PaceMakers: Predicting Average Heart Rate for Bike Rides

Esteban Murillo, Patrick Williams, Sarah Parsons

December 6, 2019

1 Abstract

With more and more people concerned about their health in a world full of sedentarism, it is important to keep track of exercise data so that individuals can understand and analyze in a way that is most beneficial to them. Thanks to novel approaches and the ease of access to *data mining* techniques, information regarding sports and general well-being is more attainable than ever before. While most ongoing research focuses on either performance or well-being, this research attempts to balance both. This paper discusses several methods that were implemented that predict *heart rate* successfully. Moreover, with access to cycling data from indoor and outdoor environments, a cluster analysis was performed on the data to distinguish between the different types of activities. It is shown that this research can easily be used by both cyclists that are working to improve their performance, and people that desire to be more physically active.

2 Background and Prior Work

Research regarding sports and well being is not new. However, due to technology being more accessible and easier to understand, this topic has been researched now more than any other time. Moreover, because of *data mining* and *machine learning* in general, extracting relevant metrics related to sports is easier than ever before.

There are two very different areas in physical activity. First, there is research that focuses on predicting metrics strictly related to performance, such as (4; 1). Second, there is research that centers on physical activity data, which naturally impacts more people.

For instance, in (4) their main focus is on *power*, a significant metric related to cycling performance. Similar to the research in this paper, they collected data from numerous athletes, ranging from amateurs to professionals, and tracked all kinds of important measurements. For their research they used more data and also followed a more structured form. In order to derive the final results, the authors divided the tests into three different stages. The first corresponded to the classical FTP test, described by Dr. Coggan, which is defined as the maximum amount of power that a cyclist can sustain for approximately one hour. Then they followed with an *intermittent power (IP)* test. This consisted of 20 high-intensity intervals, each lasting 45 seconds with a 15 second recovery period in between. The final stage was the race itself in which the individuals competed for approximately 70 minutes.

They conducted all of the experiments with eleven (11) male competitive *cross-country mountain bike (XCO-MTB)* cyclists. According to researchers, all of them were in good shape; free of injury, and at the time of experimentation were physically active, having four or more weekly sessions. Also, it is worth mentioning that since all three stages of the examination were physically demanding, at least 72 hours passed between every stage. Even though agreements were reached between the test individuals and authors, the data was not released to the public.

They conclude that both *FTP* and *IP* represent good methods at determining the actual outcome of the race. They show that using *IP* in the linear regression minimized the error by a small factor, thus making *IP* preferable over *FTP* for these types of predictions.

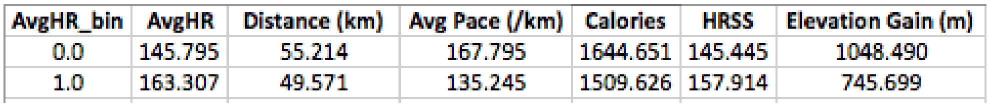
Another paper that focuses on performance is (1). It follows similar lines as our research, using relevant features from both amateur and professional cyclists. Their goal is to show that: “engine matters, but the tuning is fundamental”, meaning that regardless of how well-fitted an athlete was born, if no hours are put into training then everything else is meaningless. They backup their data with very cycling-specific metrics, such as *power*, *average climbing speed*, and structured *heart rate zones*, which is a way of dividing the range of an individual's heart rate by creating sub-ranges.

It is worth mentioning that they collected data from multiple athletes, something that most likely provided an advantage at the time of classification and clustering. The yielded results for their clustering are very precise. They performed a *3-Means* cluster analysis, and consequently, the activities were successfully classified. The first cluster contained activities performed by 'low trained' riders, that did not exercise often. The second cluster consisted of activities by riders that exercised on a more regular basis. Lastly, the third cluster was formed by activities performed by professional athletes. As mentioned, in order to classify each activity correctly, more cycling related metrics were taken into account than those chose for this project.

Finally, (5) also focuses on *heart rate* but from a different perspective. It centers itself on daily physical activity instead of sporting competitions. Their used methods differ to the methods outlined in this paper in that they use *Evolutionary Neural Networks* to make their predictions.

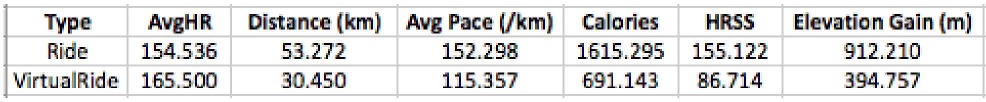
3 Data and Methods

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.



AvgHR_bin	AvgHR	Distance (km)	Avg Pace (/km)	Calories	HRSS	Elevation Gain (m)
0.0	145.795	55.214	167.795	1644.651	145.445	1048.490
1.0	163.307	49.571	135.245	1509.626	157.914	745.699

Figure 1: Average Metrics Grouped By AvgHR Bin



Type	AvgHR	Distance (km)	Avg Pace (/km)	Calories	HRSS	Elevation Gain (m)
Ride	154.536	53.272	152.298	1615.295	155.122	912.210
VirtualRide	165.500	30.450	115.357	691.143	86.714	394.757

Figure 2: Average Metrics Grouped By Ride Type (VirtualRide and Ride)

Using the Distance, Elevation Gain, Avg Pace, and HRSS, several methods - moving average (MA), binary classification, and various regression models, 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 Logistic Regression and Random Forest require a categorical target variable for classification (binary), the AvgHR was transformed into a binary target variable (AvgHR_bin) with labels of 0 or 1. Using a threshold of 154 bpm, 0 corresponded 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 of the full dataset, with 52.8 percent of 'high' samples and 47.2 percent of 'low' samples. A few descriptions of the average feature values grouped by high/low heart rates and for Rides versus VirtualRides are provided in Figures 1 and 2. Also, a scatterplot of the time series data for each of the 309 daily activities is shown in Figure 3.

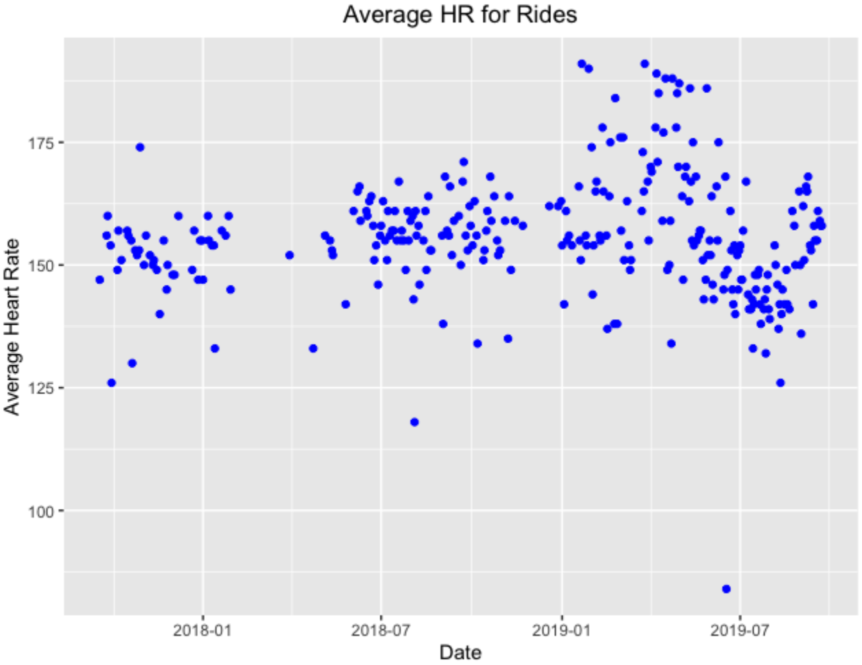


Figure 3: Average Heart Rates Plotted Over Time

Furthermore, in an effort to understand the differences between virtual rides and outdoor rides, analysis was performed on each subset of the data separately. To do so, separate data frames were built for virtual rides and outdoor rides. The same analysis was performed on both datasets to determine which method performed best at predicting the average heart rate for each type of ride.

A simple and cumulative moving average algorithm, two binary classification models, and three regression models were implemented to forecast and predict the average heart rate of each bike ride. For each method, an optimal window size was determined and a test set was predefined. Defined as the minimum number of data points necessary to predict one day in the future, the window size was chosen by iterating over every possible window size and comparing the accuracies for each iteration. Furthermore, the optimal look ahead value, or maximum number of days to be predicted given a window size, was also determined by running each model across all possible look ahead values and assessing their performance. With this information, it was possible to determine the capabilities of each model and the quantity of data in proportion to the total dataset needed to accurately predict the average HR for a specific number of daily bike activities.

3.1 Moving Average Algorithms

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 this 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 remains the same size, until all the corresponding predictions are made. 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* 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 and a confusion matrix was built (see Results).

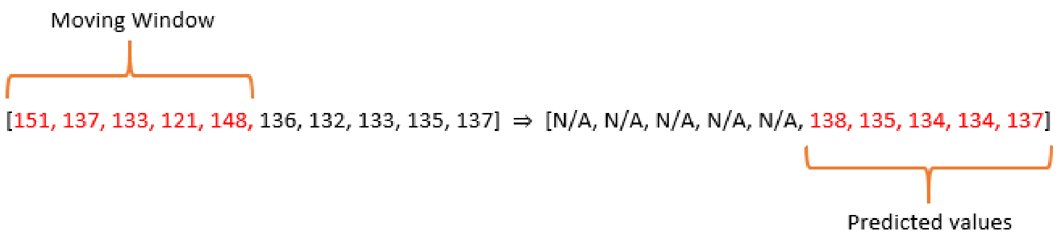

Moving Window
[151, 137, 133, 121, 148, 136, 132, 133, 135, 137] ⇒ [N/A, N/A, N/A, N/A, N/A, 138, 135, 134, 134, 137]
Predicted values

Figure 4: Example of Moving Average Implementation

To implement the classification and regression models, 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 and was made up of previous records, while the test set consisted of the remaining records 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.

3.2 Classification Methods

For binary classification of AvgHR (0/1) for outdoor, virtual, or all types of bike rides, Logistic Regression and Random Forest were implemented. With the train and test sets defined, sklearn's Logistic Regression model was configured (with default parameters) and fit to the train set.(3) 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 while the oldest data point was dropped from the train set (see example below). Similarly, the same process was used when implementing sklearn's Random Forest Classifier with 100 estimators.(2) In addition to finding the optimal window size, the optimal look ahead value was also computed using a comprehensive process for testing each ahead value.

In order to compare virtual and outdoor rides, datasets were configured with daily activities for each type of ride. Given the relatively small set of virtual rides (14) in the total dataset (309 records), the threshold for low and high AvgHR values in the virtual rides dataset was modified to 163 to maintain an even distribution of classes. Moreover, it is worth noting that the size of the virtual ride dataset greatly impacts the significance of the analysis. With less data, testing the model can present issues and less reliable results, as compared to those for the outdoor rides (295 records). Despite these limitations, the virtual and outdoor ride datasets were trained and tested with LR, yielding decent results (see Results). Due to time constraints, both classification models were used to analyze the outdoor and virtual datasets, but not regression models.

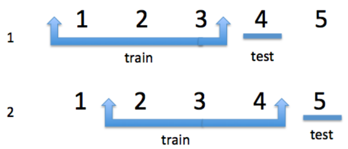


Figure 5: 'Leave One Out' Cross-Validation Process for SMA, Classification, and Regression Models

3.3 Regression Models

For regression, Linear Regression (LinR), Gradient Boosting Regression (GB), and Random Forest Regression (RFR) were implemented. To configure, the default parameters for all of these sci-kit learn algorithms were used (500 estimators, max depth of 4, 2 minimum samples split, learning rate of 0.01, and 'ls' loss function for the Gradient Boosting model; 100 estimators for the RFR model; and no parameters for LinR). Similar to the classification models, the optimal window size and optimal look ahead value for these models were determined by iterating over all possible window sizes and look ahead values. By identifying the parameters with the lowest mean-squared error, the best window size and look ahead value were selected. For regression, although each AvgHR prediction was a continuous value, in order to compare them with the classification models consistently, the predicted values were converted to binary values using the initial heart rate threshold and the accuracy was computed for each model.

3.4 Clustering Model

In working with the data, it was observed that there were some possibly significant differences between the values for outdoor rides and virtual rides. In order to test this, a 2-means clustering was performed to divide the selected workouts into two clusters and assess the results. The three features chosen were Distance, Workout Duration, and Calories Burned. These features were selected manually because a trend was observed that outdoor rides tended to be longer in duration, meaning the distance and calories were often higher than for virtual rides. Initially, this was considered unnecessary because there were very few indoor ride workouts recorded, but after adding more recent workout data to the dataset, a k-Means clustering algorithm was run on 361 workouts.

4 Results

For cumulative LR and SMA, as shown in Figure 4, regardless of the window size, the accuracy of SMA remained lower than LR except for the window size range of 300 to 304, at which point the accuracy of SMA exceeded LR until converging to 1 at window size 305. These initial results provided insights into the sophistication of classification models compared to the more rudimentary moving average algorithms. Thus, other more robust models, such as Random Forest, Gradient Boosting Regression, and Random Forest Regression, were expected to produce similar results as LR.

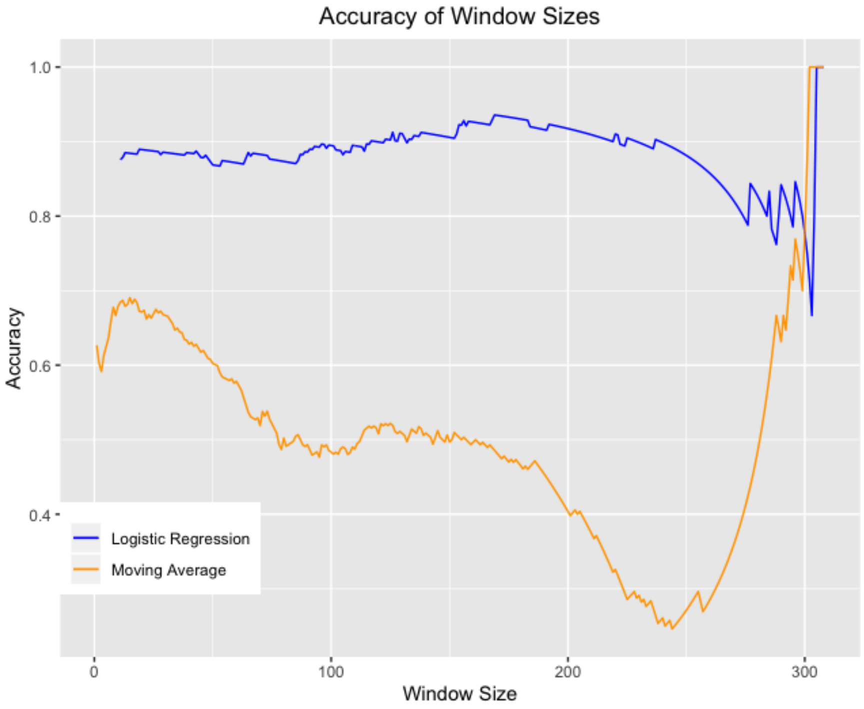
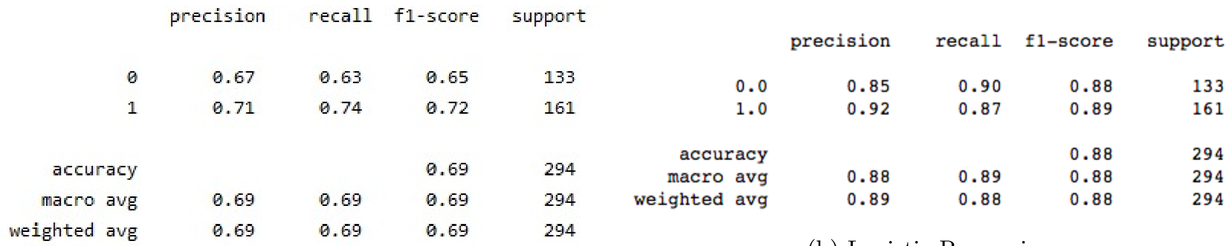


Figure 6: Accuracy per Window Size for Cumulative Logistic Regression and Simple Moving Average, Window Size: [0, 309]

At the conclusion of the first phase of the project, it was discovered that the optimal window size was not required to be the same for each model. As such, the metrics and confusion matrices provided below for cumulative LR and SMA were helpful in initial development, but did not provide as much insight as expected. However, including these results is important given their significance to the project as a crucial turning point. With this knowledge, further work was completed to determine the best optimal window sizes for each model.

In the first phase of the project, simple and cumulative moving averages were tested, as well as Logistic Regression with a cumulative 'Leave One Out' cross-validation process. As shown in Fig. 6, the best accuracies for cumulative Logistic Regression (LR) and SMA was at window sizes 169 and 301, respectively, with accuracies of 0.94 and 0.88 for LR, SMA.



```
              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
```

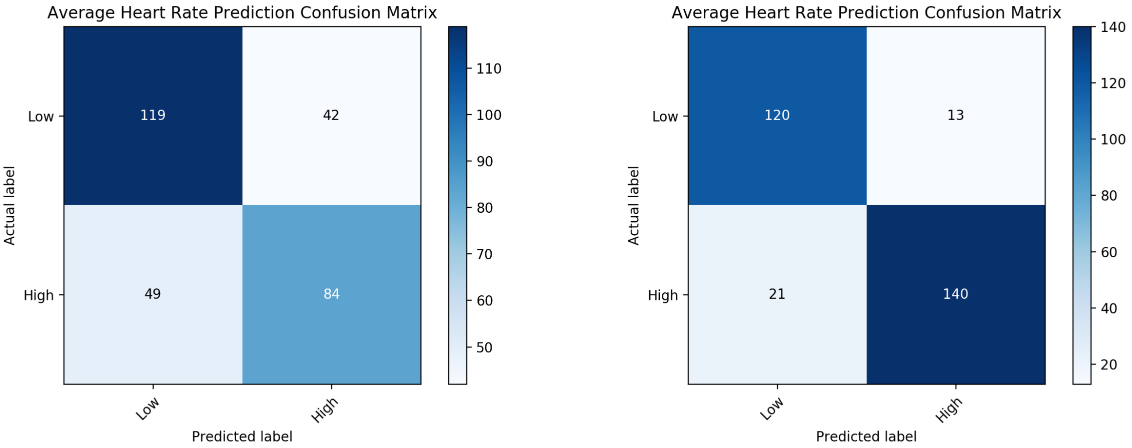
```
              precision    recall  f1-score   support

         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

(a) Simple Moving Average

Figure 7: Precision, Recall, F1 for SMA and LR (Cumulative) (window size = 15)

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

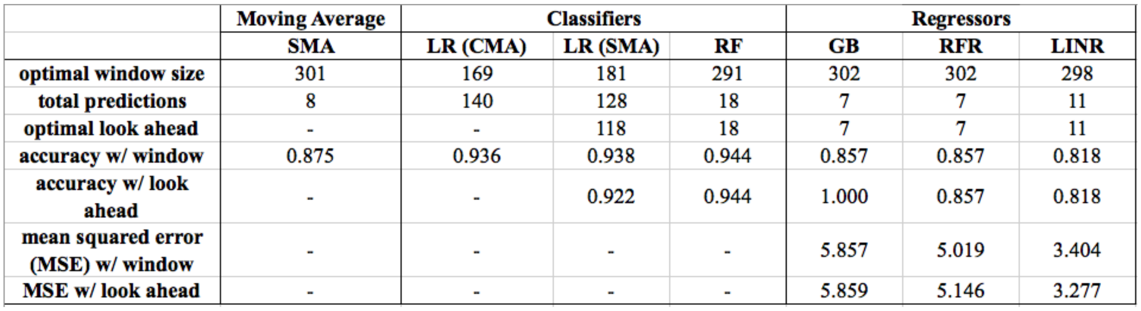


(a) Simple Moving Average Confusion Matrix

(b) Logistic Regression (Cumulative) Confusion Matrix

Figure 8: Confusion Matrices for Simple Moving Average and Logistic Regression Implementations

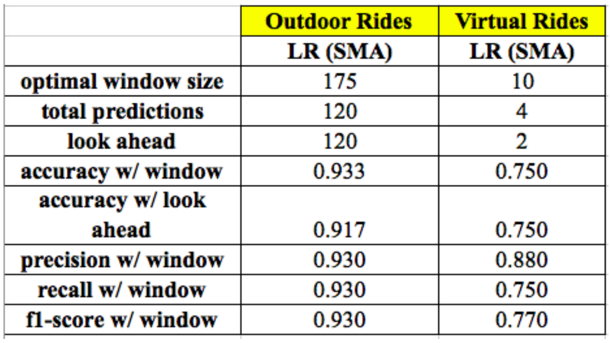
The results shown in Figure 9 provide a summary of the optimal window size, optimal look ahead value, and performance metrics (accuracy and mean-squared error) for all of the models. As shown, the classifiers yielded smaller optimal window sizes and slightly higher accuracies, all above 90%, than the regression models. Logistic Regression appeared to dominate Random Forest, as the optimal window size was smaller for LR and consequently, the look ahead value was larger for LR. Additionally, the accuracies for LR were very similar to those of RF. As a result, LR was determined to be a better model than RF for predicting low/high average heart rate for bike rides. Using Linear Regression (LinR) as a baseline for the regression models, it was determined that the mean-squared error for the optimal window size and optimal look ahead value was better for LinR but the accuracy score was slightly less than that of Gradient Boosting and Random Forest Regression (82% versus 86%). Furthermore, the optimal window size for LinR was slightly smaller than GB and RFR, which made the algorithm more appealing than the other regression models.



	Moving Average	**Classifiers**			**Regressors**		
	SMA	**LR (CMA)**	**LR (SMA)**	**RF**	**GB**	**RFR**	**LINR**
optimal window size	301	169	181	291	302	302	298
total predictions	8	140	128	18	7	7	11
optimal look ahead	-	-	118	18	7	7	11
accuracy w/ window	0.875	0.936	0.938	0.944	0.857	0.857	0.818
accuracy w/ look ahead	-	-	0.922	0.944	1.000	0.857	0.818
mean squared error (MSE) w/ window	-	-	-	-	5.857	5.019	3.404
MSE w/ look ahead	-	-	-	-	5.859	5.146	3.277

Figure 9: Results for Moving Average, Classification, and Regression Models

When comparing virtual rides to outdoor rides using Logistic Regression, due to the limited number of virtual rides, the results were not as valuable as possible had the datasets contained adequate data for both ride types. As shown in Figure 9, the accuracy for predicting the AvgHR_bin for outdoor rides was much more reliable than for virtual rides (0.93 versus 0.75 for outdoor versus virtual). Although the optimal window size appears to be significantly smaller for virtual rides, this was most likely the case due to the fact that the dataset was incredibly small with only 14 records.



	Outdoor Rides	**Virtual Rides**
	LR (SMA)	**LR (SMA)**
optimal window size	175	10
total predictions	120	4
look ahead	120	2
accuracy w/ window	0.933	0.750
accuracy w/ look ahead	0.917	0.750
precision w/ window	0.930	0.880
recall w/ window	0.930	0.750
f1-score w/ window	0.930	0.770

Figure 10: Results for Virtual and Outdoor Rides Using Logistic Regression

The results of the 2-means clustering on outdoor rides and virtual rides are shown in Table 1.

Cluster	1	2	Total
Outdoor Rides	239	90	329
Virtual Rides	32	0	32
Total	271	90	361

Table 1: Distribution of Outdoor and Virtual Rides Among 2-Means Clustering Results

While all of the virtual rides were contained in one cluster, that same cluster also contained most of the outdoor rides. The second cluster contained many of the longer-distance outdoor rides, which confirmed the observation that the virtual rides were shorter than many outdoor rides. It was concluded that the selected features may not have been significant enough to completely separate the two types of rides, and that it is also possible there was not enough data on virtual rides to successfully cluster the two types of rides.

5 Conclusion and Future Work

In conclusion, in an effort to predict the average heart rate as a binary and continuous target variable for daily outdoor and virtual bike rides, the trained classifiers, regressors, and moving average models for this research performed relatively well. Although the optimal window size for simple moving average cross-validation consistently was smaller for classification algorithms than regression models, both types of models achieved accuracies above 85%. When selecting one algorithm over others to predict average HR, the results favored Logistic Regression due to its ability to predict heart rate for more days at a time, its high accuracy of 94%, and its relative speed when processing large volumes of data. It was determined that regression models also produced accurate results, but were not as suitable for predicting heart rate for a large time frame of future dates. Furthermore, when clustering virtual and outdoor rides, despite the limited amount of virtual ride data, it was shown that there appears to exist a distinction between virtual and outdoor rides, as all virtual rides were grouped into the same cluster. These results provide a gateway into better understanding the impact of exercise metrics on performance and the intricacies of various types of exercises, such as indoor and outdoor cycling.

In order to fully analyze the capabilities of each of the models used in this research, a next step includes predicting a different variable from the dataset, such as the distance traveled, number of calories burned, or average pace. By doing so, further insights can be gained and applications of this research can be expanded. Although an accuracy of approximately 94% was achieved by LR for predicting average heart rate, the standard deviation of AvgHR was only 12.5, and therefore, had little variation. Consequently, altering the focus to a different exercise metric may serve as an improved gauge of the model performances. Furthermore, given the conclusions from the 2-means clustering of outdoor and virtual rides, it could be possible to perform feature selection given a set of relevant clustering features.

References

[1] Cintia, P., Pappalardo, L., and Pedreschi, D. ”engine matters”: A first large scale data driven study on cyclists’ performance. In *2013 IEEE 13th International Conference on Data Mining Workshops* (Dec 2013), pp. 147–153.

[2] Koehrsen, W. *An Implementation and Explanation of the Random Forest in Python*, 2018 (accessed December 5, 2019).

[3] Li, S. *Building A Logistic Regression in Python, Step by Step*, 2017 (accessed December 5, 2019).

[4] Miller, M. Validity of using functional threshold power and intermittent power to predict cross-country mountain bike race outcome. *Journal of Science and Cycling 3*, 1 (2014).

[5] Xiao, F., Yuchi, M., Ding, M., and Jo, J. A research of heart rate prediction model based on evolutionary neural network. In *2011 International Conference on Intelligent Computation and Bio-Medical Instrumentation* (Dec 2011), pp. 304–307.