CAPSTONE PROJECT

INSURANCE COST-HEALTHCARE PROJECT

PROJECT NOTES 2

—

VIDYA V

—

PGPDSBA.O.2023B

CONTENTS

LIST OF FIGURES

LIST OF TABLES

BACKGROUND

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

Steps done at the last stage:

- Exploring the given dataset, understand the various predictor variables and their natures
- 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
- Elimination of non-significant variables by RFE
- Exploring the data by segmentation, identify segments and relate them to insurance cost

Steps at Current Stage:

- 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, R Square, MAE and MAPE
- 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
- Interpretation of model and business implications

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

MODEL BUILDING AND INTERPRETATION

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.

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 test set:

Model	RMSE	MAE	MAPE	R^2
Linear Regression	3353.23	2705.54	0.15	0.94
Ridge Regression	3353.24	2705.53	0.15	0.94
Lasso Regression	3353.23	2705.54	0.15	0.94
LDA	3326.37	2570.28	0.12	0.17
Elastic Net Regression	5786.46	4631.22	0.27	0.84
Random Forest Regression	3118.31	2505.53	0.12	0.95
SV Regression	14265.68	11726.16	0.79	0
AdaBoost Regression	3259.43	2679.64	0.15	0.95
Gradient Boosting Regression	3014.01	2433.56	0.12	0.96
XGB RF Regression	3045.59	2451.01	0.12	0.95
kNN Regression	7209.08	5650.15	0.32	0.75

Table.1. Model performance on test set

Model Interpretation:

- Linear, Ridge and Lasso Regression:
  - The performance metrics (RMSE, MAE, MAPE, R^2) for these linear regression models are nearly identical, suggesting that regularization techniques like Ridge and Lasso did not significantly impact their performance
  - The high R^2 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 R^2 (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 R^2 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 with the third best RMSE (3118.31) among all models
  - The R^2 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 R^2 value of 0 indicates that the model does not capture any variance in the target variable

- AdaBoost Regression:
    - AdaBoost Regression performs well with moderate RMSE and MAE values and a respectable R^2 value (0.95)
- Gradient Boosting Regression:
    - Gradient Boosting Regression demonstrates superior performance with the lowest RMSE (3014.01) and MAE (2433.56) among all models
    - The high R^2 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 R^2 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

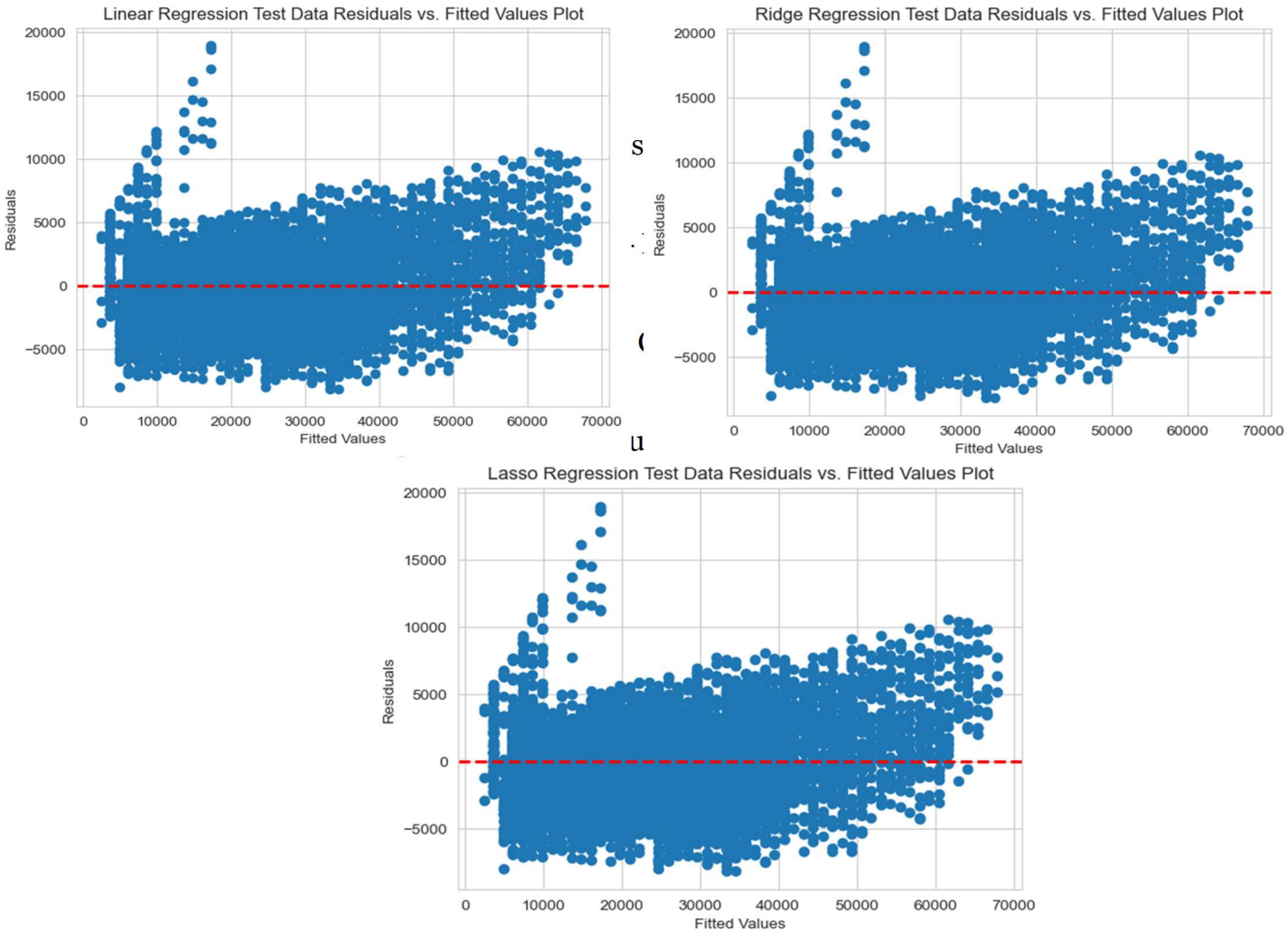


Fig.1. Linear, Ridge and Lasso Regression Residual plots

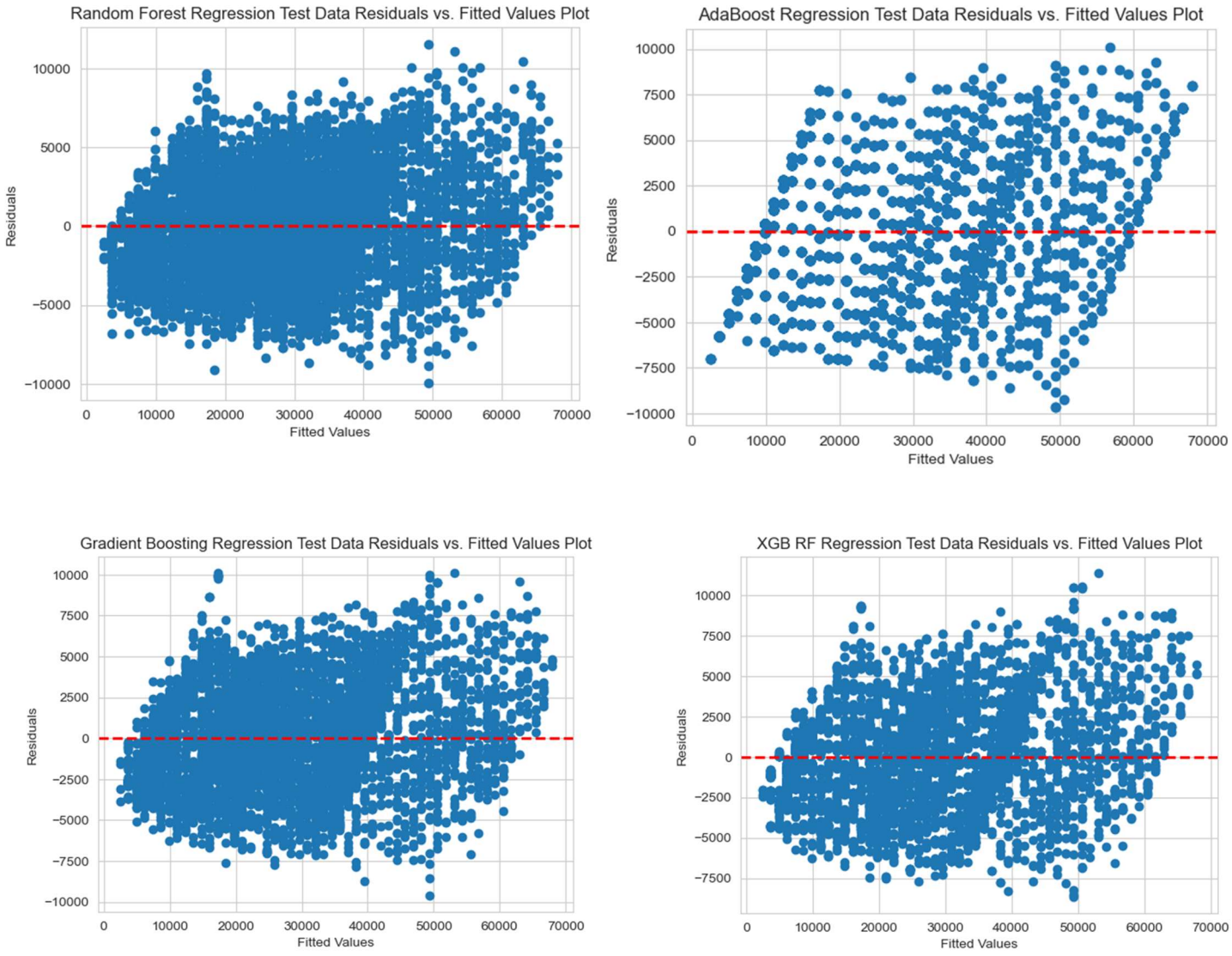


Fig.2. Better performing Models- Random Forest, AdaBoost, Gradient Boost and XGBRF Regression Residual plots

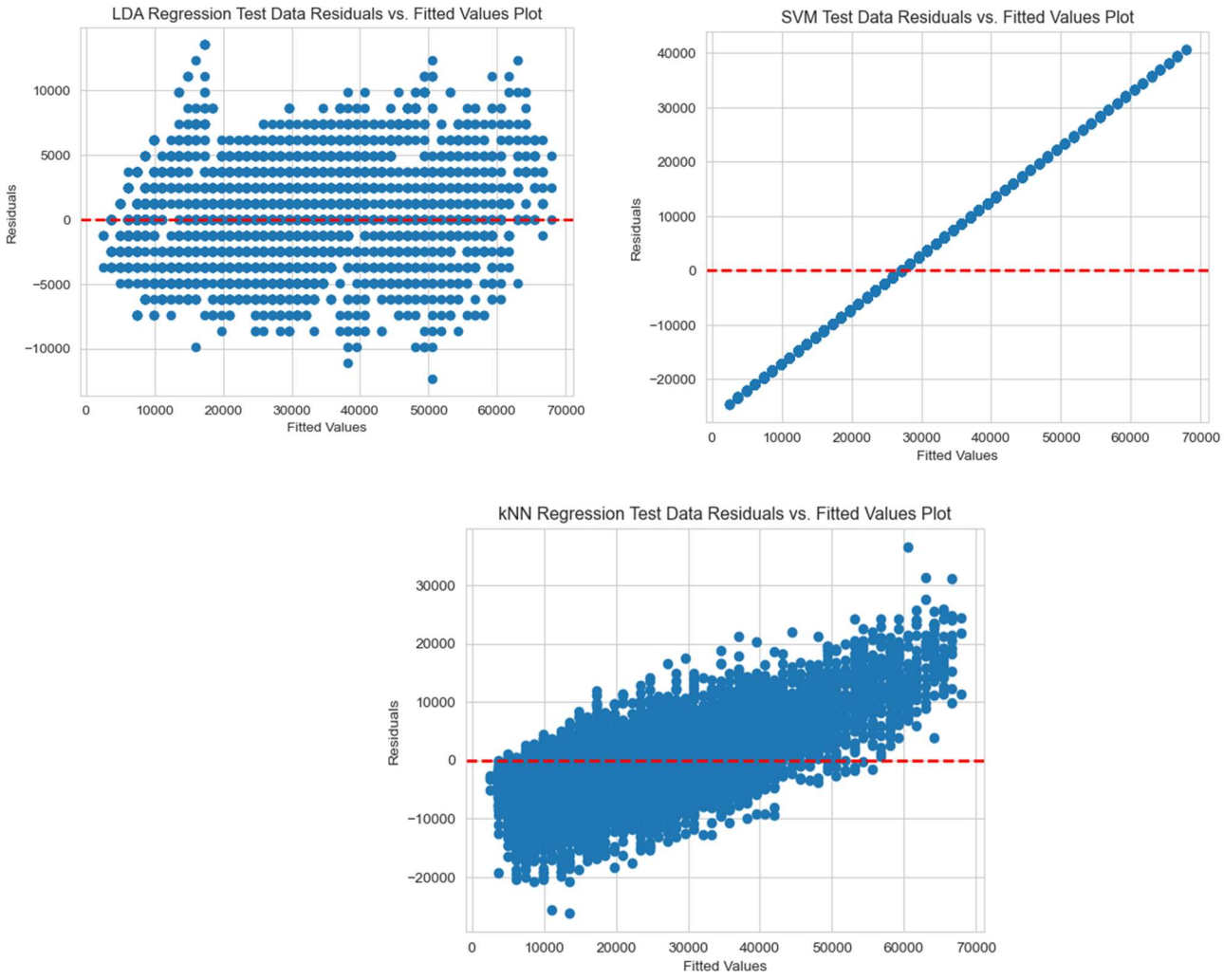


Fig.3. Poor performing models- LDA, SVR and kNN Regression residual plots

Summary:

- Random Forest Regression, Gradient Boosting Regression, and XGBoost RF Regression stand out as top-performing models with low RMSE and high R^2 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

MODEL TUNING

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:

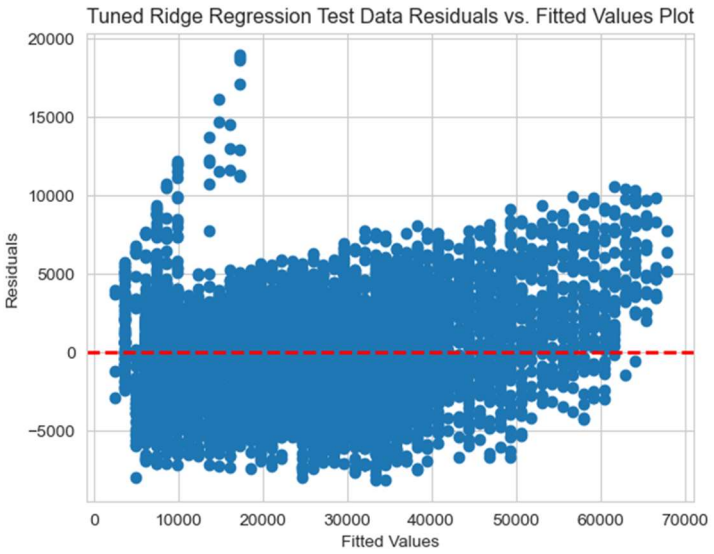
- Hyperparameter Tuning:
  - The primary focus was on hyperparameter tuning, a crucial step to optimize model performance. RandomizedSearchCV and GridSearchCV was employed for efficient exploration of hyperparameter spaces:
    - In the case of linear models (Linear Regression, Ridge, Lasso), tuning 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

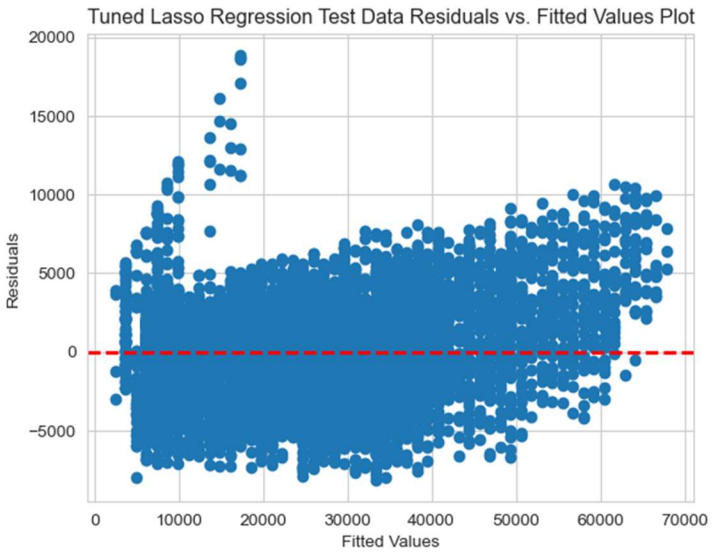
- Cross- Validation:
  - Cross-validation was utilized during the hyperparameter tuning process to ensure robustness and reliability of model performance assessment. This involved splitting the training data into multiple folds (3 or 5), training the model on different subsets, and evaluating its performance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

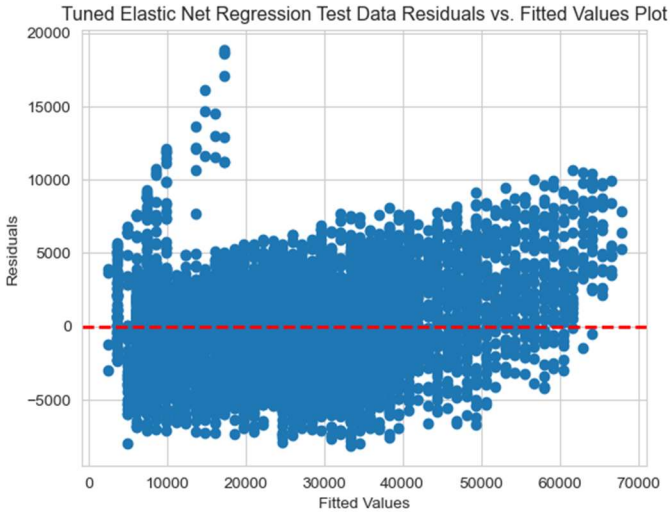
Model Performance on test set after tuning:

Model	RMSE	MAE	MAPE	R^2
Tuned Ridge Regression	3353.27	2705.48	0.15	0.94
Tuned Lasso Regression	3353.5	2705.62	0.15	0.94
Tuned EN Regression	3353.5	2705.62	0.15	0.94
Tuned AB Regression	3229.27	2614.64	0.14	0.95
Tuned GB Regression	3013.97	2430.37	0.12	0.96
Tuned XGBRF Regression	10243.4	8393.92	0.56	0.49
Tuned RF Regression	3062.76	2462.16	0.12	0.95

Table.2. Model performance on test set after tuning







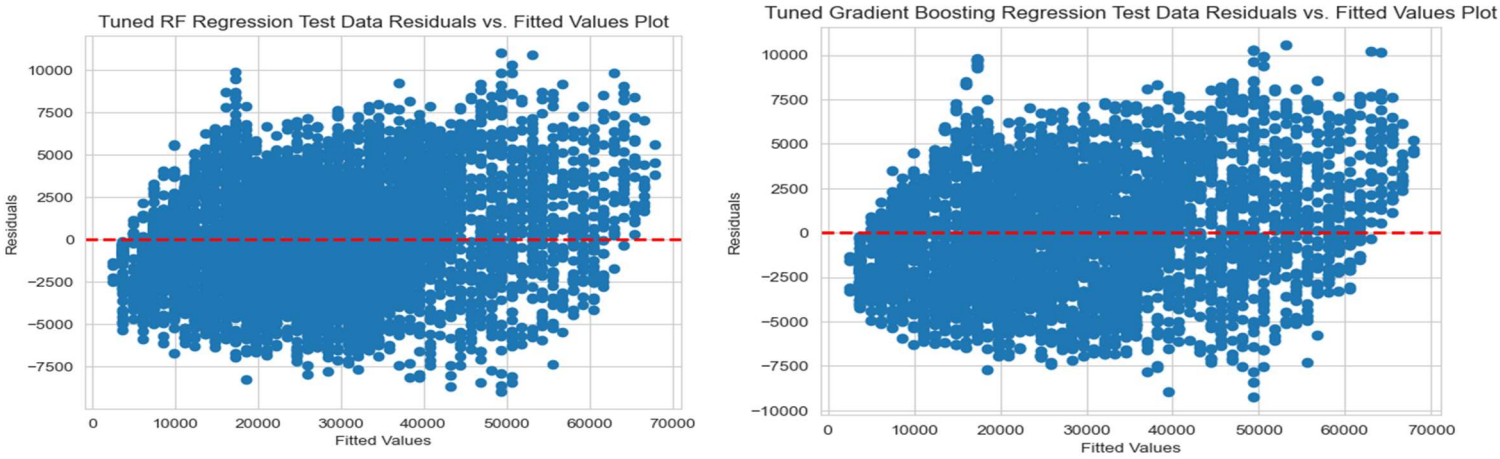


Fig.4. Residual plots of models after tuning

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 (R^2 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 R^2 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 continues to be a robust performer with a low RMSE (3062.76) and high R^2 (0.95).
  - This model retains its strength in predictive accuracy and interpretability
- Gradient Boosting Regression:
  - Hyperparameter tuning for Gradient Boosting Regression leads to substantial improvements.
  - The tuned model achieves the lowest RMSE (3013.97) and MAE (2430.37) among all models. The R^2 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 R^2 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 R^2 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 R^2 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