Bapatla Engineering College: Bapatla – 522102
(Autonomous)
Department of Computer Science and Engineering

Optimizing Data-Driven Healthcare Cost Prediction Using XGBoost ML Algorithm

Presented by

M.Naga Supraja[Y21ACS499]

M.Sai Vardan[Y21ACS503] K.Durga Manohar Reddy[Y21ACS482]

K.Sekhar Venkata Prasad[Y20ACS470]

Batch No – B1

Under the Guidance of

Mr.N.Srikanth, M.Tech.,(Ph.D.)

Assistant Professor

Agenda

- Outline of the Implementation Code
- Input(s) and Output(s) of the Project
- Analysis of the Result
- Conclusion and Future Scope

Outline of the Implementation Code

- **Importing Libraries:** The first step in the implementation involves importing essential Python libraries required for data processing, visualization, machine learning, and web deployment.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

- **Data Loading:**Loaded the health insurance dataset using pandas for further processing.

```
df = pd.read_csv("/content/Medical_insurance.csv")
```

- **Data Preprocessing:** Cleaned the data and encoded categorical variables to prepare it for modeling.

```
df['sex']= label_encoder.fit_transform(df['sex'])
df['smoker']= label_encoder.fit_transform(df['smoker'])
df['region']= label_encoder.fit_transform(df['region'])
```

- **Train-Test Split**: Split the dataset into training and testing sets to evaluate model performance.

```
from sklearn.model_selection import train_test_split
x = df[["age", "bmi", "smoker", "region", "children", "sex"]]
y = df['charges']
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

- **Model Training & Predicting**:Trained the XGBoost model using the training dataset and selected parameters.

```
from xgboost import XGBRegressor
from sklearn.metrics import r2_score,mean_squared_error,mean_absolute_error
xgb_r = XGBRegressor(objective ='reg:squarederror', n_estimators = 10, seed = 123)

xgb_r.fit(x_train,y_train)# fitting the data
xgbrpred=xgb_r.predict(x_test)#predicting the price
```

- **Model Evaluation**:Assessed the model's performance using R2 score.

```
r2_score=r2_score(y_test,xgbrpred)
print("R2 Score[Testing]:",r2_score)
```

```
R2 Score[Testing]: 0.8890125853010048
```

- **Parameter Tuning:** Identify the best parameters and train the model to get high accuracy.

```
from xgboost import XGBRegressor

best_model = XGBRegressor(
    objective='reg:squarederror',
    random_state=42,
    n_estimators=700,
    max_depth=5,
    learning_rate=0.1,
    gamma=0.1,
    subsample=0.7,
    colsample_bytree=0.9
)
```

```
best_model.fit(x_train,y_train)
x_train_predict=best_model.predict(x_train)
x_test_predict=best_model.predict(x_test)
```

- **Model Evaluation after parameter tuning:** Assessed the model's performance using R-squared, RMSE, and MAE metrics.

```
r2_score2=r2_score(y_test,x_test_predict)
print("R2 Score[Testing]:",r2_score2)
```

```
R2 Score[Testing]: 0.9517504529165639
```

Input(s) of the project

- **Age**:Age of the individual (in years)
- **Sex**:Gender of the individual (male or female)
- **BMI**:Body Mass Index (numeric value based on height and weight)
- **Children**:Number of dependent children
- **Smoker**:Smoking status (yes or no)
- **Region**:Residential region (northeast, northwest, southeast, or southwest)

Output(s) of the Project

- **Insurance Premium Amount(per year):** The model predicts the health insurance cost for each individual by analyzing the relationship between various input features such as age, gender, BMI, number of children, smoking status, and region. By learning patterns from historical data, the model is able to estimate the expected insurance premium that a person is likely to be charged based on their personal and lifestyle attributes.

Analysis of the Result

- **Comparative Model Performance:** The project evaluated several machine learning regression models, including Linear Regression, Decision Tree, Random Forest, and XGBoost. While all models were trained on the same dataset, XGBoost clearly outperformed the rest in terms of prediction accuracy and error metrics.
- **Accuracy and Generalization:** The XGBoost model achieved the highest R^2 score (95.17%), indicating a strong ability to explain the variance in the dataset. This high score reflects its effectiveness in generalizing well on unseen data, making it suitable for real-world prediction tasks.
- **Error Metrics Insight:** Compared to other models, XGBoost reported the lowest MAE (1263.52) and RMSE (2721.28). These results indicate that its predictions are more consistent and closer to the actual values, significantly reducing the risk of under- or over-estimating insurance premiums.
- **Importance of Input Features:** Feature importance analysis conducted by the XGBoost model revealed that smoking status, BMI, and age are the most influential features impacting medical insurance charges. This insight helps insurers prioritize key risk factors when determining premium amounts.

Conclusion

- **Model Performance:** XGBoost effectively predicted health insurance prices with high accuracy, outperforming simpler models like Linear Regression.
- **Key Features:** The most significant factors influencing insurance charges include age, BMI, and smoking status, as determined by the model's feature importance.
- **Practical Application:** The predictive model can be used by insurance companies to estimate health insurance premiums based on individual characteristics, improving pricing accuracy.
- **Evaluation Metrics**: The model demonstrated strong performance, with evaluation metrics like R-squared and RMSE indicating a good fit to the data.
- **Real-World Impact:** The model's results can be directly applied to real-world insurance pricing systems, aiding in more personalized and accurate premium calculation.

Future Scope

- **Integration with Real-Time Health Data:** The model can be connected to hospital databases or health-tracking devices (e.g., Fitbit, Apple Health) to fetch real-time health metrics. This would enable dynamic insurance pricing based on current health trends rather than static user inputs.
- **Incorporation of Deep Learning Models:** Future work can explore advanced models like Artificial Neural Networks (ANN) or hybrid deep learning architectures that might outperform XGBoost in larger, more complex datasets.
- **Feature Engineering:** Explore additional features (e.g., family history, pre-existing conditions) to enhance prediction accuracy.
- **Feedback-Driven Retraining:** The model can be periodically updated using real-world feedback, such as final premium values or appeal outcomes, allowing it to adapt to changing market trends and customer behaviors.
- **Expansion to Other Insurance Domains:** While this project focuses on health insurance, the same predictive framework can be extended to other types such as life insurance, motor insurance, and travel insurance by training on domain-specific datasets..

Thank You