ran.