**Weekly Report**

**Week 1**

**Date:** 11/09/2020 – 11/13/2020.

**Total time contributed**: 12 hours

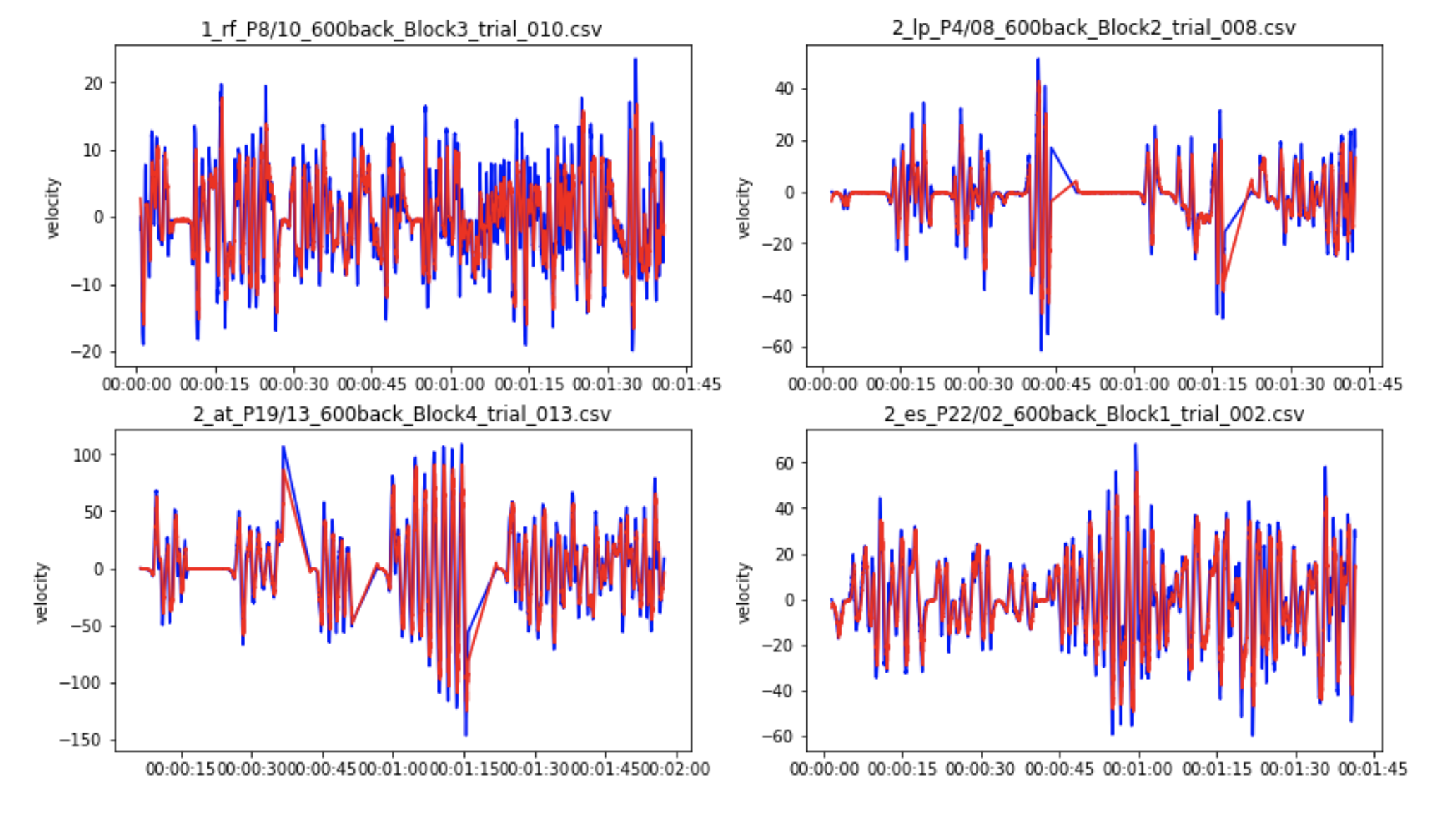
**Total time allowed**: 10 hours

In this week, I mainly did data cleaning and data manipulation work to understand the data.

**Data manipulation**

The data includes 34 people two day’s experiments. There are 1360 csv data sets in total. I tried several methods to load all the data using python. First I tried used two simple for loops to read the data, however, it took me so much time to run it. Then I used a list and concatenate all the csv data sets into one big data set. I added a new column called ‘peopleTrialKey’ which could identify the data from which people in which trial and in which day. Also I included a datetime format column called ‘datetimeNew’, which could be used to do date-related manipulation in the future work. Another new column I added is called ‘calculated\_vel’, which is the new velocity and derived from current velocity. The formula is ‘calculated\_vel’ = (Position\_t - Position\_t-1)/ (Seconds\_t - Seconds\_t-1). The reason I did this is because there’s a delay of the machine while it records the velocity. The delay is around 30-40ms. In the end I wrote the new big dataset into a csv file called ‘data\_all.csv’.

Since we have a calculated velocity, we need to validate it. So I did some visualization on both velocities. I randomly selected some ‘peopleTrialKey’. Below are the graphs. The bule line is original machine velocity, while the red one is the calculated velocity.



Both machine velocity and calculated velocity are quite similar over all, even though there might be some cases in machine velocity which have big gap compare to calculated velocity.

**Data exploration**

Since there’s a column called ‘trialPhase’, which represents the status of the machine. ‘1’ means under the machine control, usually means the machine is resetting after a crash. We don’t know those data, so I **excluded** all of them when I do the future work. ‘3’ means under human control. ‘4’ means there’s a crash happening.

Since there’s another criteria which is if the position is beyond 60 degree or - 60 degree, we say there’s a crash. The first criteria I mentioned above when the ‘trialPhase’ = ‘4’, there’s a crash. However, when I used the 60 degree criteria, I found there’s a difference from the ‘trialPhase’ criteria. The conclusion is that: there are 4300 crash events based on +- 60 degree criteria; and there are 6251 crash events based on ‘trialPhase’. We can see there’s a huge difference. The reason is because there’s tiny bias of the machine when recording the position and the degree of that position. In the end, we decided to use the **‘trialPhase’ criteria.**

There are also some cases that two consecutive crashed happened during a very short time. Since we will take the previous time point data as our features, we need to know how is the distributions of crash events in different time scales. Below is a graph showing the distribution:

The x-axis is the time scale between two consecutive crashes, while the y-axis is the total number of crash events. We can see the majority of crash events have enough time scale (2 seconds).



In the following work, I excluded all the crash events that have less than 2 seconds time gap. There are 12 crash events are removed.

**Next Step:** continue do data manipulation and start to do feature engineering.

**Week 2**

**Date:** 11/16/2020 – 11/20/2020.

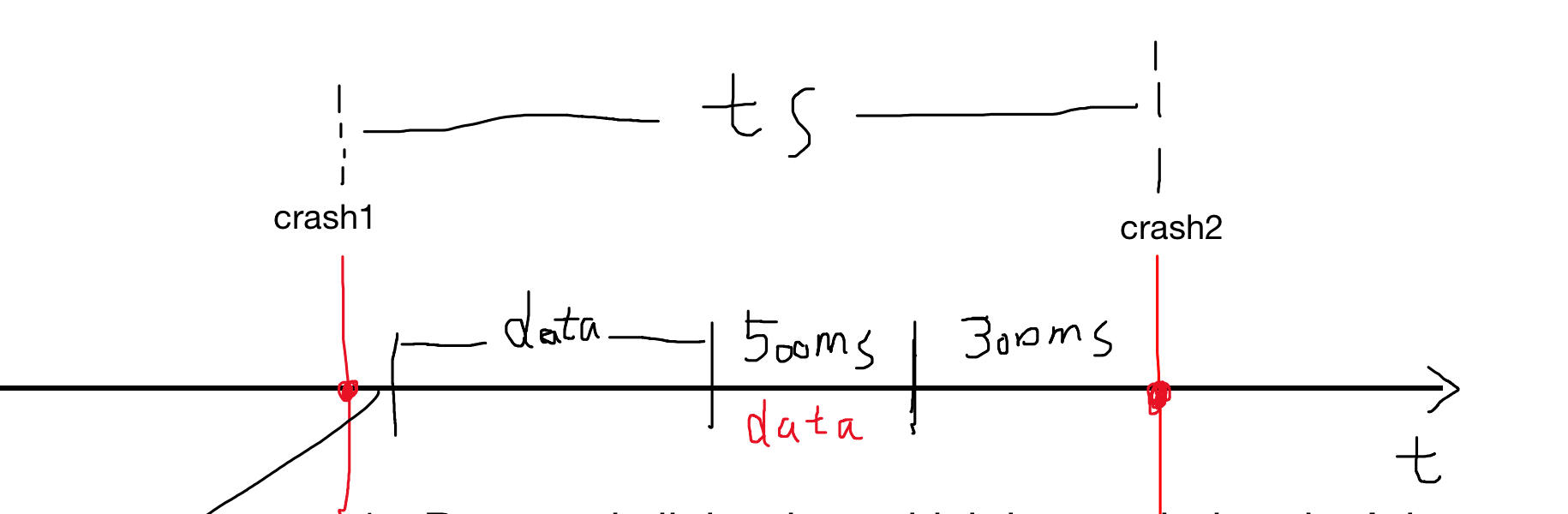
**Total time contributed**: 8 hours

**Total time allowed**: 10 hours

This week, I continued doing data manipulation and also start to do feature engineering.

**Data manipulation & feature engineering**

After last week’s work, I have already got somehow cleaned data. Then I tried to create the data for training which includes the features and labels. When label equals to 1, it represents there’s a crash, when label equals to 0, it represents non-crash events. In terms of the features part, below is a graph I drew.



We first decided to predict 300 ms in advance if there’s a crash happen or not. So we leave the data 300 ms close to a crash event out. Then we take time scale = 500 ms as our feature for that crash event. For non-crash events, I took all the data between two crash events except for the 300 ms close to a crash and 500 ms that used as crash event feature (**positive**). In the graph, there’s a time line called ‘data’, it represents the data we will used for non-crash events (**negative**). However, for non-crash event, we decided to use a sliding window to get the features, with 500 ms time scale as features and 300 ms as jumping step. In this way, our number of crash events (**positive**) only have 6251cases, while our number of non-crash events (**negative**) is way more than our positive cases. So our data is very **imbalanced**. It’s not good for our model. And we need to utilized techniques to deal with this issue.

After we get all the features and labels, I found that the sampling rate is quite different for both crash and non-crash events. I drew a sampling rate distribution of crash event in below graph.



We can see that the sampling rate is different, which could be an **issue** for our future LSTM model. So we need to make all samples’ sampling rate the same, so that we could feed into our model.

**Next step:** solve the imbalanced data issue and also the inconsistent sampling rate.

**Week 3**

**Date:** 11/23/2020 – 11/27/2020.

**Total time contributed**: 8 hours

**Total time allowed**: 10 hours

In this week, I solved the two issues I mentioned last week. One is the imbalanced data, one is the inconsistent sampling rate.

**Imbalanced data**

I first used a under sampling method to solve the imbalanced data. To be more specific, since the number of crash events is around 6000, I randomly chose 6000 non-crash events among all the negative cases. In this way, we could have a balanced data set. However, the drawback of this method is we will miss lots of information on non-crash event side. So it’s not a good idea.

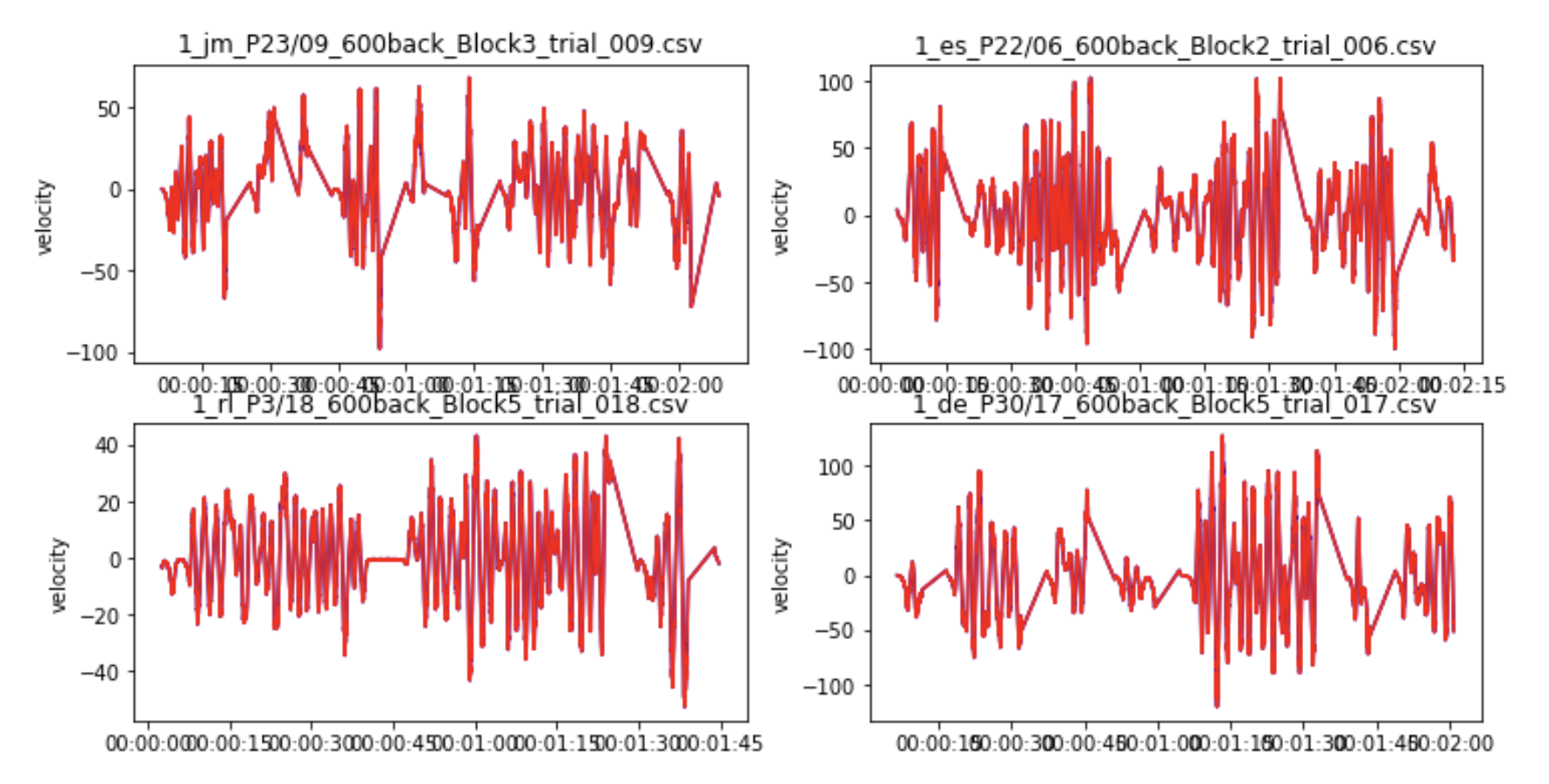
Another method which is corresponding to under sampling method is called over sampling. This method is to copy all the crash events till the total number of it reaches the same number of non-crash events. This method’s defects are there are so many copies.

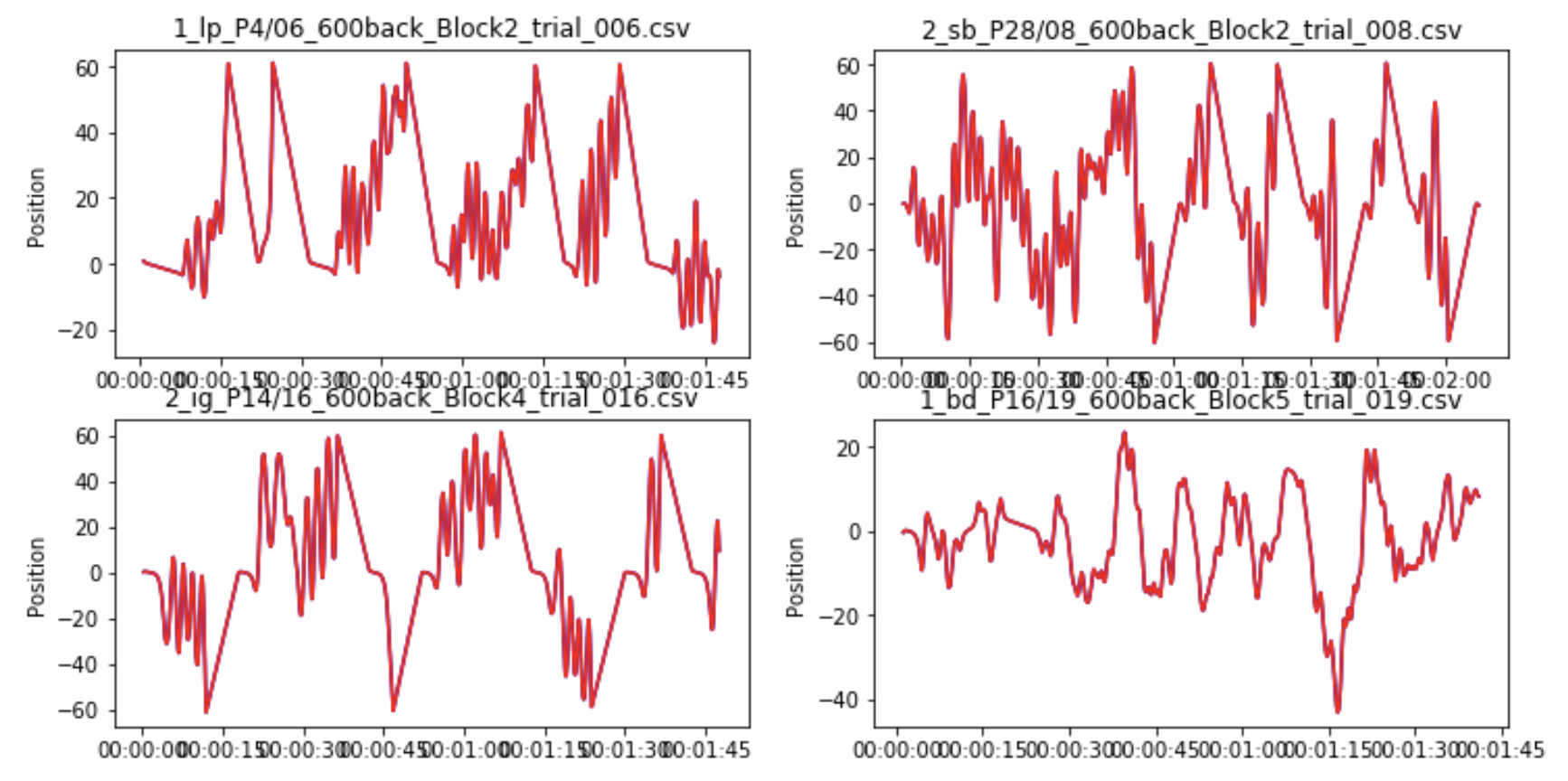
The last method I tried is based on our model. We can assign different weights to different classes. Usually we would assign higher weights to positive cases, since we value it much more. I decided to try 1:1, 1:10, 1:50, 1:100 ratios.

**Inconsistent sampling rate**

To solve this problem, we usually use resampling and interpolation methods.

**Pandas resampling:** there’s a function in pandas used for time series data called resample and interpolate. Using this method, we could have a very good fit before and after resampling. Below are some visualizations included resampled velocity and positions. I randomly select 4 different ‘peopleTrialKey’. Blue stands for original data, and red stands for resampled data.





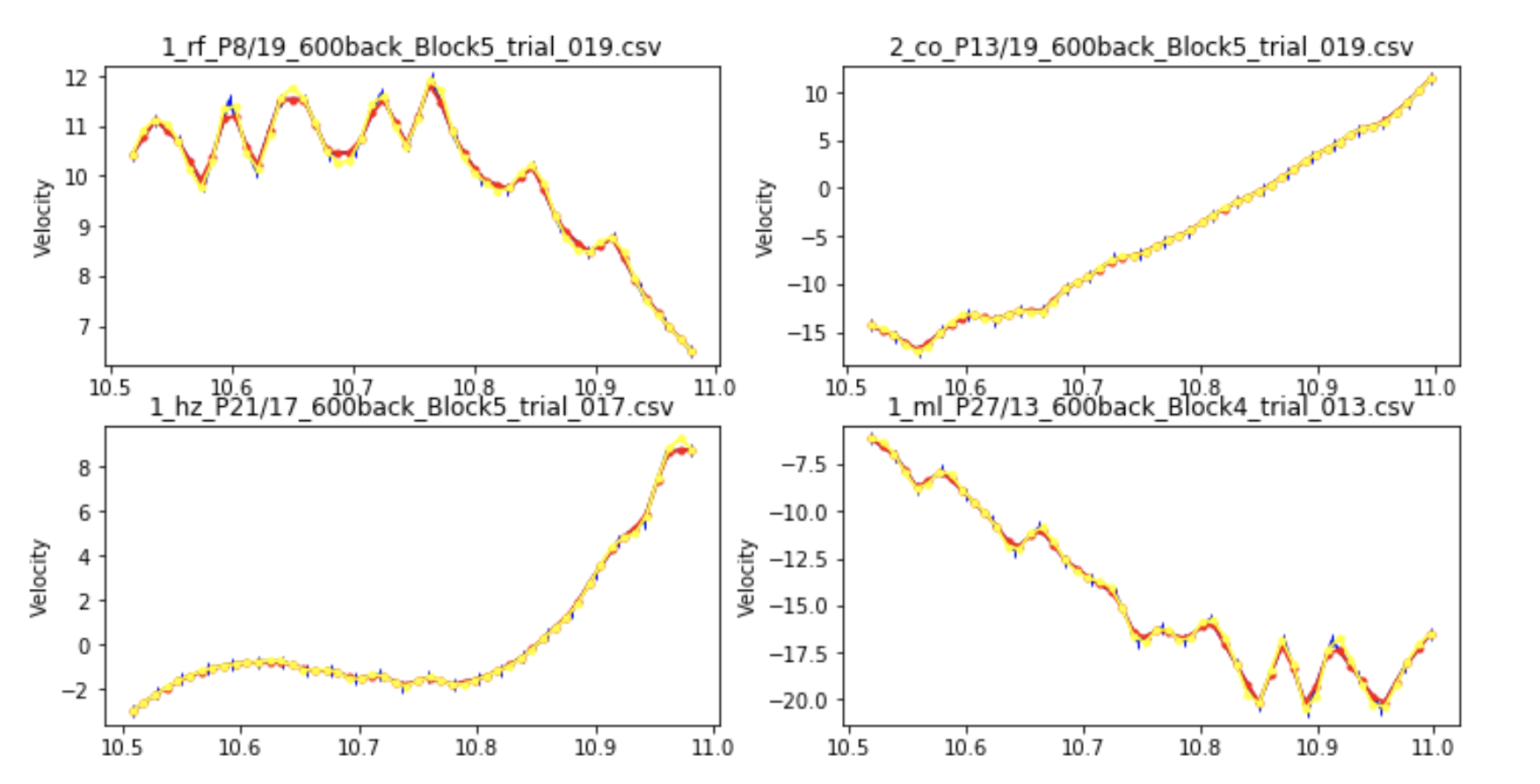
Even though this method can have a very good matched between before and after data, the drawback of this is it still cannot solve the different sampling rate issue.

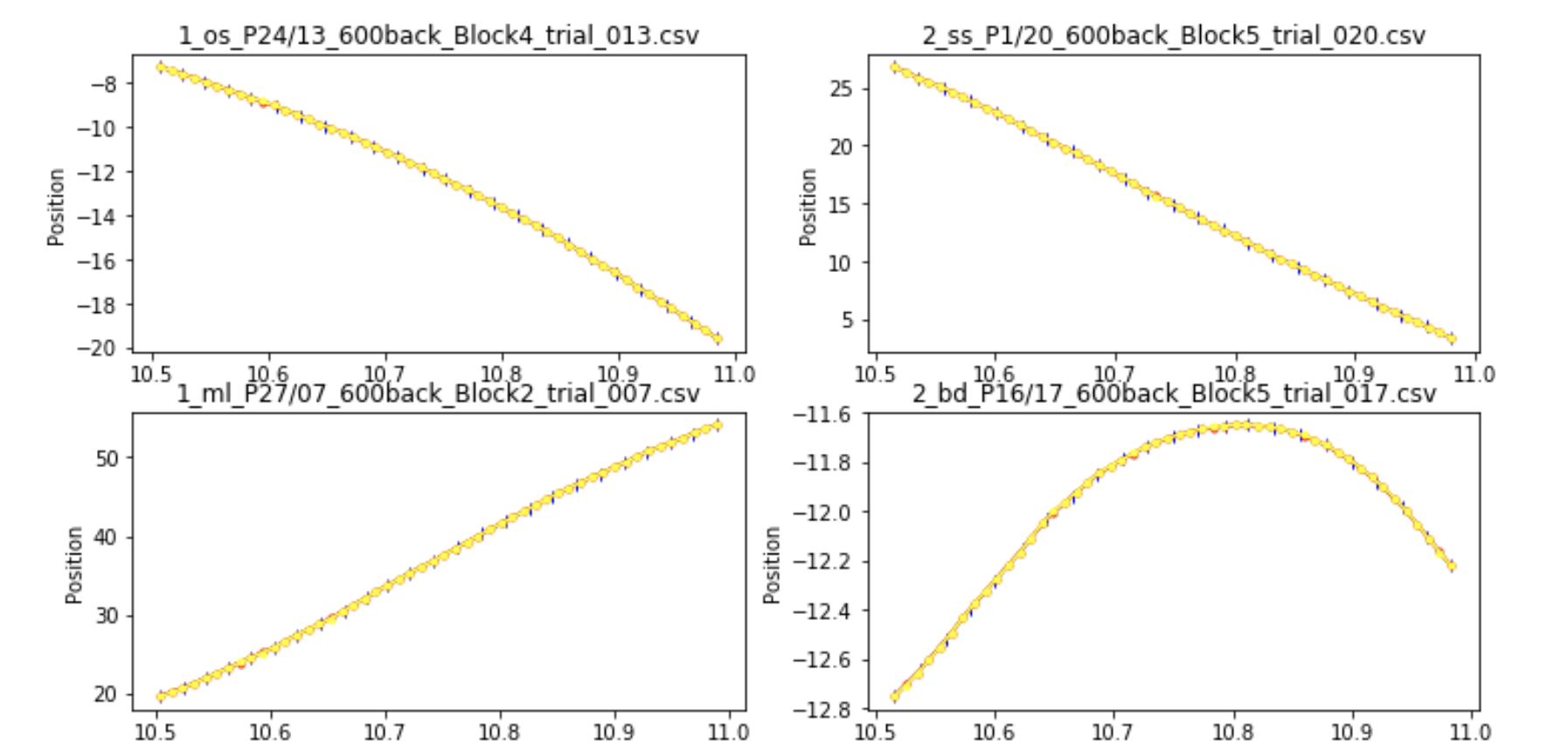
Because:

e.x.1: from 1.003 seconds to 1.500 seconds, there would be 498 sampling rate;

e.x.2: from 1.005 seconds to 1.500 seconds, there would be 496 sampling rate.

**Numpy & scipy resampling:** this methods can sovle the different sampling rate issue since numpy.linspace could help us have a fixed length. **Here we choose sampling rate of 50**. Another thing to be noticed is that we could use different method while do resampling. For example, we could use ‘linear’, or ‘quadratic’. To identify which method is better, I drew both. Below are graphs of before and after resampling. Blue is the original data, red stands for ‘linear’ resampling method, and yellow line stands for ‘quadratic’ resampling method.





This method works very well too! And one benefit of this is that we could make the sampling rate fixed. So I decided to use this method while do feature engineering. Also I tried both **linear** and **quadratic** interpolate methods, but there's **no big difference**, even though the quadratic is a little bit smoother. In the end, I choose the linear method to do interpolation.

**Week 4**

**Date:** 11/30/2020 – 12/04/2020.

**Total time contributed**: 2 hours

**Total time allowed**: 10 hours

This week, since it’s my final week, so I didn’t work too much. However, I spent some time building up the model and ran some experiments.

The model I used is the long and short term memory (LSTM), which is designed for time series analysis.

**Next step:** run the experiments and analyze the results.

**Week 5**

**Date:** 12/07/2020 – 12/11/2020.

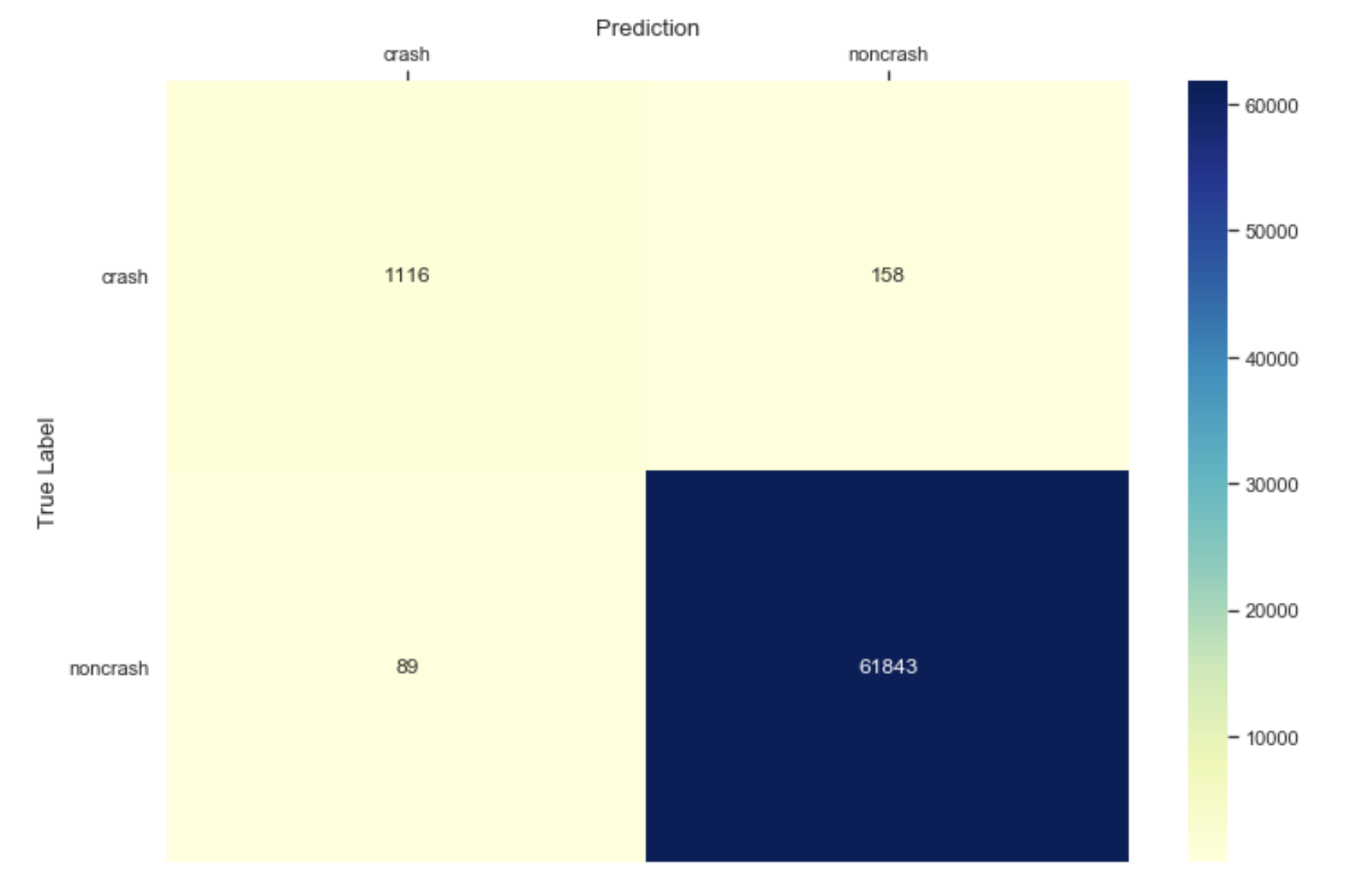
**Total time contributed**: 12 hours

**Total time allowed**: 20 hours

This week, I spent time to build up the model and run lots of experiment.

I tried different weights while running the experiments. In terms of the model, I first used a LSTM layer with 128 units, followed by a dropout layer with rate = 0.5. Then there are two dense layers. One is with 128 units and activation function is ‘relu’. One is with output shape = 1 and a ‘softmax’ activation. The loss I used is 'categorical\_crossentropy' and optimizer is 'adam'.

After I got the results, I drew the confusion matrix for our results. Later on, I calculated accuracy, precision and recall.



Below are the summarized results.

|  |  |  |  |
| --- | --- | --- | --- |
| **Class\_weights** | **Accuracy** | **Precision** | **Recall** |
| 1:1 | 99.61% | 92.61% | 87.60% |
| 1:10 | 99.28% | 92.61% | 87.60% |

**Next step:** try both calculated velocity and original velocity in the models to see if there’s difference among results. Also, assign more class\_weights to do comparison.

**Week 6**

**Date:** 12/14/2020 – 12/18/2020.

**Total time contributed**: 10 hours

**Total time allowed**: 20 hours

This week, I removed all the machine control data and ran the experiments again with weights of {1:1, 1:10, 1:50, 1:100}. Also, I wrote reports for the previous weeks.

The results are showing below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Velocity type** | **Class\_weights** | **Accuracy** | **Precision** | **Recall** |
| Calculated | 1:1 | 99.59% | 90.65% | 95.53% |
| 1:10 | 98.80% | 70.44% | 99.83% |
| 1:50 | 98.67% | 68.37% | 99.83% |
| 1:100 | 96.72% | 46.62% | 100% |
| Original | 1:1 | 99.66% | 92.08% | 96.27% |
| 1:10 | 99.66% | 90.49% | 98.43% |
| 1:50 | 98.45% | 64.83% | 99.83% |
| 1:100 | 97.35% | 51.87% | 99.92% |

From the results we can see that, both calculated and original velocity feature performed well when class weight is 1 to 1. However, when the class weight is 1 to 10, the precision of calculated velocity performed much worse than original one. Another thing I noticed is when the class weights increase, the performance is becoming worse, except for the 1 to 10 weight with original velocity.

Here comes to our conclusions:

1. The class weight 1 to 1 is a good choice for both feature types.

2. Increase the weights is not a good idea except for the original features till 1 to 10 weight.

Issues: I am having difficulty using GPU on HPCC to run the experiments.

Next step: increase the time scale.