Joshua Mannion and Lawton Manning
CSC 673 Data Mining
November 13, 2020

**Data Mining Assignment 5**

---

The objective of this assignment is to calculate the importance of data using Shapley values and leave-one-out (LOO) and to explore the differences between high value points and low values points [3]. We have modified the existing code available on GitHub to complete the different experiments.

# 1 Data

The "brunello" (human) genomics data set used in this assignment was provided by the professor on the class canvas site in comma separated values format [1]. The four nucleotide frequencies and sixteen dinucleotide frequencies were added as features for every sample. Then the large data set was sampled to create a test data set of 5000 total sequences, 2500 in class 0 and 2500 in class 1 and a train data set of 100 total sequences, 50 in class 0 and 50 in class 1. The attributes included in these data sets are the sequence, single nucleotide frequencies, dinucleotide frequencies, and the class label. (table 1).

Table 1: Description data sets

| **Name** | **# Attributes** | **# Observations** | **# Class 0** | **# Class 1** |
|---|---|---|---|---|
| brunello_train | 22 | 100 | 50 | 50 |
| brunello_test | 22 | 5000 | 2500 | 2500 |
| cycling_train | 9 | 1937 | 1112 | 825 |
| cycling_test | 9 | 25 | 10 | 15 |

The second data set used in this assignment is the cycling data set collected and analyzed by [2]. Given its small size, no samples were removed from the data set, and only correlated features were removed for classification purposes.

# 2 Experiment 1

The code on GitHub was modified to calculate Shapley and LOO values for two different data sets, several algorithms, and two metrics (Tables 2, 3). The data sets used are

described in Table 1. For the brunello data, logistic and gradient boost classifier algorithms were used to calculate along with both metrics, accuracy and AUC. For the cycling data, naive bayes and linear SVC were used along with AUC and accuracy, with the exception of using AUC on SVC. We were unable to calculate AUC scores when using the linear SVC algorithm due to a library API issue in scikit-learn. We calculated the Shapley and LOO values for all the 100 samples in brunello_train. The samples used to evaluate those 100 values were the first 4000 from brunello_test. The remaining 1000 data points from brunello_test data set (about 20%) were held out for evaluating model performance in experiments 2, 3, and 5.

For the cycling data set, a subset of the cycling_train data set consisting of 193 points (about 10%) of the data set was used to calculate Shapely values. Of the training data, 1162 points (about 60%) were used in evaluating models created from the 193 Shapely value points. The last 30% was held out for evaluating Shapely value removal in fig. 4, fig. 6, and table 5.

Table 2: Shapley Values

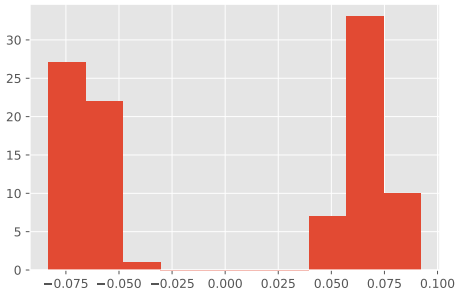
| **Data** | **Algorithm** | **Metric** | **Min** | **Max** | **Mean** | **Variance** |
|---|---|---|---|---|---|---|
| brunello | Logistic | Accuracy | -4.20e-03 | 5.90e-03 | 8.10e-04 | 4.39e-06 |
| brunello | Logistic | AUC | -1.08e-01 | 9.74e-02 | -5.00e-04 | 5.20e-03 |
| brunello | Gradient Boost | Accuracy | -8.20e-03 | 6.80e-03 | 1.10e-03 | 5.34e-06 |
| brunello | Gradient Boost | AUC | -8.09e-02 | 5.04e-02 | -5.00e-04 | 8.00e-04 |
| cycling | Naive Bayes | Accuracy | -3.96e-03 | 3.47e-03 | 5.87e-04 | 1.80e-06 |
| cycling | Naive Bayes | AUC | -1.57e-02 | 1.32e-02 | -2.41e-04 | 2.93e-05 |
| cycling | Linear SVC | Accuracy | -1.41e-02 | 5.33e-03 | 6.39e-04 | 9.32e-06 |

Table 3: LOO Values

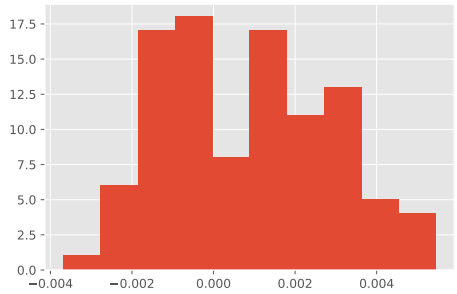
| **Data** | **Algorithm** | **Metric** | **Min** | **Max** | **Mean** | **Variance** |
|---|---|---|---|---|---|---|
| brunello | Logistic | Accuracy | -6.70e-03 | 4.10e-03 | -1.30e-03 | 7.12e-06 |
| brunello | Logistic | AUC | -4.68e-02 | 4.45e-02 | -2.00e-04 | 1.20e-03 |
| brunello | Gradient Boost | Accuracy | -1.50e-02 | 1.80e-02 | 6.00e-04 | 4.19e-05 |
| brunello | Gradient Boost | AUC | -7.38e-02 | 2.28e-02 | -1.71e-02 | 2.00e-04 |
| cycling | Naive Bayes | Accuracy | -2.58e-03 | 1.72e-03 | -1.69e-04 | 6.19e-07 |
| cycling | Naive Bayes | AUC | -5.64e-03 | 3.08e-03 | -7.81e-04 | 2.88e-06 |
| cycling | Linear SVC | Accuracy | -6.02e-02 | 3.17e-01 | 3.58e-02 | 6.20e-03 |

Then using the calculated Shapley values, histograms were created for each combination of runs ( fig. 1,  fig. 2). The AUC metric clearly separates the brunello samples into low value points and high value points. The accuracy metric and logistic regression combination appear to create two classes of samples without as much separability when using the AUC metric (fig. 1 (b)). Whereas the combination of the gradient boost classifier and accuracy metric produce a population of values that look normally distributed with a little shift to the right ( fig. 1 (d)).

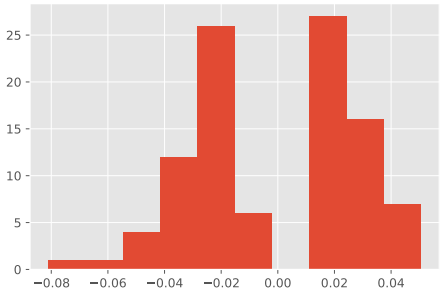
For the cycling data set in fig. 2, the AUC and accuracy metrics did very similarly in distribution for the Naive Bayes classifier. For Linear SVC in fig. 2 (c), the values are much less varied and skewed to be right above 0.



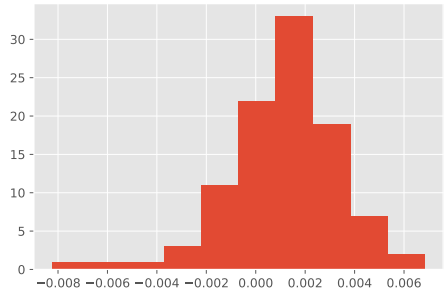

(a) Logistic - AUC




(b) Logistic - Accuracy

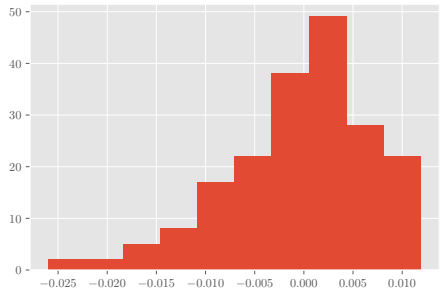



(c) Gradient Boost - AUC

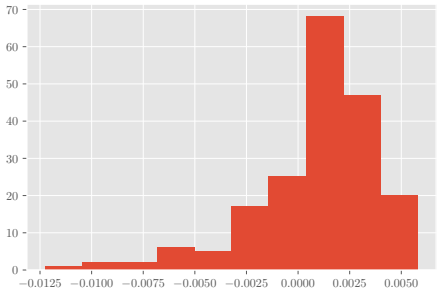



(d) Gradient Boost - Accuracy

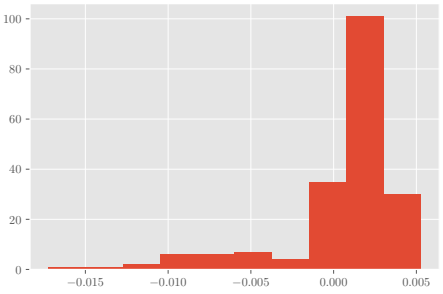
Figure 1: Histograms of TMC estimated Shapley values for the brunello data set. The number of samples is on the y-axis and the Shapley values are on the x-axis. (a) shows the distribution of Shapley values when using the logistic regression algorithm and AUC metric. (b) shows the distribution of Shapley values when using the logistic regression algorithm and accuracy metric.(c) shows the distribution of Shapley values when using the gradient boosting classifier algorithm and AUC metric. (d) shows the distribution of Shapley values when using the gradient boosting classifier algorithm and accuracy metric.




(a) Naive Bayes - AUC




(b) Naive Bayes - Accuracy

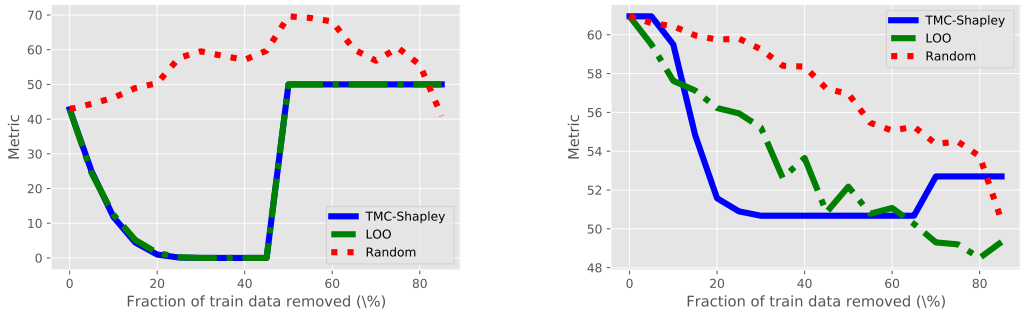



(c) Linear SVC - Accuracy

Figure 2: Histograms of TMC estimated Shapley values for cycling data set. The number of samples is on the y-axis and the Shapley values are on the x-axis. (a) shows the distribution of Shapley values when using the naive bayes algorithm and AUC metric. (b) shows the distribution of Shapley values when using the naive bayes algorithm and accuracy metric.(c) shows the distribution of Shapley values when using the linear SVC algorithm and accuracy metric.

## 3 Experiment 2

The objective of experiment 2 is to evaluate the removal of highly valuable data points. To do this, we adapted the code and progressively removed more and more valuable data points by percentage and tested the performance of the the various combinations of algorithms and metrics.



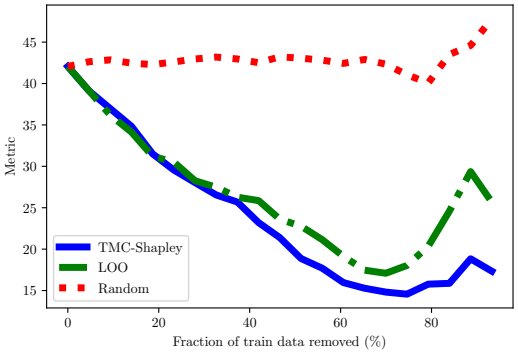

(a) Logistic - AUC

(b) Logistic - Accuracy

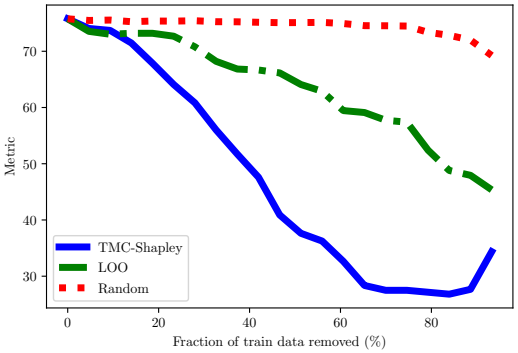
(c) Gradient Boost - AUC

(d) Gradient Boost - Accuracy

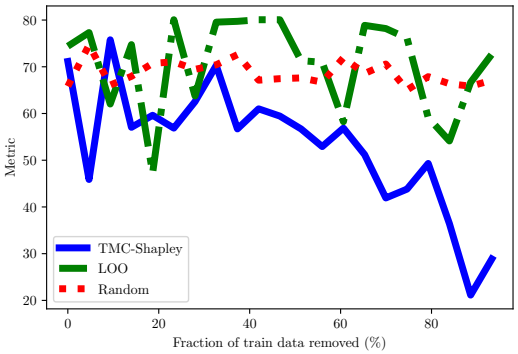
Figure 3: Removing high value points from the brunello data. The value of the metric, accuracy or AUC, in (%) is on the y-axis. The percent of data removed is on the x-axis (%). Red is random, blue is using TMC estimated Shapley values, and green is using LOO values. (a) is the experiment using the logistic regression algorithm and AUC metric.(b) is the experiment using the logistic regression algorithm and accuracy metric. (c) is the experiment using the gradient boost classifier algorithm and AUC metric. (d) is the experiment using the gradient boost classifier and accuracy metric.




(a) Naive Bayes - AUC




(b) Naive Bayes - Accuracy




(c) Linear SVC - Accuracy

Figure 4: Removing high value points from the cycling data. The value of the metric, accuracy or AUC, in (%) is on the y-axis. The percent of data removed is on the x-axis (%). Red is random, blue is using TMC estimated Shapley values, and green is using LOO values. (a) is the experiment using the naive bayes algorithm and AUC metric.(b) is the experiment using the naive bayes algorithm and accuracy metric. (c) is the experiment using the linear SVC algorithm and accuracy metric.

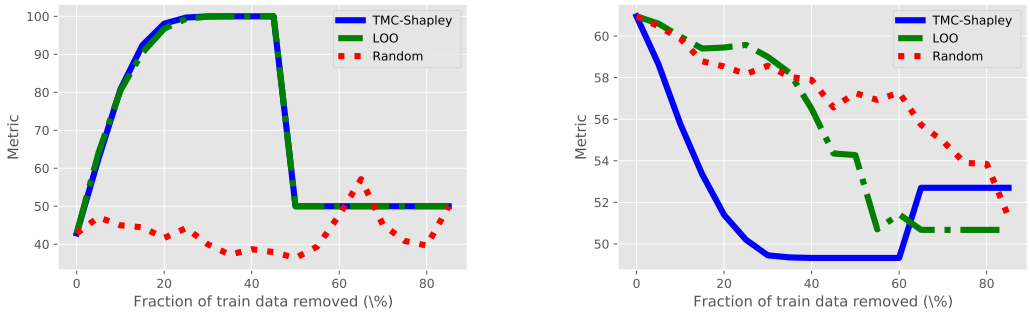
Figure 3 (d) shows the best performance of TMC Shapely as the curve drops steeply and is the most steady of all the combinations. Figure 3 (a) and (c) show similar spikes around 50% of data removed. This jump indicates that models no longer have class separability which makes sense since at this point most of the high value data points have been removed and so the models are no longer well performing in separating classes.

Figure 4 (b) also shows the best performance of TMC Shapely compared to LOOV. The accuracy jumps down quickly and the last small jump after removing at least 80% of the data points can be attributed to the small number of data points left to train the classifier.

For Linear SVC in fig. 4 (c), the curve is more sporadic for both Shapely and LOOV. This indicates that the Shapely values are not very accurate predictors in this experiment.

# 4 Experiment 3

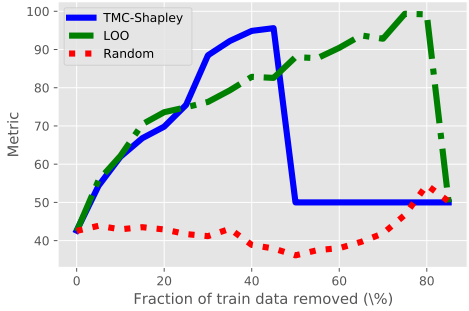
The objective of experiment 3 is to evaluate the removal of least valuable data points. To do this, we adapted the code and progressively removed more and low value data points by percentage and tested the performance of the the various combinations of algorithms and metrics.




(a) Logistic - AUC




(b) Logistic - Accuracy




(c) Gradient Boost - AUC

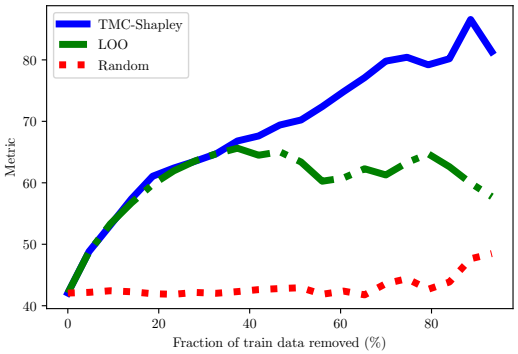

(d) Gradient Boost - Accuracy

Figure 5: Removing low value points from the brunello data. The value of the metric, accuracy or AUC, in (%) is on the y-axis. The percent of data removed is on the x-axis (%). Red is random, blue is using TMC estimated Shapley values, and green is using LOO values. (a) is the experiment using the logistic regression algorithm and AUC metric.(b) is the experiment using the logistic regression algorithm and accuracy metric. (c) is the experiment using the gradient boost classifier algorithm and AUC metric. (d) is the experiment using the gradient boost classifier and accuracy metric.

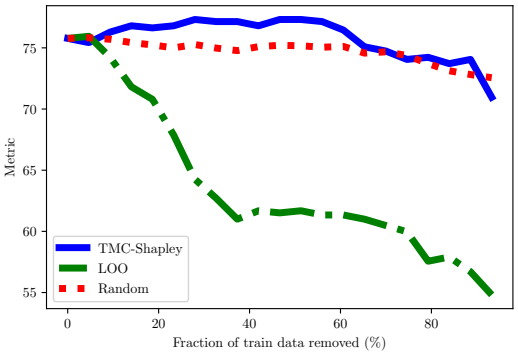
Figure 5 (d) shows the best performance of TMC Shapely as the curve drops slowly and is the most steady of all3the combinations. Figure 5 (a) and (c) show similar drops from nearly 100% to around 50% of data removed. This drop indicates that models no longer have class separability; however, this doesn't make much sense since at this point most of the low value data points have been removed and so the models are should continue to perform better like LOO in Figure 5 (c).

Figure 6 (b) shows the most steady performance as low value data points are removed. In this experiment, flat or slightly positive curves are very desirable. In all combinations of

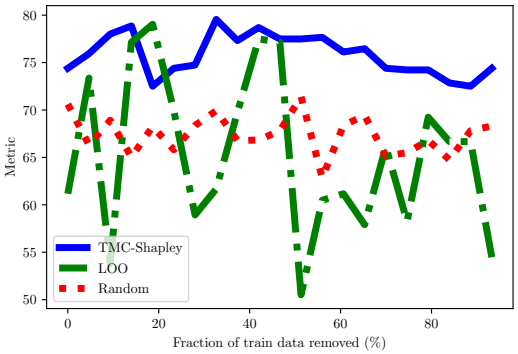
metrics and classifiers on the cycling data set, the Shapely values seem to perform well in predicting which values should be removed to gain accuracy. As discussed in the previous section, this is not true for removing high value data points.




(a) Naive Bayes - AUC




(b) Naive Bayes - Accuracy




(c) Linear SVC - Accuracy

Figure 6: Removing low value points from the cycling data. The value of the metric, accuracy or AUC, in (%) is on the y-axis. The percent of data removed is on the x-axis (%). Red is random, blue is using TMC estimated Shapley values, and green is using LOO values. (a) is the experiment using the naive bayes algorithm and AUC metric.(b) is the experiment using the naive bayes algorithm and accuracy metric. (c) is the experiment using the linear SVC algorithm and accuracy metric.

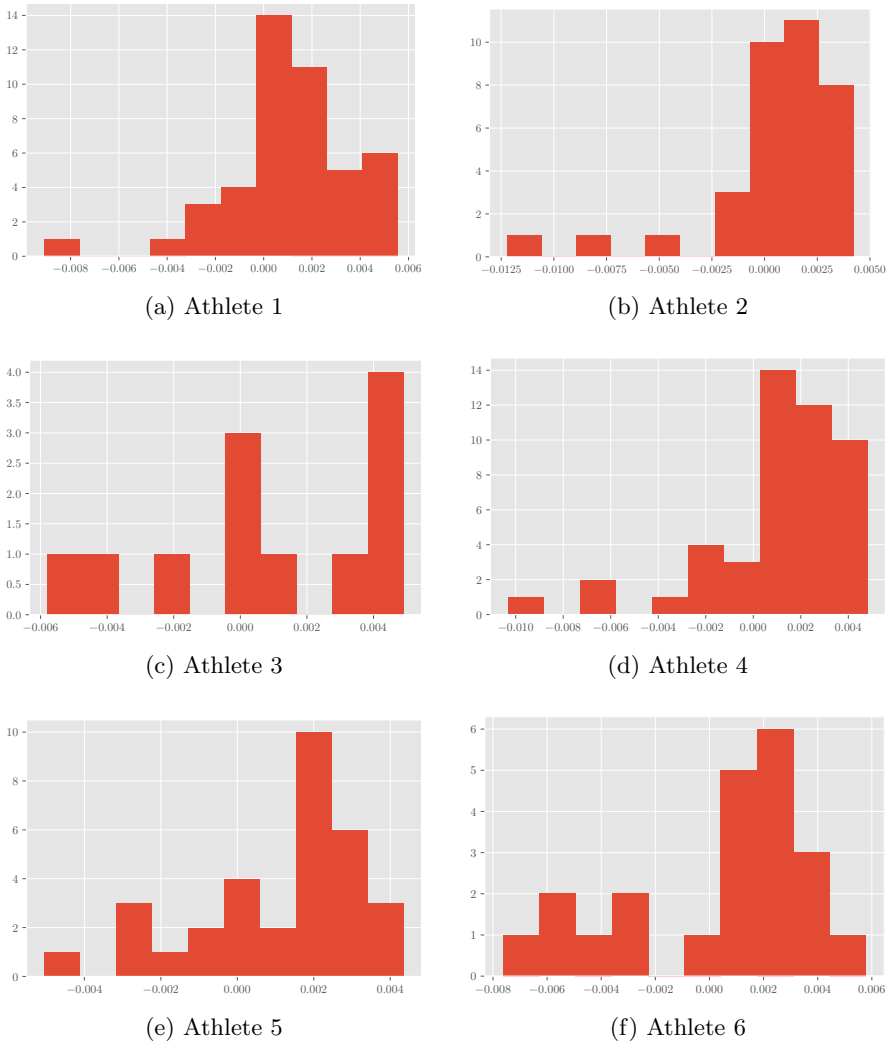
## 5 Experiment 4

In this experiment, the 193 Shapely values were grouped by athlete. The Shapley values came from the Naive Bayes classifier and accuracy metric. Table 4 shows the summary

statistics on these groupings. There is some variance between the athletes but the average shapely value for each athlete is a positive value.

| **Athlete** | **Count** | **Min** | **Max** | **Mean** | **Variance** |
|---|---|---|---|---|---|
| 1 | 45 | -9.11e-03 | 5.56e-03 | 1.07e-03 | 2.68e-03 |
| 2 | 35 | -1.22e-02 | 4.23e-03 | 5.20e-04 | 3.31e-03 |
| 3 | 12 | -5.79e-03 | 4.91e-03 | 8.38e-04 | 3.42e-03 |
| 4 | 47 | -1.03e-02 | 4.81e-03 | 1.09e-03 | 3.05e-03 |
| 5 | 33 | -5.04e-03 | 4.35e-03 | 1.13e-03 | 2.20e-03 |
| 6 | 22 | -7.63e-03 | 5.79e-03 | 3.12e-04 | 3.56e-03 |

Table 4: Shapley value statistics for the Naive Bayes classifier and accuracy metric for the cycling data set. These values are grouped by athlete id in the cycling data set. By viewing the mean Shapley value for each athlete, we can assess the importance of that athlete in the classifier. The most important athletes by Shapley value should be 1, 4, and 5. The least important should be athlete 6.



(a) Athlete 1

(b) Athlete 2

(c) Athlete 3

(d) Athlete 4

(e) Athlete 5

(f) Athlete 6

Figure 7: Histogram Shapley values grouped by athlete. The Shapley values are the same ones from table 2. The most valuable athletes of 1, 4, and 5 have skewed histograms towards positive values. The least valuable athlete 6 has more negative Shapley values although this effect is only slight.

| Athlete | Withholding Accuracy (%) | Change from Normal (%) |
|---|---|---|
| 1 | 76.12 | 0.17 |
| 2 | 72.16 | -3.78 |
| 3 | 76.12 | 0.17 |
| 4 | 75.09 | -0.86 |
| 5 | 75.60 | -0.34 |
| 6 | 75.77 | -0.17 |

Table 5: Witholding accuracies for each athlete. The witholding accuracy was computed by training on all of the 193 Shapely value data points with the exception of those for the withheld athlete. The resulting model was scored on the holdout data. The change from normal is the percent change from the accuracy computed by training on all the Shapely value data points.

Table 5 shows that the highest valued athlete should be athlete 2. However, table 4 shows that the highest average valued athlete in Shapely value is a close tie between athletes 1, 4 and 5. In that same table, athlete 2 is actually one of the least valued athletes. From section 5, it is true that athlete 2 has relatively few negatively value data points. However, it is surprising that the drops in accuracy do not seem to be well aligned to the average value for each athlete. This method of removing highly valued groups may only be worth considering when the average value is far different than others. The average values computed in table 4 were quite similar and all had positive means and so may not be uneven enough for removing one of them to make a difference.

# 6 Experiment 5

The objective of experiment 5 is to acquire a new data set. To do this, a random forest regression model was fit using the 100 TMC estimated Shapley values calculated in experiment 1 from the gradient boosted classifier and accuracy metric. This particular combination of algorithm and metric was chosen because the Shapley values produced from this combination had the best performance in experiments 2 and 3 for the brunello data. Then using the newly fit random forest regression model, we predicted on our "new" data set of 4000 samples, the same 4000 in experiments 2 and 3 used to evaluate the points. The predicted Shapley values ranged from -3.49e-04 to 3.92e-03 with a mean of 8.18e-04 and variance of 7.81e-07 (Figure 8). The new data set is saved and provided in the zip file as "new_data.csv". It has 4000 samples and 23 attributes including the predicted Shapley values.

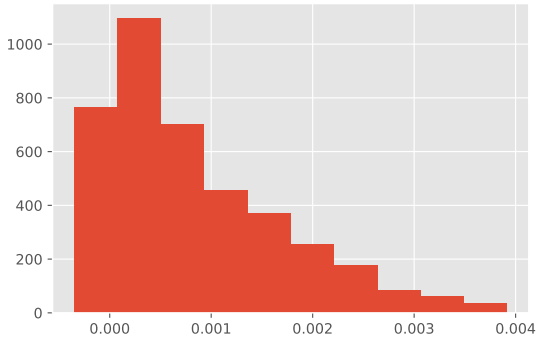
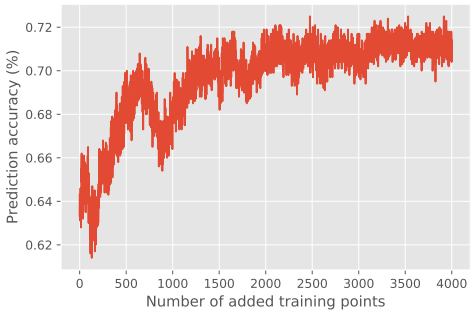
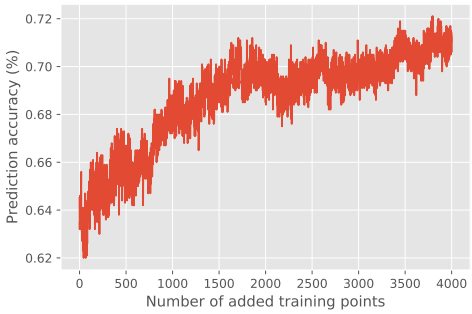



Figure 8: Histogram of predicted Shapley values. The y-axis is the number of samples and the x axis is the predicted Shapley value.

With the newly acquired data of 4000 samples and predicted Shapley values, the data was added sequentially from least to most valuable and most to least valuable (Figure 9). The model performance in accuracy was determined after testing using the same data from experiments 2 and 3, the last 1000 samples from brunello_test. In Figures 9, one can see a more pronounced/ sharp increase in accuracy when adding the high Shapley value points (a) compared to a minimal initial increase in accuracy when adding the low Shapley value points (b). These results show that the shapley values are determining high value points for the new data but once we add more and more training data, the prediction accuracy starts to converge around 70% for both high value and low value addition.




(a) Adding high value training points




(b) Adding low value training points

Figure 9: Measuring the value of samples from the newly acquired data. The y-axis is the prediction accuracy (%) and the x-axis is the number of training points added. (a) The 4000 samples for the new data were added sequentially to the training data from highest value to lowest value, training and testing at each step to get the model performance in accuracy. (b) Similarly, the 4000 samples for the new data were added sequentially to the training data from lowest value to highest value, training and testing at each step to get the model performance in accuracy.

## References

[1] N. Bhagwat and N.Khuri. Predicting targets for genome editing with long short term memory networks. *Transactions on Computational Science & Computational Intelligence, Springer Nature – Advances in Computer Vision and Computational Biology*, pages 1–14, 2020.

[2] Esteban Murillo Burford, Sarah Parsons, and Natalia Khuri. Data-driven prediction of cycling performance.

[3] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. *arXiv preprint arXiv:1904.02868*, 2019.