Medical Cost Prediction

A Project Proposal by Team Sage

Madhura Prashant Vaidya, Rajkumar Baskar, Shreeyash Amit Yende

Abstract

One of the very basic needs almost every individual needs access to is the Healthcare offered by their country. As such, a considerable portion of money gets spent on the insurance, which could become a potential problem if everyone has to pay same monthly premium. This project aims to determine a fair price that can be set by the insurance provider so that they are both affordable and curated to the needs of the individuals based on their medical history. In this project, a wide variety of regression techniques will be utilized to predict the cost, and their metrics will be compared.

Introduction

The United States' national health expenditure (NHE) grew 5.8% to $3.2 trillion in 2015 (i.e., $9,990 per person), which accounted for 17.8% of the nation's gross domestic product (GDP) [1]. With this increasing growth and the constant need for access to primary healthcare, predicting the healthcare cost for individuals as accurate as possible is one of the major challenges that could benefit more than one stakeholder, not limited to just the medical insurers [2]. These forecasts would also greatly benefit the insured candidates since they could plan for their yearly medical costs that could arise due to unforeseeable incidents and choose the insurance type in advance wisely. In this project, for medical cost prediction, we will be exploring in detail some of the regression models for prediction to evaluate best the medical cost for individuals based on their age, medical conditions, and other commonly available factors.

Proposed Project

The main objective of this project is to predict the premium for an individual based on readily available information that could be utilized to find a pattern from the previous years' data. Here, we use the Kaggle dataset for insurance forecast, consisting of about 1338 records, each with seven features [3]. The dataset also contains few missing values, which will involve data cleanup and normalization before use.

The seven features include – age, sex, BMI (body mass index), children count, smoking status, region, and premium/charges [4]. The goal would be to accurately predict the charges for unseen data based on the given information. For this purpose, the available data would be divided into training and test set, with 80% of it being utilized for training and the rest 20% for testing.

The basic regression model for linear regression [5] will be evaluated for the initial cost prediction, along with other different combinations of regression models to compare metrics. Also, the Naive Bayes model would be used to analyze the performance based on priors. This project could also be extended to the use of Deep Neural Networks (DNNs), and Convolutional Neural Networks (CNNs) based on the timeline.

References

1. *National Health Expenditure Data*. (2016). Retrieved from The Centers for Medicare & Medicaid Services (CMS): https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata
2. Agarwal, N., & Lahiri, B. (2014). *Predicting healthcare expenditure increase for an individual from medicare data*. Retrieved from Proceedings of the ACM SIGKDD Workshop on Health Informatics: https://scholar.google.com/scholar_lookup?journal=Proceedings+of+the+ACM+SIGKDD+Workshop+on+Health+Informatics&title=Predicting+healthcare+expenditure+increase+for+an+individual+from+medicare+data&author=C+Lahiri&author=N+Agarwal&publication_year=2014&
3. Choi, M. (2018). *Medical Cost Personal Datasets (insurance.csv)*. Retrieved from Kaggle: https://www.kaggle.com/datasets/mirichoi0218/insurance
4. Morid, M. A., Kawamoto, K., Ault, T., Dorius, J., & Abdelrahman, S. (2018). *Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation*. Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977561/#r1-2731392
5. Uzila, A. (2021). *Medical Cost Prediction*. Retrieved from Towards Data Science: https://towardsdatascience.com/medical-cost-prediction-4876e3449adf