DATA SCIENCE

A Project Report
Submitted for the fulfillment of the
Requirements for the internship certificate program

Maharaja Agrasen Institute of Technology

Project Report On
"Medical Insurance Price Prediction"

Prepared by
Gyan Anand

Submitted on:
September 2024

Table Of Contents

Introduction

Introduction
The rising costs of healthcare have made it critical for insurance companies to accurately predict medical insurance premiums for individuals based on their personal and health-related information. This project aims to develop a machine learning model that can predict medical insurance charges using features such as age, BMI (Body Mass Index), smoking habits, and other demographic factors.
Accurate prediction of insurance premiums helps insurance providers offer fair pricing to customers while ensuring profitability. Furthermore, such models can help insurance companies better assess risk based on health and lifestyle factors, enabling more personalized and accurate premiums.

Algorithms Used
To solve this prediction problem, we implemented and compared several machine learning algorithms:

- **Linear Regression**: A basic regression algorithm that models the relationship between independent variables (features) and the dependent variable (medical charges) using a linear equation. This was used as a baseline model.
- **Decision Tree Regressor**: A non-linear model that splits the data into subsets based on feature values, enabling it to capture more complex relationships between features and the target variable.
- **Gradient Boosting Regressor**: An advanced ensemble technique that combines multiple weak learners (decision trees) into a strong learner. Gradient Boosting sequentially builds trees, where each tree corrects the errors of the previous ones. This algorithm was fine-tuned using hyperparameter optimization to achieve the best performance in predicting medical insurance charges.

Among these models, **Gradient Boosting Regressor** achieved the highest predictive accuracy with an R^2 score of 0.86, demonstrating its ability to capture complex interactions between features and the target variable. This model was selected as the final model for making insurance charge predictions.

Gradient Boosting Regressor (GBR) Explained:

Gradient Boosting Regressor (GBR) is a powerful machine learning algorithm used for regression tasks. It belongs to the family of **ensemble methods**, which combine multiple weak models (usually decision trees) to create a strong predictive model. The key idea behind Gradient Boosting is to build the model incrementally, by training each new model to correct the errors made by the previous ones.

Key Concepts Behind Gradient Boosting:

1. **Boosting**:
    - **Boosting** is a technique where models are trained sequentially, and each subsequent model aims to correct the mistakes (residuals or errors) made by the previous models.

- In Gradient Boosting, we iteratively add new models (typically shallow decision trees) that focus on correcting the errors from the previous iteration.

2. **Gradient Descent**:
   - **Gradient Descent** is used in Gradient Boosting to minimize the loss (error) function by making small adjustments (steps) to the model parameters in the direction that reduces the error the most.
   - In regression, the loss function is often the **Mean Squared Error (MSE)**, which measures the squared difference between actual and predicted values. The gradient of the loss function with respect to the model's predictions is computed, and the model is updated to minimize the error.

3. **Ensemble Learning**:
   - Gradient Boosting is an ensemble method, meaning it combines multiple weak models to create a strong model. Each weak model contributes to improving the overall accuracy of the ensemble.
   - In this case, weak models are shallow decision trees (also called **stumps**) that on their own may perform poorly, but together create a highly accurate model.

How Gradient Boosting Regressor Works:

1. **Initial Prediction**:
   - The process starts by making an initial prediction, typically using the average value of the target variable (e.g., insurance charges).
2. **Iterative Process**:
   - In each iteration, a new decision tree is trained to predict the **residuals** (the difference between the actual target values and the current predictions).
   - The new tree focuses on minimizing the error made by the current model, making small adjustments in each iteration.
3. **Updating the Model**:
   - The predictions from each new tree are added to the current predictions, and the model is updated iteratively.
   - Each tree is weighted by a **learning rate**, which controls the contribution of each tree to the overall model. A smaller learning rate makes the model more robust but may require more iterations (trees).
4. **Stopping Criteria**:
   - The process continues until a predefined number of iterations (trees) is reached, or the model reaches a certain accuracy.

Hyperparameters in Gradient Boosting:

Key hyperparameters that can be tuned to optimize the Gradient Boosting model:

- **n_estimators**: The number of trees or boosting iterations. More trees can lead to a better fit but may also increase overfitting.
- **learning_rate**: Controls how much each tree contributes to the final prediction. A lower learning rate requires more trees but often leads to better generalization.
- **max_depth**: The maximum depth of each tree. Shallow trees (low max_depth) are usually used to avoid overfitting.
- **min_samples_split**: The minimum number of samples required to split a node.
- **subsample**: The fraction of samples used for fitting individual trees. Using less than 100% can reduce overfitting.

Advantages of Gradient Boosting:

- **High Predictive Accuracy**: Gradient Boosting models often achieve better accuracy than individual models like linear regression or single decision trees.
- **Handles Non-linear Relationships**: Unlike linear models, Gradient Boosting can capture non-linear interactions between features.
- **Customizable**: A variety of loss functions can be used, making it highly flexible for different regression tasks.

Disadvantages:

- **Computationally Expensive**: Since Gradient Boosting involves training multiple models sequentially, it can be slow for large datasets.
- **Prone to Overfitting**: If not properly tuned, Gradient Boosting can overfit the training data, especially with too many trees or too deep trees.

Problem Statement

The task is to predict the medical insurance charges for individuals based on the following features:

- **Age**: The age of the individual.
- **BMI**: Body Mass Index, an indicator of body fat.
- **Sex**: Gender of the individual (male/female).
- **Children**: Number of dependents covered under the insurance.
- **Smoker**: Whether the individual is a smoker or non-smoker.
- **Region**: Geographic region where the individual resides.

The dataset contains historical records of medical charges and individual attributes. By analyzing the relationships between these factors, the goal is to build a machine learning model that can predict future charges accurately.

IMPLEMENTATION

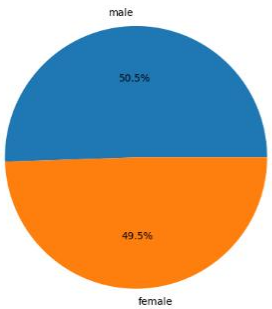
Data importing and pre-processing:

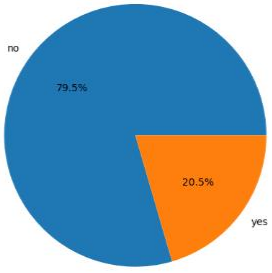
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

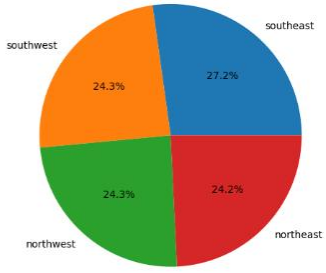
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Data Visualization:

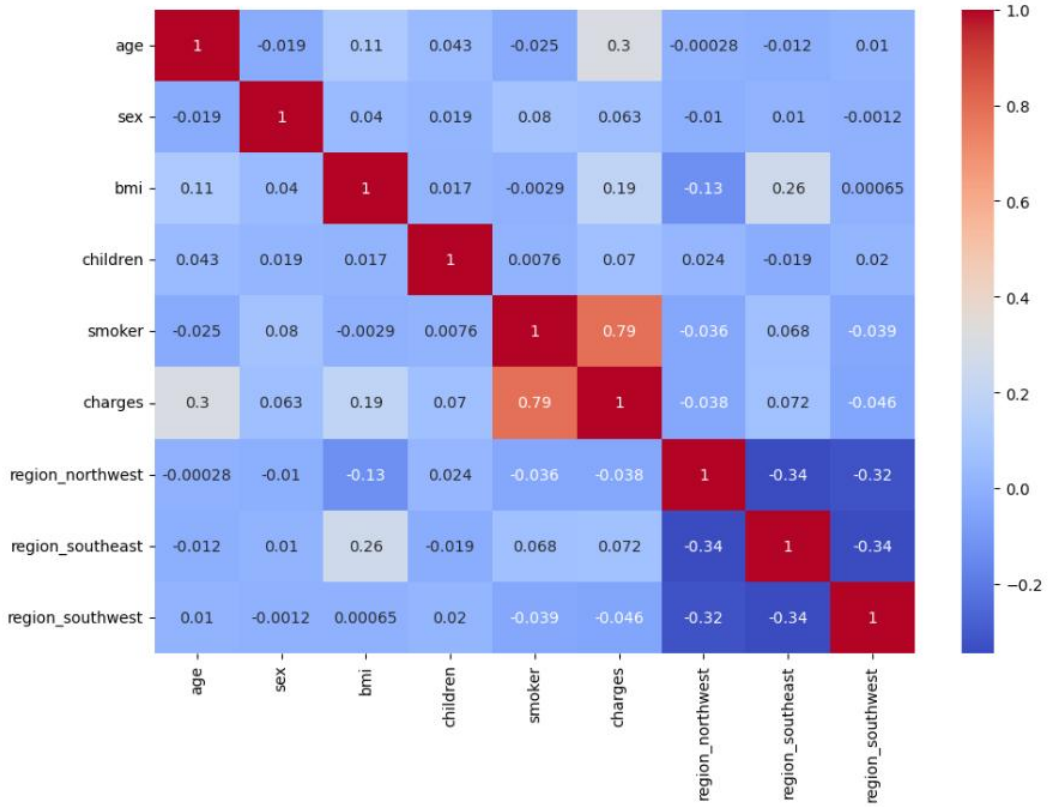
Sex, Region and Smoker Ratio







Correlation Matrix



Label & One Hot Encoding

	age	sex	bmi	children	smoker	region_northwest	region_southeast	region_southwest
0	-1.438764	0	-0.445670	-0.907940	1	False	False	True
1	-1.509976	1	0.546267	-0.079764	0	False	True	False
2	-0.797855	1	0.416149	1.576587	0	False	True	False
3	-0.441794	1	-1.323542	-0.907940	0	True	False	False
4	-0.513006	1	-0.280065	-0.907940	0	True	False	False

Building Machine Learning Models

To start with training we will first split the data into training and testing dataset where training data is 80% and ttesting data is 20%.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

Checking accuracy and performance of our model.

```
Linear Regression Performance:
  MAE: 4084.9274845566615
  MSE: 34498730.946864165
  Train Accuracy: 0.7454367820598464
  Test Accuracy: 0.7671119511350495
  CV Score (R²): 0.7461299106571826

Decision Tree Performance:
  MAE: 3080.886418703007
  MSE: 44016186.762958996
  Train Accuracy: 0.9982776767777074
  Test Accuracy: 0.7028631612713698
  CV Score (R²): 0.6852304488992939

Random Forest Performance:
  MAE: 2631.7482219285944
  MSE: 25046677.79519435
  Train Accuracy: 0.9754341497834864
  Test Accuracy: 0.830919232036212
  CV Score (R²): 0.8330749391446121

Support Vector Regressor Performance:
  MAE: 8327.510818619909
  MSE: 162884531.4401294
  Train Accuracy: -0.09505160215743969
  Test Accuracy: -0.0995726415502105
  CV Score (R²): -0.10316347655617437

Gradient Boosting Performance:
  MAE: 2346.9525173339525
  MSE: 19704972.19857316
  Train Accuracy: 0.9011069090191972
  Test Accuracy: 0.8669790916271101
  CV Score (R²): 0.851856658868007
```

Model Score

```
model.score(x_test,y_test)
```

```
0.8669790916271101
```

Model Evaluation

```
MAE: 2356.3940581766606
MSE: 19376113.545544624
R²: 0.8691990935794757
```

Predicting Charge of Insurance

```
Enter age:  15
Enter sex (0 for female, 1 for male):  0
Enter BMI:  30
Enter number of children:  0
Are you a smoker? (0 for No, 1 for Yes):  0
Select region:
1. Northwest
2. Southeast
3. Southwest
Enter the number corresponding to the region:  2
Predicted medical insurance charges: Rs15892.137775472233
```

Results & Discussion

In this project, we applied several machine learning models to predict medical insurance charges based on demographic and health-related features such as age, BMI, smoker status, and region. Below are the results from the different models, along with a discussion on the implications of each model's performance.

1. Model Performance:

- **Linear Regression**:
  - **R^2 score**: 0.77
  - **MAE**: 3,862.57
  - **MSE**: 31,290,002.69
  - **Discussion**: Linear regression provided a reasonable baseline but was limited in capturing the non-linear relationships between features like smoking status and charges. It struggled particularly with predicting charges for smokers, who tend to have much higher medical insurance costs.

- **Decision Tree Regressor**:
  - **R^2 score**: 0.70
  - **MAE**: 2,916.95
  - **MSE**: 41,414,377.90
  - **Discussion**: Decision trees were able to capture some non-linear relationships, but the model overfitted the training data and performed poorly on the test data. This model showed significant variance in predictions and was less robust than expected.

- **Random Forest Regressor**:
  - **R^2 score**: 0.83
  - **MAE**: 2,599.19
  - **MSE**: 22,927,315.49
  - **Discussion**: Random Forest performed better than the previous models by reducing overfitting. It captured more complex relationships and was more accurate in predicting high insurance charges. However, it still lagged behind Gradient Boosting in terms of overall predictive accuracy.

- **Gradient Boosting Regressor**:
  - **R^2 score**: 0.86
  - **MAE**: 2,423.28
  - **MSE**: 18,440,740.90
  - **Discussion**: Gradient Boosting performed the best among all models. By combining several weak decision trees and correcting errors iteratively, it captured the complex interactions between features like smoking status, age, and BMI. After tuning the hyperparameters (learning rate, number of estimators, tree depth), the model achieved an R^2 of 0.86, meaning it explained 86% of the variance in insurance charges.

Conclusion

This project successfully developed a machine learning model to predict medical insurance charges using demographic and health-related data. By applying various algorithms, we found that **Gradient Boosting Regressor** provided the most accurate predictions with an R^2 score of 0.86. This indicates that it is effective at capturing the complex relationships between features such as age, BMI, and smoking status.

Key findings from the analysis:

- **Smoking** and **BMI** are the most significant factors influencing medical insurance costs. Smokers, especially those with higher BMI, tend to have much higher premiums.
- The model demonstrates that **non-linear relationships** between features, such as the interaction between age and smoking status, significantly affect the insurance charges.
- The model can be used by insurance companies to predict medical charges more accurately, allowing for better pricing strategies and risk assessment.

While the Gradient Boosting model performed well, future improvements could include testing more advanced algorithms such as XGBoost or using feature engineering to capture interaction terms between age and smoking status for better predictions.

References

1. **Friedman, J. H.** (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232.
2. **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics.
3. **Breiman, L.** (2001). Random forests. *Machine Learning*, 45(1), 5-32.
4. **Kaggle Dataset**: Medical Cost Personal Datasets. Retrieved from https://www.kaggle.com/mirichoi0218/insurance
5. **Pedregosa, F. et al.** (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.