# Data Mining Project Milestone 3 - CSC691

Esteban Murillo, Patrick Williams, Sarah Parsons

November 18, 2019

## 1 Introduction

Using a dataset of exercise statistics for a given individual, this project aims to analyze and compare the effectiveness of various algorithms at predicting the trend of statistics from recent values. Given an effective model, an athlete can view predictions of statistics for any number of future workouts. This is useful to assess if they are on pace to meet a statistical goal by a certain time, and to determine areas where improvement is needed. An explanation of methods for gathering and pre-processing data, and a presentation of the analyses of two models on an isolated statistic, average heart rate, are provided. A Logistic Regression model was chosen to expand this work to evaluate the effectiveness of the Moving Average model and to include more features to provide more robust predictions.

## 2 Data

The data used for this project consists of exercise statistics measured for daily indoor (virtual) and outdoor bike rides over a span of two years. In total, there are 309 records with 14 virtual rides and 295 outdoor bike rides. The data was recorded using a mobile application, Strava, and was retrieved using a browser plug-in, 'Elevate for Strava', that consolidates all data in Strava for exportation to a .csv file. For each ride, a measurement of average heart rate (AvgHR (bpm)), distance traveled (Distance (km)), elevation gained (Elevation Gain (m)), average pace (Avg Pace (/km)), calories burned (Calories), and heart rate stress score (HRSS) were recorded. HRSS corresponds to a score based on the maximum heart rate that an individual can sustain for up to an hour. In simple terms, it is the equivalent of assigning a score to the heart rate for a given activity. These six statistics were chosen due to their ease of interpretation and appropriate format (continuous, real values) for the project. The overall average heart rate throughout the dataset is 155 bpm, while the AvgHR for virtual rides is 165.5 bpm and for outdoor rides is 154.5 bpm. Furthermore, for each ride, the date, time, and type of ride were stored for reference to help inform the order and type of activity of each event in the time series.

```
1 df.groupby('AvgHR_bin').mean()
```

| AvgHR_bin | AvgHR | Distance (km) | Avg Pace (/km) | Calories | HRSS | Elevation Gain (m) |
|---|---|---|---|---|---|---|
| 0.0 | 145.794521 | 55.214384 | 167.794521 | 1644.650685 | 145.445205 | 1048.490411 |
| 1.0 | 163.306748 | 49.571166 | 135.245399 | 1509.625767 | 157.914110 | 745.698773 |

Figure 1: Data Grouped By AvgHR Bin

```
1 df.groupby(['Type']).mean()
```

| Type | AvgHR | Distance (km) | Avg Pace (/km) | Calories | HRSS | Elevation Gain (m) |
|---|---|---|---|---|---|---|
| Ride | 154.535593 | 53.271525 | 152.298305 | 1615.294915 | 155.122034 | 912.209831 |
| VirtualRide | 165.500000 | 30.450000 | 115.357143 | 691.142857 | 86.714286 | 394.757143 |

Figure 2: Data Grouped By VirtualRide and Ride

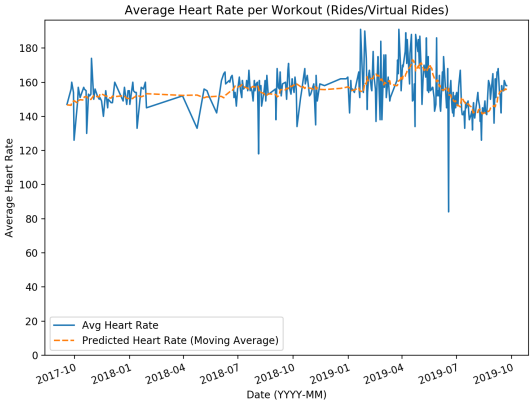



Figure 3: Actual and Predicted Average Heart Rates, Using Simple Moving Average

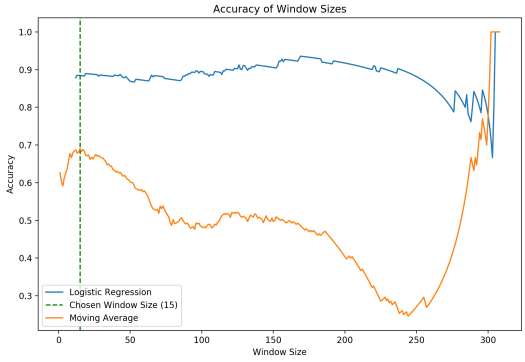



Figure 4: Accuracy of Window Sizes for Both Implementations

Using the Distance, Elevation Gain, Avg Pace, and HRSS, two methods - moving average (MA) and logistic regression (LR), were implemented to forecast and predict the AvgHR, respectively. To prepare for these methods, the features were pre-processed and the data was cleaned. All rows without an AvgHR value were removed, all features were converted to numbers, Avg Pace was converted to seconds, and the datetime field was parsed to form two separate columns of Date and Time. Finally, given that LR requires a categorical target variable for classification (binary or multinomial), the AvgHR was transformed into a binary target variable (AvgHR bin) with labels of 0 or 1. Using a threshold of 154 bpm, 0 corresponds to a 'low' heart rate and 1 to a 'high' heart rate. 154 bpm was chosen as the threshold in order to maintain the closest even sample distribution of labels across the two classes, with 52.8 percent of 'high' samples and 47.2 percent of 'low' samples. A few descriptions of the average feature values for high/low heartrates and for Rides/VirtualRides are provided in Figures 1 and 2.

# 3 Methods Approaches

A moving average (MA) algorithm and a logistic regression (LR) model were implemented to forecast and predict the average heart rate of each bike ride. For both methods, an optimal window size was determined and a test set was predefined. The window size was chosen by iterating over every possible window size and comparing the accuracies for MA and LR. Due to the large discrepancies in accuracies for both algorithms at different window sizes (see Fig. 4), selecting the optimal window presented a challenge. As shown in Fig. 4, the best accuracy for both models was at window size 15 (accuracies: 0.69, 0.88 for MA, LR) and from 305 to 308 (accuracy: 1.0). Given that a larger window size results in fewer predictions, in order to implement a valuable model, 15 was selected as the optimal window size.

The *moving average* algorithm was used to forecast the average heart rate for a given window size. It is worth noting that two different versions of the algorithm were implemented: simple and moving. The basic idea behind both algorithms is to use a specific portion of the data as the training set (window) in order to predict newer values in the future. For the simple MA (SMA), the focus is placed on a particular window of data points to make the next prediction and the window continues to increment the list of data points, but remain the same size, until all the corresponding predictions are made. A graph of predicted and actual average heart rates using SMA is shown in Fig. 3. In a similar way, the cumulative version uses the window, but grows as it accumulates data points with each iteration through the list. It was determined that the *simple moving average* version worked best for the data. Below is an example of how the algorithm works. The accuracy, precision, recall, and f1-score were computed for a window size of 15 (see Preliminary Results), and a confusion matrix was built (see Evaluation Proposal).

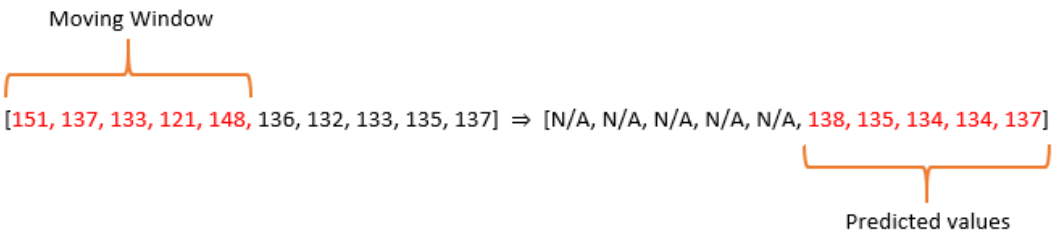



Figure 5: Example of Moving Average Implementation

To implement LR, after pre-processing a set of features and cleaning the data, a feature extraction method (RFE, or recursive feature elimination) from sci-kit learn (sklearn) was used to determine the significance and ranking of each feature. First, the training set was defined to be the size of the optimal window and was made up of previous records, while the test set consisted of the remaining records that were recorded most recently. Given the train and test sets, the output of RFE assigned all selected features a ranking of 1, and thus, all features were kept. Next, the sklearn statsmodels Logit function was used to implement a logistic regression model using the full dataset to better understand the coefficient and p-value of each feature. Based on the result summary, the p-value for Calories was greater than 0.05, but all other features had a p-value equal to 0. Consequently, Calories was removed from the training set.

With the train and test sets defined, sklearn's Logistic Regression model was configured (with default parameters) and fit to the train set. To predict each test instance, a quasi-'Leave One Group Out' cross-validation process was implemented to include each prior test instance in the train set for the next iteration of predictions (see example below). The accuracy, precision, recall, and f1-score of the model were computed (see Preliminary Results), and a confusion matrix was constructed (see Evaluation Proposal).

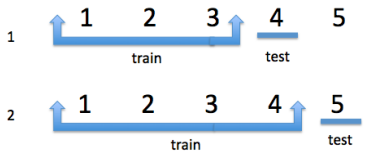



Figure 6: Cross-Validation Process for LR

# 4 Preliminary Results

For LR and MA with a window size of 15, the accuracy is 0.88 and 0.69, respectively, for 294 predictions, and the following precision, recall, and f1 scores were computed. It is clear that the Logistic Regression model outperforms the Moving Average model, as the accuracy, precision, recall, and f1-score for LR are better than those for MA. Likewise, as shown in Figure 4, regardless of the window size, the accuracy of MA remains lower than LR except for the window size range of 300 to 304, at which point the accuracy of MA exceeds LR until they converge to 1 at window size 305.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.67 | 0.63 | 0.65 | 133 |
| 1 | 0.71 | 0.74 | 0.72 | 161 |
| accuracy | | | 0.69 | 294 |
| macro avg | 0.69 | 0.69 | 0.69 | 294 |
| weighted avg | 0.69 | 0.69 | 0.69 | 294 |

(a) Moving Average

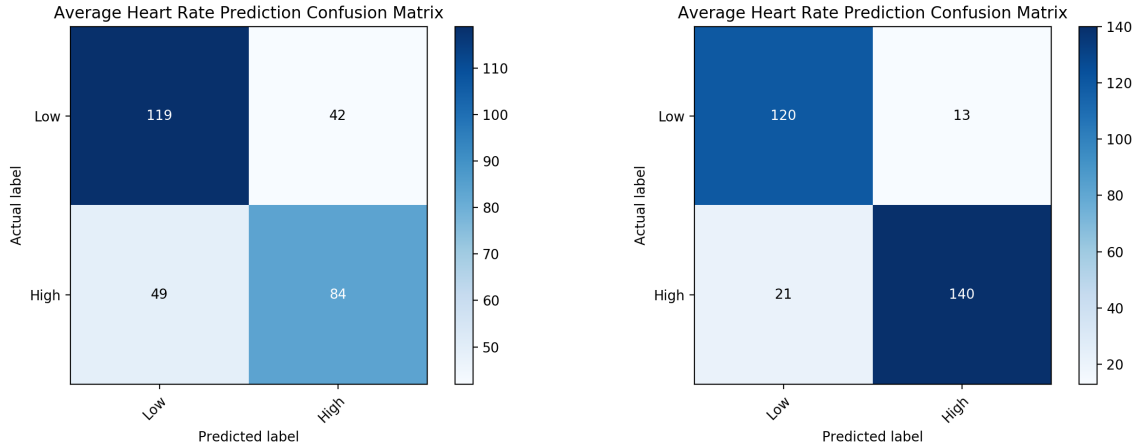
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.85 | 0.90 | 0.88 | 133 |
| 1.0 | 0.92 | 0.87 | 0.89 | 161 |
| accuracy | | | 0.88 | 294 |
| macro avg | 0.88 | 0.89 | 0.88 | 294 |
| weighted avg | 0.89 | 0.88 | 0.88 | 294 |

(b) Logistic Regression

Figure 7: Precision, Recall, F1 for both Implementations (window size = 15)

### 4.1 Evaluation Proposal

The confusion matrices for both MA and LR with a window size of 15 are shown below. As shown, the results for LR are better than MA. Based on these preliminary results, it is apparent that of the two models, LR should be used for predicting the average heart rate for indoor and outdoor bike rides given the specified feature set.




(a) Moving Average

(b) Logistic Regression

Figure 8: Confusion Matrices for Moving Average and Logistic Regression implementations

## 5 Next Steps

In order to fully analyze the capabilities of moving average, a next step is to implement an exponential moving average model for bike rides. Then a conclusion can be drawn on whether or not LR is more sufficient than all moving average algorithms for predicting average heart rate. Moreover, an additional next step includes predicting a different variable from the dataset, such as the 'Type' of exercise, for all records in the dataset including runs, workouts, and rides. Although an accuracy of approximately 0.88 was achieved by LR for predicting average heart rate, the standard deviation of AvgHR was only 12.5, and therefore, had little variation. Furthermore, the demand for heart rate predictions may not be as valuable as that of exercise type. By accurately predicting the type of exercise, the model could be more robust and more applicable to real-world use cases. Finally, given the diverse methods of modeling the dataset, not only is it desirable to sustain these implementations for evaluation purposes, it is also preferred that the code be merged for easier consumption and future maintenance.