

***“PROJECT REPORT”***

***COURSE: ARTIFICAL INTELLEGENCE***

***Group Members:***

***Asma Hafiz Mir Nawaz, 2112106***

***Syeda Mahnoor Hasan, 2112135***

***Respected Teacher:***

***Usama Khalid Shiekh***

***Wali Muhammad Khubaib***

***June 2024***

|  |  |  |
| --- | --- | --- |
| **Sr.no** | **Topics** | **Page no** |
| 1 | *Summary* | 02 |
| 2 | *Purpose* | 03 |
| 3 | *Comparative analysis* | 03 |
| 4 | *Conclusion* | 06 |
|  |  |  |

**Summary:**

A key challenge for the insurance industry is to charge each customer an appropriate premium for the risk they represent. The ability to predict a correct claim amount has a significant impact on insurer's management decisions and financial statements. Predicting the cost of claims in an insurance company is a real-life problem that needs to be solved in a more accurate and automated way. Several factors determine the cost of claims based on health factors like BMI, age, smoker, health conditions and others. Insurance companies apply numerous techniques for analysing and predicting health insurance costs.

**Purpose:**

The goal is to develop accurate and automated models that can predict the cost of insurance claims for individual customers. By doing so, insurance companies can make informed management decisions, set appropriate premiums, and better manage their financial statements.

Comparative analysis:

Calculate and print the R-squared (R2) and Mean Absolute Error (MAE) for each model.

**Linear Regression (LR):**

1. **Model Type**:
   * Assumes a linear relationship between features and target variable.
2. **Computational Efficiency**:
   * **Low Complexity**: Fast to train due to its simplicity.
   * **Scalability**: Efficient with large datasets, but performance can degrade with high-dimensional data.
3. **Assumptions**:
   * **Linearity**: Assumes linear relationship between predictors and response.
   * **Independence**: Assumes independent errors.
4. **Handling non-linearity**:
   * **Limited**: Poor at capturing non-linear relationships unless using polynomial or interaction terms.
5. **Accuracy:**

**R2 score (R Sqaured): 0.74**

**MAE(Mean absolute error): 4071**

**Support Vector Regression (SVR)**

1. **Model Type**:
   * **Kernel-based**: Can use various kernels (linear, polynomial, RBF) to handle non-linearity.
2. **Computational Efficiency**:
   * **Moderate to High Complexity**: Computationally intensive, especially with large datasets or high-dimensional spaces.
   * **Scalability**: Can be slow and memory-intensive with large datasets.
3. **Assumptions**:
   * **No Strong Assumptions**: Does not assume a specific distribution for errors or a linear relationship.
4. **Handling non-linearity:**
   * **Strong**: Excellent at capturing complex non-linear relationships through kernel functions.
5. **Accuracy:**

**R2 score (R Sqaured): -0.103**

**MAE(Mean absolute error): 8291**

**Random Forest Regressor (RF)**

1. **Model Type**:
   * **Ensemble**: Combines multiple decision trees to form a robust model.
2. **Computational Efficiency**:
   * **Moderate Complexity**: Training can be computationally expensive due to the ensemble of trees.
   * **Scalability**: Can handle large datasets well, but training time increases with the number of trees.
3. **Assumptions**:
   * **No Strong Assumptions**: Does not assume linearity, normality, or independence of errors.
4. **Handling non-linearity**:
   * **Strong**: Capable of modelling complex interactions and non-linear relationships.
5. **Accuracy:**

**R2 score (R Sqaured): 0.96**

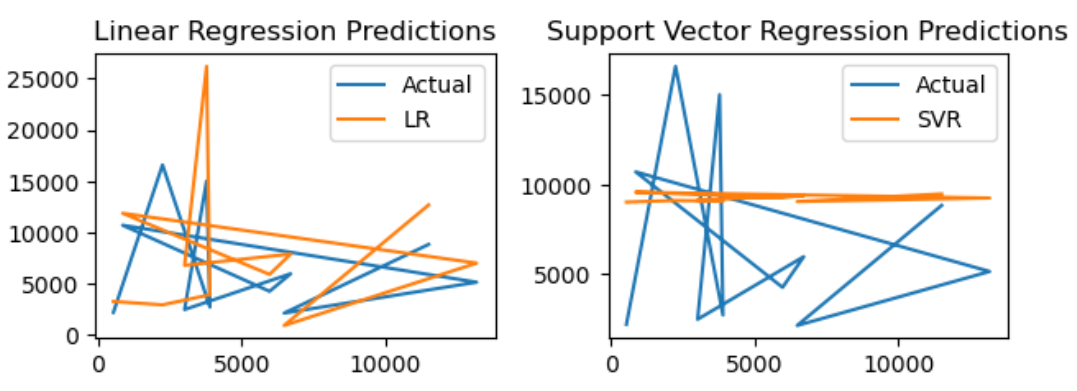
**MAE(Mean absolute error): 452.7**

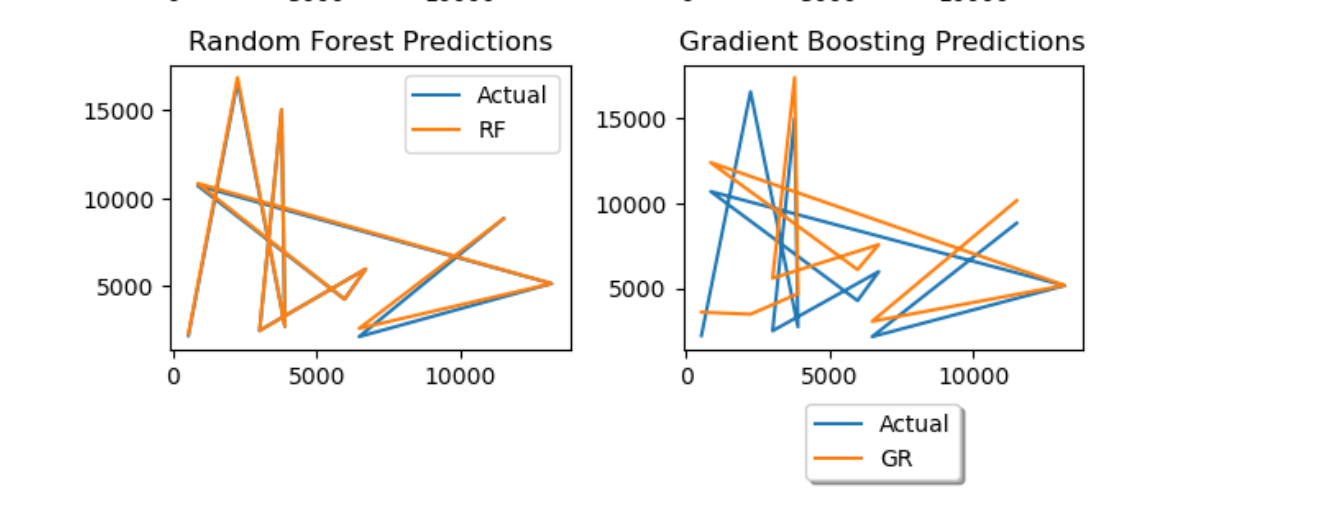
**Gradient Boosting Regressor (GBR)**

1. **Model Type**:
   * **Ensemble**: Sequentially builds models that correct errors of the previous models.
2. **Computational Efficiency**:
   * **High Complexity**: Training is computationally intensive due to the sequential nature and complexity of the boosting algorithm.
   * **Scalability**: Can be slow and resource-intensive, especially with a large number of boosting rounds.
3. **Assumptions**:
   * **No Strong Assumptions**: Flexible and does not require assumptions about the data distribution or linearity.
4. **Handling Non-linearity**:
   * **Very Strong**: Excellent at capturing intricate patterns and non-linear relationships in the data.
5. **Accuracy:**

**R2 score (R Sqaured): 0.86**

**MAE(Mean absolute error): 2666.23**





**Conclusion:**

In this project, we aimed to predict health insurance claims using various regression models. We used a dataset containing features such as age, BMI, sex, hereditary diseases, city, and job title to predict the claim amount. The following models were trained and evaluated:

* Linear Regression (LR)
* Support Vector Regression (SVR)
* Random Forest Regressor (RF)
* Gradient Boosting Regressor (GR)

The **Random Forest Regressor (RF)** and **Gradient Boosting Regressor (GR)** generally performed better than Linear Regression (LR) and Support Vector Regression (SVR) in terms of both R-squared score and Mean Absolute Error. This indicates that ensemble methods, which combine multiple decision trees, provide a more accurate prediction for this dataset. Overall random forest have the best performance of the health insurance data set with high value of r squared and lower value of mean absolute error.