A logo of university of new haven

Description automatically generated

**Title**

**DSCI: 6002 -1 - Intro to Data Science – Fall 2023**

**THE COST EFFECTIVENESS ANALYSIS AND PREDICTION FOR HEALTH CARE**

**Team – 16**

* **Sathish Gandhi**, Parasuram Reddy
* **Ayushi**, Jar
* **Mounika**, Aithagoni
* **Sujith**, Kankanala

**GitHub**

[ayushijar/Health-insurance-cost-predictor (github.com)](https://github.com/ayushijar/Health-insurance-cost-predictor/tree/main)

**Contents**

Title…………………………………………………1

Abstract…………………………………………..2

Contents………………………………………….3

Business Challenge………………………….4

Solutions………………………………………….4

Methodology………………………………...….5

Data Preparation…………………………..….6

Data visualization……………………………..6

Modeling…………………………………………..7

Evaluation…………………………………………7

Deployment………………………………………7

Model User Interface…………………………8

Future Improvements……………………….8

**Contents**

**Abstract:**

This project endeavors to develop an advanced health insurance prediction model, employing a holistic methodology encompassing data preparation, sophisticated modeling techniques, and thorough evaluation. Through meticulous data curation and feature engineering, the dataset is refined to address missing values, outliers, and class imbalances. Machine learning algorithms are then strategically applied in the modeling phase, with a focus on optimizing performance and ensuring interpretability. The seamless deployment of the model considers scalability and real-time applicability, while continuous monitoring strategies uphold its reliability. Evaluation metrics, cross-validation, and ethical considerations guide a comprehensive assessment, ensuring the model's accuracy, generalizability, and fairness. The outcomes of this project aim to contribute to informed decision-making in health insurance, enhancing financial planning, transparency, and the overall patient experience.

**Abstract**

**Business challenge**

* In order to maximize financial planning, increase transparency, and enhance the patient’s experience, how can we properly anticipate and estimate the future health insurance cost for a patient before their admittance to a hospital?
* Individuals vary in both their body types and their living cultures. Therefore, depending on these circumstances, the diseases that affect them or the cost of treating them also differ. For instance, since smokers have a higher chance of developing a chronic illness, their medical insurance premium may be higher than that of a healthy individual free of bad habits.

**Solutions**

* Data insights from an analytical dashboard that may be sent to healthcare and insurance providers.
* Utilize advanced data analytics and predictive modeling techniques to analyze historical patient data and identify patterns in healthcare utilization.
* An artificial intelligence (AI) based prediction model that projects future insurance costs.
* Establish a feedback mechanism to gather input from patients regarding the accuracy of cost estimates and their overall satisfaction with the financial planning process.

**Objective**

**Methodology**

We have used CRISP-DM Methodology, The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology serves as a comprehensive framework for guiding data mining and machine learning projects. In its iterative process, it begins with understanding the business objectives, emphasizing the importance of aligning data analysis with overarching business goals. Below are the various steps which we have taken while building the Model.

A diagram of data processing

Description automatically generated

1. Collect the hospital package pricing dataset.
2. Explore and understand the data.
3. Clean the data.
4. Perform engineering and preprocessing to prepare for the modeling step.
5. Select the suitable predictive model and train it with the data.
6. Deploy the model on a live server and integrate it into a web application to predict the pricing in real time.
7. Monitor the model in production and iterate.

**Methodology**

**Data preparation**

We meticulously curated and refined the dataset to ensure its suitability for analysis. This involved addressing missing data through imputation methods, handling outliers, and performing necessary transformations, including feature engineering and normalization.

A screenshot of a computer

Description automatically generated

**Data Visualization**

A graph of different colored dots

Description automatically generated

A close-up of a color chart

Description automatically generated

**EDA**

Here are few EDA Visualization Results.

A screen shot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

**EDA Results**

**Modeling**

In the data modeling phase, we selected and trained machine learning models tailored to our health insurance prediction objectives. Rigorous evaluation and optimization ensured their effectiveness, and meticulous planning preceded the deployment of the chosen model. Ongoing monitoring strategies were implemented to maintain reliability and relevance.

**Regression Evaluation**

As per the EDA we have got few results for the accuracy. Based on accuracy, we can conclude that random forest has the highest accuracy and decision tree classifier is overfitting.

|  |  |
| --- | --- |
| **Regression Method** | **Accuracy** |
| Linear Regression Train | 0.7443271565246132 |
| Lasso Train | 0.7443253454766152 |
| Ridge Train | 0.7442924539952123 |
| Decision Tree Train | 1.0 |
| Random Forest Train | 0.9753696022085427 |

**Model Evaluation**

In evaluating our health insurance prediction model, we meticulously selected metrics, cross-validated for stability, and prioritized interpretability and ethical considerations. Continuous improvement strategies ensure ongoing relevance.

**Deployment**

Random forest model for prediction model was deployed using flask, prioritizing scalability and real time applicability. Continuous monitoring ensures its ongoing reliability and effectiveness in the operational environment.

**Modeling, Evaluation & Deployment**

**Predicted Value**

**Input values**

**Initial Interface**

**User Interface**

After the Data preparation we have created a model using Regression. Here are the Images of the User Interface which we created.

A screenshot of a computer screen

Description automatically generated A screenshot of a computer screen

Description automatically generated

A screenshot of a computer error

Description automatically generated

**Model UI**

**Future Improvements**

* Important advancements for the project include obtaining a variety of datasets, putting strong data cleaning practices into place, investigating novel characteristics, and testing out different machine learning models to increase accuracy.
* SAPH values and other explain ability strategies will be used to improve transparency. Priorities include addressing unbalanced data, ongoing observation, and prompt updates.
* Important usability factors include a user-friendly interface, privacy, and security, as well as a smooth integration of the model with healthcare systems.
* Fairness and other ethical issues will be dealt with beforehand, and continued success depends on staying current with machine learning and healthcare advances.

**Future**