CAPSTONE PROJECT

INSURANCE COST-HEALTHCARE PROJECT

PROJECT REPORT

—

VIDYA V

—

PGPDSBA.O.2023B

CONTENTS

1.	**Introduction of the business problem**	**4**
2.	**EDA and Business Implications**	**6**
3.	**Advanced EDA and Data Pre-Processing**	**14**
4.	**Model Building and Validation**	**23**
5.	**Model Tuning**	**26**
6.	**Appendix**	**30**

LIST OF FIGURES

Name	Page No.
Fig.1. Inspection of target variable	7
Fig.2. Correlation heatmap of numerical Variables	8
Fig.3. Boxenplot - regular_checkup_last_year vs insurance_cost	8
Fig.4. Boxenplot- adventure_sports vs insurance_cost	9
Fig.5. Boxenplot - covered_by_any_other_company vs insurance_cost	9
Fig.6. Boxenplot- weight_change_in_last_one_year vs insurance_cost	10
Fig.7. Boxenplot – weight_cat vs insurance_cost	10
Fig.8. Heatmap- regular_checkup_last_year vs adventure_sports vs insurance_cost	11
Fig.9. Missing Data Visualization	14
Fig.10. Boxplot of numerical variables before outlier treatment	16
Fig.11. Boxplot of numerical variables after outlier treatment	16
Fig.12. Elbow plot to identify optimum number of clusters	18
Fig.13. Insurance cost distribution across clusters	19
Fig.14. Weight Distribution across clusters	19
Fig.15. Fat percentage Distribution across clusters	20
Fig.16. Clusters and years of insurance with us	20
Fig.17. Catplot- Weight vs clusters	21
Fig.18. Catplot- Insurance cost vs clusters	21

LIST OF TABLES

Name	Page No.
Table.1. Numerical fields descriptive summary	6
Table.2. Categorical fields descriptive summary	7
Table.3. Cluster profiling	18
Table.4. Model performance on train set	23
Table.5. Model performance on test set	24
Table.6. Model performance on train set after tuning	27
Table.7. Model performance on test set after tuning	27
Table.8. Features in the dataset and their importance	29

INTRODUCTION OF BUSINESS PROBLEM

Need for the study:

Health Insurance is one of the fields of focus in recent times. With the rise in diseases and treatment costs, more and more people are inclined towards securing a health insurance policy that covers all expected and unexpected medical costs. Insurance companies, thus need a meticulous evaluation of various parameters to determine the premium, so that the risk is rightly assessed and priced appropriately. Incorrect or inaccurate predictions might lead to losses for the companies and insufficient coverage might lead to client dissatisfaction.

Problem Statement:

The dataset here contains various lifestyle such as the amount of exercise, smoking habits etc., and medical parameters like heart diseases history and weight etc., and competitive parameters like whether or not they have been covered by other insurance companies and the target variable is the insurance cost, determined by the predictors.

Objective/ Business Opportunity:

- Parametric Evaluation:
  - To identify and understand the correlation between the various predictors, and the target variable, and to estimate the significance of the predictors on the target variable.
- Risk Assessment:
  - To explore the parameters and compare them against the insurance cost, and identify potential areas of risk, if any, and to develop strategies to avert the same
- Cost Prediction:
  - To build prediction models and optimize their performance by tuning
  - Identify the best model for the given dataset by means of accuracy scores and RMSE scores

Process:

- Exploring the given dataset, understand the various predictor variables and their nature
- Performing necessary cleaning and treatment to make the data optimal for analysis
- Performing Exploratory Data Analysis by breaking down the dataset into Uni, Bi and Multivariate combinations and seeking insights from observations
- Exploring the data by segmentation, identify segments and relate them to insurance cost
- Building various predictive models and evaluating their performance against the train and the test sets
- Tuning the models in order to achieve optimum performance
- Comparing and evaluating the performance of models on train and test sets by metrics like RMSE and accuracy
- Identifying the best model for the dataset and the problem statement
- Fitting the model to the test data and determining the insurance costs for the same

Software Used/ Tools Used:

- Jupyter notebook- Python Kernel
- Numpy version 1.24.4
- Pandas Version 1.4.4
- Seaborn Version 0.13.0
- Matplotlib version 3.5.2

EDA AND BUSINESS IMPLICATIONS

Dataset Information:

- This dataset contains 25000 applicants' information from ID 5000-24999 along with their insurance costs
- There are 24 variables in the dataset, including the target variable
- There are no time/date related variables in the given dataset
- The variables of the dataset can be classified as below:
  - Personal demographic information:
    - Age, Gender, Location, Occupation, Applicant ID
  - Biomedical indices:
    - Weight, Cholesterol Level, Fat Percentage, BMI, Average glucose level
  - Lifestyle Parameters:
    - Exercise, smoking and alcohol habits, Walking related counts, Whether involved in Adventurous sports
  - Medical History:
    - Heart disease history, Other major disease history, number of regular checkups and doctor visits in the past year, year in which the person was last admitted, weight change in the past year
  - Insurance company related variables:
    - Number of years of insurance with us, Whether covered by other insurance companies
  - Target Variable:
    - Insurance cost
- Thus, a comprehensive coverage of most of the parameters that could be involved are captured

Descriptive Summary of data:

Variable	Minimum Value	Mean	Median	Maximum	Skewness
Daily_avg_steps	2034	5215.89	5089	11255	0.91
Age	16	44.92	45	74	0.01
Avg_glucose_level	57	167.53	168	277	-0.01
BMI	12.3	31.39	30.5	100.6	1.05
Weight	52	71.61	72	96	0.11
Fat_percentage	11	28.81	31	42	-0.36
Insurance_cost	2468	27147.41	27148	67870	0.33

Table.1. Numerical fields descriptive summary

Variable	Number of Unique Values	Modal Value	Frequency of modal value
Years_of_insurance_with_us	9	3	11.96%
Regular_checkup_last_year	6	0	60.96%
Adventure_sports	2	0	91.83%
Occupation	3	Student	40.68%
Visited_doctor_last_1_year	12	2	34.68%
Cholesterol_level	5	150 to 175	35.05%
Heart_decs_history	2	0	94.54%
Other_major_Decs_history	2	0	90.18%
Gender	2	Male	65.68%
Smoking_status	4	Never Smoked	37.00%
Location	15	Bangalore	6.97%
Covered_by_any_other_company	2	N	69.67%
Alcohol	3	Rare	55.00%
Exercise	3	Moderate	58.55%
Weight_change_in_last_1_year	7	4	20.30%

Table.2. Categorical fields descriptive summary

EDA:

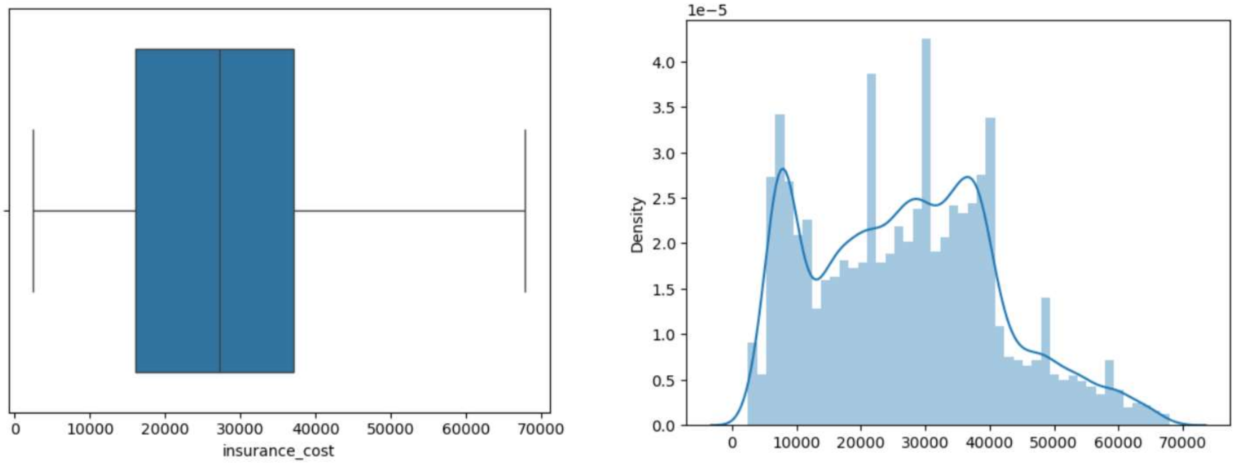


Fig.1. Inspection of target variable

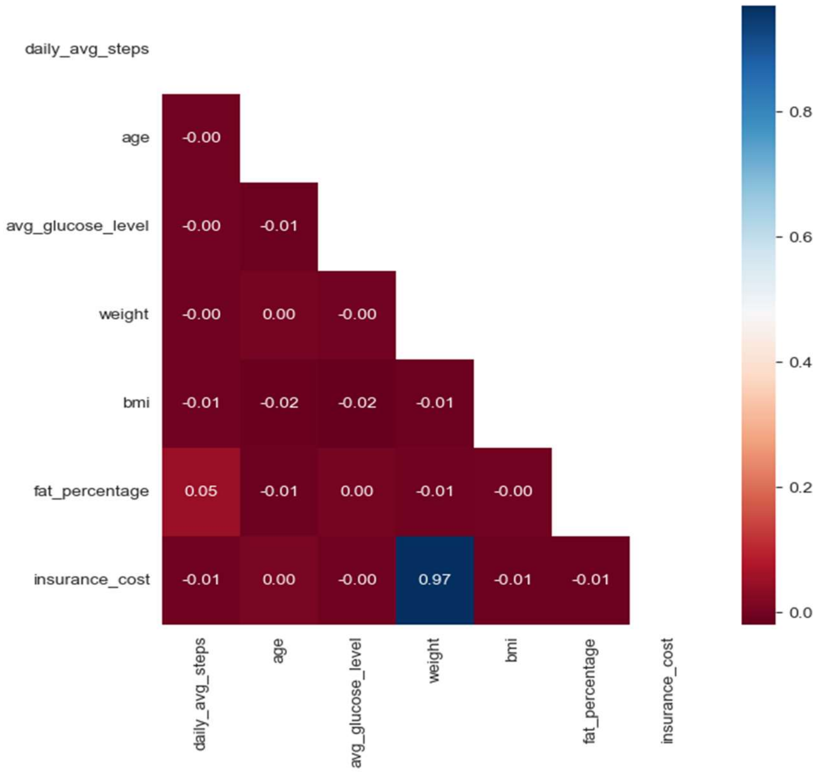


Fig.2. Correlation heatmap of numerical Variables

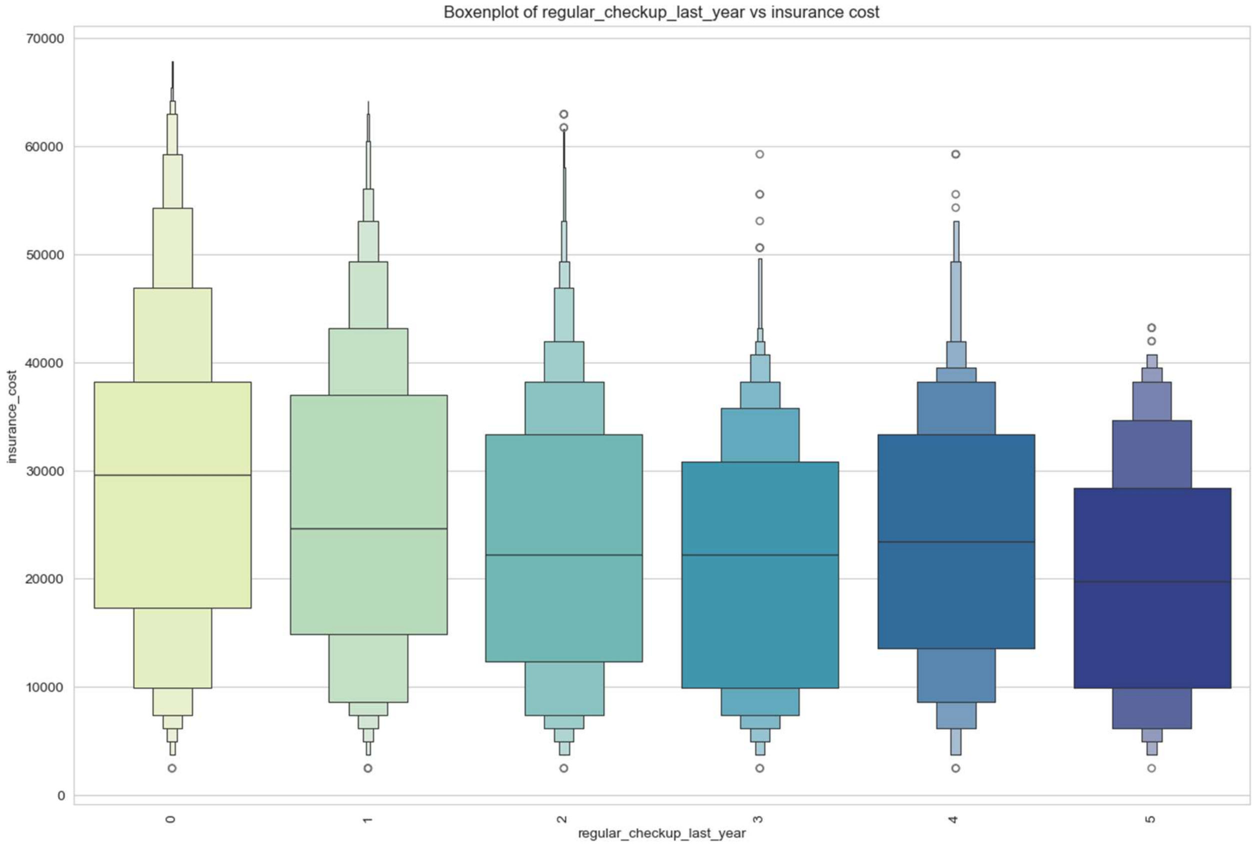


Fig.3. Boxenplot - regular_checkup_last_year vs insurance_cost

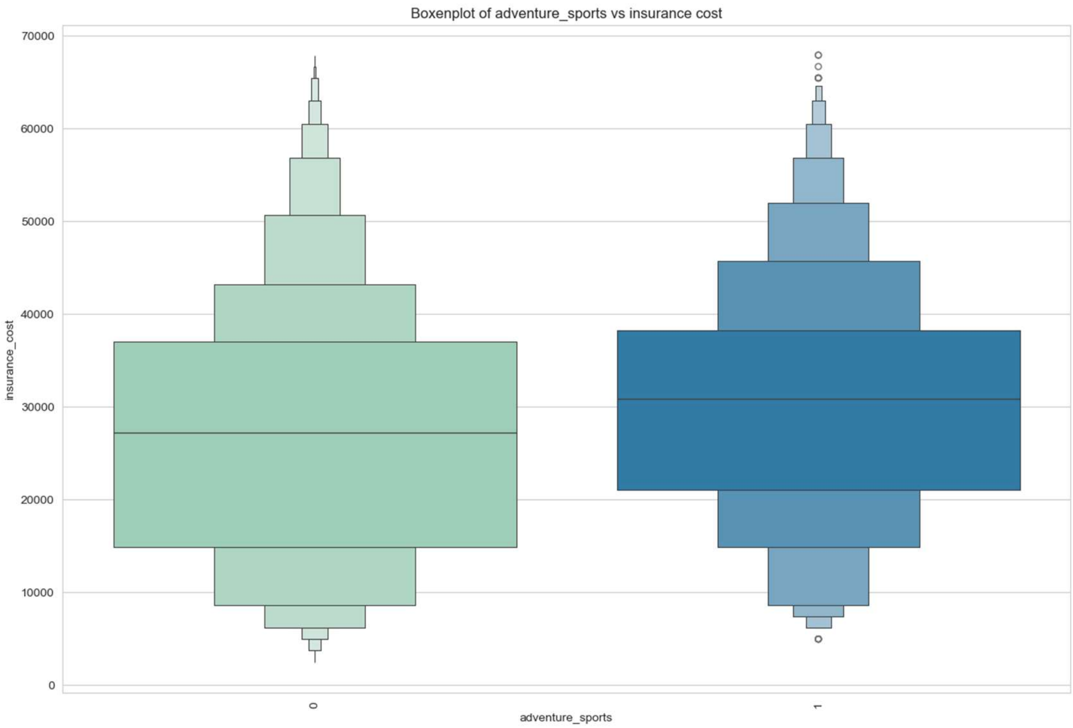


Fig.4. Boxenplot- adventure_sports vs insurance_cost

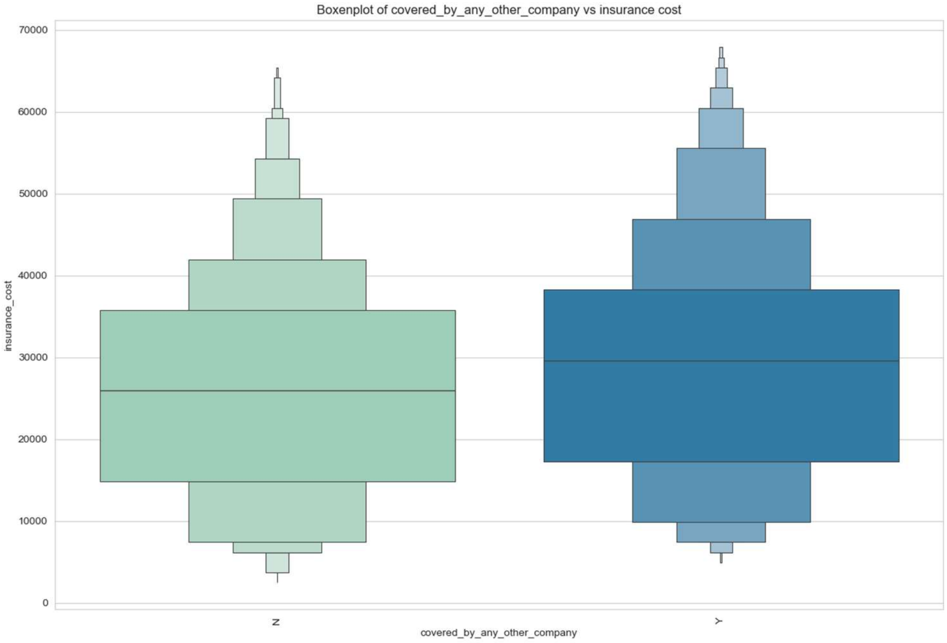


Fig.5. Boxenplot - covered_by_any_other_company vs insurance_cost

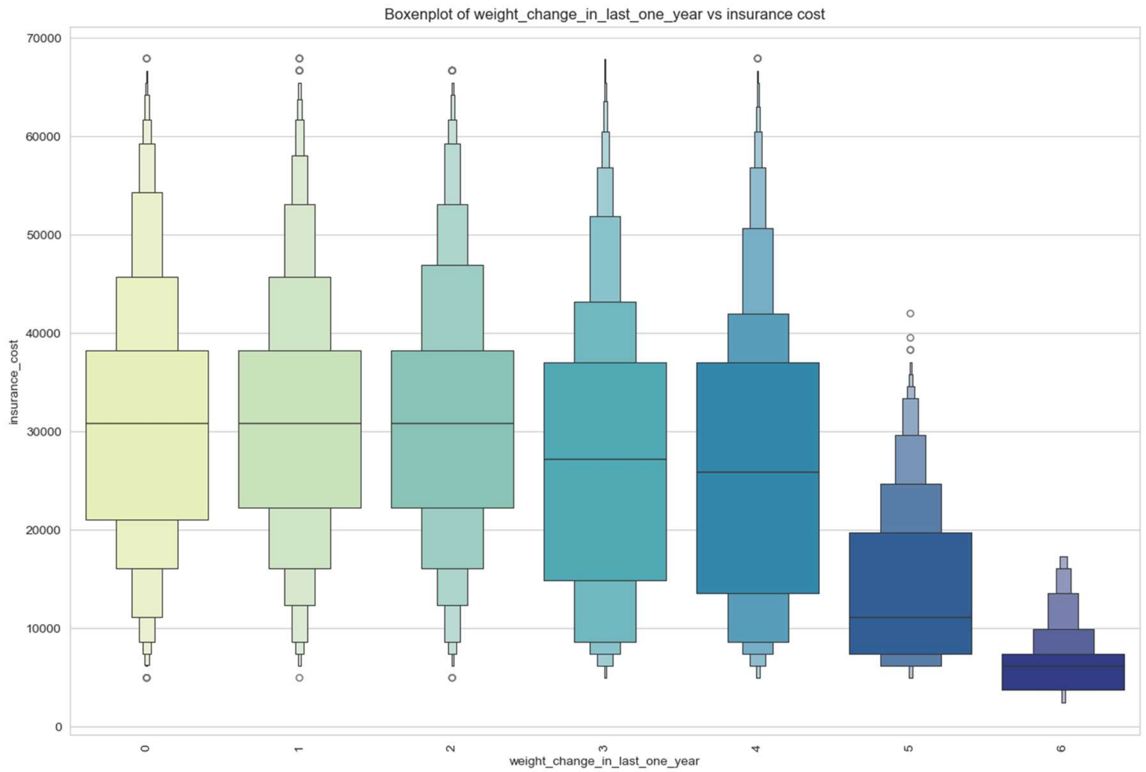


Fig.6. Boxenplot- weight_change_in_last_one_year vs insurance_cost

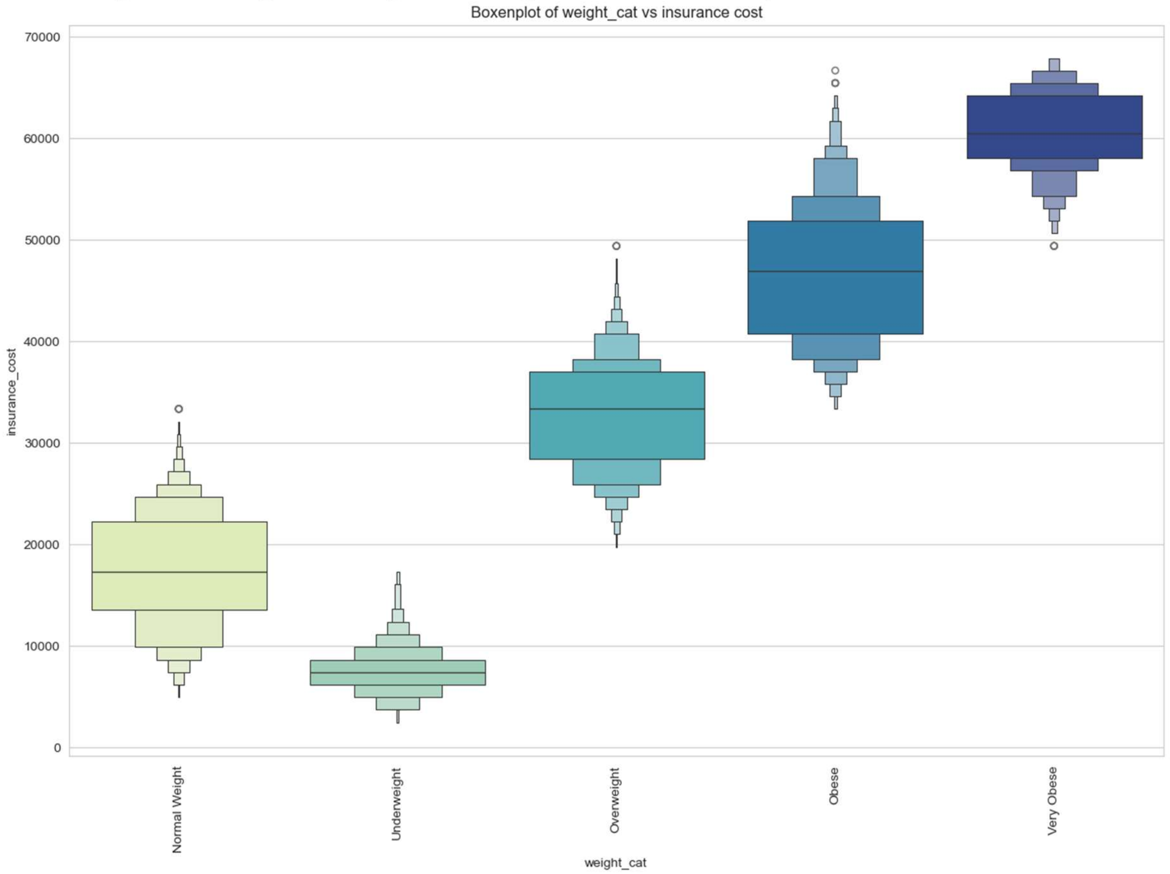


Fig.7. Boxenplot – weight_cat vs insurance_cost

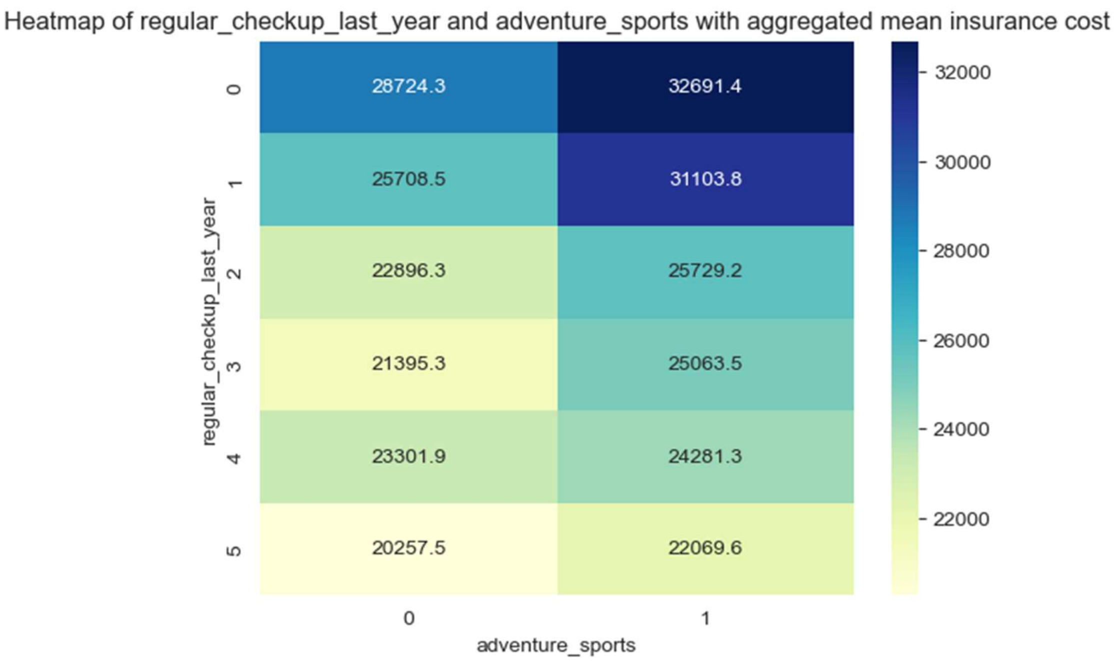


Fig.8. Heatmap- regular_checkup_last_year vs adventure_sports vs insurance_cost

Observations:

- **Univariate Analysis:**
  - The target variable is insurance cost, having values ranging from 2468 to 67870, with the median value being 27148, having only 54 unique values
  - Most of the people have not had any regular checkup last year
  - Columns with outliers: daily_avg_steps
  - age,avg_glucose_level have a uniform distribution
  - weight variable has a multimodal distribution
  - fat_percentage has 32 unique values ranging from 11-42. Hence the distribution is multimodal
  - Most of the people seem to have visited the doctor 2-4 times last year
  - Most of the people have either 125-150 or 150-175 cholesterol level. Those of them who have higher cholesterol levels are very low in count
  - There are 15 unique locations in the dataset, where people are from, and they are almost equally distributed amongst these locations.
  - Most of the people have reported Rare or no alcohol consumption. Very few have reported daily alcohol consumption
  - Most of the people have reported moderate exercise
  - Most of the customers have reported a weight change of 3-4 in the past year
  - Most of the customers are classified as lightly active, i.e. they take 4500-6000 steps everyday
  - Most of the people are overweight, i.e. have weight ranging from 70-80. Very few are classified as very obese i.e. >90.

- Most of the people have High body fat.ie, body fat_percentage ranging from 30-40
- Based on the classification of insurance cost, most of the people belong to the 4th tier, i.e., 30000-40000
- Among the categorical variables, Gender, adventure_sports, covered_by_other_company, heart_decs_history, other_decs_history, are binary in nature
- Less than 10% have had other major diseases, and less than 5% have had heart diseases
- About 8% of the people engage in adventure sports
- About 30% are covered by other insurance companies

- **Bivariate Analysis:**
  - From the pairplots of numerical variables, a linear relation can be observed only between weight and insurance cost
  - Those who have had 0 regular checkups last year have the highest median insurance cost
  - Those who have had 5 or more than 5 regular checkups have the lowest median insurance cost
  - Those who engage in adventure sports have higher median insurance cost than those who don't
  - Those who have coverage from other insurance companies tend to have a higher median insurance cost than those who don't
  - As the weight change in the past year increases, the median insurance cost decreases
  - As the weight category increases, median insurance cost also increases
  - 2801 people have less than 0 years insurance with us and are not being covered by any other company, which means they are new to the health insurance field
  - Around 13469 people have not had a regular checkup last year and are not into adventure sports. This is more than half the dataset size.
  - Around half the population have visited the doctor last year and have not had any regular checkups
  - Most of the people who have not had regular checkup last year have a cholesterol level below 175
  - More than half the population have had no major disease or heart disease and have not had regular checkup- these might be at risk of unknown medical issues. This might be a potential risk.
  - Around 10242 people have not had regular checkup and are not covered by any other company
  - 20% of the people have reported rare or daily intake of alcohol, but have not had regular checkup
  - 5851 people are overweight, and have not had regular checkup. This might be a potential risk.
  - 6562 people have high body fat and have not had regular checkup.
  - No student has a cholesterol level over 175
  - Business people are prone to high cholesterol levels 150-225

    - Business people have reported the highest number of heart diseases, followed by students
    - People with very High body fat, i.e. fat_percentage from have cholesterol level greater than 200
    - Almost half the females have reported an Unknown smoking status
    - Weight clearly seems to have a positive correlation with insurance cost
- **Multivariate Analysis:**
    - People who are covered by other insurance companies, and have lesser number of years of insurance with us tend to have higher insurance costs
    - People who are into adventure sports and have no regular checkups have high insurance cost
    - People who have not had regular checkups, but have visited doctor for 5 times have the highest mean insurance cost
    - People who have heart and other diseases, but have had 5 or more regular checkups have low insurance cost
    - People who do not consume alcohol, and have regular checkups have low insurance cost
    - People who have a cholesterol level of 225-250 and have had heart disease have higher insurance cost
    - When people belong to the very obese weight category, the insurance cost tier is 5 or above, i.e. insurance cost is greater than 50000.
    - When they are underweight, their insurance tier is 2 or below, i.e., insurance cost is less than 20000.

Business Recommendations based on EDA:

- Insurance plans and pricing can be customized, to cater to each individual's profile, rather than having a pre-determined price for similar groups of people
- Encouraging regular checkups by providing concessions and penalizing higher rates by slapping additional charges
- Around 3000 people are new to insurance. These people can be targeted for add-ons or upgradation of insurance packages
- Rather than the broad categorization of occupation, a much deeper categorization can be done to identify potential occupational hazards, and design mitigative policies
- Rewarding people with good exercise and physical activity by lowering of pricing or providing additional coverage for the same price
- Health parameters and ailments are in some cases gender specific. This needs to be addressed by designing completely different strategies for the genders
- Medical and health related parameters are useful only if they are updated on a regular basis. This is also one of the reasons to mandate regular health checkups. Tie-ups with hospitals can be arranged, and master health checkups can be provided at concessional rates for policy holders

ADVANCED EDA AND DATA PRE-PROCESSING

Removal of unwanted variables:

- The ID variable in this dataset is ‘applicant_id’
- It has 25000 continuous values ranging from 5000 to 24999
- As this variable is a unique ID, it was useful in determining that there were no duplicates in the given dataset
- It was also used to ensure that there were no gaps in the data
- However, this variable is no longer needed for further exploration as it does not have any kind of correlation with any of the other variables.
- Hence, this variable is removed and is not a part of further analysis
- Thus, after this step, the dataset shape is (25000,23)

Missing Value Treatment:

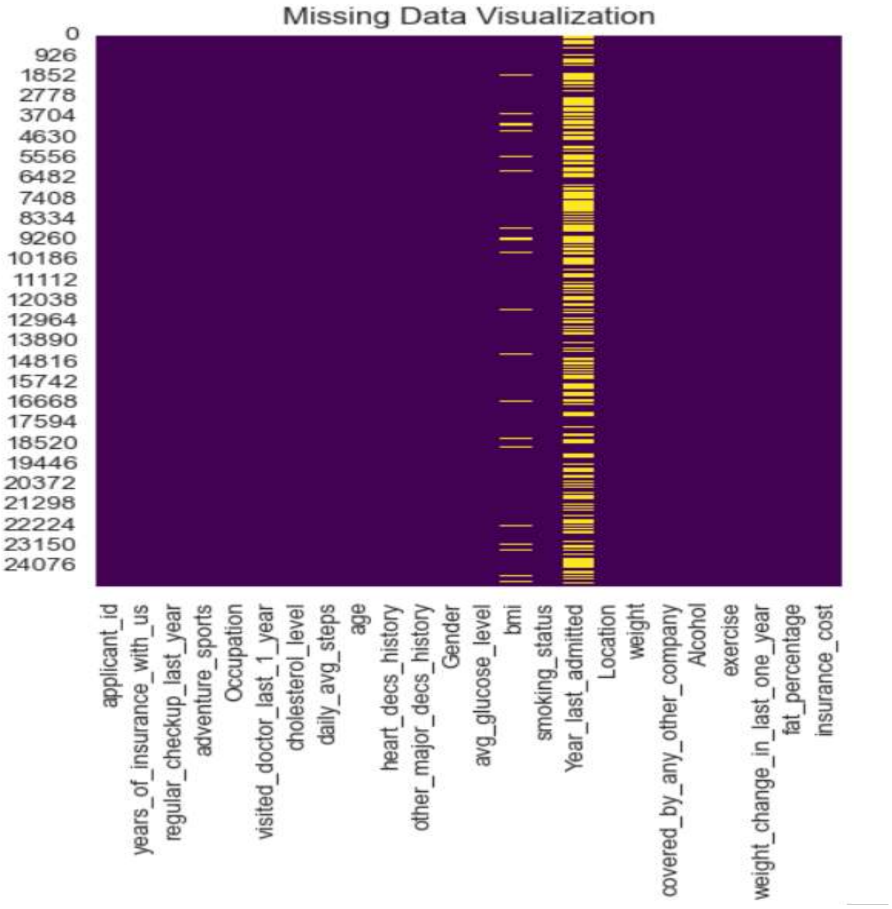


Fig.9. Missing Data Visualization

- From the above figure, we find that there are missing values in the fields 'year_last_admitted' and 'bmi'
- In the 'year_last_admitted' field has 11881 null values, which is about 47.5% of the size of the field
- The bmi field has 990 missing values, which is about 3.96 the size of the field
- 47.5 % is a very high percentage, and in a business scenario, efforts have to be made to obtain the actual data
- However, in this case, a decision has to be made whether to drop this variable or impute missing values
- The field in general has a high correlation with the target variable and a couple of predictor variables
- A t-test was done to ascertain the significance, and resulted in a p-value<0.05, which indicated that the variable is significant
- Hence, instead of making the data entirely synthetic, a better option would be to drop the column entirely
- Also, the 'bmi' field has very few missing values (3.96%). Thus, here as well, the better option would be to just drop the rows that contain null values in the field instead of imputing
- Hence, dropping rows with null values in this field and dropping the field' year_last_admitted' entirely will result in 8.13% data loss, which is a small percentage
- Hence, after the missing value treatment by dropping the field 'year_last_admitted' and dropping the rows that contained null values in the 'bmi' field, the shape of the dataset is (24010,22)
-

Data Balancing:

The given problem statement involves building predictive models to estimate the insurance cost, which is a numerical target variable. Hence, data balancing is not required as it is applicable only to classification models

Outlier Treatment and Scaling:

- After the EDA, in order to make the data ready for modelling, scaling and outlier treatments were done
- For scaling, StandardScaler() function in sklearn preprocessing library was used
- For outlier treatment, Whiskerization method was used. That is, the outliers were cut off at the whiskers by bringing the outlier values to 1.5 times the inter-quartile range of the distribution of the said variable
- This method helps maintain the distribution of the variables better than other methods
- Only one column, i.e., daily_avg_steps had outliers and was treated

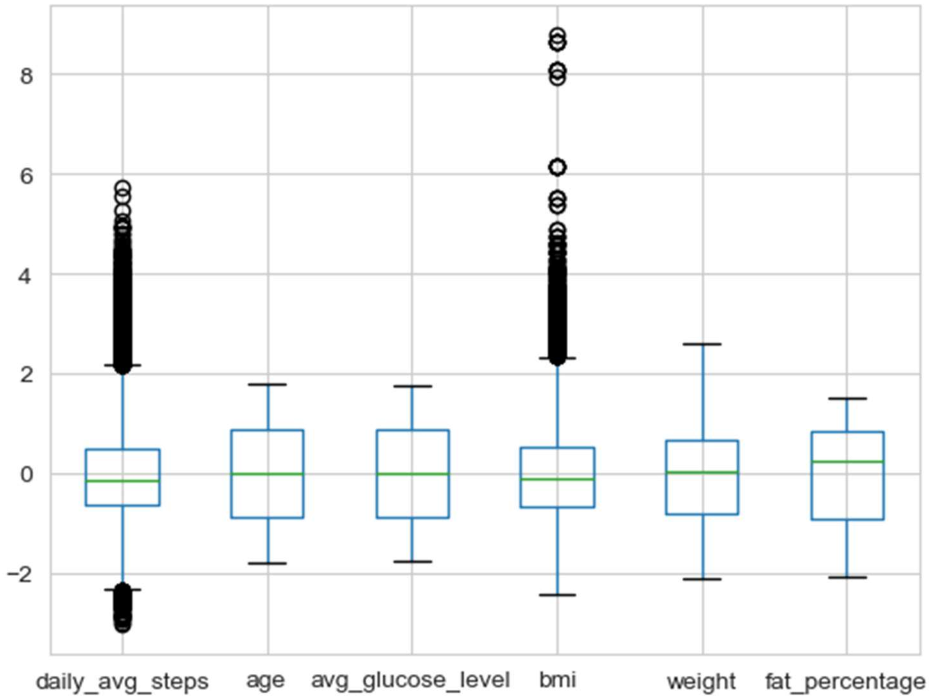


Fig.10. Boxplot of numerical variables before outlier treatment

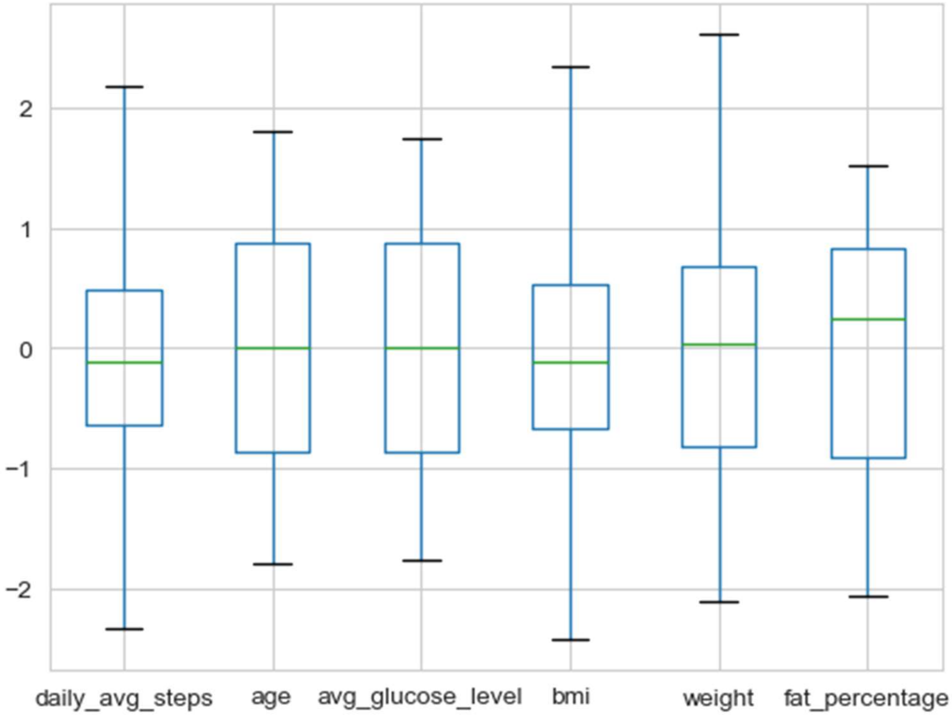


Fig.11. Boxplot of numerical variables after outlier treatment

Encoding Data:

- In order to enable further predictive modelling, the categorical columns needed to be converted to numerical variables
- Hence one-hot encoding was performed on columns- Location, Occupation, smoking_status, Alcohol, Gender, covered_by_any_other_company and exercise
- Label Encoding was done on cholesterol_level variable
- After encoding the shape of the data is (24010,40)

Need for feature elimination:

It is established that the target variable is the insurance cost. The EDA results established a linear relationship between weight and insurance cost. It was also established that there were correlations between the regular checkup, adventure sports, years of insurance variables and insurance cost. Other than these, the EDA was inconclusive. Out of 42 variables, only a handful seemed significant in the context of the target variable. Hence, there was a need to eliminate the non-significant variables. Since the dataset contained both categorical and continuous predictors, Recursive Feature Elimination was used and 15 significant predictors were selected out of 43 for model building.

Feature Elimination:

- The given dataset contains 40 fields after encoding, which includes one target variable
- From EDA, it was established that only a very few parameters, like weight have a significant influence on the target variable
- Certain variables, like location, for example, has too many categories, and after encoding got split into 11 different variables
- This is a huge increase in dimensionality, and is deemed unnecessary and cumbersome for further analysis
- Hence, there was a need to eliminate the insignificant variables in order to build efficient models
- Thus, Recursive Feature elimination was done based on a Random Forest Regressor model and 15 top features were identified
- The Random Forest model was chosen for this purpose, because it is a non-parametric model, which has the least assumptions involved

Clustering:

After RFE, an attempt was made to identify possible clusters from the selected features. For this:

- KMeans Clustering was done for the identified features
- After optimizing the number of clusters by silhouette scores and elbow plots, the optimum number of clusters were identified to be 3

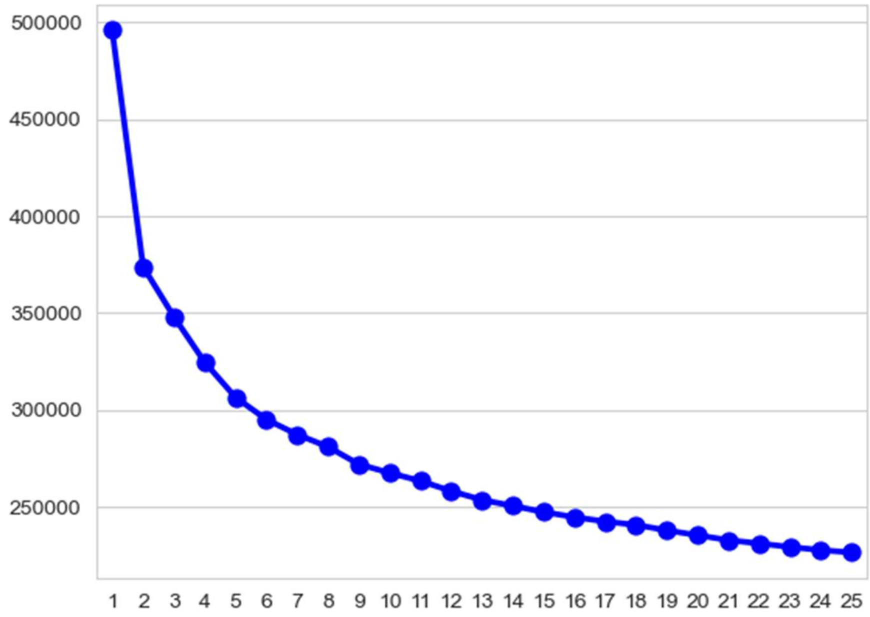


Fig.12. Elbow plot to identify optimum number of clusters

The clusters thus obtained were profiled as given in the table below:

Variables	**Cluster 0**	**Cluster 1**	**Cluster 2**
Weight (median)	67	72	75
Daily_avg_steps (median)	5112.5	5094.5	5058
Years_of_insurance(mode)	5	0	8
Bmi (mean)	31.36	31.52	31.24
Insurance_cost (median)	19744	27148	30850

Table.3. Cluster profiling

Observations:

- From the above table, we find that the most differentiating factors between the three clusters is the median weight
- Cluster 0 has the lowest median weight (67) and cluster 2 has the highest median weight (75)
- The median insurance cost, being proportional to weight, also reflects the same pattern, increasing from 19744 in cluster 0 to 30850 in cluster 2
- The ‘daily_average_steps’ is the highest in cluster 0 and lowest in cluster 2
- However, BMI is the lowest in cluster 2 and highest in cluster 1
- Also, Cluster 2 contains the most loyal customers, despite high median insurance cost and Cluster 1 contains the newest of the customer base
- All other parameters seemed to overlap with each other across clusters and hence are not tabulated

Advanced EDA based on Clustering:

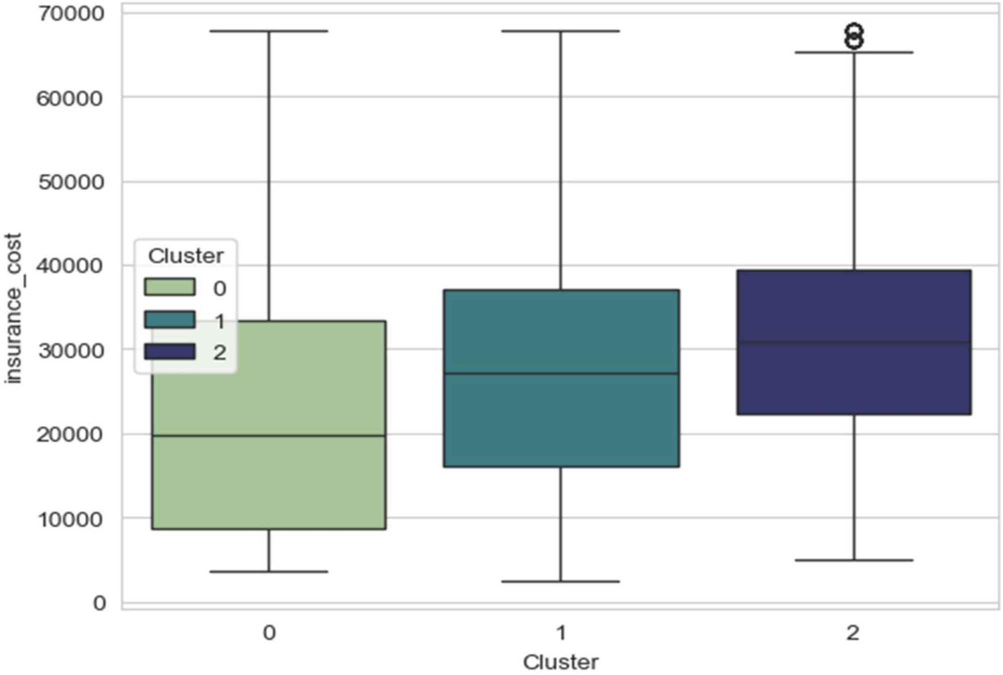


Fig.13. Insurance cost distribution across clusters

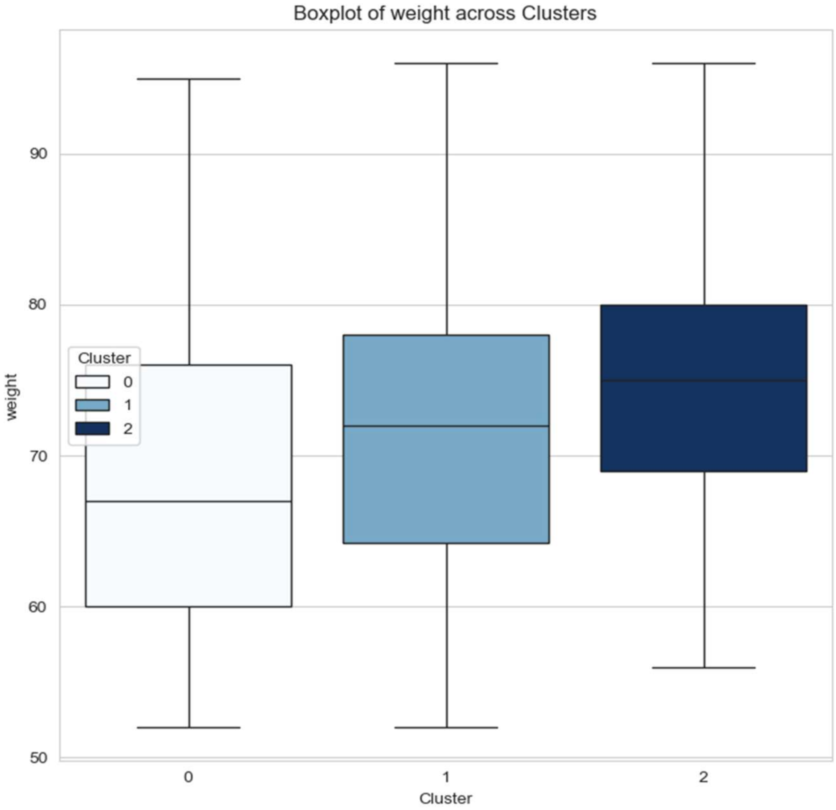


Fig.14. Weight Distribution across clusters

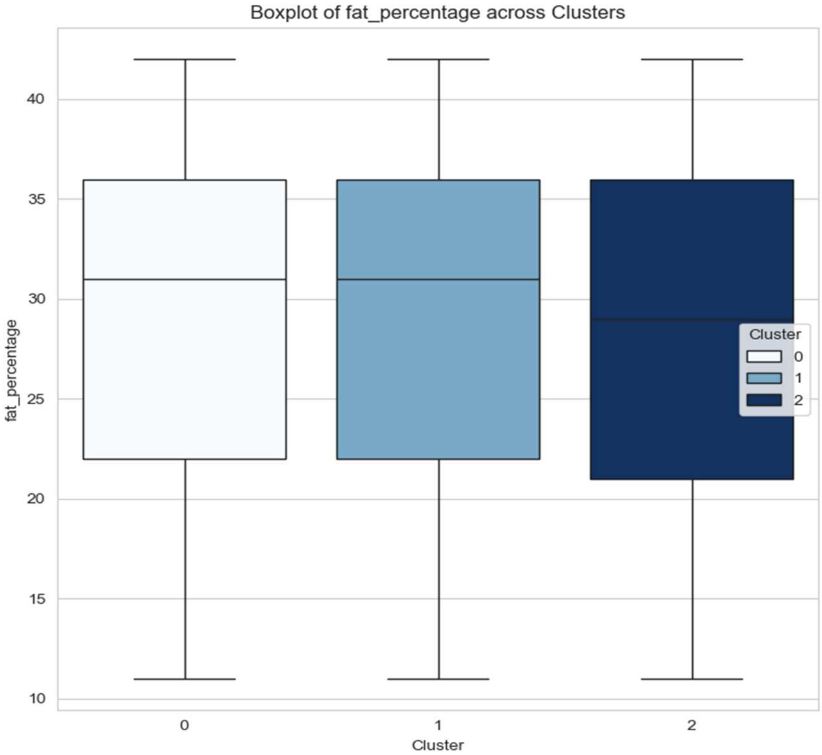


Fig.15. Fat percentage Distribution across clusters

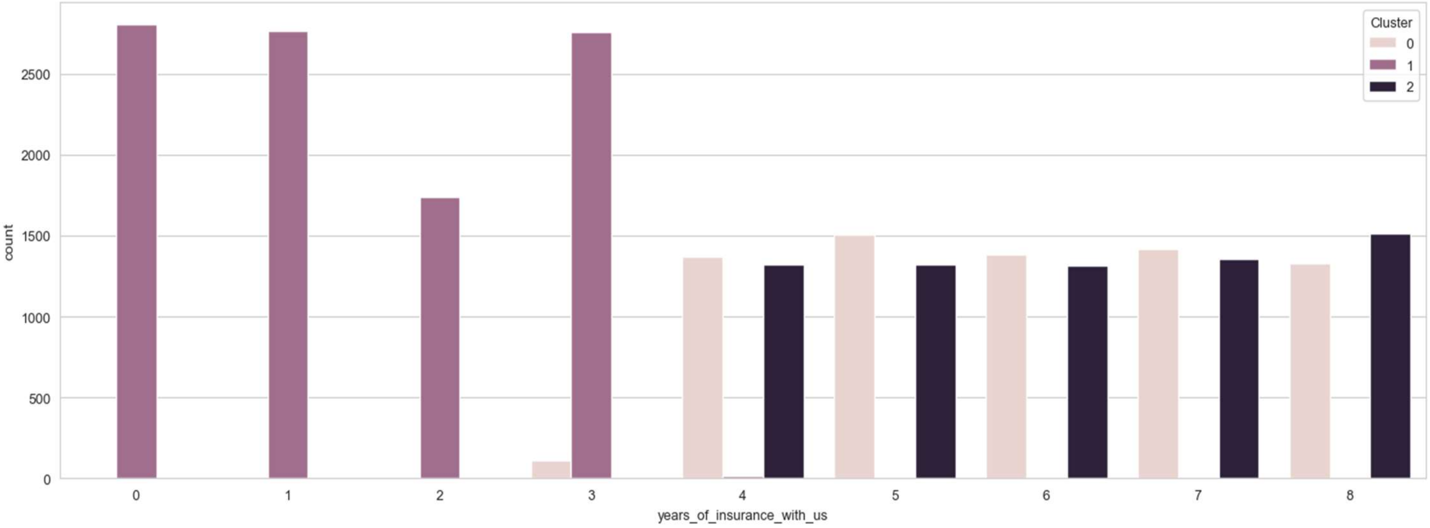


Fig.16. Clusters and years of insurance with us

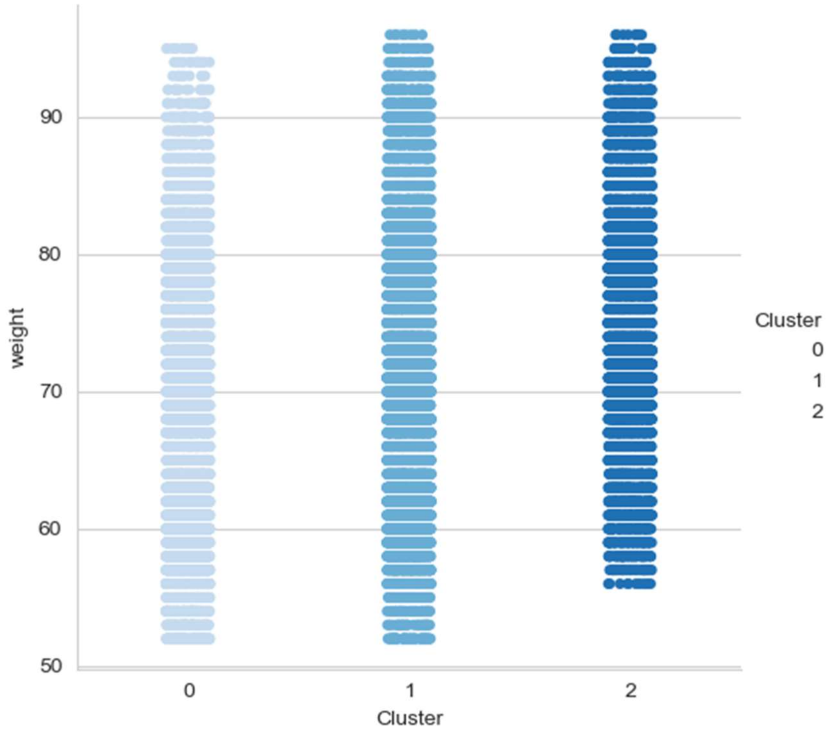


Fig.17. Catplot- Weight vs clusters

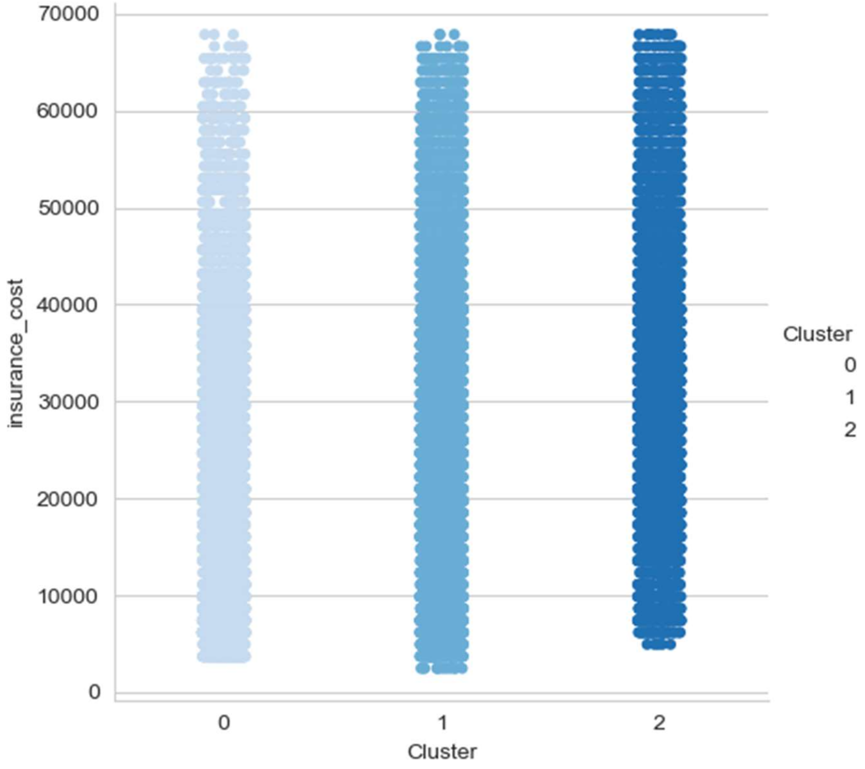


Fig.18. Catplot- Insurance cost vs clusters

Business Insights from EDA:

- **Identification of Business Opportunities:**
    - Around 3000 people in the dataset are entirely new to the health insurance stream. Hence, they can be targeted and possible add-ons to existing policies can be sourced
- **Risk Estimation:**
    - From the EDA, it is evident that the situations where the risk is high lead to higher insurance costs
    - That is, when people do not have regular checkups, or participate in adventure sports, or have a higher weight or an unknown smoking status, their insurance costs tend to be higher
    - This is a sound policy, which follows risk aversion
    - However, there are certain steps that can be recommended.
    - Encouraging the people to have regular checkups might lead to the updation of certain biomedical indices, like cholesterol level, which need to be updated frequently
    - This also avoids surprises relating to undiagnosed medical issues
    - This can be accomplished by rewarding the people with regular checkups by lowering the insurance costs, and penalizing those that don't
    - Also, emphasizing the disclosure of smoking_status can lead to better pricing, and risk mitigation
- **Identification of significant variables:**
    - Based on the EDA, the following variables were found to be significant in determining the insurance costs:
        - Weight
        - Regular checkups
        - Participation in Adventure sports
        - Fat percentage
    - These can be considered while designing new policies, and other insignificant variables like location could be dropped, or may be aggregated to form more meaningful and significant variables

MODEL BUILDING AND VALIDATION

Model Building:

Both parametric and non-parametric models were used for the same

- Parametric Models:
  - Linear Regression
  - Linear Discriminant Analysis
  - Elastic Net Regression
  - Ridge Regression
  - Lasso Regression
- Non-Parametric Models:
  - Random Forest
  - AdaBoost Regression
  - Gradient Boosting Regression
  - XGBoost Regression
  - kNN Regression

Performance Metrics:

Evaluation metrics such as

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE) and
- R squared value

were employed to quantify model performance. These metrics provided an understanding of predictive accuracy, error and explanatory power of the models.

Model Performance on train and test sets:

Model	RMSE	MAE	MAPE	R^2
Linear Regression	3371.46	2722.17	0.15	0.94
Ridge Regression	3371.46	2722.15	0.15	0.94
Lasso Regression	3371.46	2722.14	0.15	0.94
LDA	3315.99	2532.24	0.12	0.19
Elastic Net Regression	5772.59	4621.97	0.28	0.84
Random Forest Regression	1173.32	931.70	0.04	0.99
SV Regression	14027.18	11593.09	0.79	0.04
AdaBoost Regression	3284.24	2717.89	0.16	0.95
Gradient Boosting Regression	2992.48	2416.04	0.12	0.96
XGB RF Regression	3019.66	2429.63	0.12	0.96
kNN Regression	5173.40	4072.35	0.22	0.87

Table.4. Model performance on train set

Model	RMSE	MAE	MAPE	R²
Linear Regression	3353.28	2705.05	0.15	0.94
Ridge Regression	3353.29	2705.04	0.15	0.94
Lasso Regression	3353.28	2705.05	0.15	0.94
LDA	3331.49	2572.85	0.12	0.17
Elastic Net Regression	5786.39	4631.28	0.27	0.84
Random Forest Regression	3111.31	2499.73	0.12	0.95
SV Regression	13981.45	11491.7	0.77	0.04
AdaBoost Regression	3259.43	2679.64	0.15	0.95
Gradient Boosting Regression	3015.24	2434.97	0.12	0.96
XGB RF Regression	3045.55	2452.21	0.12	0.95
kNN Regression	6274.12	4945.9	0.27	0.81

Table.5. Model performance on test set

Model Interpretation:

- Linear, Ridge and Lasso Regression:
  - The performance metrics (RMSE, MAE, MAPE, R²) for these linear regression models are nearly identical, suggesting that regularization techniques like Ridge and Lasso did not significantly impact their performance
  - The high R² value (approximately 0.94) indicates that these models explain a substantial amount of the variance in the target variable
- Linear Discriminant Analysis:
  - LDA shows competitive performance with lower RMSE and MAE compared to linear regression models
  - However, the lower R² (0.17) suggests that LDA may not capture the variability in the target variable as effectively
- Elastic Net Regression:
  - Elastic Net Regression shows a higher RMSE, MAE, and MAPE compared to linear regression models, indicating a less accurate prediction
  - The R² value (0.84) suggests that the model explains a significant portion of the variance but falls short compared to other models
- Random Forest Regression:
  - Random Forest Regression demonstrates strong performance in the train dataset, but only the third best in the test set RMSE (3111.31) making it an overfit model
  - The R² value (0.95) indicates a high level of explained variance, making it a robust choice for accurate predictions
- Support Vector regression:
  - SV Regression exhibits the highest RMSE, MAE, and MAPE values, suggesting poorer predictive performance compared to other models
  - The R² value of 0.04 indicates that the model does not capture any significant variance in the target variable
- AdaBoost Regression:

- AdaBoost Regression performs well with moderate RMSE and MAE values and a respectable R² value (0.95)

- Gradient Boosting Regression:
  - Gradient Boosting Regression demonstrates superior performance with the lowest RMSE (3015.24) and MAE (2434.97) among all models
  - The high R² value (0.96) suggests excellent explanatory power, making it the best-performing model
- XG Boost RF Regression:
  - XGBoost RF Regression shows competitive performance with a slightly higher RMSE and MAE compared to Gradient Boosting Regression
  - The R² value (0.95) indicates a strong ability to explain variance in the target variable
- Knn Regression
  - The performance of the kNN Regressor was also poor, and is the second poorest performing model in terms of all the metrics

Summary:

- Random Forest Regression, Gradient Boosting Regression, and XGBoost RF Regression stand out as top-performing models with low RMSE and high R² values
- Linear regression models and ensemble methods like AdaBoost also show competitive performance
- Support Vector Regression, Elastic Net Regression and KNN Regression exhibit comparatively lower predictive accuracy

Choice of Models for tuning:

Based on the performance in Table.1, the following models were identified for tuning:

- Ridge Regression and Lasso regression:
  - The regularization of Linear Regression did not result in better performance.
  - Hence, these were chosen to be tuned in the hope for a better performance
- Elastic Net Regression:
  - Ideally the elastic net regression combines the regularization factors of ridge and lasso, and hence is bound to perform well
  - However, in the given dataset, despite better performances by Ridge and Lasso, EN model performed very poorly.
  - Hence, it was chosen to improve the performance
- Random Forest, AdaBoost, Gradient Boost and XGBRF Regression:
  - These were the top performing models without tuning.
  - Hence, these were chosen in order to obtain better results

Tuning Process:

- Hyperparameter Tuning:
  - The primary focus was on hyperparameter tuning, a crucial step to optimize model performance. RandomizedSearchCV and GridSearchCV was employed for efficient exploration of hyperparameter spaces:
    - In the case of linear models (Linear Regression, Ridge, Lasso), tuning involved adjusting regularization strength and related parameters
    - For tree-based models (Random Forest, AdaBoost, Gradient Boosting, XGBoost RF), parameters like the number of estimators, maximum depth, and learning rate were tuned
    - For Random Forest, Gradient Boost Regressor, and XGBRF Regressor, hyperparameters were randomly sampled from specified distributions to avoid exhaustive searches while allowing for comprehensive coverage
    - For AdaBoost regressor hyperparameters were obtained from grid search

- Cross- Validation:
  - Cross-validation was utilized during the hyperparameter tuning process to ensure robustness and reliability of model performance assessment. This involved splitting the training data into multiple folds (3 or 5), training the model on different subsets, and evaluating its performance on the remaining data

Model Performance on train and test sets after tuning:

Model	RMSE	MAE	MAPE	R²
Tuned Ridge Regression	3371.46	2722.09	0.15	0.94
Tuned Lasso Regression	3372.01	2722.23	0.15	0.94
Tuned EN Regression	3372.01	2722.23	0.15	0.94
Tuned AB Regression	3233.60	2630.26	0.14	0.95
Tuned GB Regression	2935.19	2639.40	0.12	0.96
Tuned XGBRF Regression	10246.34	8453.68	0.58	0.49
Tuned RF Regression	2482.17	1980.02	0.10	0.97

Table.6. Model performance on train set after tuning

Model	RMSE	MAE	MAPE	R²
Tuned Ridge Regression	3353.32	2705	0.15	0.94
Tuned Lasso Regression	3353.41	2705.27	0.15	0.94
Tuned EN Regression	3353.41	2705.27	0.15	0.94
Tuned AB Regression	3219.21	2600.65	0.14	0.95
Tuned GB Regression	3017.47	2434.18	0.12	0.96
Tuned XGBRF Regression	10242.48	8393.32	0.56	0.49
Tuned RF Regression	3061.93	2462.18	0.12	0.95

Table.7. Model performance on test set after tuning

Interpretation of Model Performance after tuning:

- Ridge and Lasso Regression:
    - The tuned Ridge and Lasso Regression models exhibit similar performance to their untuned counterparts. Hyperparameter tuning did not result in significant improvements.
    - These models maintain a high level of explanatory power (R² approximately 0.94) with consistent RMSE, MAE, and MAPE
- AdaBoost Regression:
    - The tuned AdaBoost Regression model shows improvement across all metrics compared to its baseline. Notably, there is a reduction in RMSE and MAE, suggesting enhanced predictive accuracy.
    - The R² value also increases to 0.95, indicating a better fit to the data
- Random Forest Regression:
    - The tuned Random Forest Regression model maintains consistent performance with the baseline. It also continues to be an overfit model even after tuning
    - It continues to be a robust performer with a low RMSE (3061.93) and high R² (0.95).
    - This model retains its strength in predictive accuracy and interpretability
    - However, it is extremely overfitting, with the train performance in ideality, while the test performance is on par with that of the other models
- Gradient Boosting Regression:

    - Hyperparameter tuning for Gradient Boosting Regression leads to improvements.
    - The tuned model achieves the lowest MAE (2430.37) among all models. The R² value of 0.96 signifies exceptional explanatory power and a better overall fit to the data
    - It still remains the best performing model on the dataset
- XG Boost RF Regression:
    - The tuned XGBoost RF Regression model, despite having a higher RMSE compared to the baseline, experiences a significant drop in MAE. However, the R² value decreases to 0.49, suggesting a reduction in explanatory power
    - Hence, the model performance can be considered to have declined after tuning

Summary:

- Ridge, Lasso, and Elastic Net Regression models show limited sensitivity to hyperparameter tuning in this context, possibly due to the nature of the regularization techniques applied
- AdaBoost Regression benefits from tuning, demonstrating improved accuracy and a better fit to the data
- XGBoost RF Regression, while achieving competitive accuracy, seems to have a potential trade-off between accuracy and model interpretability
- Gradient Boosting Regression stands out as the top performer, showcasing the most significant enhancements in predictive accuracy and explanatory power

Best Performing model and interpretations:

- Model Performance and accuracy:
    - The tuned Gradient Boosting Regression model demonstrates exceptional predictive accuracy, as evidenced by its low Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The R² value of 0.96 indicates that the model explains a significant portion of the variability in the target variable.
    - These metrics suggest that the model provides highly accurate predictions, minimizing errors and maximizing the fidelity of the predictions to the actual outcomes
- Explanatory power and feature importances:
    - The high R² value implies that the model effectively captures the relationships and patterns within the dataset. In a business context, this means that the selected features (by RFE) have a significant impact on the predicted targets.

Business Implications:

- Enhanced Decision Making:
    - The accurate predictions provided by the Gradient Boosting model can empower decision-makers within the business.
    - Pricing strategies can be devised based on the model as the model's precision allows for more informed and effective decision-making.
- Customer Segmentation and Targeting:
    - The model's ability to identify and leverage important features can inform targeted marketing and customer segmentation strategies.
    - Customizing marketing efforts based on the identified influential factors can enhance customer engagement and satisfaction
- Risk Management:
    - The predictive capabilities of the model can contribute to risk assessment and management

- For example, the regular health checkup was found to be an important feature, and most of the people have not had any regular health checkup the previous year. This is an area of potential risk

Business Insights:

The below table gives the result of the feature importances derived from the tubed gradient boosting model:

Features	Importance
weight	0.99307726
covered_by_any_other_company_Y	0.002269823
regular_checkup_last_year	0.001879571
weight_change_in_last_one_year	0.000557934
daily_avg_steps	0.000480724
years_of_insurance_with_us	0.000433986
avg_glucose_level	0.000404933
age	0.000339663
bmi	0.000303353
fat_percentage	0.000118858

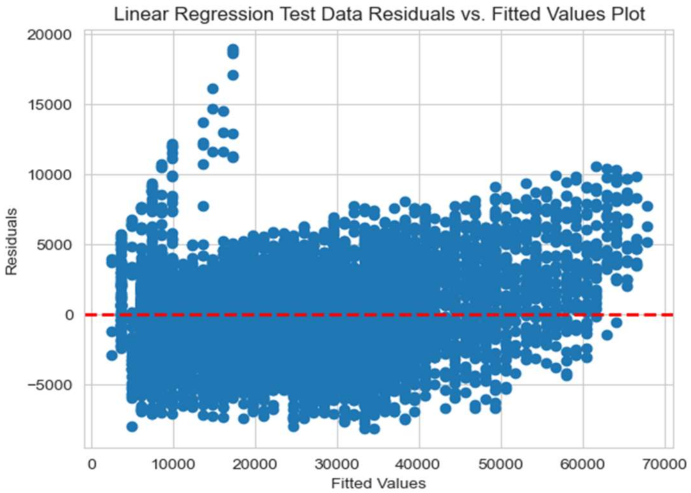
Table.8. Features in the dataset and their importance

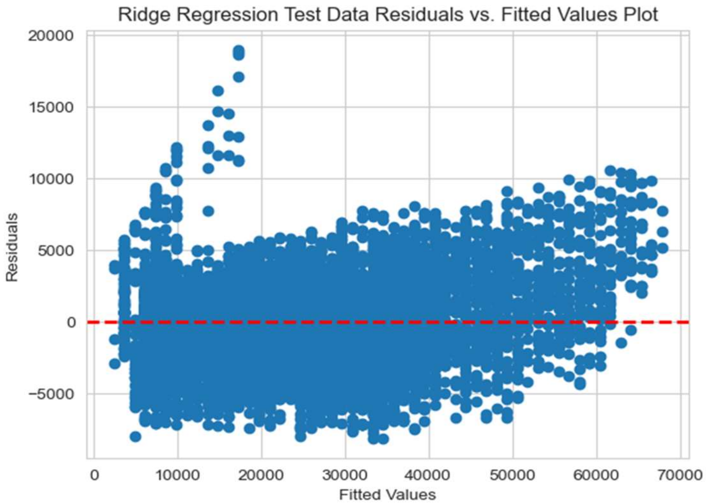
- Weight is the most significant parameter, outranks all other variables
- Influence of dual coverage- 29% have dual coverage
- Importance of regular health checkups – 58% have 0 regular health checkups
- Weight change and daily average steps – most are lightly active
- Customer loyalty is also a factor in the cost determination
- Medical parameters take a backseat, though present
- Age, BMI and fat percentage also play a subtle role
- Location and gender do not seem to have a role

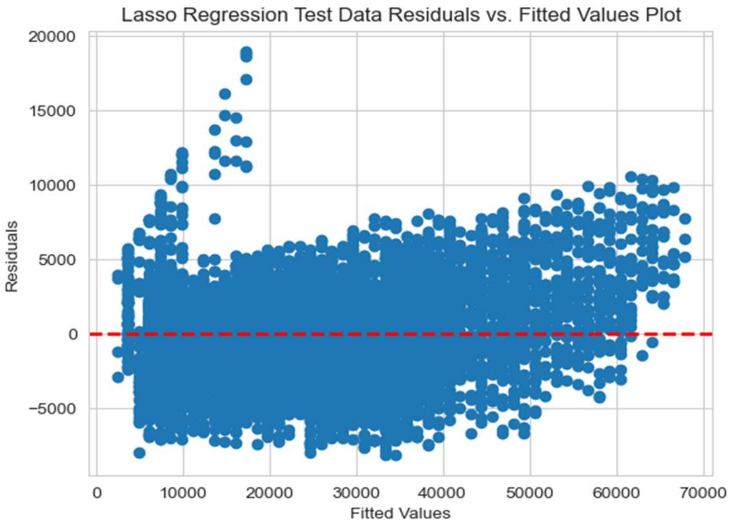
Business Recommendations:

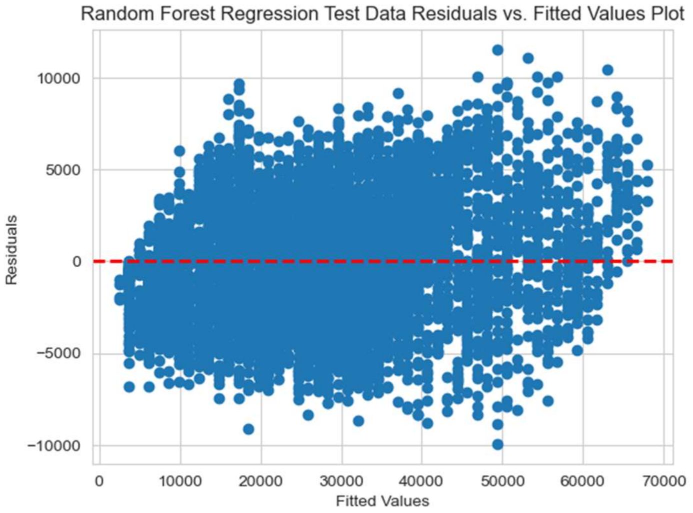
- Establishing the significance of weight among the customers and rewarding healthy weight loss
- Investigating the people with dual insurance and providing a more comprehensive package
- Penalizing irregular checkups more strenuously in order to avoid potential risks
- Rewarding customer loyalty by add-ons or increased coverage
- Medical indicators need to be updated more frequently and the premium should be revised accordingly
- This could be made more dynamic, either by regular data collection through customer outreach programs or by teaming with fitness apps, for example
- From the current structure, history of diseases do not seem to be significantly impacting the insurance cost. There is a need to investigate this
- Designing specialized policies after segmentation based on age, gender and other features
- Personalized packages for every single applicant can be thought of, rather than a general range or pre-determined cost
- Other biomedical factors also can be incorporated in the study to build a robust and comprehensive model

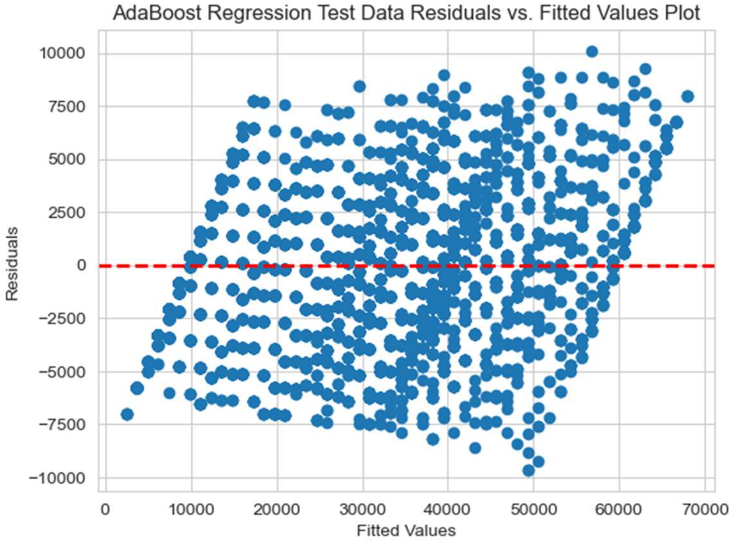
APPENDIX

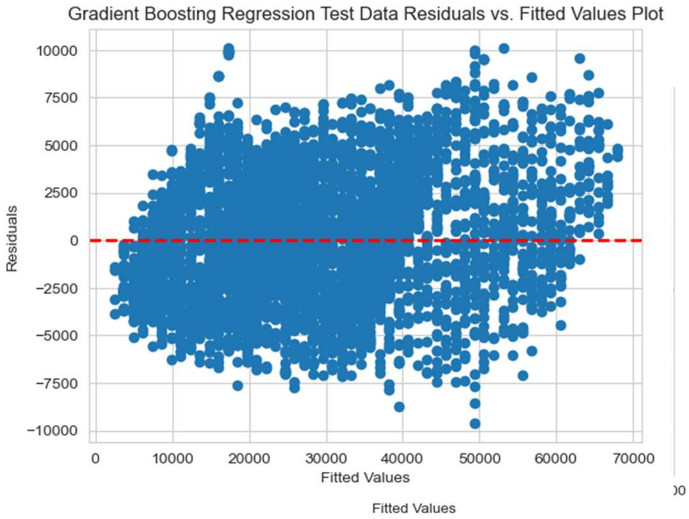

Linear Regression Test Data Residuals vs. Fitted Values Plot

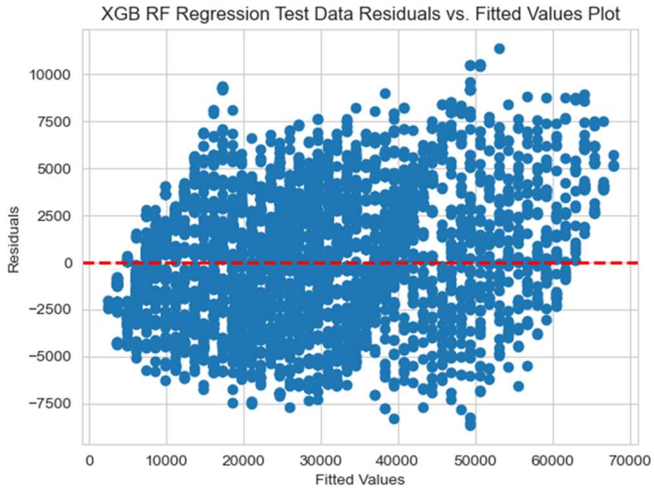

Ridge Regression Test Data Residuals vs. Fitted Values Plot

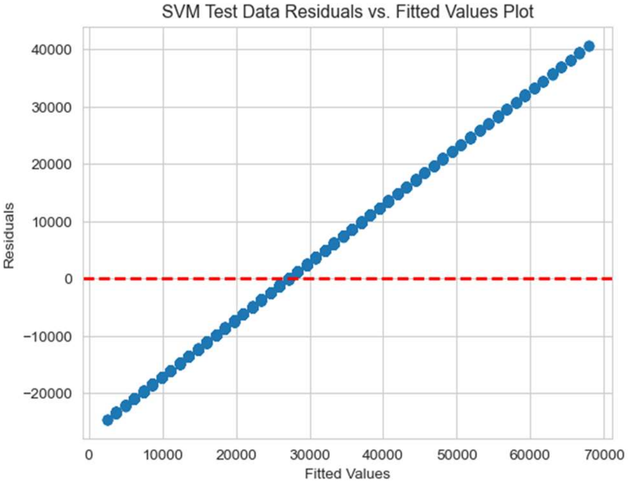

Lasso Regression Test Data Residuals vs. Fitted Values Plot

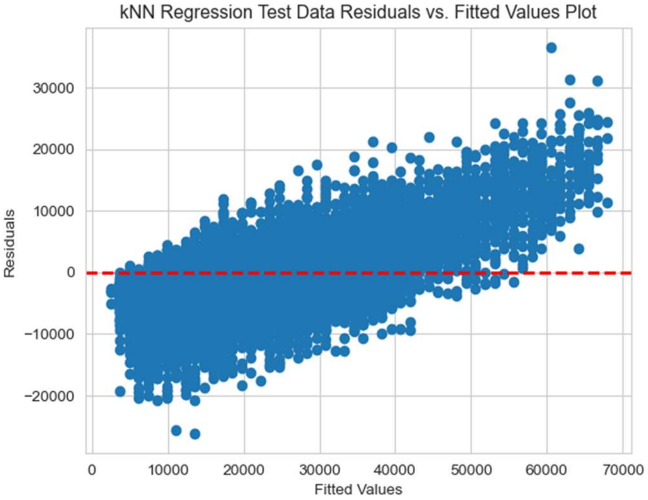

Random Forest Regression Test Data Residuals vs. Fitted Values Plot
Residuals
10000
5000
0
−5000
−10000
Fitted Values
0
10000
20000
30000
40000
50000
60000
70000

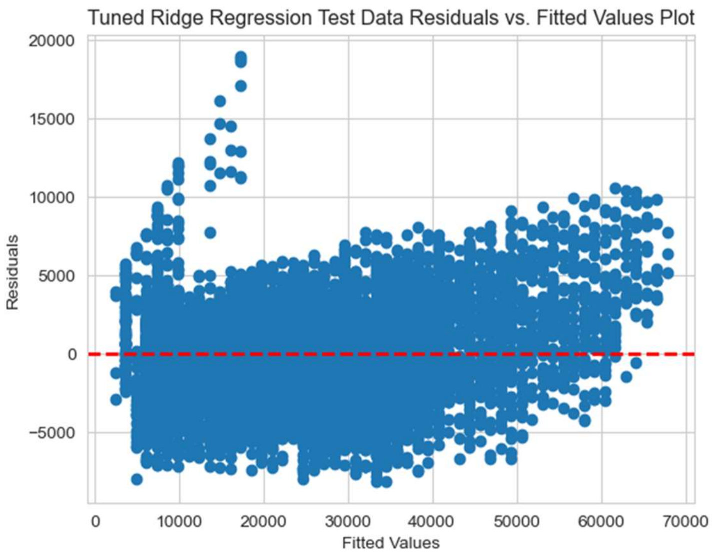

AdaBoost Regression Test Data Residuals vs. Fitted Values Plot
Residuals
10000
7500
5000
2500
0
−2500
−5000
−7500
−10000
Fitted Values
0
10000
20000
30000
40000
50000
60000
70000

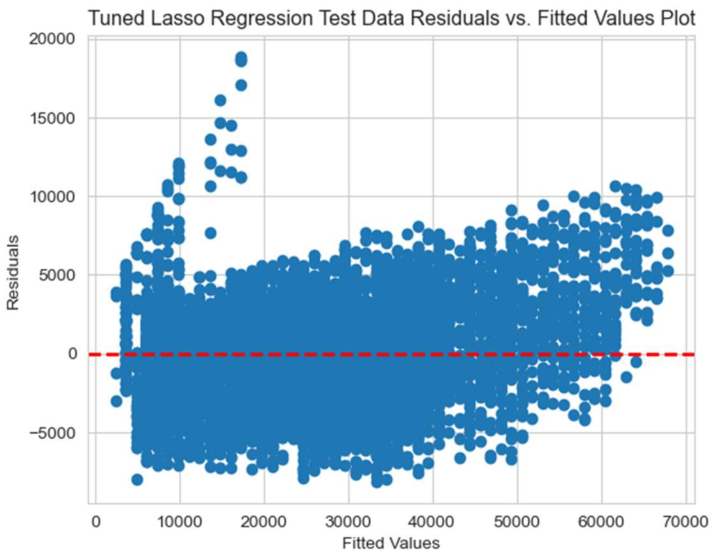

Gradient Boosting Regression Test Data Residuals vs. Fitted Values Plot
Residuals
10000
7500
5000
2500
0
−2500
−5000
−7500
−10000
Fitted Values
0
10000
20000
30000
40000
50000
60000
70000
Fitted Values
)0

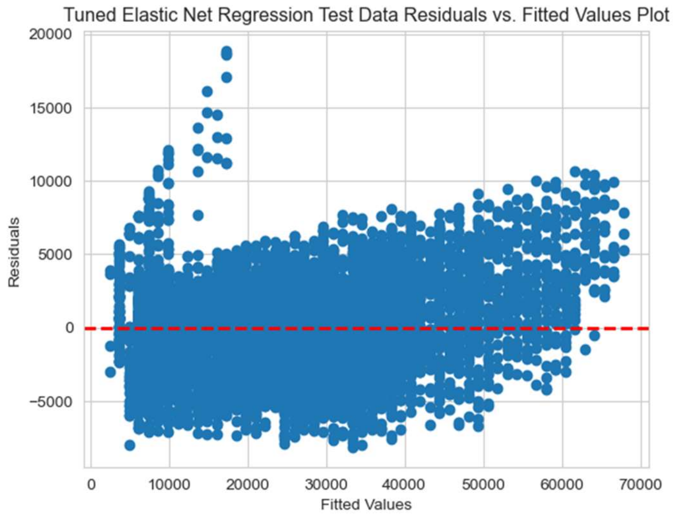

XGB RF Regression Test Data Residuals vs. Fitted Values Plot
Residuals
10000
7500
5000
2500
0
−2500
−5000
−7500
Fitted Values
0
10000
20000
30000
40000
50000
60000
70000

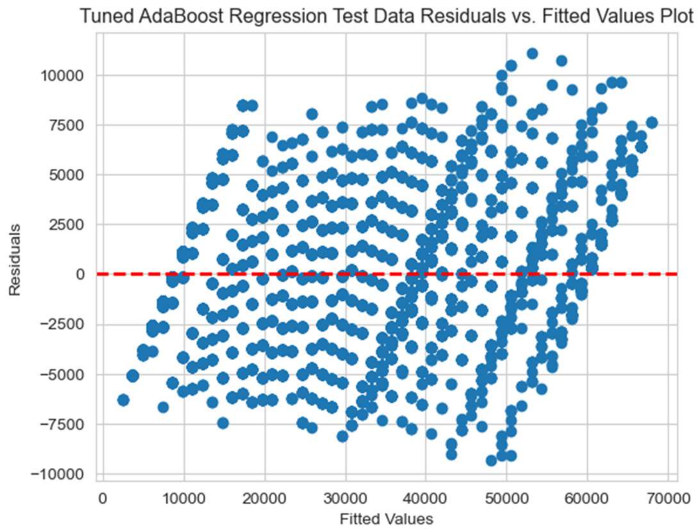

SVM Test Data Residuals vs. Fitted Values Plot
Residuals
Fitted Values
40000
30000
20000
10000
0
−10000
−20000
0
10000
20000
30000
40000
50000
60000
70000

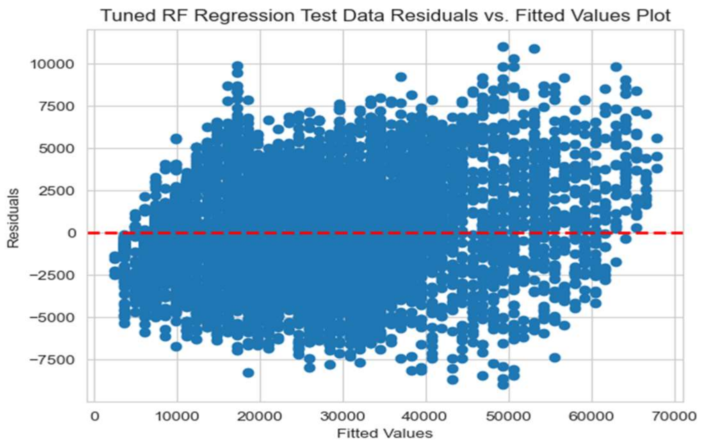

kNN Regression Test Data Residuals vs. Fitted Values Plot
Residuals
Fitted Values
30000
20000
10000
0
−10000
−20000
0
10000
20000
30000
40000
50000
60000
70000

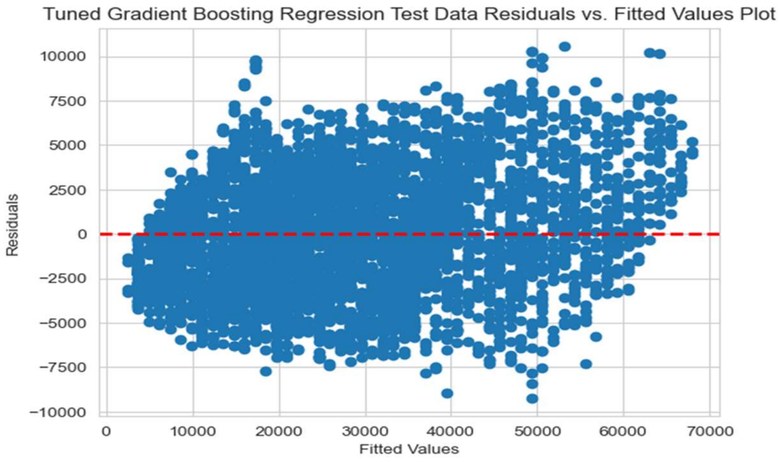

Tuned Ridge Regression Test Data Residuals vs. Fitted Values Plot


Tuned Lasso Regression Test Data Residuals vs. Fitted Values Plot


Tuned Elastic Net Regression Test Data Residuals vs. Fitted Values Plot


Tuned AdaBoost Regression Test Data Residuals vs. Fitted Values Plot


Tuned RF Regression Test Data Residuals vs. Fitted Values Plot


Tuned Gradient Boosting Regression Test Data Residuals vs. Fitted Values Plot