Title

University of New Haven

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

Contents

Contents

Title……………………………………………….1
Contents………………………………………..2
Abstract…………………………………………3
Business Challenge…………………………4
Solutions………………………………………..4
Methodology………………………………….5
Data Preparation……………………………6
Data visualization…………………………..6
Visualization Results………………………7
Modeling…………………………………………8
Evaluation………………………………………8
Deployment……………………………………8
Model User Interface………………………9
Future Improvements…………………….10
Conclusion……………………………………11

Abstract

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.

Objective

Business challenge

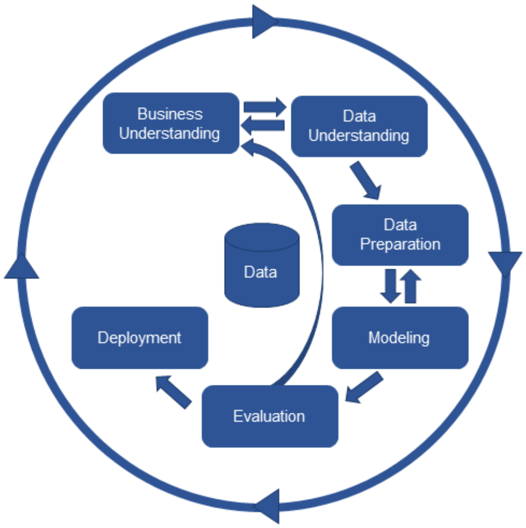
- In order to maximize financial planning, increase transparency, and enhance the patient's experience, how can we properly anticipate and estimate the future health insurance cost for a patient before their admittance to a hospital?
- Individuals vary in both their body types and their living cultures. 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.

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.

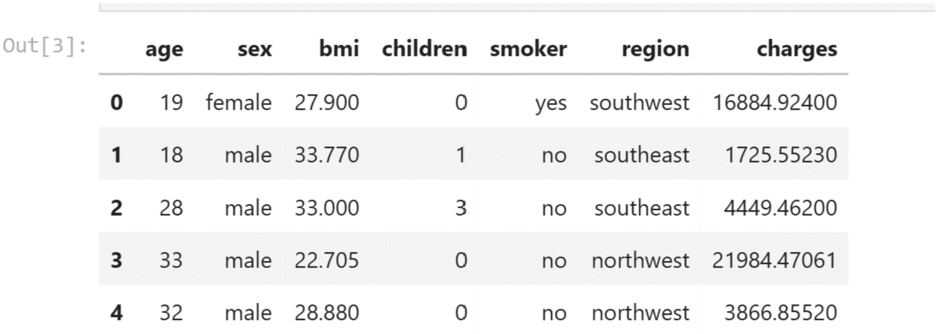


1. Collect the hospital package pricing dataset.
2. Explore and understand the data.
3. Clean the data.
4. Perform engineering and preprocessing to prepare for the modeling step.
5. Select the suitable predictive model and train it with the data.
6. Deploy the model on a live server and integrate it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

EDA

Data preparation

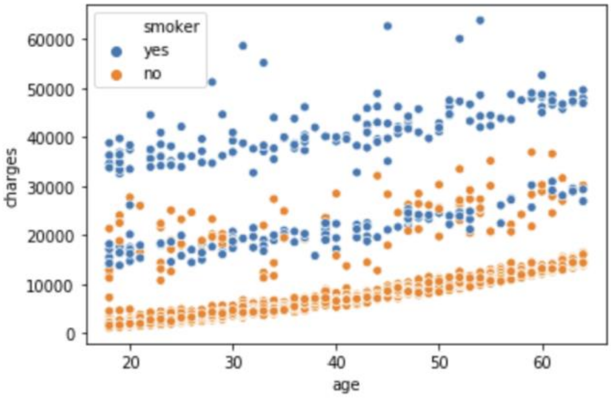
We meticulously curated and refined the dataset to ensure its suitability for analysis. This involved addressing missing data through imputation methods, handling outliers, and performing necessary transformations, including feature engineering and normalization.

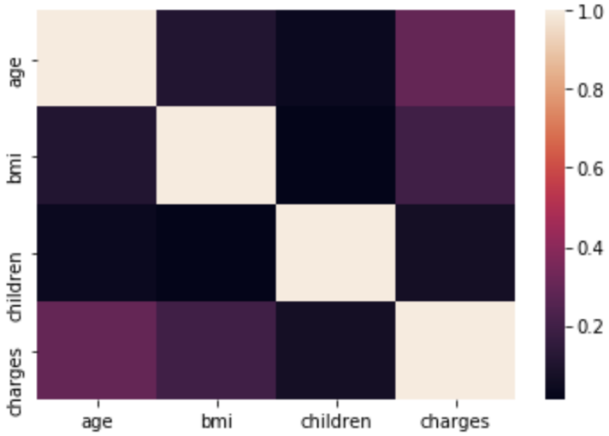


```
Out[3]:
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

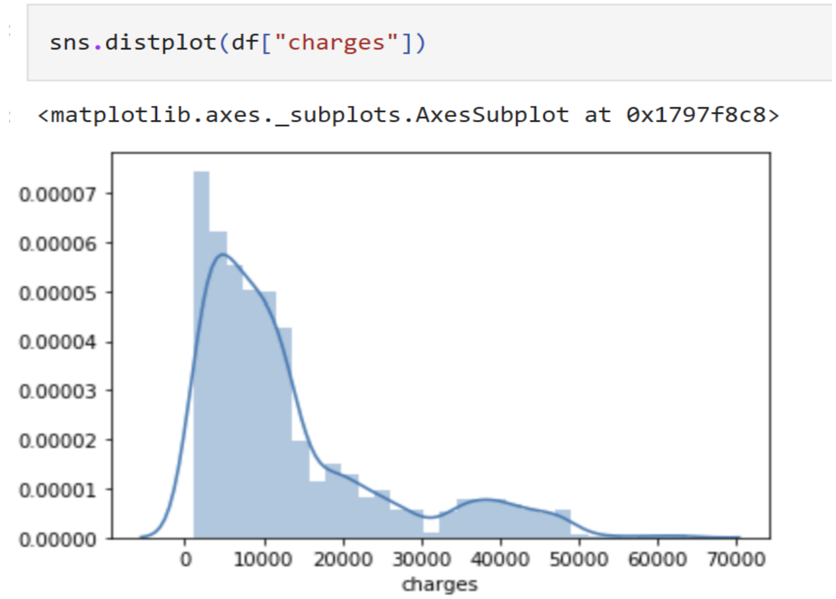
Data Visualization





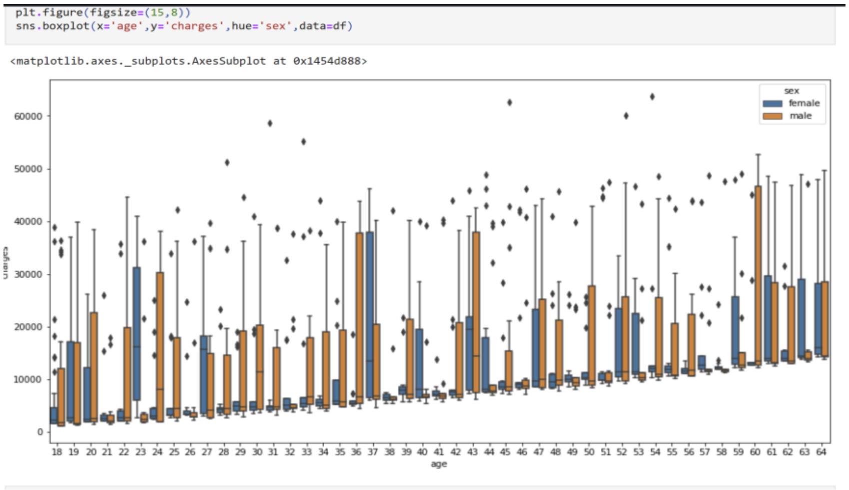
EDA Results

Here are few EDA Visualization Results.



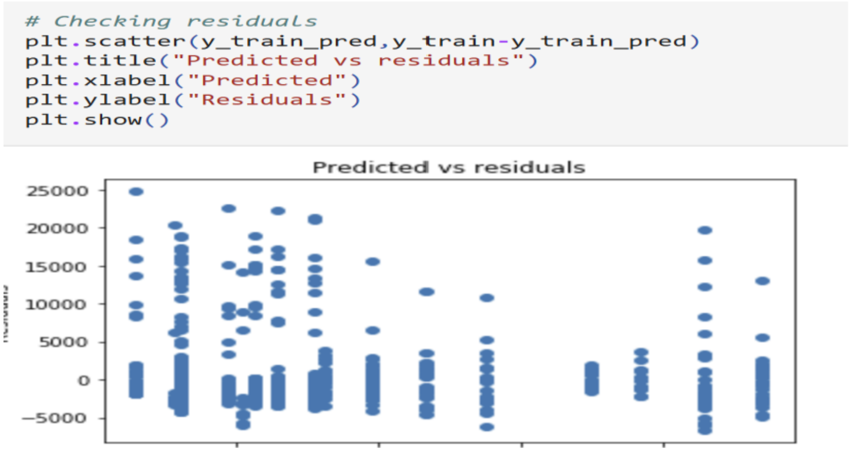
```
sns.distplot(df["charges"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1797f8c8>
```



```
plt.figure(figsize=(15,8))
sns.boxplot(x='age',y='charges',hue='sex',data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1454d888>
```



```
# Checking residuals
plt.scatter(y_train_pred,y_train-y_train_pred)
plt.title("Predicted vs residuals")
plt.xlabel("Predicted")
plt.ylabel("Residuals")
plt.show()
```

Modeling, Evaluation & Deployment

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

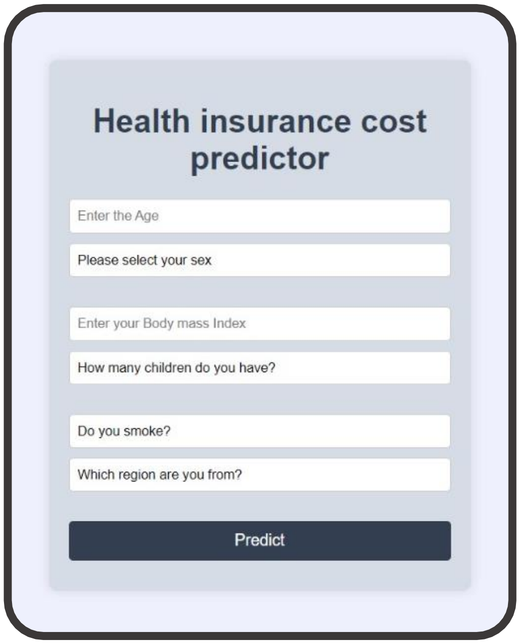
Random forest model for prediction model was deployed using flask, prioritizing scalability and real time applicability. Continuous monitoring ensures its ongoing reliability and effectiveness in the operational environment.

Model UI

User Interface

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

Initial Interface



Health insurance cost predictor

Enter the Age

Please select your sex

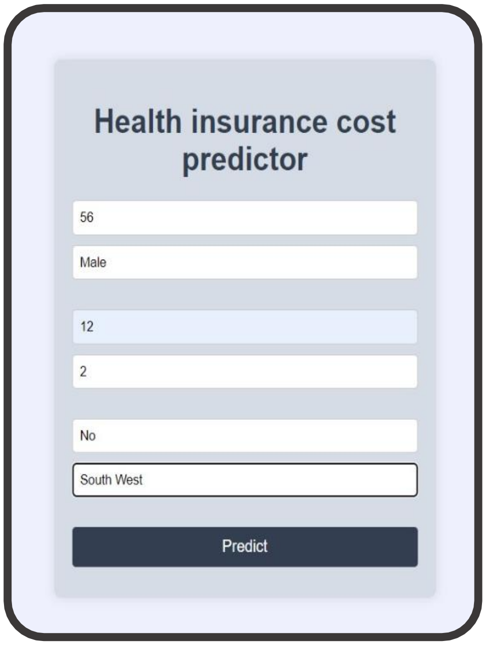
Enter your Body mass Index

How many children do you have?

Do you smoke?

Which region are you from?

Input values



Health insurance cost predictor

56

Male

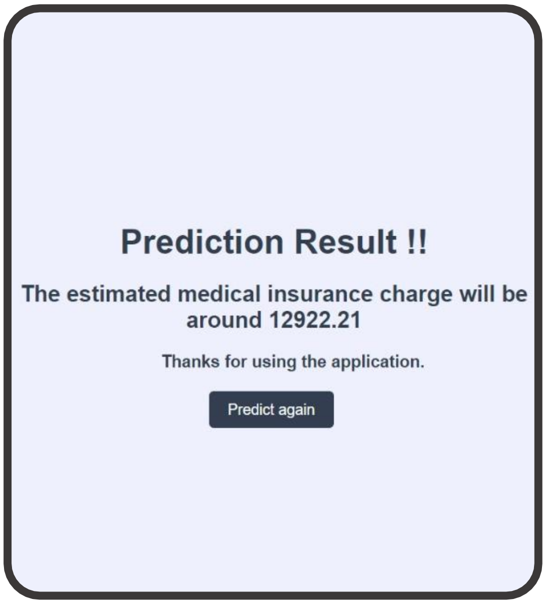
12

2

No

South West

Predicted Value



Prediction Result !!

The estimated medical insurance charge will be around 12922.21

Thanks for using the application.

Future

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.

Conclusion

In conclusion, our comprehensive efforts to develop a health insurance prediction model have successfully navigated the intricate stages of data preparation, modeling, and evaluation. The refinement of the dataset, achieved through meticulous curation and feature engineering, establishes a robust foundation for our machine learning algorithms. These algorithms, carefully selected and optimized, demonstrate high accuracy and reliability, as evidenced by rigorous evaluation metrics and cross-validation results. The model's deployment into the operational environment prioritizes scalability, interpretability, and real-time applicability, ensuring its seamless integration into dynamic healthcare settings. Continuous monitoring strategies uphold its reliability, addressing the dynamic nature of healthcare trends and allowing for timely interventions.

As we reflect on this project, it stands as a significant contribution to the evolving landscape of health insurance prediction. By enhancing financial planning accuracy, transparency, and the overall patient experience, our model aligns with the industry's changing needs. The thorough evaluation phase, considering ethical considerations, interpretability, and real-world applicability, positions our model not just as a predictive tool but as a strategic asset in shaping a more transparent, efficient, and patient-centric healthcare ecosystem. The continuous improvement mechanisms established ensure the model's adaptability, maintaining its relevance as healthcare landscapes evolve and ensuring lasting positive impacts on decision-making in health insurance.