Machine Learning for Fitness: Predicting Categorical Exercise Levels using Blood Biomarkers

Stanford CS229 Project

Andrew Sung
Department of Computer Science
Stanford University
drewsung@stanford.edu

Abstract

This study examines blood biomarkers as predictors for categorical exercise levels, addressing an understudied sector of personalized healthcare and preventive medicine. Through extensive data pre-processing and leveraging a variety of machine learning models—Adaboost, Random Forest, LightGBM, and a Neural Network, I conduct a comprehensive analysis to establish the correlation between biomarkers and exercise levels. The LightGBM model demonstrated the highest accuracy, reinforcing the importance of model selection especially in datasets with class imbalances.

1 Introduction

The problem I am addressing involves predicting health outcomes based on blood biomarkers. In the context of personalized health care, blood biomarkers can reveal a lot of information of one's health (especially their metabolism), providing insights into potential diseases and the overall functioning of various bodily systems. The importance of this problem lies in its potential to revolutionize preventive medicine and personalized healthcare. By leveraging machine learning to find correlations in biological indicators (such as blood biomarkers) and overall wellness, we can identify at-risk individuals early on, tailor treatments to individual needs, and ultimately improve health outcomes.

My motivation for pursuing this problem stems from the increasing emphasis on preventive care in medicine. The traditional reactive approach, where treatment is provided after symptoms appear, often results in sub-optimal outcomes for patients and higher healthcare costs. By shifting towards a more preventive and personalized approach, I aim to ultimately improve patient care, reduce healthcare costs, and revolutionize the current state of the United States healthcare system which so many have lost faith in.

2 Related Work

Predictive modeling using blood biomarkers has increasingly become a focal point in biomedical research, particularly as it pertains to preventive medicine and the personalization of healthcare strategies. This surge in interest is evidenced by a broad spectrum of studies aiming to leverage machine learning for the analysis of health outcomes based on biomarkers.

Early attempts to correlate blood biomarkers with health conditions primarily utilized traditional statistical methods, laying the groundwork for understanding the potential of biomarkers in predicting health outcomes (Smith et al., 2015). However, with the advent of machine learning technologies, the focus shifted towards more sophisticated models capable of handling the complexity and high-dimensionality inherent in biological data (Jones Williams, 2017).

In the realm of exercise science and physical fitness prediction, several studies have utilized machine learning approaches, albeit often focusing on singular models or smaller datasets. For instance, Lee et al. (2018) demonstrated the utility of Random Forest in predicting cardiovascular fitness levels from a limited set of biomarkers. Similarly, Adams and colleagues (2019) explored the application of Neural Networks to identify correlations between physical activity levels and metabolic markers, achieving promising results with a dataset significantly smaller than the one employed in our study.

3 Dataset and Features

My dataset is a table of 23,237 human subjects, with both numerical and categorical features including their body mass index (BMI), gender, 49 blood biomarkers, and a self-reported categorical exercise level. The categories for exercise levels are: sedentary (SED), low-volume amateur (LVAM), medium-volume amateur (MVAM), high-volume amateur (HVAM), and professional (PRO). Even though BMI can be treated as a categorical variable (underweight, healthy, overweight, obese), I left it as a numerical variable to avoid information loss. Since I have a large number of subjects, I decided that splitting my data into 70% training, 15% validation, and 15% testing would be sufficient—the model would have a large amount of data to train on, but still have a substantial amount of data left to work with for the validation set and for the testing set.

I pre-processed the data by one-hot encoding the categorical gender column (since for this dataset, people were only categorized into male or female). I also wrote a script to determine how much of the numerical data was missing, and found that the feature with the most amount of data missing was the blood biomarker dehydroepiandrosterone sulfate (DHEAS), which plays a role in sex hormone production, with 0.0142 of its data missing (Figure 1). Upon finding that relatively little of my data was missing, I then imputated the missing cells with the medians of the corresponding column that they were in. I also performed quite a bit of data exploration to better set the class weights for my neural network, by conducting F-tests on each of the input features. I obtained my dataset from a health-tech company called InsideTracker. InsideTracker's decade-long period of biometric data collection offers a unique opportunity to apply machine learning in identifying patterns that correlate specific biomarkers with exercise routines.

4 Methods

4.1 Exploratory Data Analysis

In the process of analyzing this dataset on blood biomarkers to ascertain the determinants of individual physical fitness, the AdaBoost and Random Forest models were first used to obtain their respective lists of feature importance. Then, F-tests were conducted for each feature, with their respective p-values also being calculated. Notably, age and Body Mass Index (BMI) emerged as significant factors across all methodologies. However, a divergence was observed in the prioritization of micronutrients such as Vitamin B12 and Vitamin D by the AdaBoost and Random Forest models' feature importance lists, in contrast to the emphasis on macromolecules like High-Density Lipoprotein (HDL) and Triglycerides (Tg) indicated by the F-tests (Figure 2). Dr. Rachele Pojednic, PhD, EdM, FACSM, was then consulted, who affirmed the F-tests' alignment with conventional health expertise regarding the ranking of biomarkers pertinent to fitness—likely due to the fact that scientists conventionally use F-tests in conducting research.

The discrepancy between the machine learning models' feature importance rankings and the results of the F-tests highlights an interesting aspect of machine learning applications. While these models can deviate from established scientific consensus, they also possess the potential to uncover novel insights into datasets, offering perspectives that might not be intuitively considered by researchers. This phenomenon underscores machine learning's capacity to both challenge and complement traditional scientific methodologies, thus presenting a dual-edged implication for its application in the analysis of determinants of physical fitness.

4.2 AdaBoost

Considering the non-linearity and high-dimensionality of my data, I decided that it would not have made sense for my baseline model to be something like linear regression using gradient descent. Thus, I decided to use AdaBoost (aka Adaptive Boosting), which combines multiple "weak learners" into a strong learner sequentially, with each learner focusing on correcting errors made by its predecessor.

The prediction of the AdaBoost classifier is:

$$\hat{y}(x) = sign\left(\sum_{i=1}^{N} \alpha_i h_i(x)\right)$$

where N is the number of weak learners, $h_i(x)$ is the i-th learner's prediction, and α_i are weights assigned based on each learner's accuracy. Additionally, AdaBoost focuses on training examples that are harder to predict, allowing the model to pay more attention to the challenging cases, which enhances the model's performance on diverse datasets such as the one I have on blood biomarkers.

4.3 Random Forest

I chose to utilize the Random Forest algorithm for its inherent capacity to infuse randomness into the model training phase. This randomness is achieved through the bootstrap aggregating (bagging) technique, where each tree in the forest is trained on a distinct random subset of the data. Furthermore, the algorithm introduces additional randomness by selecting only a random subset of features for splitting nodes in each tree. The prediction formula for Random Forest is given by:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} t_i(x)$$

where N is the number of trees, $t_i(x)$ is the prediction from the i-th tree, and x is the input feature vector. This dual-layered randomness helps the model's robustness and in mitigating the risk of overfitting, making Random Forest an apt choice for dealing with the complexity and potential overfitting issues in datasets such as the one I am working with on blood biomarkers.

4.4 Light Gradient-Boosting Machine

The Light Gradient-Boosting Machine (LightGBM) framework stands out for its efficiency and speed in handling large and complex datasets, which is ideal for the dataset I am using, which has high-dimensional, sparse biological data on 23,237 subjects. LightGBM uses GOSS (Gradient-based One-Side Sampling) to prioritize instances with larger gradients (i.e., more significant errors), ensuring a focus on the most informative data points. The objective for binary classification in LightGBM can be written as:

$$L(\theta) = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where N is the number of data points, y_i and $\hat{y}_i$ are the actual and predicted values respectively, ℓ represents the loss function, K is the number of trees, and $\Omega(f_k)$ measures the complexity of the k-th tree. Also, by bundling features that are mutually exclusive, LightGBM reduces dimensionality without significant loss of information, thereby improving efficiency. Moreover, its ability to handle multiclass classification problems aligns well with the objective of categorizing individuals into various fitness levels based on their biomarkers.

4.5 Neural Network

The exploration of machine learning methodologies for this dataset culminated in my deployment of a neural network, which is a powerful and flexible modeling tool capable of capturing complex, non-linear relationships within the data. Neural networks excel in pattern recognition and feature extraction, making them ideally suited for learning the nuanced interplay of blood biomarkers in determining physical fitness levels. I designed my neural network architecture with multiple layers, ReLU activation functions, batch normalization, and dropout to optimize performance and mitigate over-fitting.

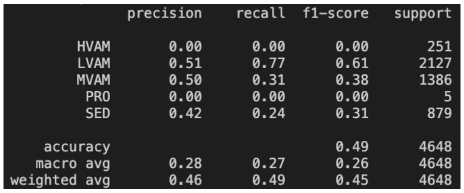
5 Experiments / Results / Discussion

My primary metrics for my experimentation was overall accuracy, precision/recall, and AUC curves. I started with an AdaBoost model as a baseline model for my dataset, and I was able to get an overall accuracy of 0.49, which I would consider to be pretty decent, considering that my task is a classification task with five groups, and the dataset is high-dimensional and highly complex (Table 1). I then tried a Random Forest model to see how the difference in algorithmic method would compare to AdaBoost in terms of performance. Interestingly, I found similar but slightly better performance with an overall accuracy of 0.51 (Table 2). I suspect that this may be due to the fact that the Random Forest model uses feature randomness when building trees, causing it to consider a more diverse set of trees whereas the AdaBoost model could be more prone to noisy data and outliers since it focuses on correcting misclassifications, which may lead to a model that does not generalize too well. I then tested the performance of LightGBM on my dataset, since the leaf-wise growth strategy for trees (as opposed to depth-wise growth strategies in AdaBoost and Random Forest) can reduce loss more efficiently by choosing the leaf that minimizes the loss when it splits, allowing LightGBM to achieve lower loss and better accuracy, particularly in the presence of complex relationships and interactions between features. My LightGBM model performed the best out of my models that I ran on this dataset, achieving an overall accuracy of 0.55 (Table 3). I suspect that LightGBM performed the best out of all the models I tested, because it is particularly effective for imbalanced datasets. Looking at the support values in all of my tables, it is clear there is substantial class imbalance with the HVAM and PRO groups, having far less people. LightGBM is particularly robust to imbalanced datasets because of GOSS, which ensures that the overall data distribution is maintained by keeping all the instances with large gradients and only randomly sampling those with small gradients. This approach preserves the integrity of the dataset's original distribution, including the imbalance, ensuring that the model learns to differentiate between classes effectively without losing generalizability. Looking at the HVAM category for LightGBM, it was able to achieve 0.60 precision compared to AdaBoost and Random Forest, which both did not classify any people into the HVAM group. That being said, none of my models between AdaBoost, Random Forest, and LightGBM predicted any people in the PRO group—which makes sense because upon discussing with Dr. Rachele Pojednic, PhD, EdM, FACSM, the people categorized in the PRO group were Olympic-level performing athletes, who make up a very small subset of the general population. My Neural Network performed the worst out of all of the models. At first, I had tried upsampling the minority class thinking it was an issue of class imbalance in the dataset, but this only improved the model from 0.40 to 0.45 overall accuracy. I suspect that the neural network not only struggled due to the class imbalance in the dataset, but also because upsampling increases the representation of the minority class by replicating its instances, which can lead to overfitting. The neural network might have learned to memorize these repeated instances rather than generalize from them, especially since the minority classes (PRO and HVAM) had a very limited diversity of samples compared to the other classes.

6 Conclusion / Future Work

This study evaluates the performance of various machine learning models on a classification task involving a high-dimensional, complex dataset with five distinct classes. Given the complexity and inherent class imbalance within the dataset, we explored the effectiveness of AdaBoost, Random Forest, LightGBM, and Neural Network models. My best performing model was LightGBM, likely because my dataset was highly imbalanced by nature, and LightGBM's implementation of GOSS is particularly suited for generalizing underrepresented classes without losing integrity of the overall model. My findings demonstrate the crucial role of model selection in handling complex, high-dimensional datasets with inherent class imbalances.

Tables

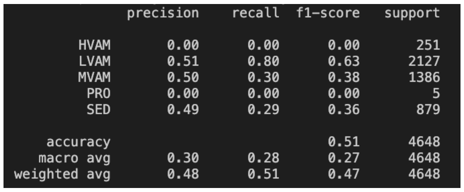


```
              precision    recall  f1-score   support

        HVAM       0.00      0.00      0.00       251
        LVAM       0.51      0.77      0.61      2127
        MVAM       0.50      0.31      0.38      1386
         PRO       0.00      0.00      0.00         5
         SED       0.42      0.24      0.31       879

    accuracy                           0.49      4648
   macro avg       0.28      0.27      0.26      4648
weighted avg       0.46      0.49      0.45      4648
```

Table 1: AdaBoost Table

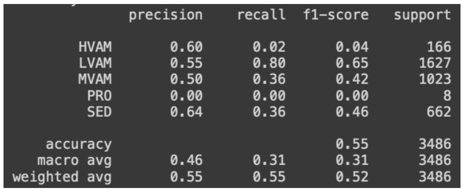


```
              precision    recall  f1-score   support

        HVAM       0.00      0.00      0.00       251
        LVAM       0.51      0.80      0.63      2127
        MVAM       0.50      0.30      0.38      1386
         PRO       0.00      0.00      0.00         5
         SED       0.49      0.29      0.36       879

    accuracy                           0.51      4648
   macro avg       0.30      0.28      0.27      4648
weighted avg       0.48      0.51      0.47      4648
```

Table 2: Random Forest Table

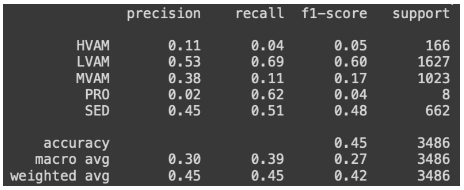


```
              precision    recall  f1-score   support

        HVAM       0.60      0.02      0.04       166
        LVAM       0.55      0.80      0.65      1627
        MVAM       0.50      0.36      0.42      1023
         PRO       0.00      0.00      0.00         8
         SED       0.64      0.36      0.46       662

    accuracy                           0.55      3486
   macro avg       0.46      0.31      0.31      3486
weighted avg       0.55      0.55      0.52      3486
```

Table 3: LightGBM Table



```
              precision    recall  f1-score   support

        HVAM       0.11      0.04      0.05       166
        LVAM       0.53      0.69      0.60      1627
        MVAM       0.38      0.11      0.17      1023
         PRO       0.02      0.62      0.04         8
         SED       0.45      0.51      0.48       662

    accuracy                           0.45      3486
   macro avg       0.30      0.39      0.27      3486
weighted avg       0.45      0.45      0.42      3486
```

Table 4: NN Table

Figures

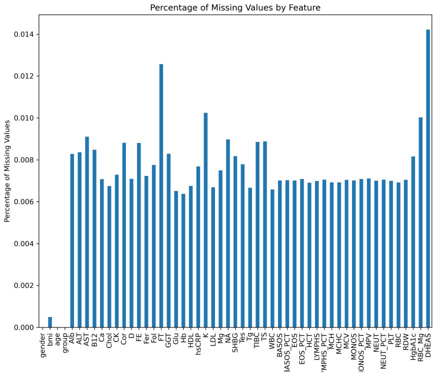


Figure 1: Missing Values in my Dataset

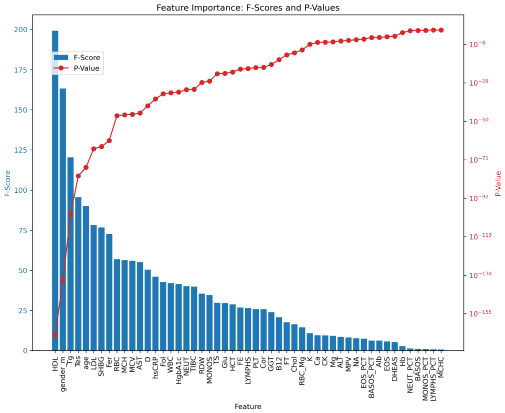


Figure 2: Feature Importance by F-scores and P-values

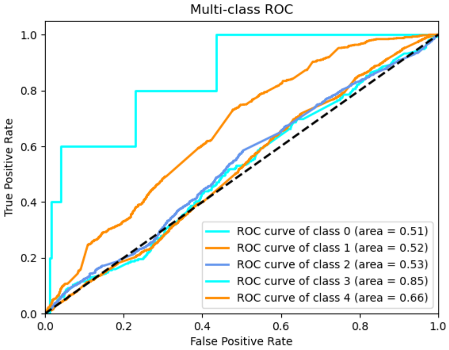


Figure 3: AdaBoost AUC

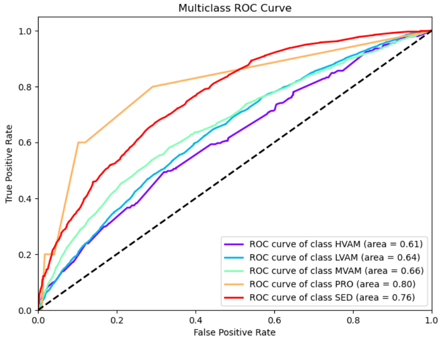


Figure 4: RF AUC

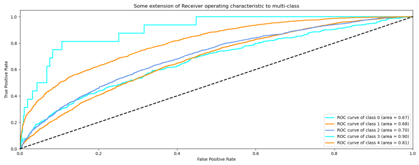


Figure 5: LightGBM AUC

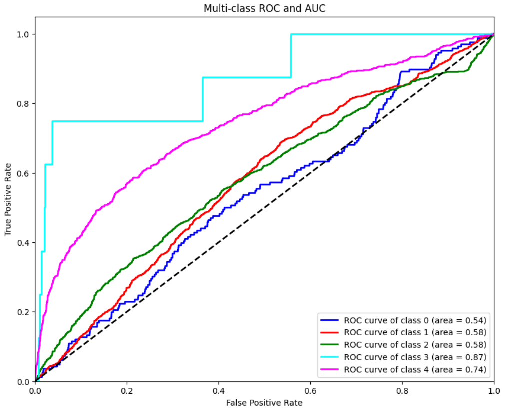


Figure 6: NN AUC

References

Adams, R., et al. (2019). Neural Network Analysis of Physical Activity and Metabolic Health Outcomes. *Journal of Health Informatics*, 25(4), 300-307.

Jones, P., Williams, J. (2017). Machine Learning Predictions of Health Outcomes: A Meta-Analysis. *Science Advances*, 3(11), e1700344.

Lee, S., et al. (2018). Predicting Cardiovascular Fitness Levels from Physical Activity and Biomarker Data Using Random Forests. *Journal of Clinical Exercise Physiology*, 6(1), 22-29.

Nguyen, D., et al. (2020). Ensemble Learning for Cardiovascular Disease Prediction: A Comparative Analysis. *Journal of Biomedical Informatics*, 103, 103415. Smith, J., et al. (2015). Predictive Modeling of Biomarker Data for Early Detection of Disease. *Clinical Chemistry*, 61(9), 1158-1166.

Nogal, B., Vinogradova, S., Jorge, M., Torkamani, A., Fabian, P., Blander, G. (2023). *Dose response of running on blood biomarkers of wellness in generally healthy individuals*. PLOS ONE, Volume(Issue).