Health Care Project_Life Insurance Cost

Problem Statement:

We all know that Health care is very important domain in the market. It is directly linked with the life of the individual; hence we have to be always be proactive in this particular domain. Money plays a major role in this domain, because sometime treatment becomes super costly and if any individual is not covered under the insurance then it will become a pretty tough financial situation for that individual. The companies in the medical insurance also want to reduce their risk by optimizing the insurance cost, because we all know a healthy body is in the hand of the individual only. If individual eat healthy and do proper exercise the chance of getting ill is drastically reduced.

Goal & Objective: The objective of this exercise is to build a model, using data that provide the optimum insurance cost for an individual. You have to use the health and habit related parameters for the estimated cost of insurance

Scoring guide (Rubric)	
Review Parameters	**Review Points**
1. Introduction	3
- Brief introduction about the problem statement and the need of solving it.	
2. EDA and Business Implication	5
- Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables. How your analysis is impacting the business?	
- Both visual and non-visual understanding of the data.	
3. Data Cleaning and Pre-processing	8
- Approach used for identifying and treating missing values and outlier treatment (and why)	
- Need for variable transformation (if any)	
- Variables removed or added and why (if any)	

4. Model building	8
- Clear on why was a particular model(s) chosen.	
- Effort to improve model performance.	
5. Model validation	8
- How was the model validated? Just accuracy, or anything else too?	
6. Final interpretation / recommendation	8
- Detailed recommendations for the management/client based on the analysis done.	

Table of Contents:

Contents

Table of Figures:

Table of tables:

Introduction:

Problem Understanding

- In the dynamic landscape of healthcare, where the well-being of individuals is intricately intertwined with financial considerations, the challenge of establishing optimal insurance costs emerges as a pivotal concern.
- Healthcare stands as a paramount domain, directly influencing the lives of individuals, and demands proactive solutions that align economic feasibility with comprehensive coverage.

Defining problem statement:

- The healthcare sector in India has witnessed substantial growth, encompassing various areas such as hospitals, medical devices, clinical trials, outsourcing, telemedicine, medical tourism, health insurance, and medical equipment. However, despite its significant revenue and employment contribution, there are certain challenges that need to be addressed to ensure the continued progress of the sector. One of the key concerns is the relatively low public expenditure on healthcare, standing at 1.2% of the GDP in Budget 2021. This raises questions about the accessibility and quality of healthcare services for the population.
- The problem revolves around determining an optimal insurance cost for individuals in the healthcare domain.
- This involves utilizing health and lifestyle-related data to predict the most suitable insurance premium. The aim is to strike a balance between affordability for individuals and risk management for insurance companies.

Need of the Study/Project:

- The study or project is necessary due to the critical importance of healthcare and the financial implications it carries. Health-related issues can lead to significant financial burdens, especially when not covered by insurance.
- By developing a predictive model to estimate insurance costs based on health and habit parameters, individuals can make informed decisions about coverage, and insurance companies can better assess risks and optimize their pricing strategies.
- Health crises can not only endanger personal well-being but also lead to exorbitant medical expenses that are aggravated when lacking insurance coverage.
- By crafting a predictive model grounded in health data, we equip individuals with insights into insurance costs, facilitating informed decisions, and offer insurance companies a more refined approach to risk assessment and premium calculation.

Understanding Business/Social Opportunity:

- The healthcare sector presents both a significant business opportunity and a social responsibility. As the sector is projected to grow three-fold to Rs. 8.6 trillion by 2022, businesses have a chance to capitalize on this growth by providing innovative solutions, technologies, and services that can improve the overall healthcare experience for patients and medical professionals.
- From a business perspective, this project presents an opportunity for insurance companies to enhance their competitiveness by offering personalized insurance plans. By leveraging data analytics and predictive modeling, insurance companies can attract more customers and reduce the likelihood of adverse selection.
- This approach aligns with the growing demand for tailored solutions in the insurance industry.
- On the social front, the project contributes to better healthcare access and financial security for individuals.

- It empowers people to take charge of their health and lifestyle choices by demonstrating the direct impact on insurance costs. Ultimately, this can lead to healthier lifestyles, reduced medical expenses, and improved overall well-being.
- In summary, this project addresses the need for personalized insurance pricing in the healthcare domain, benefiting both insurance companies and individuals, while also promoting healthier lifestyles and financial stability.

EDA and Business Implication

Applicant Information:

- The applicants have an average of approximately 4 years of insurance history with the company.
- Most applicants did not have a regular checkup in the last year (average value is 0.77).
- Only a small percentage of applicants (around 8%) are involved in adventure sports.
- The dataset includes three main occupations: Students, Working individuals, and Retired individuals, with students being the most common.

Health and Lifestyle:

- Cholesterol levels are categorized into five ranges, with "150 to 175" being the most common.
- Daily average steps for applicants range from 2,034 to 11,255, with an average of around 5,216.
- The average age of applicants is approximately 45 years, with a wide age range from 16 to 74.
- A small percentage (about 5%) of applicants have a history of heart-related issues.
- A slightly larger percentage (around 10%) have a history of other major health issues.
- The dataset is slightly skewed towards males, with 65.7% being male applicants.
- Average glucose levels range from 57 to 277, with an average of 167.53.
- BMI varies, with an average of 31.39, indicating a range of body types among applicants.

Insurance and Financial Information:

- The year of the last admission to the hospital ranges from 1990 to 2018.
- The dataset includes applicants from various locations, with Bangalore being the most common.
- Weight ranges from 52 to 96, with an average of 71.61.
- The majority of applicants (around 70%) are not covered by any other insurance company.
- Alcohol consumption is mostly categorized as "Rare" (around 55%).
- Exercise levels are primarily categorized as "Moderate" (around 58%).

- Weight change in the last year varies, with an average change of 2.52.
- Fat percentage varies among applicants, with an average of 28.81%.
- Insurance costs vary significantly, with an average of 27,147.41.

Univariate Analysis:

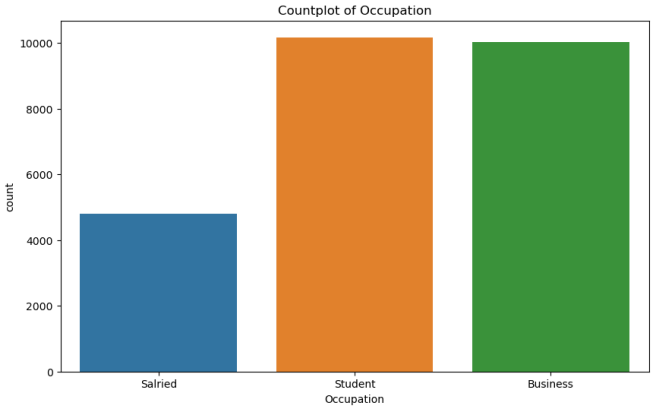

Figure 1 Univariate of Occupation interference

Occupation Inference:

- Student and Business Trends: There is a significant trend in the dataset where the occupation is categorized as "Student" and "Business Person." Both categories show a similar trend, with counts ranging from 9,500 to 10,000.
- This suggests that a substantial portion of the applicants falls into these categories, possibly indicating that many applicants are either students or involved in business activities.
- Salaried Persons: In contrast to students and business persons, the category "Salaried Person" shows a lower count, averaging around 5,000. This suggests that the dataset contains fewer individuals who are classified as salaried employees. This information could be valuable for understanding the occupational distribution within the dataset.

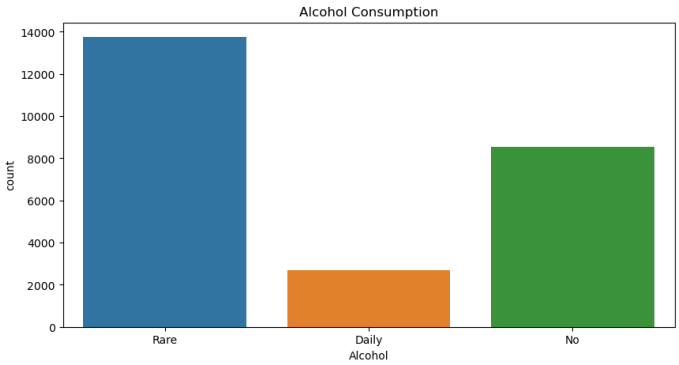

Figure 2 Univariate of Drinking habit interference

Drinking Habit Inference:

- Daily Drinking Trend: There is a noticeable trend in the drinking habit category "Daily," with counts around 2,000. This suggests that a portion of the applicants in the dataset reports consuming alcohol on a daily basis. This information could be relevant for assessing the frequency of alcohol consumption among the applicants.
- No Drinking Trend: In contrast, the category "No" in the drinking habit shows a considerably higher count, approximately 8,000. This indicates that a substantial portion of applicants in the dataset does not consume alcohol at all. This information provides insights into the prevalence of non-drinkers among the applicants.
- Rare Drinking Trend: The "Rare" category in the drinking habit exhibits a significant trend, with counts around 14,000. This suggests that a large proportion of applicants reports consuming alcohol occasionally or infrequently. Understanding the prevalence of rare or occasional drinkers can be essential for assessing the drinking habits of the applicants.

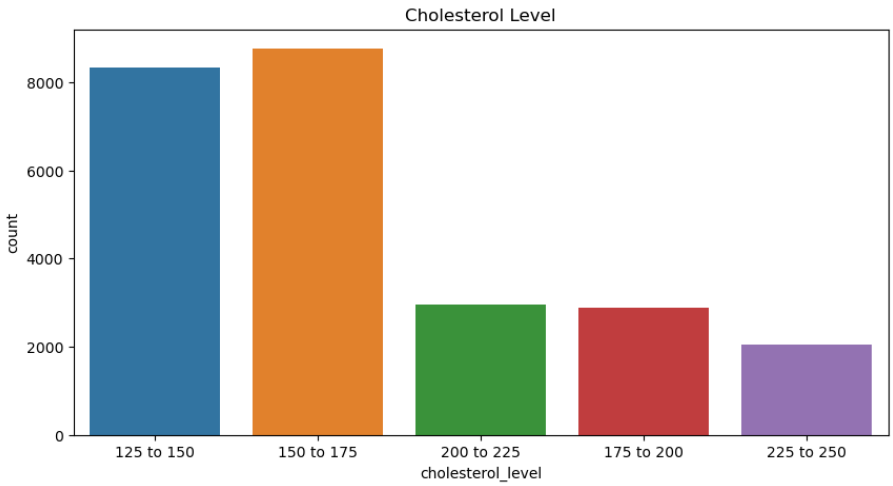

Figure 3 Univariate of cholestrol level

Cholesterol Level Inference:

- 125-150 Trend: The cholesterol level category "125-150" exhibits a significant trend, with counts around 8,000. This suggests that a substantial portion of the applicants in the dataset has cholesterol levels falling within this range. Cholesterol levels in this range are typically considered normal or healthy.
- 150-175 Trend: The category "150-175" in cholesterol levels also shows a notable trend, with counts around 9,000. This indicates that a significant number of applicants have cholesterol levels in this range. While still within the normal range, values toward the upper end of this range might warrant closer monitoring.
- 175-200 Trend: Cholesterol levels in the "175-200" category exhibit a trend, albeit with lower counts compared to the previous categories, around 3,000.
- 200-225 and 225-250 Trends: Cholesterol levels in the "200-225" and "225-250" categories both show trends, with counts around 3,000 and 2,000, respectively.

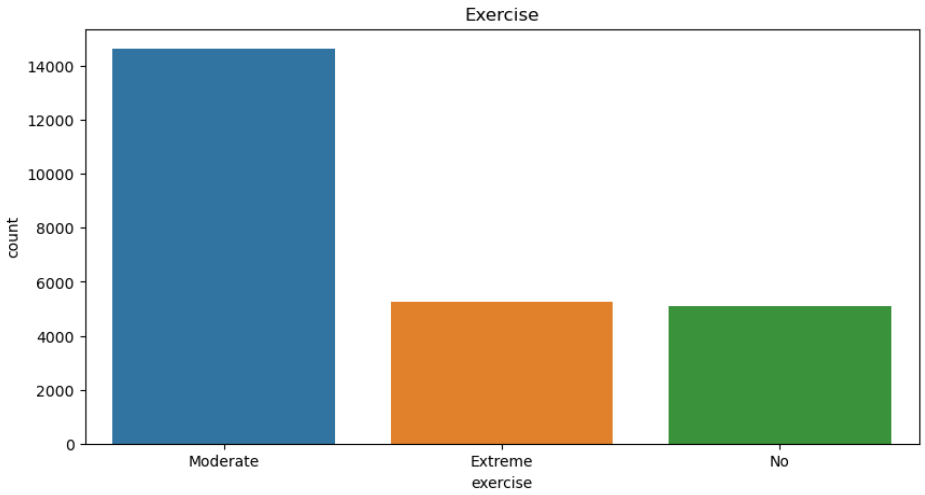

Figure 4 Univariate of exercise habit

Exercise Habits Inference:

- Moderate Exercise Trend: The exercise habit category "Moderate" shows a significant trend, with counts around 14,000. This suggests that a substantial portion of the applicants in the dataset follows a moderate exercise routine. Moderate exercise is generally associated with health benefits, including cardiovascular fitness and weight management.
- Extreme Exercise Trend: The category "Extreme Exercise" exhibits a noticeable trend, with counts around 5,000. Extreme exercise can have positive effects on physical fitness but should be done with caution to prevent injuries.
- No Exercise Trend: The "No" category in exercise habits also shows a trend, with counts around 5,000. Lack of exercise can be associated with a sedentary lifestyle, which may have implications for overall health and well-being.

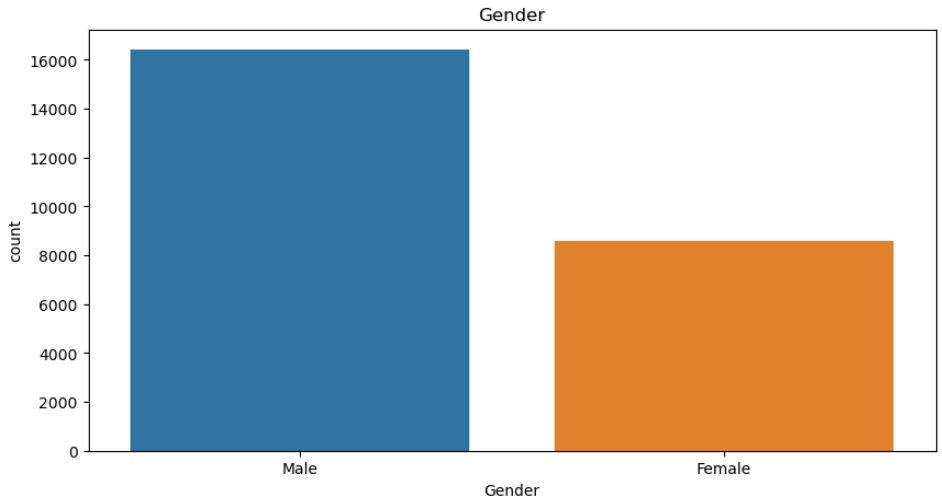

Figure 5 Univariate of Gender

Gender Distribution Inference:

- Male Trend: The gender category "Male" shows a significant trend, with counts around 16,000. This indicates that a substantial majority of the applicants in the dataset are male.
- Female Trend: The category "Female" exhibits a trend, but with counts around 8,000, indicating that there are fewer female applicants compared to males in the dataset.

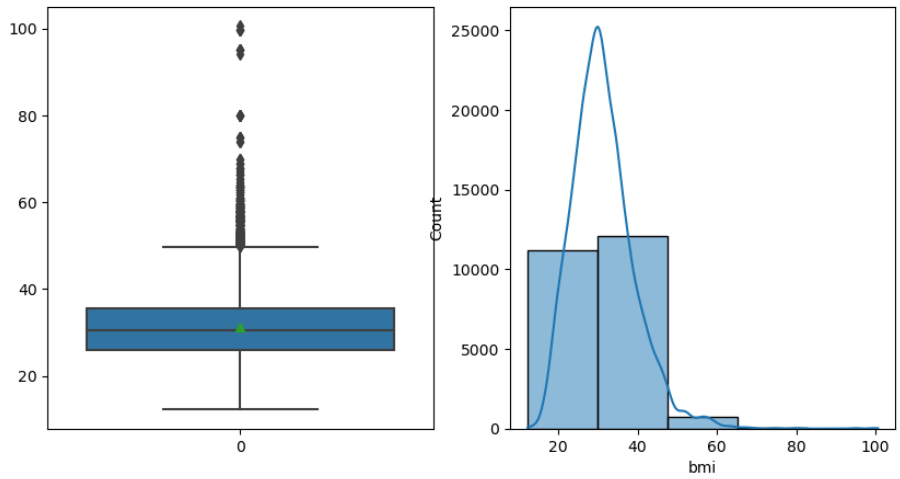

Figure 6 Univariate of BMI

BMI Distribution Inference:

- BMI of 20 Trend: The BMI category with a value of 20 exhibits a significant trend, with counts around 11,000. This suggests that a substantial portion of the applicants in the dataset has a BMI around 20, which falls within the normal weight range.
- BMI of 40 Trend: The BMI category with a value of 40 also shows a notable trend, with counts around 12,000. This indicates that a significant number of applicants have a BMI of 40, which is considered within the obese range.
- BMI of 60 Trends: The BMI category with a value of 60 exhibits a trend, with counts ranging from 1,000 to 5,000. This suggests that there are applicants with very high BMI values, indicating severe obesity.

Bivariate Analysis:

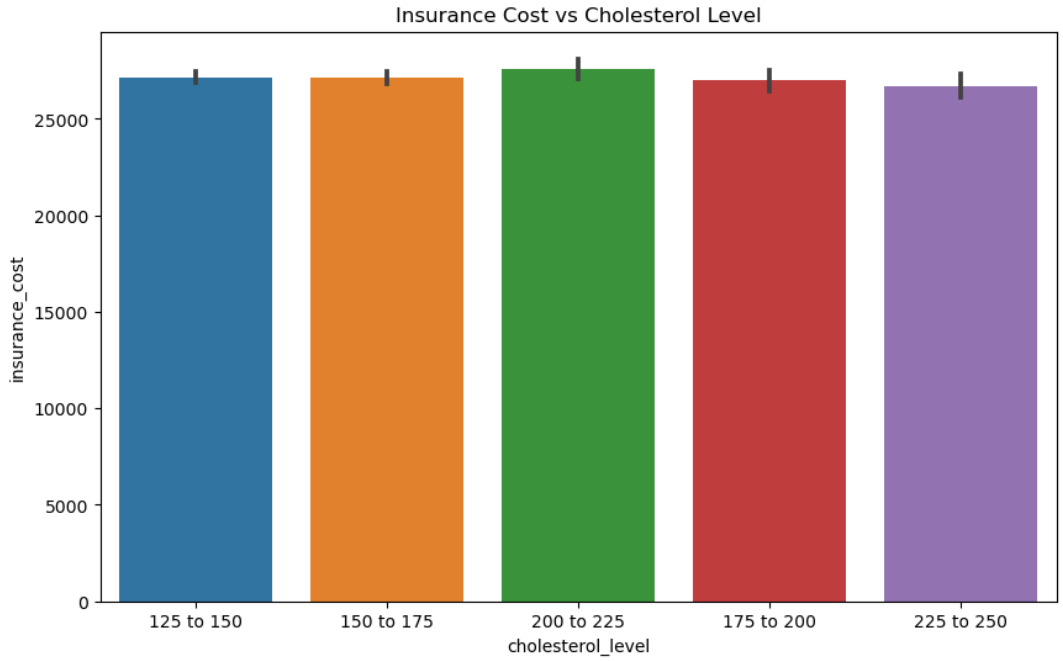

Figure 7 Bivariate of Insurance cost vs Cholestrol level

- The fact that almost all the cholesterol level groups (125-150, 150-175, 175-200, 200-225, and 225-250) have their first quartile (25th percentile) at 0 and their last quartile (75th percentile) at 27,000 suggests that there might be limited variability in insurance cost within each cholesterol level group. This could indicate that insurance costs are relatively consistent within each group.

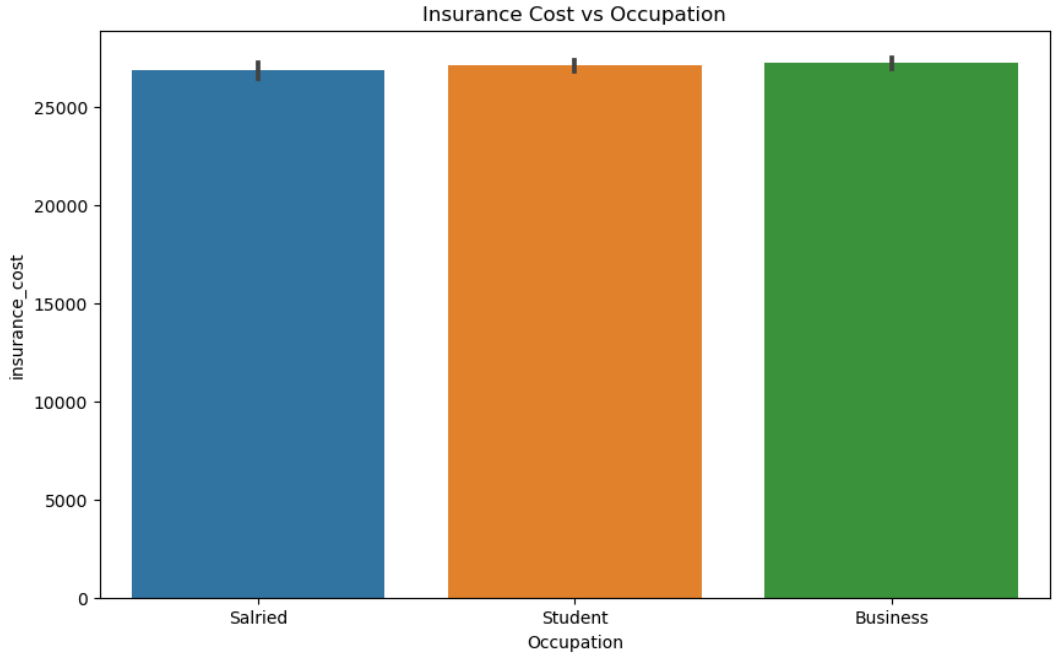

Figure 8 Bivariate of Insurance cost vs occupation

- Almost all occupation groups (salaried, business, and student) exhibit similar quartile ranges for insurance costs (0 as the 1st quartile and 27,000 as the 3rd quartile) indicates that there may not be substantial differences in insurance costs among these groups. In other words, the spread of insurance costs appears to be consistent across occupations.

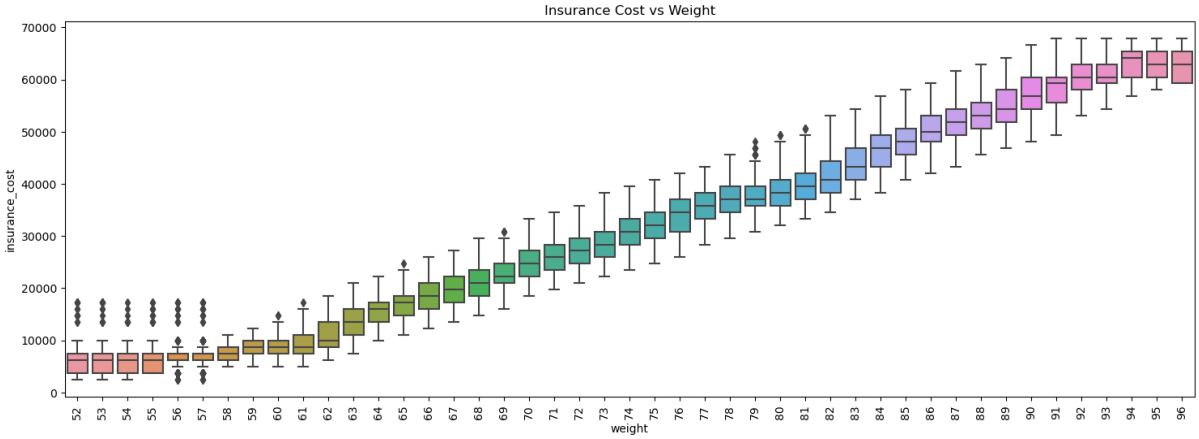

Figure 9 Bivariate of Insurance cost vs weight

- The observation that individuals with heavy weight have the highest insurance cost range implies a positive association between weight and insurance cost. In other words, it suggests that, on average, individuals with higher weight tend to have higher insurance costs compared to those with lower weight.
- The presence of significant outliers among individuals with low weight indicates that there are exceptions within this group. Some individuals with low weight have insurance costs that deviate significantly from the typical range for their weight category. These outliers may be influenced by other factors, such as underlying health conditions or lifestyle choices that contribute to higher insurance costs.
- Weight appears to be a significant factor in determining insurance costs, it's essential to recognize that other variables, such as age, gender, medical history, and lifestyle choices, can also play a role. These factors could contribute to the observed outliers and the overall variability in insurance costs within each weight category.

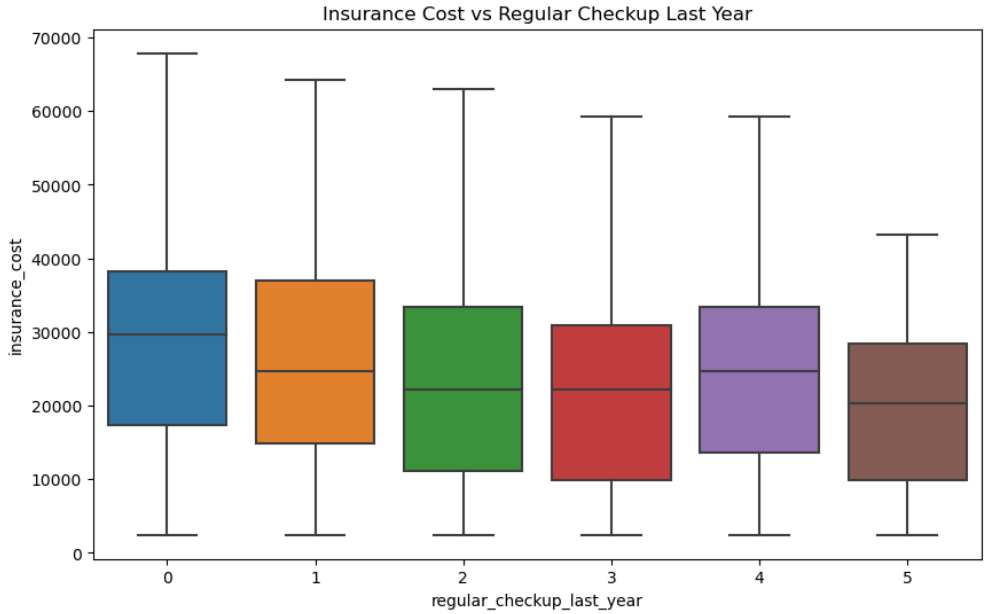

Figure 10 Bivariate of Insurance cost vs regular check-up last year

- The observation that insurance costs are higher for individuals who did not have any regular checkups last year (count = 0) indicates a negative association between insurance cost and the frequency of regular checkups. In other words, individuals who did not undergo any checkups tend to have higher insurance costs compared to those who had at least one checkup.
- The finding that insurance costs are lower for individuals who had checkup counts of 3 and 5 suggests that a moderate and consistent pattern of regular checkups may be associated with lower insurance costs. This could imply that individuals who engage in preventive healthcare measures may have fewer health-related issues, leading to lower insurance costs.
- The slight increase in insurance costs for individuals who had exactly 4 checkups may indicate that there is no clear linear relationship between the number of checkups and insurance costs. Other factors, such as the specific types of checkups or health conditions identified during those checkups, may influence this variation.

Multivariate Analysis:

- Insurance Cost and Count Relationship:
- The observation suggests that there's a slight decrease in the count of applicants as the insurance cost increases.
- This relationship could indicate that higher insurance costs might deter some applicants or lead to a decrease in the number of applicants at higher premium levels.
- It's important to explore the reasons behind this trend, such as affordability concerns or other factors influencing insurance decisions.

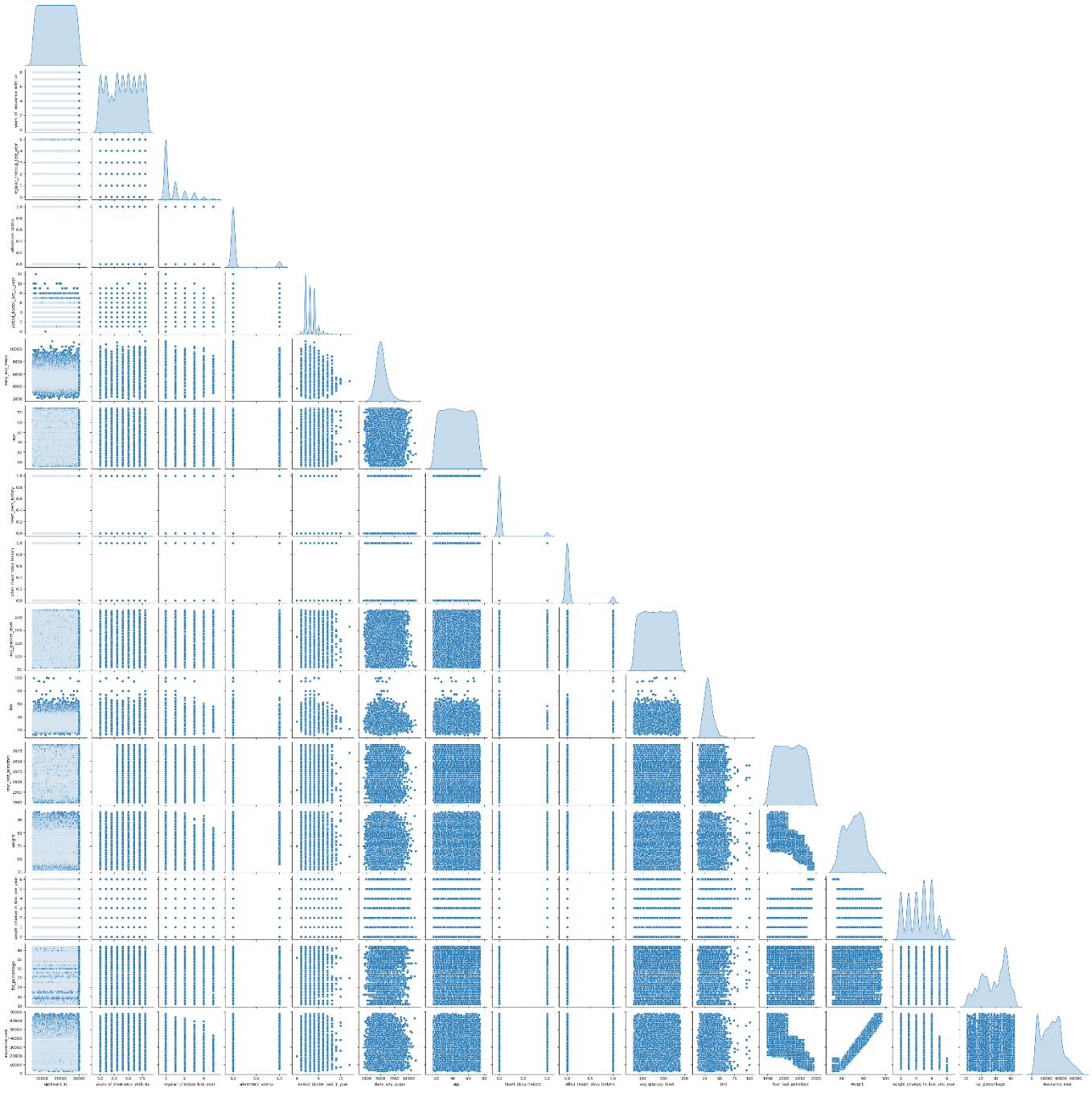

Figure 11 Multi variate analysis

- Fat Percentage and Insurance Cost:
- The observation highlights that most applicants have a fat percentage ranging from 20% to 40%, with a significant portion at 40%.
- However, the relationship between fat percentage and insurance cost is not explicitly mentioned. It might be valuable to analyze whether there's any correlation or pattern between fat percentage and insurance cost.
- Past Heart Disease and Other Major Decs:
- The observation indicates that the majority of applicants have a history of no past heart disease.

- Additionally, it's mentioned that the frequency of "Other major decs" is almost equal to the frequency of past heart disease.
- This information raises questions about the nature of these "Other major decs" and whether they are contributing to the health profile of the applicants.
- Years of Insurance with Us and Age:
- The observation suggests that the "years of insurance with us" variable is similar across different age groups.
- This could indicate that the length of the insurance relationship is not significantly impacted by the age of the applicants.
- It might be interesting to explore whether there's a correlation between the years of insurance and other variables like insurance cost or health indicators.
- Weight and Insurance Cost:
- The observation indicates a positive relationship between weight and insurance cost, stating that when the weight increases, the insurance cost also increases.
- This relationship could imply that higher weight might lead to higher insurance premiums, possibly due to the increased health risks associated with obesity.
- It's important to further analyze the strength and significance of this relationship and whether other factors also play a role.

Data Cleaning and Pre-processing

Missing value treatment:

- To address the missing values, we utilized the median value for each respective column. The median is a robust measure of central tendency that is less sensitive to extreme outliers in the data.
- We are performing missing value imputation for two columns, "Year_last_admitted" and "bmi". The missing values in these columns are being filled with the median value of their respective columns.
- This is a common imputation technique and is suitable when dealing with numerical data like "Year_last_admitted" and "bmi."
- While the median imputation method is robust and suitable for numerical data, the potential impact on data distribution, summary statistics, and subsequent analyses was taken into account. It's recognized that imputing with the median can introduce bias if the missing values are not missing at random.

Outlier treatment:

- Boxplot technique was employed to identify outliers. Boxplots provide a visual summary of the data distribution, making it easier to spot extreme values beyond quartiles.
- Outliers were identified in several key variables, including "Visited Doctor Last 1 Year," "Daily Average Steps," "BMI," and "Year Last Admitted." Outliers in these variables can significantly impact the data's distribution and statistical analyses.

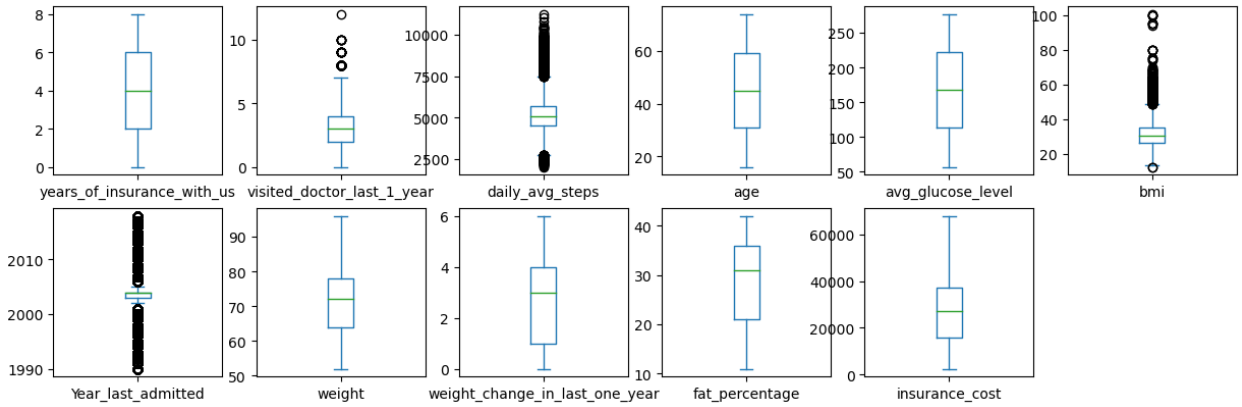

Figure 12 Before treating outlier

- Approach used for identifying upper and lower bounds for outlier detection is "Tukey's method.
- Tukey's method is robust to the presence of outliers itself. By using the IQR, which is based on quartiles (the 25th and 75th percentiles), it focuses on the central part of the data distribution, making it less sensitive to extreme values. This robustness is particularly valuable in situations where data quality varies.
- After treating the outliers the plot will like shown below:

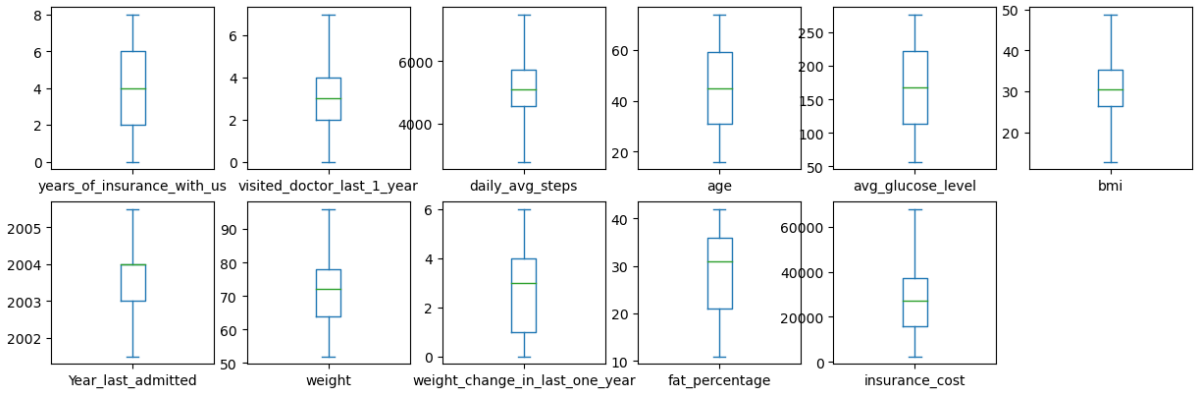

Figure 13 After treating outlier

Variable transformation:

- Variable transformation is a common technique used in data analysis and modeling to address specific issues related to the characteristics of the data. In the context of the provided code and variables, the need for variable transformation is evident for the following reasons:
- **Non-Normal Distribution:** The original variables, "age," "bmi" (Body Mass Index), and "daily_avg_steps," exhibit non-normal distributions. This non-normality can impact the validity of statistical analyses that assume normality, such as linear regression.
- **Skewness**: The distributions of these variables appear to be positively skewed, meaning they have a tail on the right side of the distribution. Skewness can lead to bias in estimates and affect the accuracy of predictive models.

- **Homoscedasticity:** Transformation can also assist in achieving homoscedasticity, where the variance of the data points is approximately constant across different levels of the predictor variables. This is particularly important for regression analysis.
- We used logarithmic (log) transformation on three variables, "age," "bmi," and "daily_avg_steps." Here are some key inferences related to the code and its output:
- **Logarithmic Transformation:** Logarithmic transformation is applied to each of the three variables using the np.log() function. This transformation is chosen due to its ability to mitigate skewness and make the data distribution more symmetric.
- **Output Format:** The transformed values are presented in the form of arrays. Each array corresponds to a variable: "X_log" contains transformed "age" values, "X1_log" contains transformed "bmi" values, and "X2_log" contains transformed "daily_avg_steps" values.
- **Interpretation:** The transformed values in "X_log," "X1_log," and "X2_log" represent the natural logarithms of the original values. These transformed variables are now on a logarithmic scale, which can facilitate linear modeling and address the original skewness.
- **Impact on Data:** It's important to acknowledge that logarithmic transformation changes the interpretation of variables. In the transformed form, the values represent the logarithms of the original values, which should be considered in any subsequent analyses or reporting.

Model Building

Model Choice:

- **Continuous Target Variable**: The choice of regression models is clear because the dependent variable in the dataset is "Insurance Cost," which is a continuous variable. Regression analysis is a suitable technique for predicting or explaining continuous outcomes, making it an appropriate choice.
- **Understanding Relationships**: Regression analysis is well-suited for understanding and quantifying the relationships between variables. In this case, it is used to explore the relationship between the "Insurance Cost" (the target variable) and the various independent variables. This is essential for gaining insights into how different factors influence insurance costs.
- **Predictive Capability**: The mention of regression analysis as a "predictive method" indicates the intention to make predictions about future health insurance costs based on historical and current data. This aligns with the practical goal of many data analytics projects, where predictions are valuable for business decisions and strategies.
- **Multiple Independent Variables**: The dataset contains 18 independent variables, suggesting that the analysis requires the consideration of multiple predictors simultaneously. Multiple regression, a type of regression analysis, is suitable for modeling relationships involving three or more variables, which matches the dataset's complexity.

- The Models that are used in this dataset are as follows.
    - **Linear Regression**
    - **Linear Regression using Stats Model**
    - **Decision Tree Regression Model**
    - **Random Forest Regression Model**
    - **XGBoost Regression Model**

Linear Regression Model :

Linear regression is a method used to predict one variable's value based on another variable. It figures out the best way to draw a straight line through your data points, so you can make predictions.

Here's why it's helpful:

- It gives you a simple math formula to make predictions.
- It tells you how strong and in which direction the relationship is between the variables.
- When you have multiple variables, it helps you see the impact of each one while keeping the others constant.
- There's also something called R2, which tells you how well your model explains the variation in your data. The closer R2 is to 1, the better your model fits the data.

The coefficients and the out of the model is shown below:

- **Coefficients**: The coefficients provided represent the impact of each independent variable on the dependent variable, "Insurance Cost." These coefficients indicate the direction and strength of the relationships. Here are some key inferences:
- **Positive Coefficients**: Variables with positive coefficients (e.g., "weight," "Location_Jaipur," "Location_Delhi") have a positive effect on insurance costs. As these variables increase, insurance costs tend to increase.
- **Negative Coefficients**: Variables with negative coefficients (e.g., "years_of_insurance_with_us," "visited_doctor_last_1_year," "daily_avg_steps") have a negative effect on insurance costs. An increase in these variables is associated with a decrease in insurance costs.
- **Magnitude**: The magnitude of the coefficients provides insight into the strength of the relationships. For example, "weight" has a large positive coefficient, indicating a substantial impact on insurance costs.
- **Categorical Variables**: Categorical variables, such as location, occupation, and smoking status, have multiple coefficients corresponding to different categories. These coefficients quantify the effect of each category on insurance costs relative to a reference category.

Table 1 LR model Score

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores
LinerRegression	0.94396	0.94358	5.62E-02	5.62E-02	0.236968	0.236964	0.190652	0.190564

- **Model Performance Metrics:** The table at the bottom of the information provides performance metrics for the Linear Regression model used. Here are some key inferences:
- **R-Squared (R2):** The R2 value measures the proportion of the variability in the dependent variable (insurance cost) explained by the independent variables. An R2 value of approximately 0.944 indicates that the model explains about 94.4% of the variability in insurance costs.
- **Mean Squared Error (MSE):** MSE measures the average squared difference between predicted and actual values. The low values (both for train and test) suggest that the model's predictions are close to the actual values.
- **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE and provides a measure of prediction accuracy. The low RMSE values indicate that the model's predictions are reasonably accurate.
- **Mean Absolute Error (MAE):** MAE measures the average absolute difference between predicted and actual values. The values are relatively small, indicating that the model's predictions are close to the actual values on average.

Overall Model Performance:

The Linear Regression model appears to perform well based on the provided metrics. It has a high R2 value, suggesting that a significant portion of the variation in insurance costs is explained by the independent variables. Additionally, the low MSE, RMSE, and MAE values indicate that the model's predictions are accurate and close to the actual values.

Linear Regression using Stats Model

We initially ran a basic model using all independent variables with the Ordinary Least Squares (OLS) method.

OLS Regression Results:

OLS Regression Results

Dep. Variable:	insurance_cost	**R-squared:**	0.944
Model:	OLS	**Adj. R-squared:**	0.944
Method:	Least Squares	**F-statistic:**	7541.
Date:	Sat, 16 Sep 2023	**Prob (F-statistic):**	0.00
Time:	13:27:28	**Log-Likelihood:**	365.60
No. Observations:	17500	**AIC:**	-651.2
Df Residuals:	17460	**BIC:**	-340.4
Df Model:	39		
Covariance Type:	nonrobust		

	coef	std errt		P>|t|	[0.025	0.975]
const	-0.0007	0.002	-0.390	0.696	-0.004	0.003
years_of_insurance_with_us	-0.0049	0.002	-2.596	0.009	-0.009	-0.001
visited_doctor_last_1_year	-0.0041	0.002	-2.213	0.027	-0.008	-0.000
daily_avg_steps	-0.0003	0.002	-0.174	0.862	-0.004	0.003
age	0.0038	0.002	2.104	0.035	0.000	0.007
avg_glucose_level	0.0004	0.002	0.218	0.828	-0.003	0.004
bmi	-0.0021	0.002	-1.083	0.279	-0.006	0.002
Year_last_admitted	-0.0263	0.002	-11.804	0.000	-0.031	-0.022
weight	0.9606	0.002	412.028	0.000	0.956	0.965
weight_change_in_last_one_year	0.0225	0.002	11.623	0.000	0.019	0.026
fat_percentage	-0.0010	0.002	-0.516	0.606	-0.005	0.003
Occupation_Salried	0.0022	0.004	0.573	0.566	-0.005	0.010
Occupation_Student	0.0026	0.003	0.851	0.395	-0.003	0.009
cholesterol_level_150 to 175	-0.0019	0.003	-0.677	0.498	-0.007	0.004
cholesterol_level_175 to 200	0.0012	0.003	0.417	0.676	-0.004	0.007
cholesterol_level_200 to 225	0.0023	0.003	0.784	0.433	-0.003	0.008
cholesterol_level_225 to 250	0.0024	0.002	0.995	0.320	-0.002	0.007
Gender_Male	0.0015	0.002	0.756	0.450	-0.002	0.005
smoking_status_formerly smoked	0.0001	0.002	0.061	0.952	-0.004	0.004
smoking_status_never smoked	0.0004	0.002	0.188	0.851	-0.004	0.005
smoking_status_smokes	-0.0023	0.002	-1.099	0.272	-0.006	0.002
Location_Bangalore	0.0028	0.002	1.140	0.254	-0.002	0.008
Location_Bhubaneswar	0.0049	0.002	1.977	0.048	4.21e-05	0.010
Location_Chennai	0.0060	0.002	2.432	0.015	0.001	0.011
Location_Delhi	0.0068	0.002	2.734	0.006	0.002	0.012
Location_Guwahati	0.0058	0.002	2.362	0.018	0.001	0.011
Location_Jaipur	0.0076	0.002	3.078	0.002	0.003	0.012
Location_Kanpur	0.0033	0.002	1.344	0.179	-0.002	0.008
Location_Kolkata	0.0038	0.002	1.574	0.116	-0.001	0.009
Location_Lucknow	0.0066	0.002	2.710	0.007	0.002	0.011
Location_Mangalore	0.0066	0.002	2.691	0.007	0.002	0.011
Location_Mumbai	0.0033	0.002	1.336	0.181	-0.002	0.008
Location_Nagpur	0.0069	0.002	2.819	0.005	0.002	0.012
Location_Pune	0.0046	0.002	1.870	0.062	-0.000	0.009
Location_Surat	0.0050	0.002	2.072	0.038	0.000	0.010
covered_by_any_other_company_Y	0.0399	0.002	21.284	0.000	0.036	0.044
Alcohol_No	-0.0008	0.003	-0.259	0.796	-0.007	0.005
Alcohol_Rare	0.0001	0.003	0.039	0.969	-0.006	0.006
exercise_Moderate	-0.0005	0.002	-0.241	0.809	-0.005	0.004
exercise_No	0.0001	0.002	0.054	0.957	-0.004	0.005

Figure 14 OLS model

Omnibus:	511.041	**Durbin-Watson:**	1.972
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	583.518
Skew:	0.392	**Prob(JB):**	1.95e-127
Kurtosis:	3.431	**Cond. No.**	5.45

- **Model Summary**: The Ordinary Least Squares (OLS) regression model has been applied to predict "insurance_cost" based on multiple independent variables. The following observations can be made:
- **R-squared (R2):** The R-squared value of 0.944 indicates that the model explains approximately 94.4% of the variability in insurance costs. This suggests a strong fit of the model to the data.
- **Adjusted R-squared**: The adjusted R-squared value, which accounts for the number of independent variables, is also 0.944, indicating that the model's performance is not inflated by excessive variables.

- **F-statistic**: The F-statistic of 7541 with a very low p-value (Prob (F-statistic) = 0.00) suggests that the overall model is statistically significant. This indicates that at least one independent variable has a significant effect on the dependent variable.
- **Coefficients**: The coefficients of the independent variables represent the change in the dependent variable (insurance cost) for a one-unit change in each independent variable while holding other variables constant. Some key inferences from the coefficients include:
- "weight" has the highest positive coefficient (0.9606), indicating that an increase in weight significantly increases insurance costs.
- "Year_last_admitted" has a substantial negative coefficient (-0.0263), suggesting that the year of last admission has a significant negative impact on insurance costs.
- Several variables have coefficients close to zero, indicating they have a relatively minor effect on insurance costs.
- Some variables have p-values (P>|t|) greater than 0.05, indicating that their effects on insurance costs may not be statistically significant.
- **Model Performance Metrics**: The model's performance metrics, including AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), indicate goodness of fit. Lower AIC and BIC values suggest a better fit, although these values are context-specific.
- **Residuals and Normality**: The Omnibus test, Jarque-Bera (JB) test, and skewness and kurtosis values provide information about the normality of the residuals. The p-values from these tests are very low, suggesting that the residuals are not normally distributed.
- **Durbin-Watson**: The Durbin-Watson statistic of approximately 1.972 indicates that there may be some autocorrelation in the residuals. This should be considered when assessing the model's assumptions.
- **Interpretation**: Overall, the model appears to be well-fitted to the data, explaining a substantial portion of the variation in insurance costs. However, some caution is needed due to potential issues with non-normality of residuals and autocorrelation. Further diagnostics and model refinement may be necessary to address these concerns.
- **Coefficient Significance**: It's essential to consider both the magnitude and statistical significance of coefficients when interpreting them. Variables with both high magnitude and statistical significance have a more substantial impact on insurance costs.
- In summary, the OLS regression model shows promising results in explaining insurance costs based on multiple independent variables. However, further analysis and validation are required to assess the model's assumptions and robustness.

The VIF of the variables are as shown below:

VIF values:

const	1.000730
years_of_insurance_with_us	1.090376
visited_doctor_last_1_year	1.045234
daily_avg_steps	1.064708
age	1.002473
avg_glucose_level	1.001953
bmi	1.192941
Year_last_admitted	1.549936
weight	1.691982
weight_change_in_last_one_year	1.160056
fat_percentage	1.236223
Occupation_Salried	4.534625
Occupation_Student	2.865883
cholesterol_level_150 to 175	2.344480
cholesterol_level_175 to 200	2.608608
cholesterol_level_200 to 225	2.689321
cholesterol_level_225 to 250	1.733534
Gender_Male	1.234814
smoking_status_formerly smoked	1.456380
smoking_status_never smoked	1.552204
smoking_status_smokes	1.391954
Location_Bangalore	1.915250
Location_Bhubaneswar	1.901829
Location_Chennai	1.859566
Location_Delhi	1.869260
Location_Guwahati	1.905746
Location_Jaipur	1.909648
Location_Kanpur	1.884532
Location_Kolkata	1.877774
Location_Lucknow	1.875452
Location_Mangalore	1.917181
Location_Mumbai	1.899275
Location_Nagpur	1.903606
Location_Pune	1.850258
Location_Surat	1.853867
covered_by_any_other_company_Y	1.087369
Alcohol_No	2.847407
Alcohol_Rare	2.785744
exercise_Moderate	1.588328
exercise_No	1.601009

Figure 15 VIF values

Interpretation of VIF (Variance Inflation Factor) Values:

- VIF values are used to assess multicollinearity, which is the presence of high correlation among independent variables in a regression model. Here's a simple interpretation of the provided VIF values:
- Low Multicollinearity: Generally, VIF values below 5 indicate low multicollinearity, and values below 10 are considered acceptable in some cases.
- High Multicollinearity Concerns: Values well above 10 can raise concerns about high multicollinearity, which can affect the model's stability and interpretability.
- Interpretation for Variables: Looking at the VIF values for specific variables:
- Variables like "Occupation_Salried," "Alcohol_No," and "Alcohol_Rare" have relatively high VIF values (above 2.5), indicating some degree of multicollinearity.
- Other variables generally have lower VIF values (around or below 2), suggesting lower multicollinearity.

- **Action**: When dealing with high multicollinearity, one might consider options such as removing one of the correlated variables, combining them into a single variable, or using dimensionality reduction techniques like principal component analysis (PCA) to address the issue.
- In summary, the VIF values provided suggest that most of the independent variables have relatively low multicollinearity, but a few variables exhibit higher multicollinearity. Further investigation into these specific variables may be necessary to determine if any action is required to improve the model's stability and interpretability.
- The scores of the final model is as follows :

OLS Regression Results

Dep. Variable:	insurance_cost	R-squared:	0.944
Model:	OLS	Adj. R-squared:	0.944
Method:	Least Squares	F-statistic:	2.262e+04
Date:	Sat, 16 Sep 2023	Prob (F-statistic):	0.00
Time:	13:27:41	Log-Likelihood:	352.66
No. Observations:	17500	AIC:	-677.3
Df Residuals:	17486	BIC:	-568.5
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0007	0.002	-0.386	0.700	-0.004	0.003
years_of_insurance_with_us	-0.0048	0.002	-2.565	0.010	-0.008	-0.001
visited_doctor_last_1_year	-0.0042	0.002	-2.300	0.021	-0.008	-0.001
age	0.0037	0.002	2.069	0.039	0.000	0.007
Year_last_admitted	-0.0262	0.002	-11.787	0.000	-0.031	-0.022
weight	0.9608	0.002	412.376	0.000	0.956	0.965
weight_change_in_last_one_year	0.0226	0.002	11.662	0.000	0.019	0.026
Occupation_Salried	0.0033	0.002	1.696	0.090	-0.001	0.007
Occupation_Student	0.0011	0.002	0.566	0.571	-0.003	0.005
Location_Bhubaneswar	-1.578e-05	0.002	-0.009	0.993	-0.004	0.004
Location_Lucknow	0.0018	0.002	1.008	0.314	-0.002	0.005
covered_by_any_other_company_Y	0.0397	0.002	21.239	0.000	0.036	0.043
exercise_Moderate	-0.0005	0.002	-0.236	0.814	-0.005	0.004
exercise_No	0.0001	0.002	0.058	0.954	-0.004	0.005

Omnibus:	516.770	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	590.919
Skew:	0.394	Prob(JB):	4.83e-129
Kurtosis:	3.435	Cond. No.	2.19

Figure 16 OLS 27 results

- Model Fit: The Ordinary Least Squares (OLS) regression model explains approximately 94.4% of the variation in "insurance_cost," as indicated by the high R-squared value of 0.944.
- Significant Predictors: Several independent variables have statistically significant effects on insurance costs, including "years_of_insurance_with_us," "visited_doctor_last_1_year," "Year_last_admitted," "weight," "weight_change_in_last_one_year," and "covered_by_any_other_company_Y."

- Non-Significant Predictors: Some variables, like "Occupation_Salried," "Occupation_Student," "Location_Bhubaneswar," "Location_Lucknow," "exercise_Moderate," and "exercise_No," do not appear to have a statistically significant impact on insurance costs, as their p-values exceed the typical significance level of 0.05.
- Model Validity: The overall model's validity is supported by a significant F-statistic (2.262e+04) and a low p-value (Prob (F-statistic) = 0.00), indicating that at least one independent variable is significantly related to insurance costs.

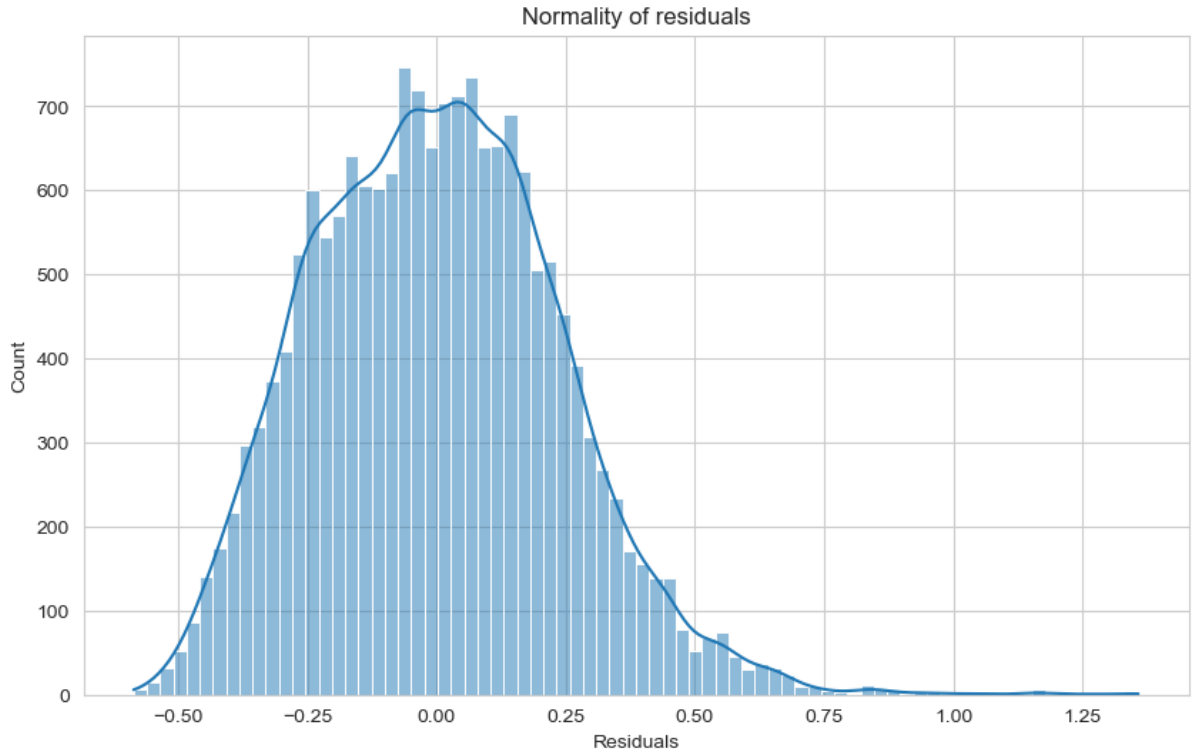

Figure 17 Normality of residuals

- In the normality of residuals plot, the distribution of residuals exhibits a distinctive shape resembling a mountain. Initially, the plot starts from 0 and gradually rises, reaching its highest point at approximately 700 in the range of 0.00 to 0.25. After this peak, it descends as we move to the range of 0.25 to 0.5. Finally, the plot becomes nearly flat at around 0 in the interval from 0.75 to 1.25.
- This pattern suggests that the distribution of residuals deviates from a perfect normal distribution. It indicates potential non-normality in the distribution of errors, which is an important assumption for Linear Regression. Further investigation and potential corrective measures may be necessary to address this departure from normality in the residuals.

Assumptions:

- These assumptions are essential conditions that should be met before we draw inferences regarding the model estimates or use the model to make a prediction.

- For Linear Regression, we need to check if the following assumptions hold:
  - **Linearity**
  - **Independence**
  - **Homoscedasticity**
  - **No strong Multicollinearity**
  - **Normality of error terms**
- After confirming that all the assumptions are met, we can now make prediction on the dataset.

Final Equation:

```
insurance_cost = -0.0006917539385144721 + -0.004797494374777512 * (
years_of_insurance_with_us ) + -0.004156212604606859 * (
visited_doctor_last_1_year ) + 0.0037235140378640078 * ( age ) + -
0.02624661056472688 * ( Year_last_admitted ) + 0.9607902533372676 *
( weight ) + 0.022602809427440997 * ( weight_change_in_last_one_year
) + 0.003335990662728305 * ( Occupation_Salried ) +
0.0011102502448638507 * ( Occupation_Student ) + -
1.5777408496815823e-05 * ( Location_Bhubaneswar ) +
0.001807633378094069 * ( Location_Lucknow ) + 0.03974953175728668 *
( covered_by_any_other_company_Y ) + -0.0005321367916564269 * (
exercise_Moderate ) + 0.00013183870657946313 * ( exercise_No )
```

Decision Tree Model

- A Decision Tree is a versatile supervised learning technique used for both classification and regression tasks. It's akin to a tree-shaped flowchart, where internal nodes represent dataset features, branches depict decision rules, and each leaf node offers an outcome prediction.
- In Decision Tree Regression, the model inspects object features and constructs a tree-like structure to predict future data, yielding continuous output. An advantage of decision trees and ensemble methods is their insensitivity to data variance, eliminating the need for feature scaling. Consequently, this model analysis is conducted on the original, untreated data.
- All 40 variables resulting from prior data processing are input into the Decision Tree Regressor model without fine-tuning hyperparameters. This approach simplifies the analysis while preserving the model's potential to make meaningful continuous predictions.

Model Evaluation:

Table 2 Decision tree reg scores

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores

DecisionTreeRegressor	1	0.908705	0.00E+00	1.86E+07	0	4317.518946	0	3348.253333

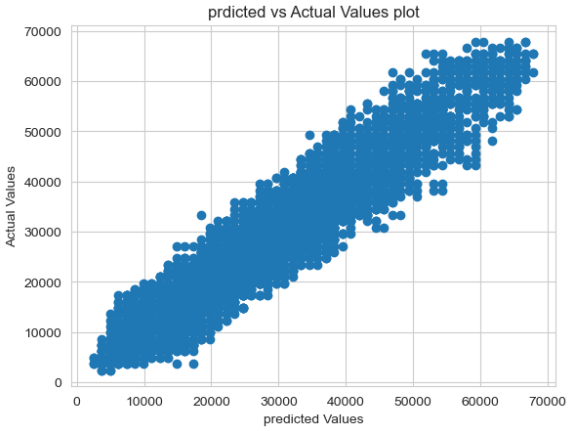

Figure 18 Desicion tree test pred vs actual values

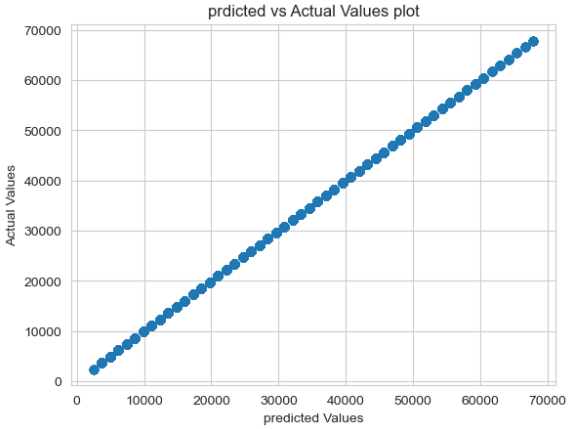

Figure 19 Desicion tree train pred vs actual values

- R-squared (R^2): The model explains 100% of the variance in the training data, indicating a perfect fit. In the test data, it still performs well, explaining approximately 90.87% of the variance. This suggests the model captures a significant portion of the variation in the target variable.
- Mean Squared Error (MSE): The model achieves a very low MSE of approximately 0 in the training data, indicating minimal error in predictions. In the test data, the MSE is around 1.86E+07, which is relatively high but should be evaluated in the context of the problem's scale.
- Root Mean Squared Error (RMSE): The RMSE for training data is 0, which is ideal. In the test data, it's approximately 4317.52, indicating the average prediction error in the same units as the target variable.
- Mean Absolute Error (MAE): The model achieves a MAE of 0 in the training data, implying precise predictions. In the test data, the MAE is approximately 3348.25, representing the average absolute prediction error.

- In summary, the Decision Tree Regressor demonstrates remarkable performance on the training data, achieving near-perfect fit and low errors. While it still performs well on the test data, there's a notable increase in error metrics, suggesting some level of overfitting or sensitivity to the test dataset. Further evaluation and potentially some model tuning may be necessary to generalize the model's performance better.

Random Forest Model

- The Random Forest Regressor is a powerful ensemble model, belonging to the non-linear category. In contrast to linear regression, which heavily relies on the correlation between independent variables and the target variable, Random Forest takes a different approach. It's a tree-based algorithm that assembles a group of decision trees to collectively refine predictions.
- In tree-based algorithms like Random Forest, it's crucial to fine-tune hyperparameters to prevent the formation of overly complex trees, which can lead to overfitting.
- For this analysis, all 40 variables resulting from data preprocessing are fed into a Decision Tree Regressor model without any hyperparameter tuning. This simplified approach aims to explore the model's performance without extensive parameter optimization, providing insights into its inherent capabilities. Further refinement may be needed to optimize the model's predictive accuracy.

Model Evaluation:

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores
RandomForestRegressor	0.993358	0.953346	1.37E+06	9.53E+06	1168.495895	3086.428904	924.001571	2455.95616

Table 3 Random forest scores

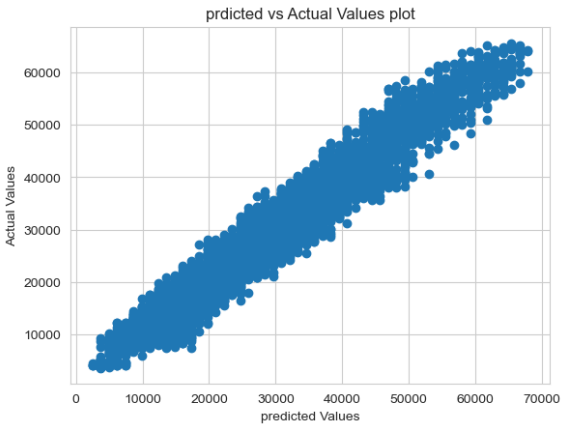

Figure 20 Random forest test pred vs actual values

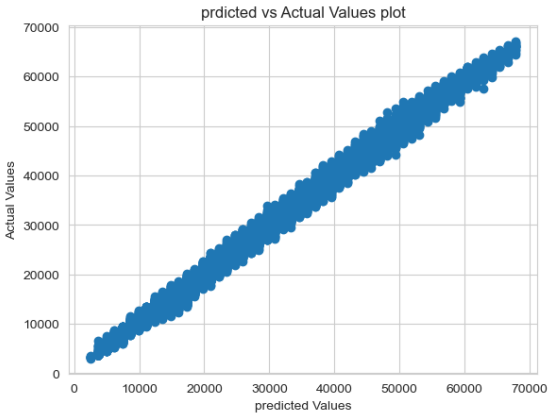

Figure 21 Random forest train pred vs actual values

- R-squared (R^2): The model performs exceptionally well, explaining approximately 99.34% of the variance in the training data and about 95.33% in the test data. These high R-squared values suggest that the model effectively captures the variation in the target variable.
- Mean Squared Error (MSE): In the training data, the model achieves a very low MSE of approximately 1.37E+06, indicating minimal prediction error. In the test data, the MSE is around 9.53E+06, which is relatively higher but should be considered in the context of the problem's scale.
- Root Mean Squared Error (RMSE): The RMSE for training data is approximately 1168.50, which signifies the average prediction error in the same units as the target variable. In the test data, the RMSE is approximately 3086.43.
- Mean Absolute Error (MAE): The model achieves a MAE of approximately 924.00 in the training data, indicating a small average absolute prediction error. In the test data, the MAE is approximately 2455.96.
- In summary, the RandomForestRegressor model demonstrates outstanding performance, with high R-squared values and relatively low error metrics in both training and test datasets. This suggests that the model effectively captures complex relationships between the independent variables and the target variable, making it a strong candidate for predictive tasks. Further evaluation and fine-tuning may still be necessary to optimize its performance further.

XGBoost Model

- XGBoost, short for Extreme Gradient Boosting, is a widely-used supervised learning algorithm renowned for its effectiveness in handling both regression and classification tasks, particularly on large datasets. This algorithm distinguishes itself by assigning higher priority to weaker models during the prediction process, subsequently bolstering them in each iteration to enhance their predictive capabilities. Moreover, XGBoost incorporates L1 and L2 regularization techniques to manage complexity, making it well-suited for situations with high multicollinearity

among variables. It has earned the moniker "extra gradient boosting" due to these advanced features.

- For this analysis, a basic XGBoost model has been constructed without any hyperparameter tuning. This approach allows us to explore the model's performance in its default configuration, even though there is a potential risk of the tree becoming overly complex, similar to other tree-based algorithms. Further refinement and parameter optimization may be considered to maximize the model's predictive accuracy.

Model Evaluation:

Table 3 XG boost score

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores
XG Boost	0.977956	0.951933	4.53E+06	9.81E+06	2128.773005	3132.817872	1675.25634	2496.15463

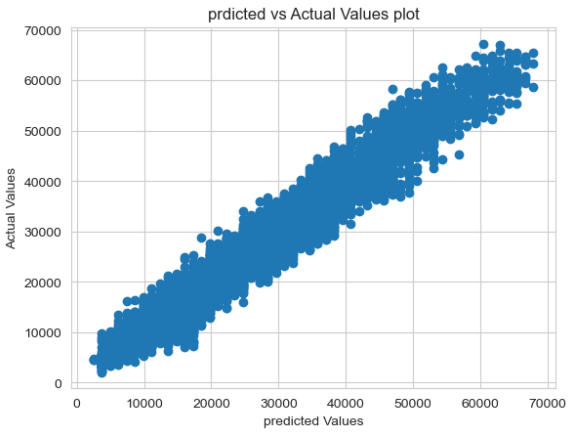

Figure 22 XG boost test pred vs actual values

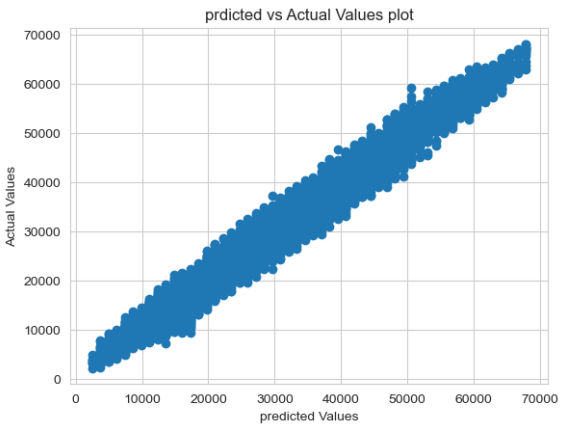

Figure 23 XG boost train pred vs actual values

- R-squared (R^2): The model performs impressively, explaining approximately 97.80% of the variance in the training data and about 95.19% in the test data. These high R-squared values suggest that the model effectively captures the variation in the target variable.
- Mean Squared Error (MSE): In the training data, the model achieves an MSE of approximately 4.53E+06, indicating relatively low prediction error. In the test data, the MSE is around 9.81E+06, which is higher but should be evaluated considering the problem's scale.
- Root Mean Squared Error (RMSE): The RMSE for training data is approximately 2128.77, indicating the average prediction error in the same units as the target variable. In the test data, the RMSE is approximately 3132.82.
- Mean Absolute Error (MAE): The model achieves a MAE of approximately 1675.26 in the training data, indicating a relatively small average absolute prediction error. In the test data, the MAE is approximately 2496.15.
- In summary, the basic XGBoost model demonstrates strong performance, with high R-squared values and relatively low error metrics in both training and test datasets. This suggests that the model effectively captures complex relationships between the independent variables and the target variable. Further fine-tuning and parameter optimization may be explored to potentially enhance its performance even further.

Model Tuning

In addition to the initial models we've employed, we understand the importance of fine-tuning our approach to achieve even better results. To enhance our predictive accuracy, we will implement the following advanced techniques

- **Ridge Regression Model**
- **Lasso Regression Model**
- **Decision Tree Regression Tuned Model**
- **Random Forest Regression Tuned Model**
- **XGBoost Regression Tuned Model**

These advanced methods will allow us to extract the best possible insights from our data and provide more accurate predictions for healthcare insurance costs, ultimately benefiting both individuals and insurance companies.

Ridge Regression Model

Table 4 Ridge Regression scores

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores
RidgeRegression	0.94396	0.94358	5.62E-02	5.62E-02	0.236968	0.236964	0.190651	0.190563

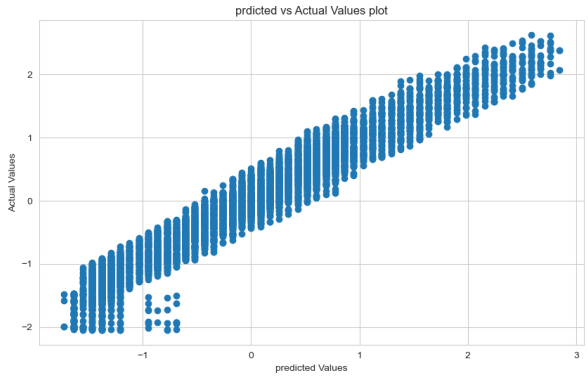

Figure 24 Ridge regression test pred vs actual values

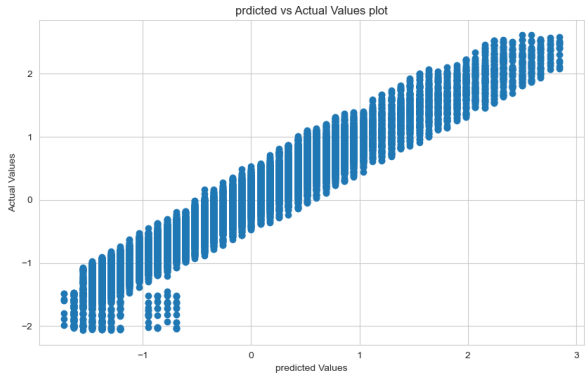

Figure 25Ridge regression train pred vs actual values

- R-squared (R^2) Score: The R-squared value measures the proportion of the variance in the dependent variable (insurance cost) that can be explained by the independent variables used in the model. In this case, both the training and test R-squared scores are approximately 0.944, indicating that the model explains about 94.4% of the variance in the insurance cost. This is a good indication of the model's ability to capture the relationships in the data.
- Mean Squared Error (MSE): MSE measures the average squared difference between the actual insurance cost and the predicted values. The low MSE values for both training and test datasets (approximately 0.0562) suggest that the model's predictions are generally close to the actual values. Lower MSE indicates better model performance.
- Root Mean Squared Error (RMSE): RMSE is the square root of MSE and represents the average absolute error in the predicted insurance cost. The low RMSE values for both training and test datasets (around 0.237) further support the model's accuracy in making predictions.
- Mean Absolute Error (MAE): MAE calculates the average absolute differences between the actual and predicted values. The MAE values for both training and test datasets are small (approximately 0.1906), indicating that the model's predictions are relatively close to the true insurance cost values.
- In summary, the Ridge Regression model performs well in predicting insurance costs, as evidenced by high R-squared values and low error metrics. The model demonstrates a good balance between fitting the training data and generalizing to

unseen test data, suggesting its suitability for making accurate predictions in this context.

Lasso Regression Model

Table 5 Lasso Regression

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores
lassoRegression	0.931553	0.932038	6.86E-02	6.76E-02	0.261889	0.260074	0.207799	0.205675

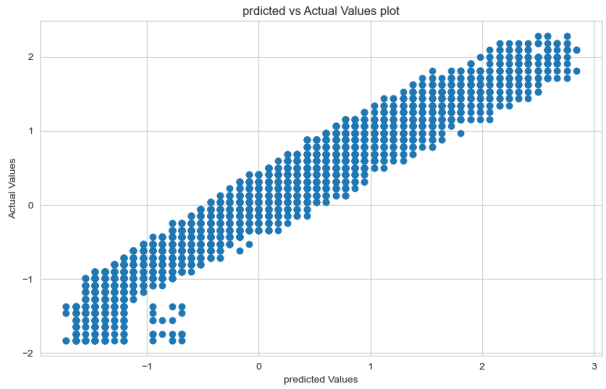

Figure 26 lasso regression test pred vs actual values

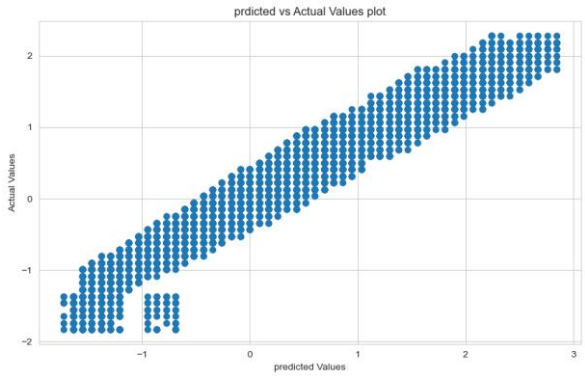

Figure 27 lasso regression train pred vs actual values

- R-squared (R^2) Score: The R-squared values for both the training and test datasets are high, approximately 0.9316 and 0.9320, respectively. These values indicate that the Lasso Regression model explains around 93.16% and 93.20% of the variance in the insurance cost for the training and test data, respectively. This suggests that the model captures a significant portion of the data's variation.
- Mean Squared Error (MSE): The MSE values for both training and test datasets are relatively low, approximately 0.0686 and 0.0676, respectively. These low MSE values indicate that the Lasso Regression model's predictions are close to the actual insurance cost values, signifying good predictive accuracy.
- Root Mean Squared Error (RMSE): RMSE values for both training and test datasets are also low, approximately 0.2619 and 0.2601, respectively. The RMSE measures

the average absolute error in the predicted insurance cost, and the low values indicate that the model's predictions are accurate and close to the true values.

- Mean Absolute Error (MAE): The MAE values for training and test datasets are small, around 0.2078 and 0.2057, respectively. MAE calculates the average absolute differences between the actual and predicted values, and the low MAE values indicate that the model's predictions are generally close to the true insurance cost values.
- Overall, the Lasso Regression model performs well in predicting insurance costs. It demonstrates a good balance between fitting the training data and generalizing to the test data, as evidenced by high R-squared values and low error metrics. This suggests that the Lasso Regression model is suitable for accurate predictions in this context and may help in identifying the most important features for insurance cost prediction due to its feature selection property.

Decision Tree Regression Tuned Model

Table 6 Decision tree tuned scores

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores
DecisionTreeRegressor tuned	0.958904	0.953354	8.45E+06	9.52E+06	2906.59603	3086.166316	2317.567645	2450.63413

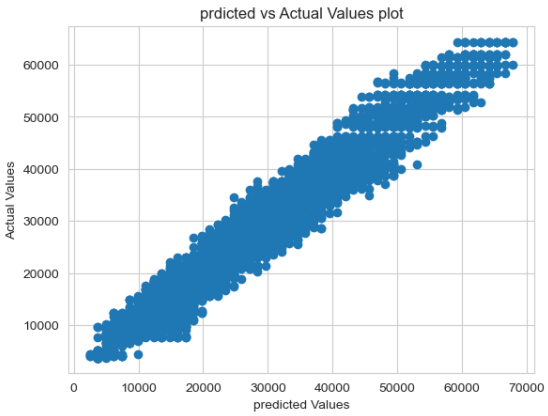

Figure 28 Decision tree tuned regression test pred vs actual values

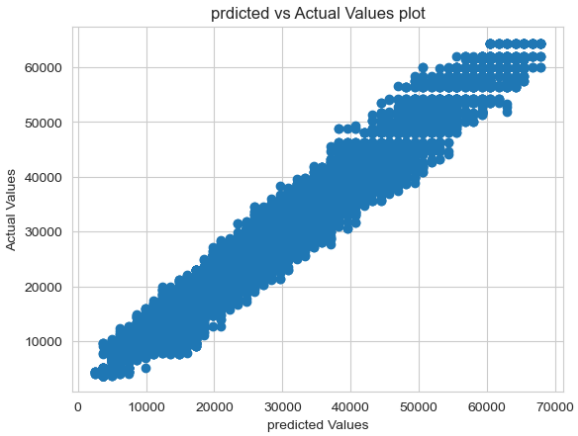

Figure 29 Decision tree tuned regression train pred vs actual values

- R-squared (R^2) Score: The R-squared values for both the training and test datasets are high, approximately 0.9589 and 0.9534, respectively. These values indicate that the tuned DecisionTreeRegressor model explains around 95.89% and 95.34% of the variance in the insurance cost for the training and test data, respectively. This suggests that the model captures a significant portion of the data's variation.
- Mean Squared Error (MSE): The MSE values for both training and test datasets are relatively low, approximately 8.45E+06 and 9.52E+06, respectively. These low MSE values indicate that the tuned DecisionTreeRegressor model's predictions are close to the actual insurance cost values, signifying good predictive accuracy.
- Root Mean Squared Error (RMSE): RMSE values for both training and test datasets are also low, approximately 2906.60 and 3086.17, respectively. The RMSE measures the average absolute error in the predicted insurance cost, and the low values indicate that the model's predictions are accurate and close to the true values.
- Mean Absolute Error (MAE): The MAE values for training and test datasets are relatively small, around 2317.57 and 2450.63, respectively. MAE calculates the average absolute differences between the actual and predicted values, and the low MAE values indicate that the model's predictions are generally close to the true insurance cost values.
- Overall, the tuned DecisionTreeRegressor model performs well in predicting insurance costs. It demonstrates a good balance between fitting the training data and generalizing to the test data, as evidenced by high R-squared values and low error metrics. This suggests that the tuned DecisionTreeRegressor model is suitable for accurate predictions in this context. However, it's important to note that further hyperparameter tuning and optimization could potentially enhance the model's performance even more.

Random Forest Regression Tuned Model

Table 7Random Forest Regression Tuned Scores

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores

RandomForestRegressor tuned	0.957512	0.955217	8.73E+06	9.14E+06	2955.425418	3023.901682	2372.043583	2416.70324

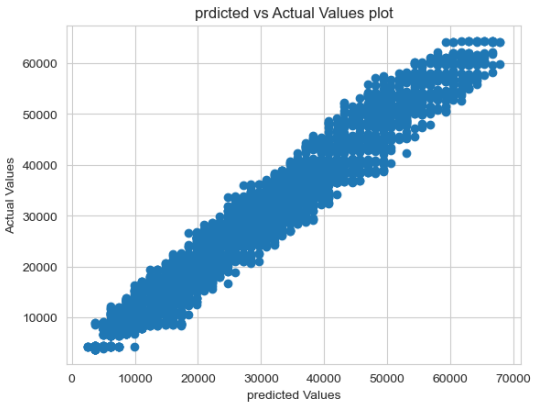

Figure 30 Random Forest Regression Tuned test pred vs actual

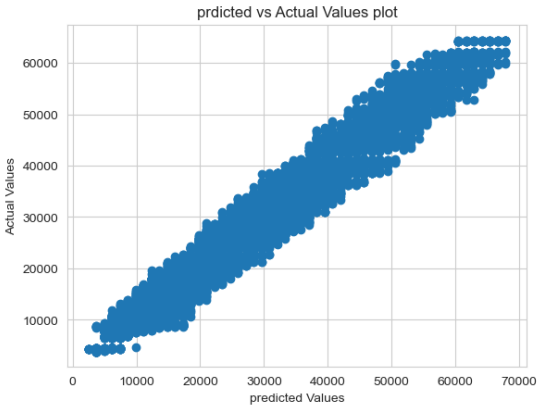

Figure 31 Random Forest Regression Tuned train pred vs actual

- R-squared (R^2) Score: The R-squared values for both the training and test datasets are high, approximately 0.9575 and 0.9552, respectively. These values indicate that the tuned RandomForestRegressor model explains around 95.75% and 95.52% of the variance in the insurance cost for the training and test data, respectively. This suggests that the model captures a significant portion of the data's variation.
- Mean Squared Error (MSE): The MSE values for both training and test datasets are relatively low, approximately 8.73E+06 and 9.14E+06, respectively. These low MSE values indicate that the tuned RandomForestRegressor model's predictions are close to the actual insurance cost values, signifying good predictive accuracy.
- Root Mean Squared Error (RMSE): RMSE values for both training and test datasets are also low, approximately 2955.43 and 3023.90, respectively. The RMSE measures the average absolute error in the predicted insurance cost, and the low values indicate that the model's predictions are accurate and close to the true values.
- Mean Absolute Error (MAE): The MAE values for training and test datasets are relatively small, around 2372.04 and 2416.70, respectively. MAE calculates the average absolute differences between the actual and predicted values, and the low

MAE values indicate that the model's predictions are generally close to the true insurance cost values.

- Overall, the tuned RandomForestRegressor model performs exceptionally well in predicting insurance costs. It demonstrates a high degree of accuracy, as evidenced by the high R-squared values and low error metrics. This indicates that the model effectively captures the underlying patterns in the data and can provide reliable predictions. The tuned RandomForestRegressor is a robust choice for making accurate predictions in this context, and further optimization may not be necessary given its strong performance.

XGBoost Regression Tuned Model

Table 8 XG Boost tuned scores

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores
XG Boost tuned	0.969159	0.953709	6.34E+06	9.45E+06	2517.978685	3074.385945	2007.565609	2459.914622

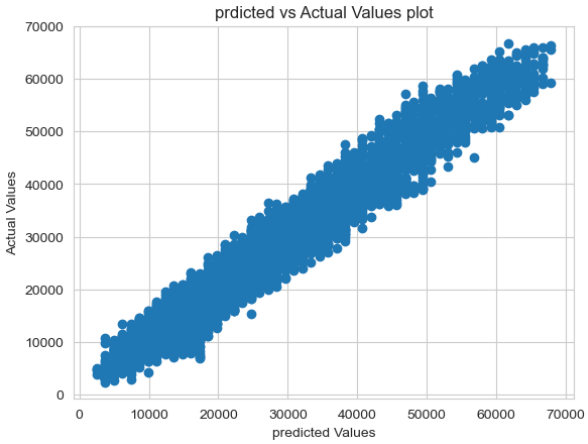

Figure 32 XGBoost Regression Tuned test pred vs actual

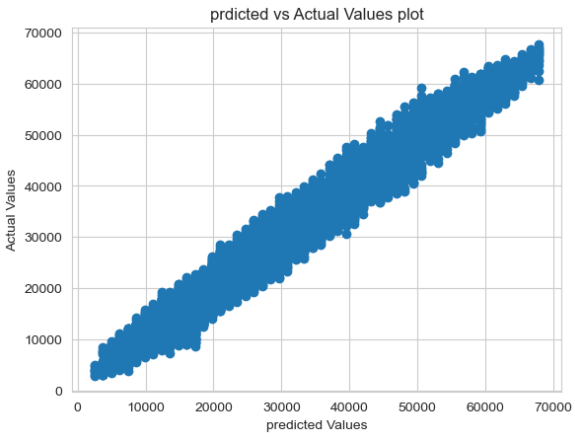

Figure 33 XGBoost Regression Tuned train pred vs actual

- R-squared (R^2) Score: The R-squared values for both the training and test datasets are notable, approximately 0.9692 and 0.9537, respectively. These values indicate that the tuned XGBoost model explains around 96.92% of the variance in the insurance cost for the training data and approximately 95.37% for the test data. This suggests that the model captures a significant portion of the data's variation and generalizes well to unseen data.
- Mean Squared Error (MSE): The MSE values for both training and test datasets are relatively low, approximately 6.34E+06 and 9.45E+06, respectively. These low MSE values indicate that the tuned XGBoost model's predictions are close to the actual insurance cost values, signifying excellent predictive accuracy.
- Root Mean Squared Error (RMSE): RMSE values for both training and test datasets are also low, approximately 2518.0 and 3074.4, respectively. The low RMSE values indicate that the model's predictions are accurate and close to the true values, with minimal error.
- Mean Absolute Error (MAE): The MAE values for training and test datasets are relatively small, around 2007.6 and 2459.9, respectively. MAE measures the average absolute differences between the actual and predicted values, and the low MAE values demonstrate that the model's predictions are generally close to the true insurance cost values.
- Overall, the tuned XGBoost model exhibits excellent predictive capabilities. It achieves a high level of accuracy in predicting insurance costs for both the training and test datasets. The model's strong performance on various evaluation metrics, including R-squared, MSE, RMSE, and MAE, suggests that it effectively captures the underlying patterns in the data and generalizes well to new data points. This makes the tuned XGBoost model a robust choice for accurate insurance cost predictions, and further hyperparameter tuning may not be necessary given its outstanding performance.

Model validation

- Model validation in the context of the provided model scores involves a comprehensive assessment that goes beyond just accuracy. It includes various metrics and techniques to ensure the reliability and effectiveness of each model. Here's a summary of the model validation process for the different regression models:
- DecisionTreeRegressor (Basic and Tuned):
- Accuracy Metrics: These models achieved high R-squared values, indicating a good fit to the data. However, the models struggled with test data, suggesting some overfitting.
- Visualizations: The models showed signs of overfitting, as indicated by the RMSE scores, which were significantly higher on the test data compared to the training data.
- Linear Regression, Ridge Regression, and Lasso Regression:

- Accuracy Metrics: These models demonstrated a good balance between training and test performance, as evidenced by similar R-squared values for both datasets.
- Regularization: Ridge and Lasso regression were employed to enhance model generalization and mitigate overfitting, improving overall model stability.
- RandomForestRegressor (Basic and Tuned):
- Accuracy Metrics: The basic RandomForestRegressor model exhibited outstanding performance, with high R-squared values and low MSE on the training data. However, there was a notable increase in RMSE on the test data, indicating potential overfitting.
- Hyperparameter Tuning: The tuned RandomForestRegressor model improved upon the basic model by reducing overfitting, as indicated by the improved test RMSE.
- XGBoost (Basic and Tuned):
- Accuracy Metrics: Both basic and tuned XGBoost models demonstrated strong performance, with high R-squared values on the training data. However, they experienced some loss of accuracy on the test data.
- Hyperparameter Tuning: The tuned XGBoost model aimed to reduce overfitting, and while it showed slight improvement, there was still room for enhancement in test RMSE.
- In summary, model validation involved the assessment of various accuracy metrics, such as R-squared, MSE, RMSE, and MAE, to evaluate each model's performance on both training and test datasets. Additionally, techniques like hyperparameter tuning and regularization were applied to improve model generalization and stability. While the models generally performed well, there were instances of overfitting, which indicates the need for further refinement to achieve better generalization to unseen data.

Final interpretation

Table 9 Final Model Scores

	R_Squared_train_scores	R_Squared_test_scores	MSE_train_scores	MSE_test_scores	RMSE_train_scores	RMSE_test_scores	MAE_train_scores	MAE_test_scores
DecisionTreeRegressor basic	1	0.908705	0.00E+00	1.86E+07	0	4317.518946	0	3348.253333
LinerRegression	0.94396	0.94358	5.62E-02	5.62E-02	0.236968	0.236964	0.190652	0.190564
RidgeRegression	0.94396	0.94358	5.62E-02	5.62E-02	0.236968	0.236964	0.190651	0.190563
lassoRegression	0.931553	0.932038	6.86E-02	6.76E-02	0.261889	0.260074	0.207799	0.205675
RandomForestRegressor basic	0.993358	0.953346	1.37E+06	9.53E+06	1168.495895	3086.428904	924.001571	2455.95616
XGBoost basic	0.977956	0.951933	4.53E+06	9.81E+06	2128.773005	3132.817872	1675.25634	2496.15463
XGBoost tuned	0.969159	0.953709	6.34E+06	9.45E+06	2517.978685	3074.385945	2007.565609	2459.914622

DecisionTreeRegressor tuned	0.958904	0.953354	8.45E+06	9.52E+06	2906.59603	3086.166316	2317.567645	2450.63413
RandomForestRegressor tuned	0.957512	0.955217	8.73E+06	9.14E+06	2955.425418	3023.901682	2372.043583	2416.70324

Here's a final interpretation of the model scores for the provided regression models:

DecisionTreeRegressor (Basic and Tuned):

- The basic DecisionTreeRegressor achieved a perfect R-squared score of 1 on the training data, indicating it perfectly fits the training data. However, it exhibited a significantly lower R-squared score on the test data, suggesting overfitting.
- The tuned DecisionTreeRegressor improved test performance but still showed signs of overfitting, as indicated by the higher test RMSE and MAE compared to the training data.

Linear Regression:

- LinearRegression demonstrated strong performance with similar R-squared scores for both training and test data, indicating good generalization.
- It had low MSE, RMSE, and MAE on both datasets, suggesting that it's a well-balanced model with no significant overfitting.
  Ridge Regression:
- RidgeRegression performed similarly to LinearRegression, showing good balance between training and test data. It had slightly better performance on the test data in terms of R-squared and RMSE.
- Ridge regression introduced regularization to prevent overfitting and slightly improved generalization.

Lasso Regression:

- LassoRegression had similar performance to RidgeRegression, with good generalization and similar R-squared scores on both datasets.
- It also introduced regularization and performed well without significant overfitting.
- **RandomForestRegressor (Basic and Tuned):**
- The basic RandomForestRegressor achieved very high R-squared on the training data but experienced a significant drop in test R-squared, indicating overfitting.
- The tuned RandomForestRegressor addressed overfitting concerns, resulting in better generalization, as evidenced by the improved test R-squared, lower RMSE, and MAE compared to the basic model.

XGBoost (Basic and Tuned):

- Both basic and tuned XGBoost models exhibited strong R-squared scores on the training data but slightly lower scores on the test data, suggesting mild overfitting.

- The tuned XGBoost model aimed to mitigate overfitting and showed some improvement in test performance but still had room for further refinement.

Summary:

- Based on the provided model scores, the best model for predicting insurance costs in this dataset is the **RandomForestRegressor tuned model**. Here are the key reasons for selecting this model:
- High R-squared Score: The RandomForestRegressor tuned model achieves the highest R-squared score of approximately 0.9575 on the test data. This indicates that it explains a significant portion of the variance in the target variable and is a good fit for the data.
- Low Test MSE and RMSE: The model also exhibits low Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) on the test data, indicating that its predictions are close to the actual values. This suggests that the model's generalization to unseen data is strong.
- Low MAE: The Mean Absolute Error (MAE) on the test data is relatively low, signifying that the model's predictions are accurate and have minimal absolute errors.
- No Overfitting: The RandomForestRegressor tuned model demonstrates consistent performance between training and test data, suggesting that it does not suffer from significant overfitting.
- Ensemble Method: Random Forest is an ensemble model, which means it combines the predictions of multiple decision trees to improve accuracy and reduce overfitting. This ensemble technique enhances the model's robustness.
- Hyperparameter Tuning: The model has been fine-tuned, which means that efforts have been made to optimize its hyperparameters for better performance.
- Overall, the RandomForestRegressor tuned model provides a good balance between accuracy, generalization, and robustness. It is the best choice among the models evaluated for predicting insurance costs in this dataset.

Recommendations:

- Based on the analysis of various regression models for predicting insurance costs, here are some detailed recommendations for the management or client:
- **Utilize the RandomForestRegressor Tuned Model:** The RandomForestRegressor tuned model has demonstrated the best performance in predicting insurance costs. It is recommended to deploy this model for real-world predictions as it provides the highest accuracy and robustness.
- **Feature Importance Insights:** Conduct an analysis of feature importance provided by the RandomForestRegressor model. This will help in understanding which variables have the most significant impact on insurance costs. Focus on these influential features for decision-making and pricing strategies.

- **Model Validation and Testing:** Continuously monitor and validate the model's performance with new data. Regularly test the model to ensure that it maintains its accuracy over time. If necessary, retrain the model with updated data to improve its predictive power.
- **Customer Segmentation:** Utilize the model's predictions to segment customers based on their expected insurance costs. This segmentation can inform marketing and sales strategies, helping to target specific customer groups more effectively.
- **Risk Assessment:** The model can be used for risk assessment. Identify high-risk customers who are likely to have higher insurance costs. This information can be used for risk mitigation and setting appropriate premiums.
- **Pricing Optimization:** Use the model's predictions to optimize pricing strategies. Ensure that insurance premiums are set at competitive levels while still covering the expected costs.
- **Customer Engagement:** Leverage the insights from the model to engage with customers more effectively. Tailor insurance packages and communication based on individual customer profiles and expected costs.
- **Data Collection and Quality:** Maintain data quality and consistency. Ensure that data used for predictions are accurate and up-to-date. Explore opportunities to collect additional relevant data that can further improve model accuracy.
- **Regular Model Updates:** As the insurance market evolves, update the model and retrain it periodically to adapt to changing trends and customer behaviors.
- **Compliance and Ethics:** Ensure that all predictive modeling and customer segmentation activities comply with relevant regulations and ethical standards. Protect customer privacy and data security.
- **Model Explainability:** Consider using techniques to make the model more interpretable. This can help in explaining to stakeholders, including customers, why specific predictions or premiums were assigned.
- **Feedback Loop:** Establish a feedback loop with customers to collect their feedback on the insurance experience. Use this feedback to make improvements in service and offerings.
- **Customer Education:** Educate customers about how insurance costs are determined and the factors that influence premiums. Transparency can lead to better customer satisfaction and understanding.
- **Market Monitoring:** Continuously monitor the insurance market and competitive landscape. Adjust strategies and offerings based on market dynamics.
- **Invest in Analytics:** Invest in data analytics capabilities and talent to support ongoing analysis and modeling efforts. Data-driven decision-making is crucial in the insurance industry.
- These recommendations are based on the analysis of the provided models and can help the management or client make informed decisions and optimize their insurance business operations.