Stony Brook University Data Science Bootcamp Capstone Project – Predicting Yearly Medical Costs Using Synthetic Patient Data

For this capstone project, synthetic patient records from Synthea will be used to create a predictive model of yearly healthcare costs. Synthea is a software mechanism that can generate realistic patient medical records used for academia, research, industry, and government applications. In this project, a dataset was generated with 100 live patients from each of the 50 states in the United States. This dataset includes 18 csv files which contain patient information linked by unique identifiers such as patient ID numbers, encounters during which the patient received care, providers, insurance payers, and claim ID numbers.

The goal of this project is to develop a predictive model for the yearly healthcare costs of patients. Cost data from 2023 for medical encounters (visits to hospitals or doctor's offices) will be the target variable, and data including number of medical encounters, number of medical procedures and their associated costs, number of medications and their associated costs, number of immunizations and their associated costs, and medical observations of the patients will be used to predict healthcare costs. This information would be useful for healthcare providers, health insurance agencies, and individuals interested in their own health and healthcare costs to understand which factors contribute toward higher or lower healthcare expenses for patients.

While developing a predictive model of healthcare costs, the data will be analyzed to understand which factors are most influential in determining healthcare costs for patients. Additionally, observations from doctor appointments including biological measurements from blood tests (cholesterol, blood glucose, etc.), height, weight, heart rate, and numerous others will be studied to determine how a patient's overall health contributes to their healthcare cost. This will assist healthcare providers in helping their patients make lifestyle changes that can improve their overall health and reduce healthcare costs.

The Data Science Method will be used to tackle this project. Now that the problem has been identified, the data tables will be imported, cleaned, and reorganized to create a coherent and relevant data set for this project. Next, the data will be explored to understand relationships between all the features and the target feature of yearly healthcare costs. Then, pre-processing and training will occur to test out different models and analyze their performance in predicting healthcare costs. Finally, a model will be selected and implemented to make predictions about future healthcare expenses. The project will then be properly documented and summarized for presentation.