Elliot Kinder & Julio A. Leanez
March 17th, 2024
ORIE 4741: Learning with Big, Messy Data
Github link: https://github.com/jleanezv/Project_ORIE4741
Professor Haiyun He

Project Proposal

Modern health care has a distinct focus on treatment rather than prevention. This approach is costly in both financial and health based ways. Health outcomes are almost always significantly better when health issues are detected and treated early with many studies showing the benefits on overall health of preventive care. Moving to the financial costs of treatment based health care, the US spends $4.5 trillion per year on healthcare which equates to $13,500 per person. While there are a number of factors that contribute to this exorbitant large amount, preventive care can reduce this cost as it has been shown to reduce expensive emergency room visits. Using this knowledge about the benefits of preventive healthcare, we pose the following question: How can risk for chronic health conditions be predicted using readily available health data?

The first step in answering this question is how do we measure risk for chronic health conditions. This is a somewhat ambiguous thing to measure as there are a variety of chronic health conditions many of which are completely unrelated. Thus, our approach is to find a metric that is clearly measurable and accurately predicts the likelihood of a variety of chronic health conditions. This led us to choose heart rate variability (HRV) as our chosen metric. HRV is where the amount of time between heartbeats fluctuates by small amounts, often fractions of a second. While this can sound like a problem, variance in heart rate is perfectly normal and part of the way our bodies adapt to changing environments. For example, in high stress situations where humans' natural fight or flight instinct kicks in, a person's heart rate will naturally increase to provide more oxygen to muscles in case of physical exertion. In contrast, when someone is calm and resting their heart rate will decrease as their body is relaxed and less oxygen is needed. With this in mind, high HRV is a sign of a healthy individual while low HRV indicates potential health problems. As HRV measures fraction of a second changes in heart rate, it requires specialized equipment to measure accurately. In medical circumstances an electrocardiogram is used to measure HRV while other devices including wearable devices offer methods of measuring HRV separate from medical care.

Low HRV indicates a number of common chronic conditions which is why we find it an important way to measure the risk of such conditions. Low HRV is a sign of the body having difficulty adapting to situations which can be a sign of high blood pressure, diabetes, depression, and a number of other conditions. Just the three chronic conditions affect the majority of Americans and pose significant health risks. High blood pressure affects 45.4% of adults and significantly raises the risk of cardiovascular disease, the leading cause of death in the US. As for

diabetes, it affects 11.6% of Americans while depression affects 8% of adults and 15% of youth in the US. These statistics show the importance of preventive and early treatment of such conditions as they are so widespread and impactful for the American public. This presents an opportunity to use HRV as a measure of risk for such conditions and allow for better, earlier treatment, which we hope to do.

The data for our project will be sourced from wearable technology, specifically devices such as the Apple Watch. These wearables have become ubiquitous, not only as tools for communication and productivity but also as monitors of health and well-being. The Apple Watch continuously tracks HRV in real-time and in real-world settings. This continuous monitoring provides a rich dataset that is more reflective of an individual's day-to-day health status than traditional, sporadic health assessments. Our primary data repository for this project will be the database of an application known as 121 Health, which has amassed a comprehensive dataset of health metrics from users who have consented to share their data for research purposes. Importantly, this dataset is 100% anonymized and complies with the Health Insurance Portability and Accountability Act (HIPAA), ensuring that individual privacy is maintained while allowing for valuable health insights to be derived from the aggregated data.

When considering the prediction of changes in HRV and, by extension, the potential onset of chronic health conditions, several data analysis models present themselves as particularly promising. Machine learning algorithms, especially those capable of processing large, complex datasets, stand out for their ability to identify patterns and correlations that may not be immediately obvious. Models such as Random Forests and Gradient Boosting Machines are well-regarded for their robustness and ability to handle the variance within big, messy data, making them suitable candidates for this task. Additionally, deep learning techniques, particularly recurrent neural networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks and the Temporal Fusion Transformer (TFT), are adept at processing sequential data, a common format in time-series data like HRV measurements. These models can learn and remember over long sequences, making them ideal for predicting changes in HRV based on historical data. Our project will explore these models' applicability and efficacy in predicting HRV changes, with the goal of identifying the most accurate and reliable method for early detection of chronic health risks. This exploration will not only contribute to the academic and practical understanding of predictive health analytics but also align with the broader objective of enhancing preventive care through technology.