***Data Engineering Team***

**RedBack Operations**

***Project Team Members***

* Vinay Ratta (Team Leader)
* Shashvat Joshi

***Brief***

Technology for measuring athletes’ health and Performance is transforming How will they get trained, compete and will manage their careers.

But how to use an athlete’s health data ethically has become a question.

The athletic world has seen the increasingly value of Technology, machine learning, advanced wireless wearable sensors to transform the moves of an athlete, to focus on muscle training to increase their chances of winning and avoid getting hurt.

This explosion of data is questioning on how do we use such personal data ethically. Data is transforming Sports- Hundreds of different metrics can be analysed today through the wearables like rings, smart fabrics.

***Data Set Details***

The data set we are using is from <https://www.garmin.com/en-AU/>

The data set compilers used information partially from one - year historical log corresponding to the years 2015.Our data set is a csv or xlsx file with information from 17379 hours with over 22 features.

The features List

|  |  |
| --- | --- |
| 1. Sport - Biking 2. ns 1 : Id 3. Start Time 4. ns1 : Total 5. ns1:DistanceMeters 6. ns1:MaximumSpeed 7. ns1:Calories 8. ns1:Value 9. ns1:Value2 10. ns1:Intensity 11. ns1:Cadence | 1. ns1:TriggerMethod 2. ns1:Time 3. ns1:LatitudeDegrees 4. ns1:LongitudeDegrees 5. ns1:AltitudeMeters 6. ns1:DistanceMeters3 7. ns1:Value4 8. ns1:Cadence5 9. ns2:Speed 10. ns2:Watts 11. ns2:AvgSpeed |

***Processing Details***

Pre-Processing: Data-cleaning and Feature Engineering

For any ML Project, pre-processing data is a crucial step. The process is usually an accumulation of Widely followed pre-processing practices, and case-specific tweaks made to the data based on judgement-calls from the person designing the model. The data model is done incorrectly can lead to ill-trained ML algorithms that provide inaccurate or sub-optimal results.

An important part of this process is Feature Engineering, which involves turning available data- features into more useful variables that can help us predict our outcome.

Let’s walk through the steps we are taking:

1. Using Prior knowledge to remove features that don't add important information, which in this case is sport, Biking.
2. Clearly, we can see that there is a positive relationship between Cadence, and watts(power).

***Cleaning the Datasets:***

The data set involves predicting the factors which will increase the chances for a Biker to win the race. For an example, it is Cadence, Watt etc.

The data we found from the Garmin Website first converted from .tcx format- Which is Training centre XML format in which the data points are grouped into laps and that each lap element has some useful data associated with it. TCX file contains altitude, distance, heart rate and Speed data.

The data we uploaded on the GitHub is in csv format which is considered as flat structure and is highly convenient because it requires fewer technical skills, and you can access files of this format with most applications. Also, csv requires less processing power as its significantly smaller than other formats.

**Identifying Columns that contain a Single Value**

Columns that have a single observation or value are probably useless for modelling. These columns or predictors are referred to zero-variance predictor as if we measured the variance (average value from the mean), it would be zero.

We can detect rows that have this property using unique () NumPy function that will report the number of Unique values in each column. But we have an issue with using Unique() method as it only works on Series and not on Data Frames. If you call the unique() method on a Data Frame, it will throw an error: Data Frame object has no attribute ‘Unique’.

In this code, then we will use Drop() method.

Consider Columns that have very few Values

In the previous section, we saw that some columns in the example dataset had very few unique values.

For example, there were columns that only had 2, 4, and 9 unique values. This might make sense for ordinal or categorical variables. In this case, the dataset only contains numerical variables. As such, only having 2, 4, or 9 unique numerical values in a column might be surprising.

We can refer to these columns or predictors as near-zero variance predictors, as their variance is not zero, but a very small number close to zero.

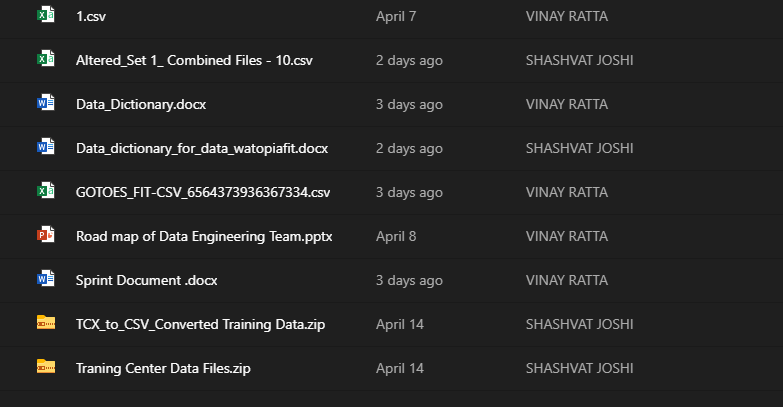
… near-zero variance predictors or have the potential to have near zero variance during the resampling process. These are predictors that have few unique values (such as two values for binary dummy variables) and occur infrequently in the data.

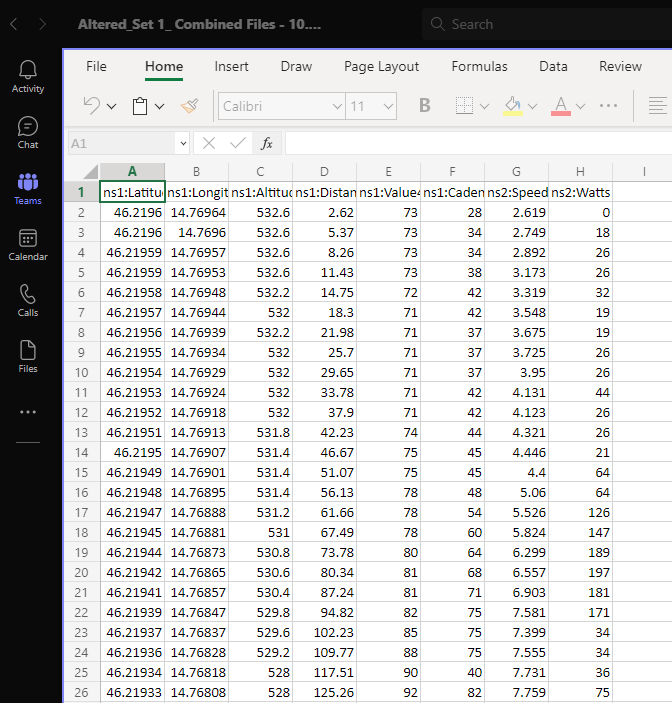
***Limitation of processing***

Due to the file size for Training Centre Dataset being around ~(5 to 6 GB) was divided into smaller sets of 10, 20 , 50 bikers each was easier processing and viewability of the data on our team members system. Data Processing and cleaning were performed on the above-mentioned smaller dataset and the results procured are provided below.

***Evidence Screenshot***

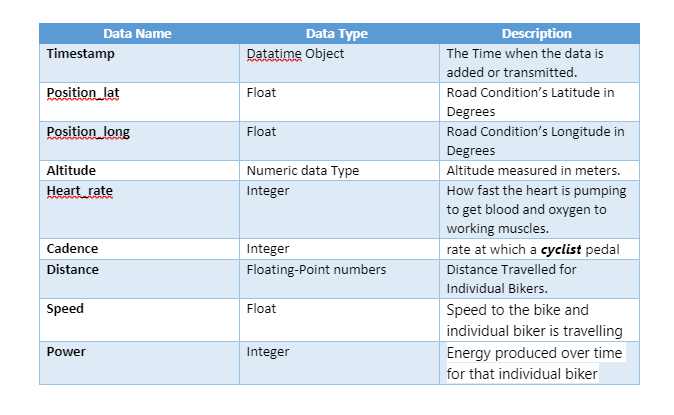
**Training Exercise Dataset**

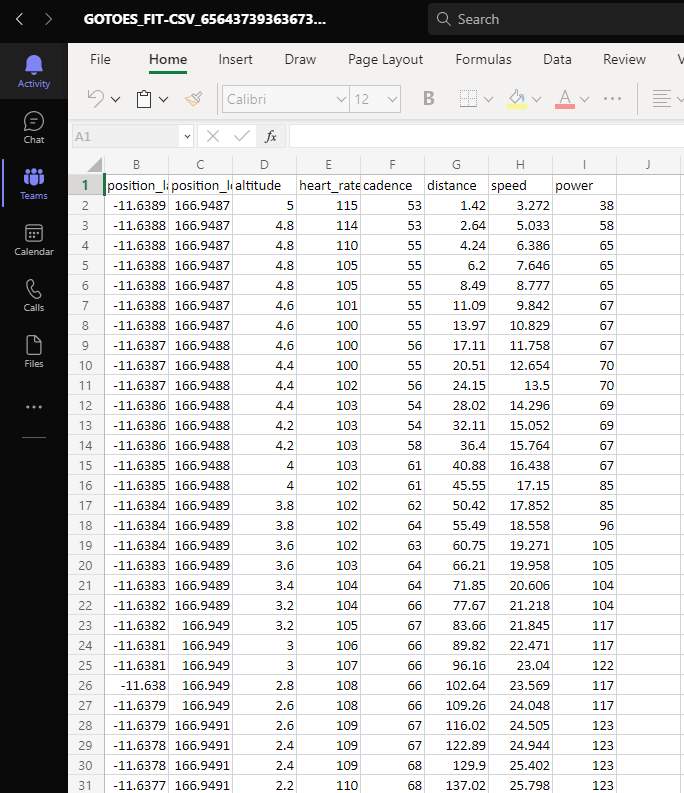




Screenshot of cleaned data set after deleting the attributes having zero-variance.

**Watopia Fit Dataset**





***Future Plans***

Next, we will be working on looking for the solution to be able to merge Data for more than 50 Bikers. For this We are looking into using PyTorch, which is an open-source Machine Learning framework to speed up the process and combine more data sets and attempting to get the balanced sample.

Apart from this dataset, we have been also working on the dataset given by our directors watopia.fit. We were able to successfully convert. fit format into .csv using online tool [Convert FIT Files to CSV (gotoes.org)](https://gotoes.org/gotoes/strava/convert_fit_files_to_csv.php)

The data was already in a very balanced format. We have already removed the attributes with Zero-variance.

***References***

|  |  |
| --- | --- |
| [1] | PyTorch*, “PyTorch Library Details”*  **[Website] -** <https://pytorch.org/> |
| [2] | Garmin Training Centre*, “Training Dataset”*  **[Website] -** <https://www.garmin.com/en-AU/>  **[Website] -** https://garmin-training-center.software.informer.com/download/#downloading |
| [3] | Browinee, Jason, (2019, February, 10), *“Training First Machine Learning Project in Python Step-By-Step”*  **[Website] -** <https://machinelearningmastery.com/machine-learning-in-python-step-by-step/> |
| [4] | Gotoes – Utilities for Strava*, “Convert a FIT file to a CSV file.”*  **[Website] -** <https://gotoes.org/gotoes/strava/convert_fit_files_to_csv.php> |