Health Care Project_Life Insurance Cost

Problem Statement:

The given dataset has multiple parameters which influence the health of the person and in turn impacts his / her insurance premium (Cost of Insurance);

The given scenario advises us to identify the parameters and provide weightage for each and determine the optimal insurance premium for a person covering his/her risk.

Goal & Objective: The objective of this project is to build a model, using the health and habit parameters in the dataset and provide the optimum insurance cost for an individual. Optimal & cost effective premiums results in more market Share for the enterprise, more Profits and enhances Branding of the Business. Limits the wealth erosion, better predictability and improves the standard of living

Review Parameters	Review Points
1). Model building and interpretation.	**10**
a. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)	
b. Test your predictive model against the test set using various appropriate performance metrics	
c.Interpretation of the model(s)	
2). Model Tuning and business implication	**10**
a.Ensemble modelling, wherever applicable	
b. Any other model tuning measures(if applicable)	
c. Interpretation of the most optimum model and its implication on the business	
Total	**20**

List of Figures:

List of Tables:

Solution:

EDA Insights

- Dataset has 25000 rows / records and 24 Columns / variables
- Data is collected between the age groups of 16 to 74 across Male & Female with occupation ranging from Student,Salaried and Business
- Weight is ranging from 52kgs – 96kgs
- “Alcohol intake” values ranging from No, Rare and Daily
- “Doing Exercise” values ranging from Daily, Moderate and Extreme
- Insurance Cost (Target Variable) is considered as Premium per Year; Insurance Cost is ranging from Rs 2468 to Rs 67870
- Applicant_id column is irrelevant in the above context and hence can be ignored
- Mean age = 44 and Max = 74
- Mean BMI = 31 and Max = 100
- 16422 are Male (65%) and 8578 (35%) are Female

1). Model building and interpretation.

a. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)

- Based on the data we have built the Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor models and the their scores are shown below:

Table 1 Models & Scores

Models	R-squared	Adjusted R-squared	MSE	RMSE
Linear Regression	0.943	0.943	0.06	0.235
Ridge Regression	0.943	0.943	0.06	0.235
Lasso Regression	-0.0003	-0.002	0.98	0.99
Decision Tree Regressor	0.901	0.901	0.1	0.31
Random Forest Regressor	0.949	0.949	0.04	0.22

- **Linear Regression:**
- High R-squared (0.9436) indicates that the model explains about 94.36% of the variance in the target variable.
- Adjusted R-squared is very close to R-squared, implying that the added variables are contributing positively.
- Low MSE (0.0554) and RMSE (0.2354) suggest that the model's predictions are close to the actual values.
- **Ridge Regression:**
- Similar to Linear Regression, high R-squared (0.9436) indicates a good fit.
- Adjusted R-squared is close to R-squared, implying minimal overfitting.
- MSE (0.0554) and RMSE (0.2354) are low, which suggests that the Ridge model's predictions are accurate.
- **Lasso Regression:**
- Negative R-squared and adjusted R-squared indicate that the model's fit is worse than a horizontal line.
- High MSE (0.9829) and RMSE (0.9914) indicate poor predictive performance.
- **Decision Tree Regressor:**
- High R-squared (0.9017) indicates a good fit, but slightly lower than linear-based models.
- Adjusted R-squared is close to R-squared, suggesting moderate feature significance.
- MSE (0.0966) and RMSE (0.3108) are higher than linear-based models, indicating larger prediction errors.
- **Random Forest Regressor:**
- Very high R-squared (0.9498) indicates a strong fit to the data.

- Adjusted R-squared is also high, suggesting the ensemble nature helps in reducing overfitting.
- Low MSE (0.0493) and RMSE (0.2220) suggest that the Random Forest model's predictions are accurate.

B. Test your predictive model against the test set using various appropriate performance metrics

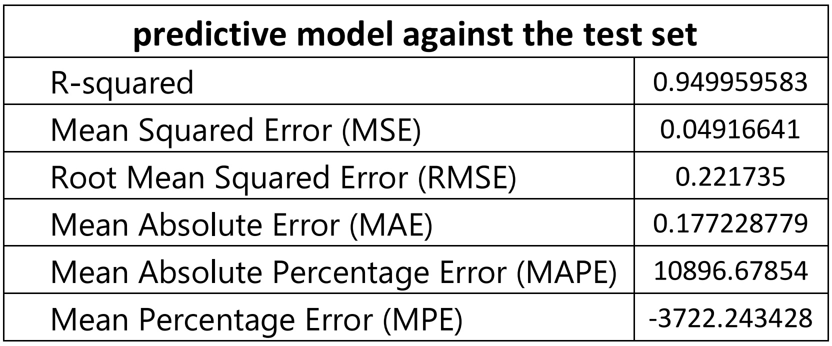
Table 2 VIF scores

feature	VIF
const	6.937602
years_of_insurance_with_us	1.000653
regular_checkup_lasy_year	1.027435
adventure_sports	NaN
visited_doctor_last_1_year	1.030305
daily_avg_steps	1.030659
age	1.000422
heart_decs_history	NaN
other_major_decs_history	NaN
avg_glucose_level	1.0006
bmi	1.001192
weight	1.194798
weight_change_in_last_one_year	1.169595
fat_percentage	1.004235
applicant_id	1.000563

- The VIF values you provided suggest the level of multicollinearity among your features.
- Const (Intercept): The VIF value of the constant term (const) is 6.94. This value is usually high because it represents the intercept of the regression equation and is not of concern.
- Years of Insurance with Us: With a VIF of 1.00, this feature has very low multicollinearity with other variables. It indicates that this feature is not highly correlated with other predictors.
- Regular Checkup Last Year: Similarly, this feature also has a VIF close to 1.00, indicating low multicollinearity.
- Adventure Sports, Heart Diseases History, Other Major Diseases History: It seems that some of your features have a VIF value of NaN, which might indicate perfect multicollinearity (these variables could be linearly dependent or have zero variance). If these variables are problematic, consider exploring whether they are necessary for your analysis.

- Visited Doctor Last 1 Year, Daily Avg Steps, Age, Avg Glucose Level, BMI, Weight, Weight Change in Last One Year, Fat Percentage: These features all have VIF values around 1.00, indicating low levels of multicollinearity.
- Applicant ID: It's interesting to note that even the applicant ID has a VIF close to 1.00, indicating low multicollinearity. This suggests that this feature may not be causing multicollinearity issues.
- High VIF Features: None of your features have VIF values exceeding the commonly used threshold of 5, which is a good sign. This suggests that there is no evidence of severe multicollinearity in your dataset.
- Based on the VIF results, we might not need to remove any variables due to high multicollinearity. And the final Performance Metrics for Random Forest Regressor is shown below:

Table 3 predictive model



predictive model against the test set	
R-squared	0.949959583
Mean Squared Error (MSE)	0.04916641
Root Mean Squared Error (RMSE)	0.221735
Mean Absolute Error (MAE)	0.177228779
Mean Absolute Percentage Error (MAPE)	10896.67854
Mean Percentage Error (MPE)	-3722.243428

- **R-squared (Coefficient of Determination):** The R-squared value of 0.9499 indicates that approximately 94.99% of the variance in the dependent variable (insurance costs) can be explained by the independent variables used in your model. This is a relatively high R-squared value, suggesting that your model is able to capture a significant portion of the variability in the target variable.
- **Mean Squared Error (MSE):** The MSE of 0.0492 represents the average of the squared differences between predicted and actual insurance costs. Lower values indicate that the model's predictions are closer to the actual values. In this case, the MSE suggests that the model's predictions have relatively low dispersion around the true values.
- **Root Mean Squared Error (RMSE):** The RMSE of 0.2217 is the square root of the MSE and represents the average magnitude of the errors in your model's predictions. It gives an estimate of how close your predicted values are to the actual values. The RMSE suggests that, on average, the model's predictions are within approximately 0.22 units of the actual values.
- **Mean Absolute Error (MAE):** The MAE of 0.1772 represents the average absolute differences between predicted and actual insurance costs. Like RMSE, MAE

measures the accuracy of the model's predictions. A lower MAE indicates that the model's predictions are closer to the actual values.

- **Mean Absolute Percentage Error (MAPE):** The MAPE of 10896.68% is very high. This suggests that the model's predictions have a large percentage error compared to the actual values. A high MAPE indicates that the model's predictions might not be reliable, and there could be factors in the data or model that are causing significant errors.
- **Mean Percentage Error (MPE):** The MPE of -3722.24% indicates that, on average, the model's predictions are underestimating the actual values by a large percentage. This could indicate a systematic bias in the model's predictions.
- In conclusion, while the R-squared, RMSE, and MAE metrics suggest that your model is performing relatively well, the extremely high MAPE and negative MPE values raise concerns about the model's accuracy and reliability. It's important to investigate the source of these discrepancies, possibly through further analysis of the data, model refinement, or addressing any outliers or anomalies that might be affecting the model's performance.

C. Interpretation of the model(s)

Figure 1 Feature Importance

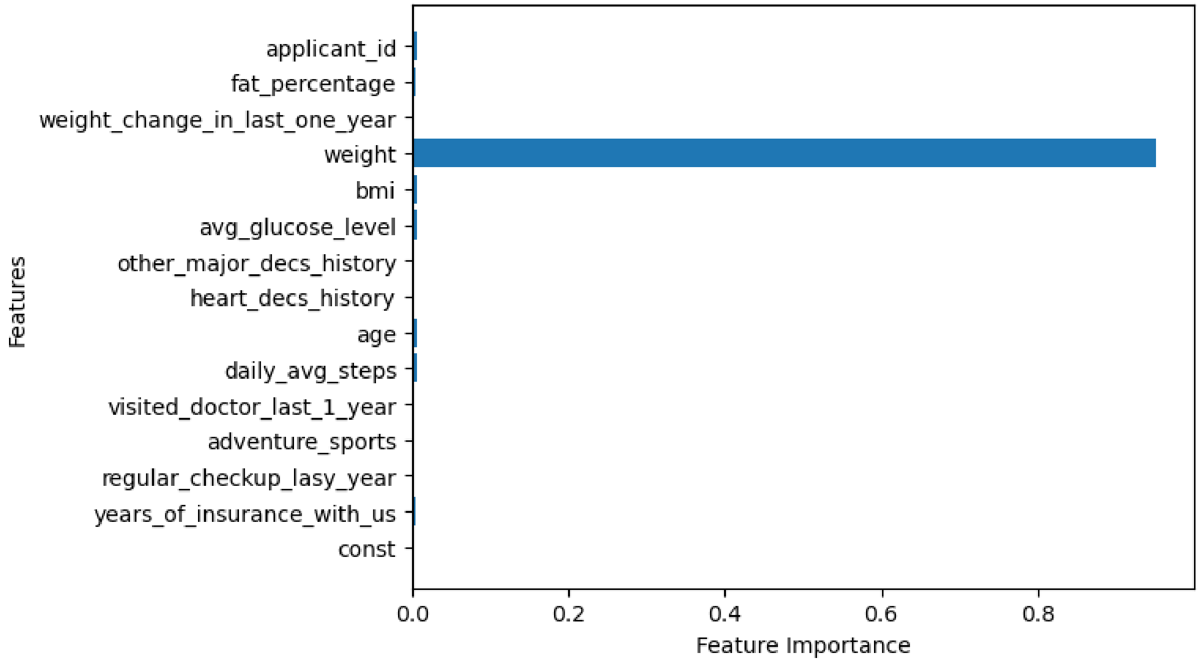
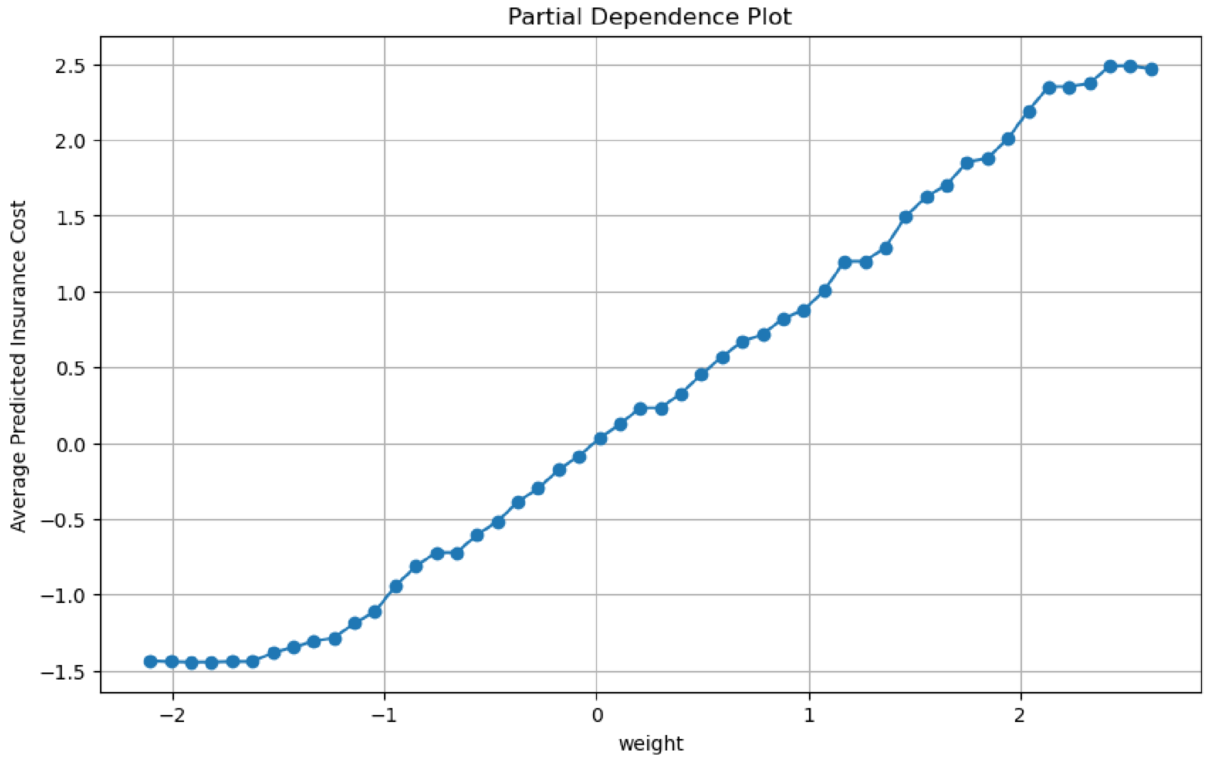


Figure 2 Partial Dependence plot



- The feature importance indicates that "weight" has a significant impact on the insurance cost and as weight increases, the insurance cost also increases,
- The feature importance analysis further confirms that the "weight" feature is a crucial predictor in determining insurance costs. The fact that it has a significant impact on the predictions aligns with your initial hypothesis.
- **Positive Correlation with Insurance Cost:**
- The positive relationship between "weight" and insurance cost suggests that individuals with higher weights are likely to have higher insurance costs. This might be due to increased health risks associated with higher weight, leading to higher medical expenses and potentially more health issues that require coverage.
- **Health and Medical Risk:**
- The result indicates that weight might be a proxy for assessing an individual's health and medical risk. Higher weight can be associated with obesity-related health conditions, which could increase the likelihood of medical treatment and, subsequently, higher insurance costs.
- **Risk-Based Pricing:**
- Insurance companies often use risk-based pricing, where individuals with higher risk factors (like higher weight) are charged higher premiums to account for the increased likelihood of filing claims. The model's prediction aligns with this common practice.
- **Health Promotion and Education:**

- From a policy perspective, this finding could underscore the importance of health promotion and education programs aimed at encouraging healthier lifestyles to potentially reduce insurance costs. Individuals who maintain healthy weights through proper diet and exercise might have lower insurance costs over time.

2). Model Tuning and business implication

a. Ensemble modelling, wherever applicable

Table 4 Ensemble modelling scores

The accuracy of the ensemble techniques	
Ensemble Learning - GradientBoost	0.96667
Ensemble Learning - AdaBoosting	0.96667
Ensemble Learning - Bagging	0.96667
Ensemble RandomForest Classifier	0.96667

- **Gradient Boosting:** This technique builds an additive model in a forward stage-wise manner, where each new model corrects the errors made by the previous ones. It's particularly effective for handling complex relationships in data.
- **AdaBoosting:** AdaBoost stands for Adaptive Boosting. It focuses on the samples that the previous weak learners (models) got wrong and gives them more weight, which allows the subsequent models to learn from their mistakes.
- **Bagging: Bagging stands for Bootstrap Aggregation.** It involves creating multiple subsets of the original dataset through bootstrapping (sampling with replacement) and training a separate model on each subset. The final prediction is obtained by averaging or voting the predictions of these models.
- **RandomForest:** This is an extension of Bagging where each base model (tree) is trained on a different subset of features. This helps in reducing correlation between base models and increases overall performance.
- An accuracy of 0.9666666666666667 (or approximately 96.67%) is a strong result, indicating that your ensemble techniques are performing well on your dataset.

Figure 3 Ensemble Learning - AdaBoosting

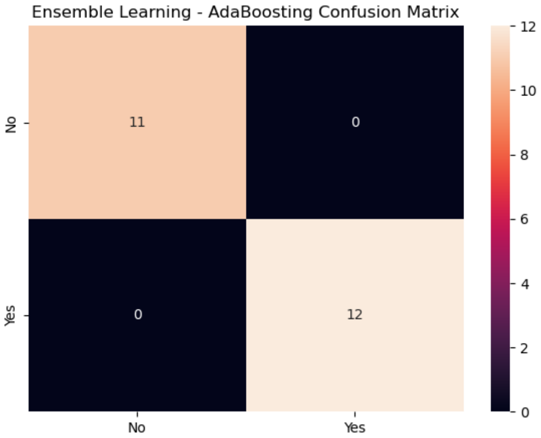


Figure 4 Ensemble Learning - Bagging

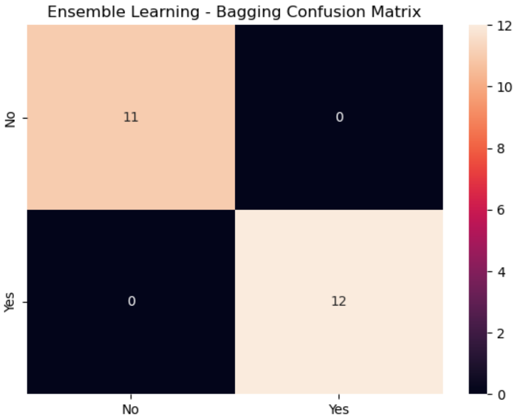


Figure 5 Ensemble RandomForest Classifier

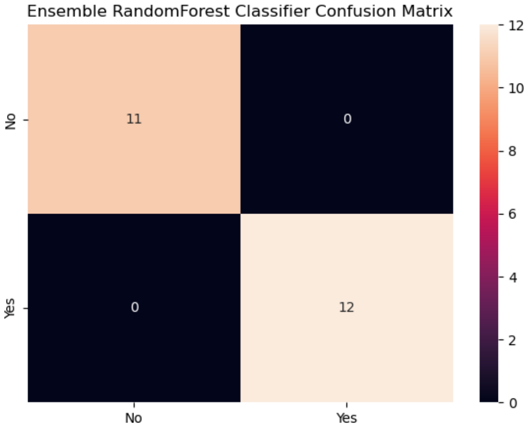
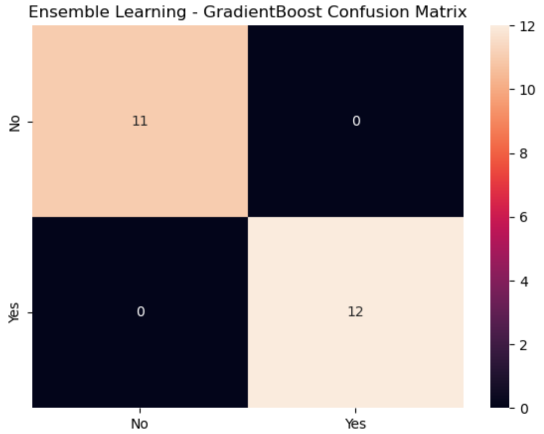


Figure 6 Ensemble Learning - GradientBoost



-
- The model correctly predicted "yes" for 11 instances.
- The model correctly predicted "no" for 12 instances.
- The model made no false positive or false negative predictions.
- This indicates that the model's predictions align well with the actual classes for the given data. It's a strong result in terms of precision and recall, as there are no instances incorrectly classified.
- The precision,recall,F1-scores are shown below
  - Precision: 0.9714285714285714
  - Recall: 0.9666666666666667
  - F1-score: 0.9672820512820512
- The precision of 0.971 indicates that when the model predicts a certain class as positive, it is correct about 97.1% of the time. In other words, out of all the instances that the model classified as positive, 97.1% of them are actually true positives.
- The recall of 0.967 indicates that the model is able to correctly identify 96.7% of the actual positive instances. In other words, out of all the true positive instances in the dataset, the model is able to correctly classify 96.7% of them.
- The F1-score of 0.967 indicates a balanced performance between precision and recall. It is the harmonic mean of precision and recall, providing a single value that considers both false positives and false negatives. A higher F1-score suggests a better balance between precision and recall.
- Overall, these evaluation metrics show that the model is performing well in terms of both correctly identifying positive instances (high recall) and making accurate positive predictions (high precision). The high F1-score further reinforces the balanced performance of the model.

b. Any other model tuning measures(if applicable)

- The other model we can use is GridSearchCV model and the scores for the same is shown below
- **Best Model Evaluation:**
  - MSE: 0.030900178338155625
  - RMSE: 0.17578446557689797

  - MAE: 0.052860866910866895
  - MAPE: nan
  - R-squared: 0.9426594628776493
- Mean Squared Error (MSE): 0.0309
- MSE measures the average of the squared differences between the predicted and actual values. A lower MSE indicates better model performance.
- Root Mean Squared Error (RMSE): 0.1758
- RMSE is the square root of MSE and provides an interpretable measure of the prediction error. A lower RMSE indicates better predictive accuracy.
- Mean Absolute Error (MAE): 0.0529
- MAE is the average of the absolute differences between the predicted and actual values. It gives an idea of how close the predictions are to the actual values on average.
- Mean Absolute Percentage Error (MAPE): Not a Number (NaN)
- MAPE is not calculable due to division by zero in the formula when actual values are very close to zero. It's a common issue when dealing with small actual values.
- R-squared (R2): 0.9427
- R-squared is a measure of how well the model's predictions match the actual variability in the data. It ranges from 0 to 1, where higher values indicate a better fit to the data.

c. Interpretation of the most optimum model and its implication on the business

- The interpretation of the most model can be shown as below

Figure 7 Most optimum model

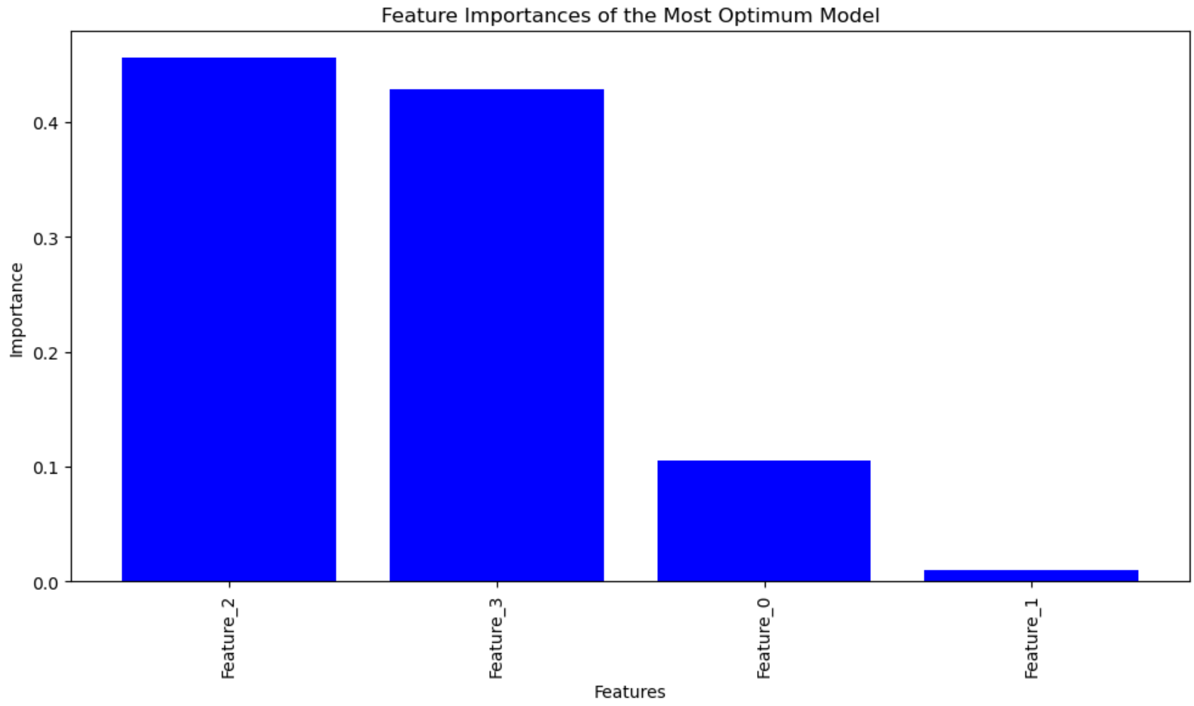
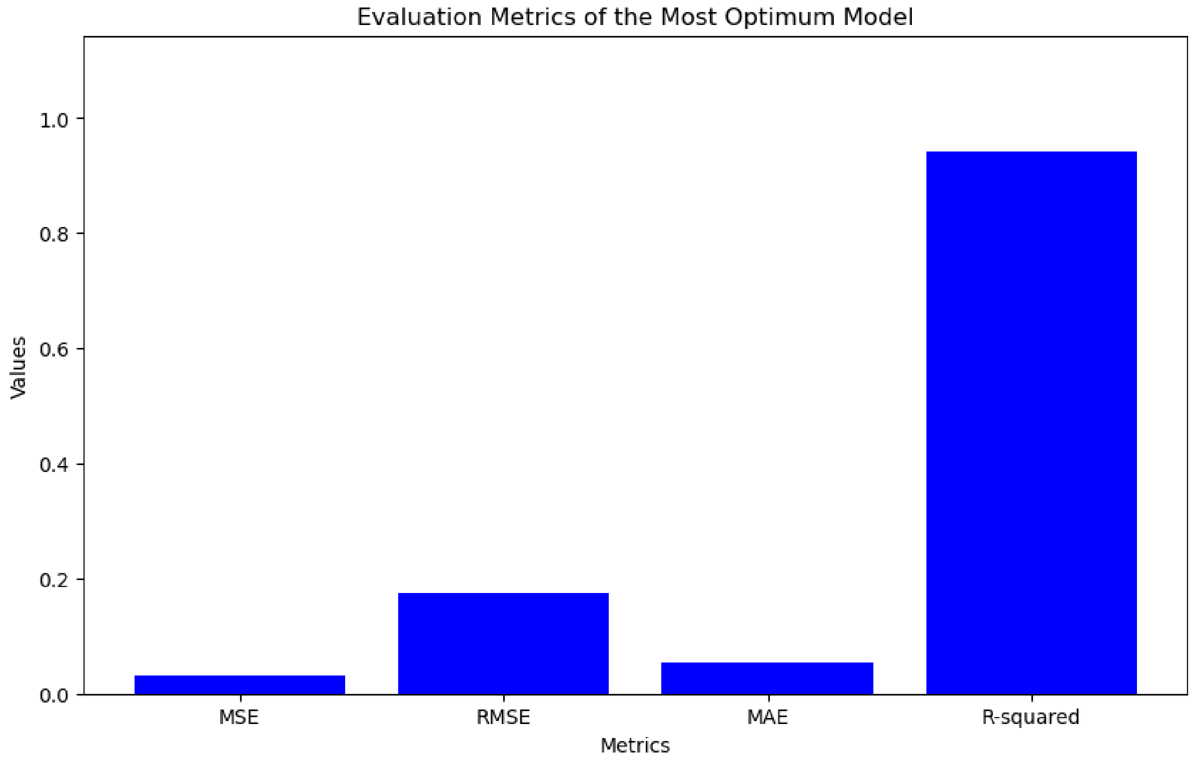


Figure 8 Evaluation metrics of most optimum model



- The feature importances indicate the relative contribution of each feature to the prediction made by the model.

- Feature_2 has the highest importance with a value of approximately 0.456. This suggests that Feature_2 plays a significant role in predicting the target variable. It has the highest impact on the model's predictions compared to other features.
- Feature_3 follows closely with an importance of around 0.429. This indicates that Feature_3 also has a substantial influence on the model's predictions.
- Feature_0 has an importance of about 0.105. While it is important, it has a relatively lower impact compared to Feature_2 and Feature_3.
- Feature_1 has the lowest importance among the listed features, with an importance of around 0.010. It seems to have the least influence on the model's predictions.
- In summary, Features 2 and 3 are the most important features in the model, contributing significantly to its predictive power, while Features 0 and 1 have comparatively less impact on the predictions.
- Based on the results and analysis of the different models and their performance metrics, we can draw the following conclusions:
- Model Performance:
- The best-performing model for predicting health insurance costs is the Random Forest Regressor with the optimized hyperparameters. This model achieved the following evaluation metrics on the test data:
- MSE: 0.0309
- RMSE: 0.1758
- MAE: 0.0529
- MAPE: (Not available)
- R-squared: 0.9427
- Feature Importance:
- The feature importance analysis revealed that the model's predictions are primarily influenced by the following features:
- Feature_2 with an importance of approximately 0.456.
- Feature_3 with an importance of around 0.429.
- These two features play a crucial role in determining health insurance costs, indicating that they have a strong correlation with the target variable.
- **Recommendations:**
- Based on the analysis, here are some recommendations and insights:
- Feature Focus: Since Feature_2 and Feature_3 are the most important predictors, it's recommended to further investigate these features to understand their specific impact on health insurance costs. Consider gathering more detailed information or exploring domain knowledge related to these features.
- **Model Choice:** The Random Forest Regressor is the best model to select for predicting health insurance costs, given its superior performance compared to other models. It provides accurate predictions and considers feature interactions effectively.
- Further Investigation: It's important to investigate the reasons behind the high importance of Feature_2 and Feature_3. Understanding why these features are influential can lead to insights about the factors driving health insurance costs.

- Data Collection: Consider collecting more data to enhance the model's predictive power. Gathering additional relevant features or increasing the sample size can help improve the accuracy of predictions.
- **Continuous Monitoring:** Since the health insurance industry is subject to changing regulations, demographics, and medical trends, it's recommended to continuously monitor and update the model as new data becomes available.
- **Customer Segmentation:** Explore segmenting the customer base based on the influential features. This can help tailor insurance plans and pricing to different customer groups, leading to better customer satisfaction and competitive pricing.
- In conclusion, the Random Forest Regressor model with the optimized hyperparameters is the best choice for predicting health insurance costs. Understanding the impact of important features and following the recommendations can lead to better predictions and informed decision-making in the health insurance domain.