Elliot Kinder (ejk232)
Julio A. Leañez (jal535)
Professor Haiyun He
ORIE 4741 Final Project, Spring 2024
Friday, May 10th, 2024

Preventive Health: Predicting daily changes in Heart Rate Variability

Problem Statement:

Health care is about 5% preventive. The US spends about $4.5 trillion dollars per year on healthcare which equates to $13,500 per person. Although there are many factors that contribute toward this exorbitant amount, preventive medicine could help reduce healthcare expenses in diseases such as cardiovascular disease.

With a desire to improve preventive care through predictive analytics, this study will focus on Heart Rate Variability (HRV) as an important metric for measuring overall resilience. HRV is a measure of the variation in time between heartbeats and is known to indicate various chronic conditions such as high blood pressure, diabetes, and depression. Wearable technology, specifically the Apple Watch, continuously tracks HRV in real-time, providing a rich dataset that reflects an individual's daily health status. The researchers in this study decided to answer the following questions:

1. Can one user wearable features to predict heart rate variability?
2. Can one determine the main contributors towards HRV increases / decreases?
3. How can one fill in the gaps of data gathered by wearables such as the Apple Watch, when the users are not wearing them?

Conclusions:

- The researchers believe that a lack of critical data led to the models created not having significant predictive results.
- The linear model trained with a quadratic loss function on a subset of features without outliers performed the best with a normalized test error of 1.21
- The classification model performed best at detecting cases where HRV does not increase (71% recall for class 0)
- Across all models created, respiratory rate was found to have a high impact, suggesting an influence on HRV
- Build a model on a dataset with multiple entities (user datasets)
- Should integrate a lot more data, such as lab test results, age, sex, and height

Dataset:

Our dataset is the product of a complex data pipeline from a longevity startup called OneTwentyOne. This company leverages Apple Health's infrastructure to sync all of its collected data to a database where the data remains anonymous and HIPAA compliant.

The fetching begins by pulling the researcher's data (found by close inspection) into 3 separate CSV files which include their sleep and raw biometric information, for a total of about 1.8 millions rows from the database tables. This data is then merged into a unified dataset by a combination of data conversion (into a map of metric to dataframe pairs), aggregation (by date), and imputation (explained below). The methods to obtain the raw data are private to the company.

Data Imputation:

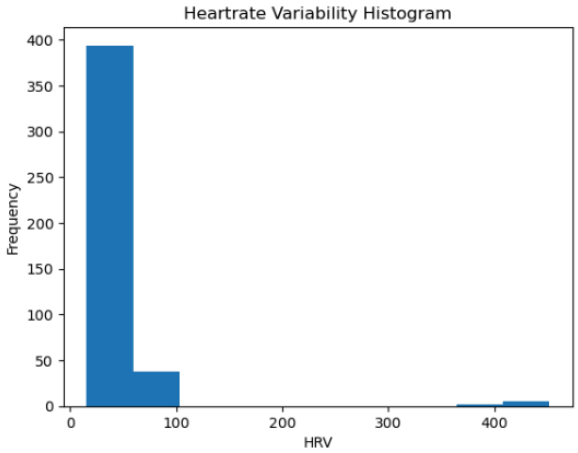
- Build a map of metric to *(spread frequency, interpolation frequency)* pairs, serving as a heuristic for the number of days forward and/or backward to spread non-null values using backward and forward filling. Additionally, the interpolation frequency indicates how many days should be covered by polynomial interpolation for each specific metric.
- If there is a way to connect two data points within the range, perform the interpolation. This leaves space for extra null values which are handled in the next step. After this, create the CSV (which has remaining null values)
- To address sparse features having a substantial impact on the models built, any feature with over 10% missing values was removed from the data set.
- To address any remaining null values, singular value decomposition was implemented to fill such values. To do so the data matrix was decomposed according to SVD and rebuilt with an estimated rank of 20. The values of this rebuilt matrix were used to fill null values in the original dataset providing a complete dataset to be used for the models.

Linear Model:

Linear regression was used as the first step in model building in the project as it provides easily interpretable results and is a flexible model that can be adjusted to highlight different features and distributions. The data with past and future HRV was not used in the linear regression in order to address some wearable devices not having the capability to record HRV. This decision allows the resulting model to generalize to people with less advanced devices. Developing a linear model followed a process of adjusting the most recent model to improve errors and thus make a stronger model. To compare the different models, the data was split into a train and test set which were used to calculate train and test error using the mean squared error. In addition to this the R squared value for the training data was used.

The initial model used was a least squares regression with a quadratic loss function on all features of the filled and normalized data set described previously. This meant that there were 30 features including a constant in the resulting model. Fitting this model and computing the errors provided some surprising results as the test error was over 3 times smaller than the

training error. This is highly unexpected, but through examining the data a potential cause was discovered of the training data having more variance than the test data. The next model iterations used a huber loss function in order to directly address this as huber loss is more robust to outliers. This proved unsuccessful however, as the test error decreased while the training error increased. To further examine the problem, the HRV values were plotted as seen to the right. Through this it was determined that the outliers around 400 are most likely errors in data collection as HRV rarely goes above 200 (https://www.rupahealth.com/post/what-is-heart-rate-variability). Thus, any HRV above 200 was dropped from the data set in successive models. The next model went back to a quadratic loss function on the full feature set without any outliers in HRV. This proved to solve the issue of training error being larger than test error showing improvement. But this introduced a new problem of the training error being nearly double the test error suggesting overfitting. To combat this problem features with less impact were dropped from the model in the next iteration leaving only 10 features and a constant for future models. A model with both quadratic loss and huber loss were fit on this smaller set of features. Both models saw the overfitting decrease but not go away though the quadratic model had more improvement. The results of all model iterations are shown in the table below.



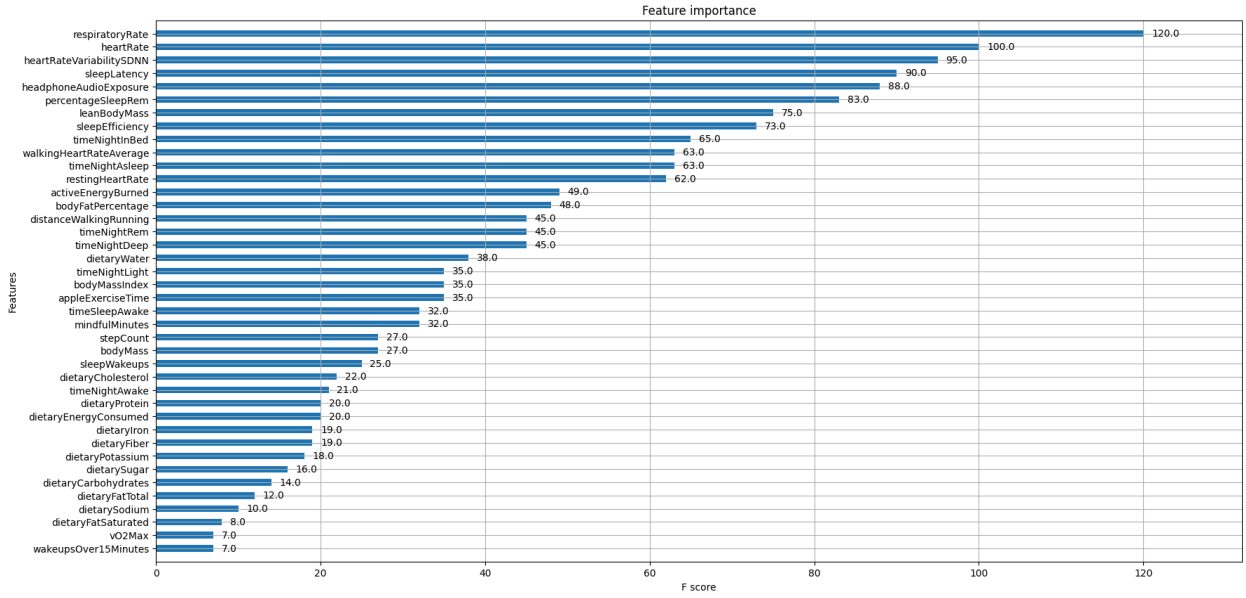
Model	R Squared	Train MSE	Test MSE
Quadratic Loss	0.23440	0.94188	0.27078
Huber Loss	-0.02690	1.24516	0.05171
Quadratic Loss without Outlier	0.17878	0.75620	1.39026
Quadratic Loss without Outlier with limited features	0.15735	0.77593	1.21431
Huber Loss without Outlier with limited features	0.13477	0.79241	1.26310

The best model can be determined through interpreting these results. The goal is a model that generalizes well so a lower mean squared error is better. This suggests that the models containing the outliers are better. However, as previously stated the train error higher than the test error is an abnormal result and this makes these models poor choices. If only the latter three models are considered the best choice is Quadratic loss with the limited features as it has the lowest test error. Examining this model further reveals some of the key features for indicating HRV include Time Night Asleep, Time Night in Bed, Dietary Fiber, Body Mass Index, and Respiratory Rate. The caveat of these findings is that despite being our best model, the model has an R squared value of 0.15735 suggesting that it does not have much predictive power and thus conclusions made from it should be treated with skepticism.

Classifier Model:

XGBoost (Extreme Gradient Boosting) was chosen for this project due to its robustness and ability to handle missing values effectively. One of its key strengths lies in its built-in mechanism to manage NaN values, making it particularly suitable for datasets that aren't perfectly complete. Furthermore, XGBoost provides comprehensive feature importance analysis, allowing us to identify the most predictive features affecting changes in HRV. By aggregating data by day and aiming to predict increases in HRV compared to the weekly average, XGBoost is well-suited because of its efficiency and accuracy in handling time series data.

To ensure optimal model performance, we used grid search with cross-validation to identify the best combination of hyperparameters. The primary parameters optimized included *n_estimators*, *max_depth*, *learning_rate*, *subsample*, and *colsample_bytree*. After obtaining the best parameters, the model was fit using this configuration, enabling accurate predictions of HRV changes.



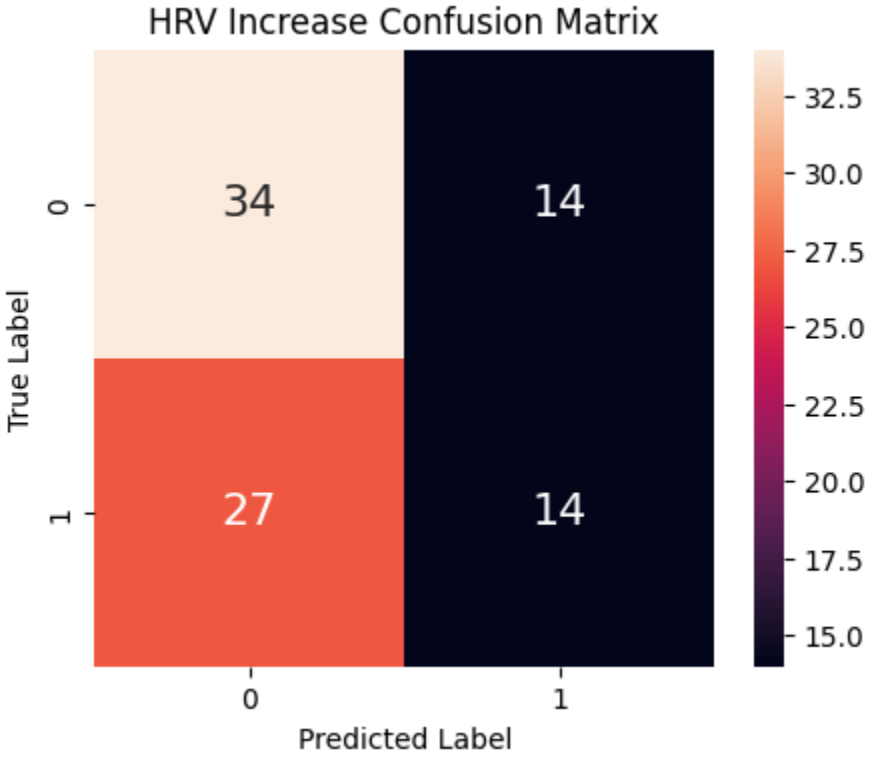
The feature importance analysis, based on F-Score, highlighted the most influential factors in predicting changes in HRV. The top features include:

- **Respiratory Rate:** Topping the list, it provides valuable insights into overall health, as fluctuations can indicate potential cardiovascular or respiratory issues.
- **Heart Rate:** A strong predictor, heart rate closely relates to HRV, offering a direct measure of cardiovascular health.
- **Heart Rate Variability (SDNN):** Directly related to HRV, SDNN measures the standard deviation of normal-to-normal intervals, reflecting autonomic nervous system activity and overall heart health.
- **Sleep Latency:** The time it takes to fall asleep, which is an important indicator of sleep quality and overall well-being.

- **Headphone Audio Exposure:** Although initially unexpected, exposure to high audio levels may correlate with stress levels or lifestyle factors affecting HRV.

Additional features, such as `percentageSleepRem`, `leanBodyMass`, and `walkingHeartRateAverage`, also demonstrated significant predictive value. This analysis reinforced the idea that wearable technology provides rich, valuable data, and that predictive models like XGBoost can yield actionable insights into health metrics associated with changes in HRV.

	precision	recall	f1-score	support
0	0.56	0.71	0.62	48
1	0.50	0.34	0.41	41
accuracy			0.54	89
macro avg	0.53	0.52	0.51	89
weighted avg	0.53	0.54	0.52	89



The model achieved an overall accuracy of 54% on the test data, with a precision of 56% for identifying no increase in HRV (class 0) and 50% for identifying increases in HRV (class 1). Recall values reveal that the model is better at detecting cases where HRV does not increase (71% recall for class 0) than cases where HRV does increase (34% recall for class 1). The resulting F1-scores are 0.62 for no increase (class 0) and 0.41 for increase (class 1).

These results indicate that the model is relatively effective at predicting when HRV does not increase but struggles to accurately identify instances where HRV does increase, as evidenced by the high false-negative rate (27 instances). The precision for both classes remains moderate, suggesting that further refinement is required to improve the model's predictive capabilities.

In summary, while the XGBoost model provides a baseline understanding of HRV changes, substantial improvements in feature engineering, data collection, and class balancing are essential to achieve a more comprehensive and accurate prediction of HRV changes, particularly in detecting increases. Beyond these improvements, advanced imputation techniques, sequential models, and incorporating clinical knowledge into feature selection and model interpretation are also essential. Ultimately, combining these strategies will provide a deeper understanding of HRV changes and lead to improved preventive health strategies.

Next Steps:

The models we have created provide a solid foundation for predicting HRV using easily attainable health data. However, the models did not have the predictive power desired to provide users with accurate information about their HRV levels and how those indicate other health outcomes. With this information, there are several next steps that should be considered to further this work and eventually create an end product that improves user's knowledge about their overall health.

The first recommendation is to pursue a more comprehensive data set to build from. The data used for this project consisted of a single user's information taken over an extended time period. This allowed a highly specific model to be made and the data set used to be dense, improving predictive power. However, this came at a loss in the ability to generalize our findings which is the end goal of this work. Thus, for future work, a larger data set containing information on a variety of people is needed. Another way to improve the data set is how the data is collected. The data being collected by Apple Watch and other similar products provides great ease in data collection, and is how the final model will be used to predict HRV in the future, but in the model building process having more accurate data is beneficial. Obviously the collection of such data provides significant time and monetary considerations. Medical equipment is necessary and will have to be collected from a large number of people over an extended period of time in order to have a large accurate data set. This would provide the benefit of an accurate model that can be generalized so such an undertaking would be worthwhile in the continuation of this work.

The second way to further this work would be to explore other model options and how they predict HRV. In this work two models were primarily focused on, linear and classifier models, but there are many other options available that could provide more predictive power. One such model is a time series model. Constructing a model using current HRV and other easily obtainable information such as exercise time, sleep quality, and body mass index to predict future HRV levels could prove more powerful than the models developed in this project. Another potential model is using decision trees to predict HRV. This would allow for non linear features to be utilized more effectively and serve to increase the predictive power of the model built.

These are just two of the many possible models that could be used to further explore the data and further work in what model should be selected and construction said model would prove beneficial.

The final recommendation to further this work is convert the model into a usable interface for users. This is a step that should be taken after a strong model has been obtained and is ready for use. There are a variety of ways such a recommendation could be implemented. The average person does not know what HRV is, much less how to interpret it as an indicator of overall health. Thus an essential part of making the model accessible to users is providing an explanation of what their individual prediction means in terms of their health. An additional part of this is indicating to users when their HRV is at a concerning level and recommending that they consult with a doctor to ensure they are healthy. A further step is to examine what features play the strongest role in the model and their personal prediction in order to present the user with actionable ways to improve their HRV. Taking these actions and making usable information available from the model will ensure that the model created has an impact on people and helps them improve their health providing meaning to the work being done in this project.

Comprehensively, the next steps recommended obtaining a larger more accurate data set, examining other model types, and providing usable information to users from the final model will further the work of this project into a meaningful end result. While our models did not provide the predictive impact aimed for, the work done is a first step in achieving the tangible results that are the goal of this project.

References:

https://www.rupahealth.com/post/what-is-heart-rate-variability
https://garden.121health.app/
https://garden.121health.app/Biometrics/heartRateVariabilitySDNN
https://xgboost.readthedocs.io/en/stable/

Additional Notes:

Github link to model code: https://github.com/jleanezv/Project_ORIE4741
Member Contributions: Project proposal and final report were done collaboratively. Elliot took the lead on the regression model while Julio led the classification model. The data imputation was shared.