Capstone Final Report

Financial Insurance Modeling Report

Executive Summary

This report outlines the analysis conducted on a financial insurance dataset to understand factors influencing customer responses to insurance offers. The goal was to identify key patterns and build predictive models to enhance customer engagement strategies.

Key Recommendations

1. **Targeted Marketing:** Focus marketing efforts on specific customer segments with higher probabilities of positive responses, as identified by the Gradient Boosting model.
2. **Policy Optimization:** Adjust insurance policies and premiums based on customer profiles and predicted responsiveness.
3. **Health Indicator Utilization:** Incorporate health indicators more prominently in customer engagement strategies to personalize insurance offers.

Introduction

Background

The client operates in the insurance industry and aims to improve customer engagement by predicting customer responses to insurance offers. The dataset includes various customer attributes, policy details, and response outcomes.

Problem Statement

The primary objective is to develop a predictive model to identify customers most likely to respond positively to insurance offers. This will enable more efficient allocation of marketing resources and personalized policy offerings.

Data Overview

Data Description

The dataset contains 50,882 records with the following key variables:

- **Customer Demographics:** City_Code, Region_Code, Accomodation_Type, Upper_Age, Lower_Age, Is_Spouse, Health Indicator.
- **Policy Details:** Holding_Policy_Duration, Holding_Policy_Type, Reco_Insurance_Type, Reco_Policy_Cat, Reco_Policy_Premium.
- **Response:** Customer's response to the recommended insurance (target variable).

Data Preparation

- **Cleaning:** Removed duplicates and handled missing values.
- **Imputation:** Replaced missing values in 'Health Indicator', 'Holding_Policy_Duration', and 'Holding_Policy_Type' with appropriate substitutes.
    - **Health Indicator:** Missing values were replaced with 'X0'.
    - **Holding_Policy_Duration:** Missing values were treated as zero.
    - **Holding_Policy_Type:** Missing values were treated as zero.
- **Transformation:** Created new features such as 'Cust_Type' (New/Old), 'Age_Conf_Interval' (Upper_Age - Lower_Age), and logarithmic transformation of 'Reco_Policy_Premium'.

Exploratory Data Analysis

Summary of Findings

- **Demographics:** Older customers (age > 40) are more likely to respond positively to insurance offers.
- **Policy Types:** Joint policies tend to have higher premiums, but individual policies have higher response rates.
- **Health Indicators:** Certain health conditions, such as 'X7', 'X8', and 'X9', correlate strongly with positive responses.

Visualizations

1. Age vs. Response

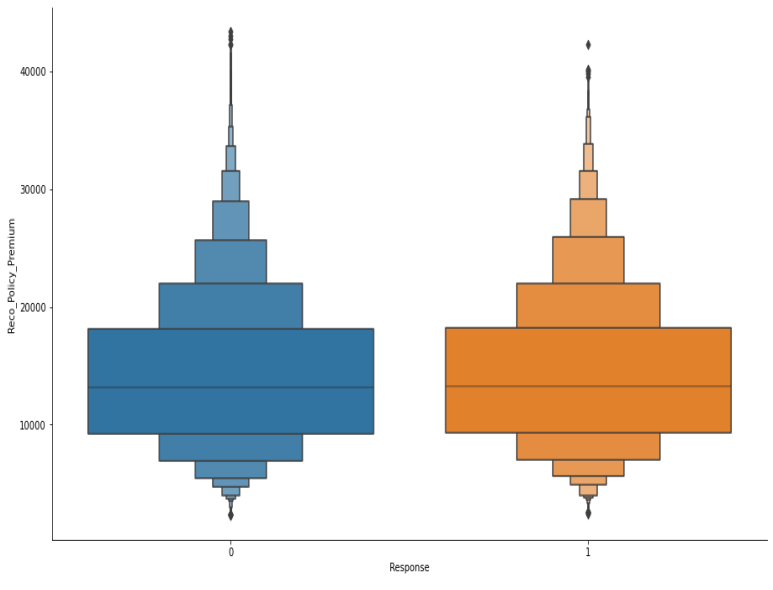
- **Description:** This box plot shows the distribution of Upper_Age and Lower_Age against the response variable.
- **Insight:** Older customers tend to show more interest in insurance offers.

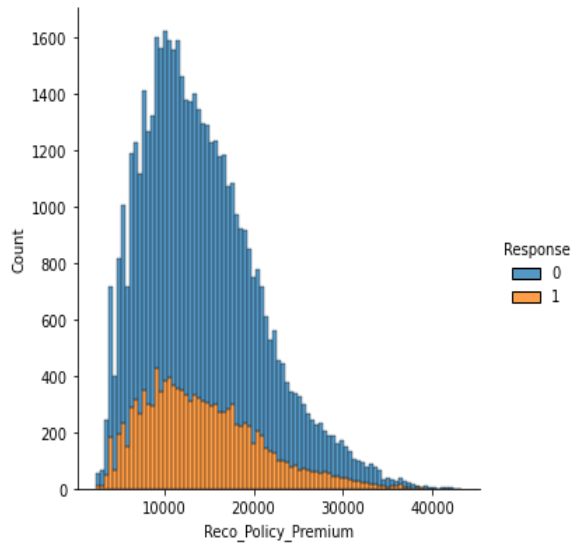
  Details:

    - **Upper Age:** The median and interquartile range of upper age for customers who responded positively vs. those who did not.
    - **Lower Age:** Similar statistics for lower age. This can highlight the age range where positive responses are more common.

2. Premium Distribution

- **Description:** This histogram and bar plot illustrate the distribution of Reco_Policy_Premium for different responses.





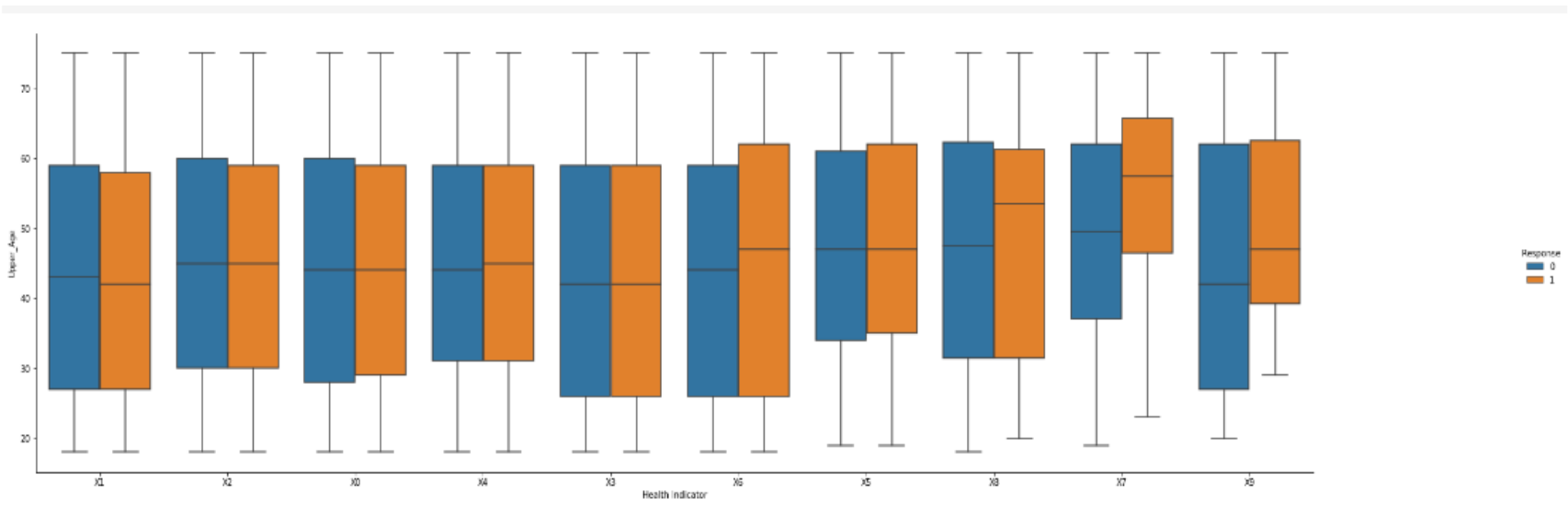
- **Insight:** Higher premiums tend to correspond with lower response rates.

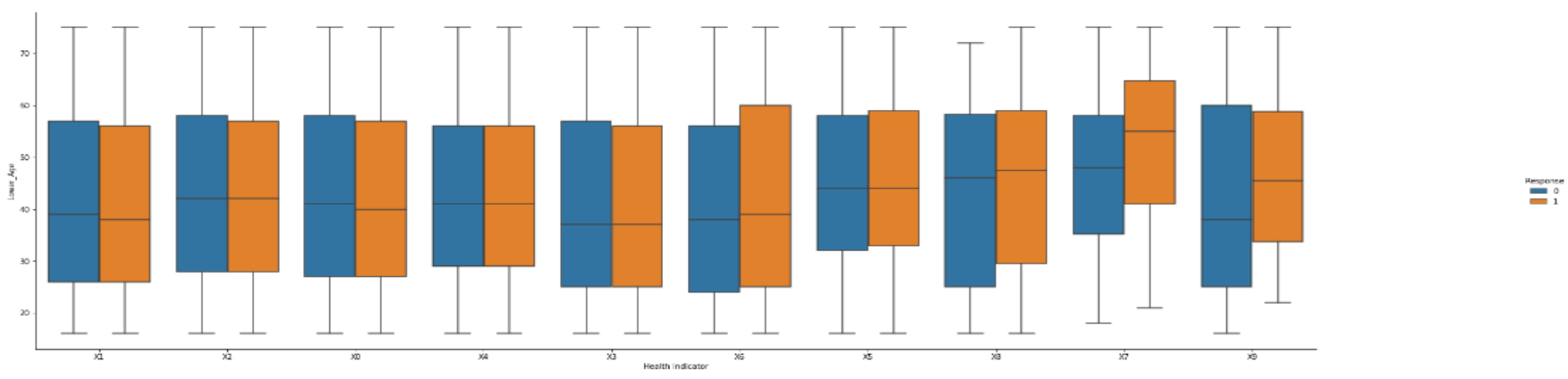
  Details:

  - **Histogram:** Shows the frequency of premiums within specified ranges, color-coded by response.
  - **Bar Plot:** Provides a clearer comparison of the average premium between responders and non-responders.

3. Health Indicators

- **Description:** This bar chart shows the distribution of Health Indicators and their impact on customer responses.





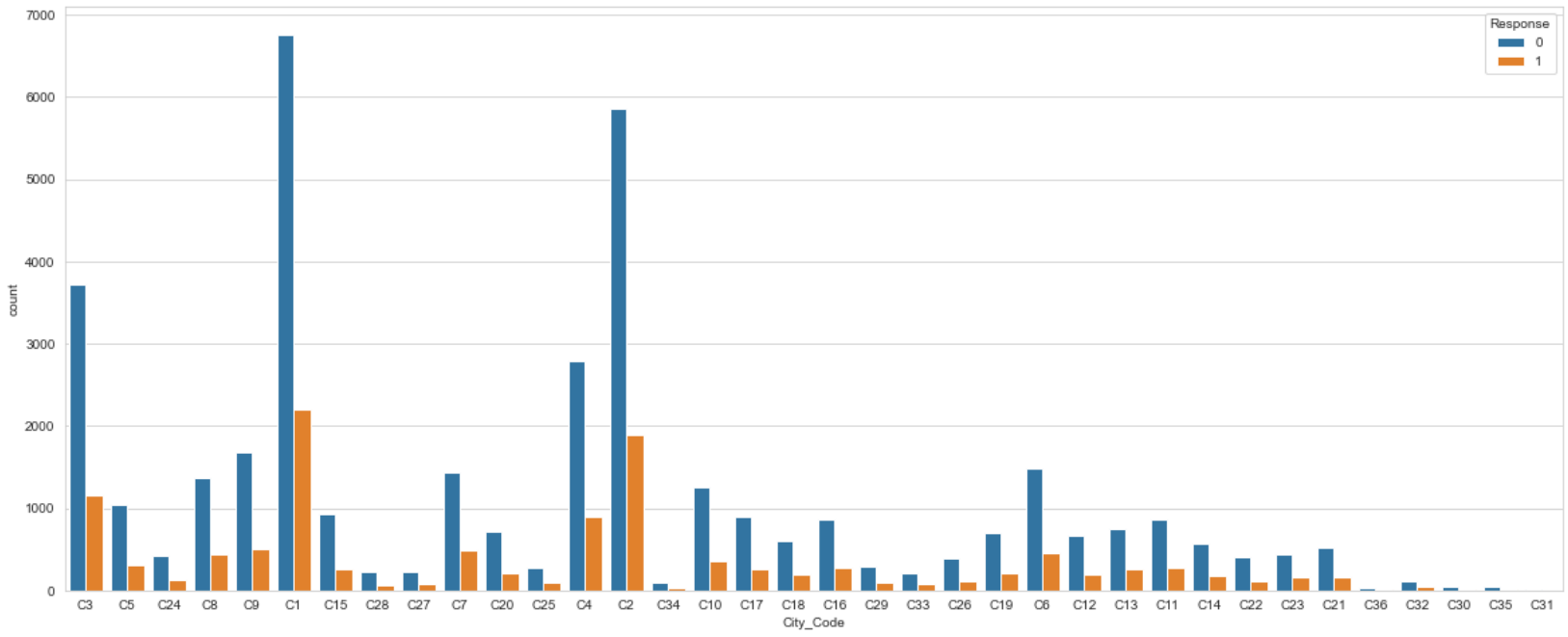
- **Insight:** Health indicators 'X7', 'X8', and 'X9' are associated with higher positive responses.

  Details:

  - **Bar Chart:** Frequency of each health indicator in the dataset.
  - **Response Overlay:** Additional layer showing the proportion of positive responses for each health indicator.

4. Response Rate by City Code

- **Description:** This bar plot displays the response rates by City_Code.



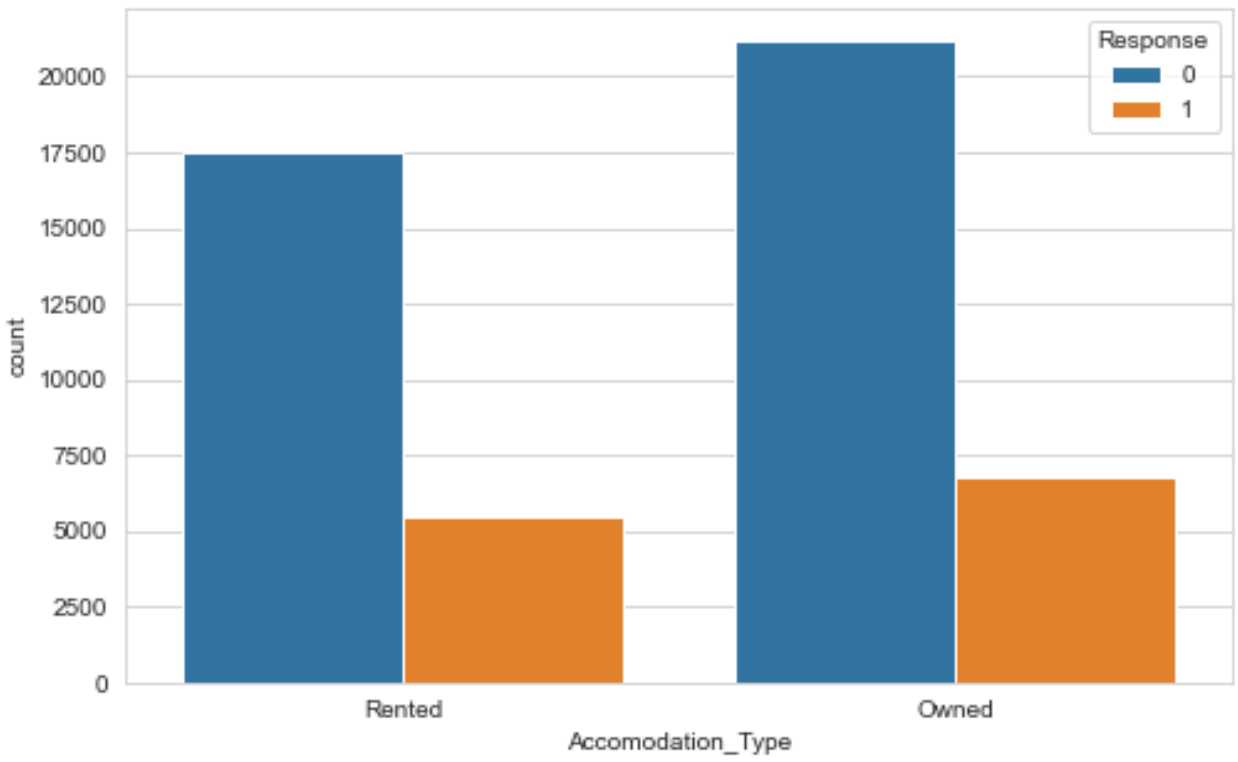
- **Insight:** Certain cities have higher response rates, suggesting geographic factors influence customer decisions.

  Details:

  - **City Codes:** Sorted by response rate to highlight top-performing and underperforming regions.
  - **Bar Heights:** Represent the percentage of positive responses in each city.

5. Response Rate by Accomodation Type

- **Description:** This bar plot shows the response rates for different accommodation types (Owned vs. Rented).



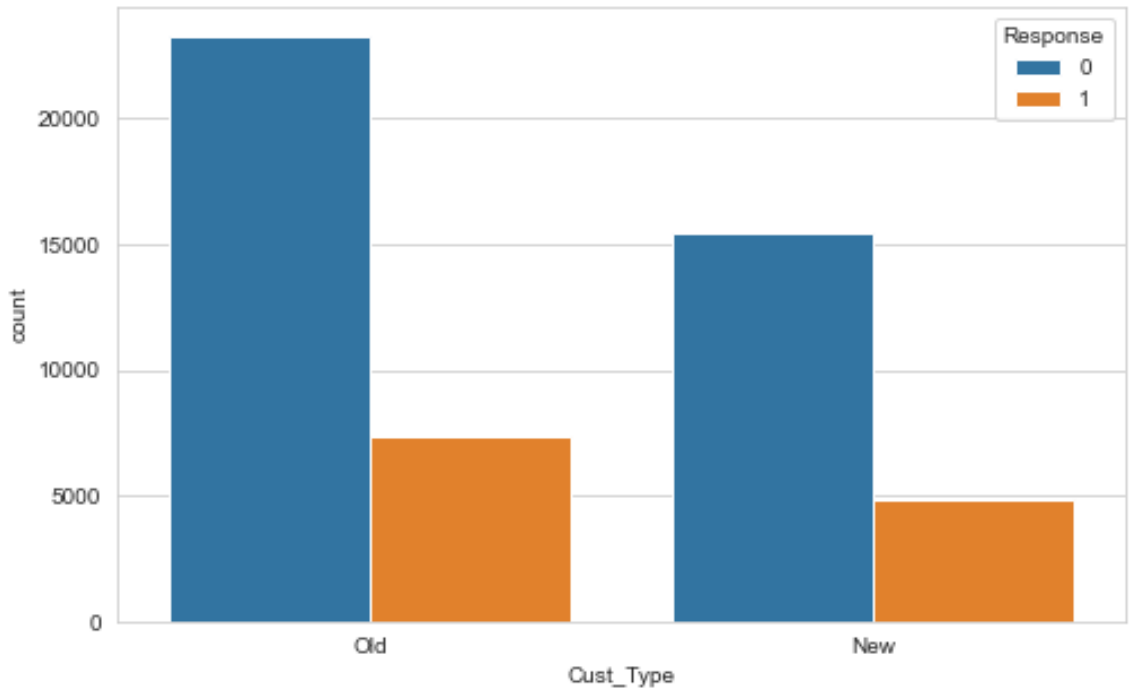
- **Insight:** Customers who own their accommodation tend to respond more positively.

  Details:

  - **Accommodation Types:** Comparison of response rates between owned and rented accommodations.
  - **Bar Heights:** Indicate the proportion of positive responses.

6. Holding Policy Duration vs. Response

- **Description:** This box plot displays the holding policy duration against the response variable.



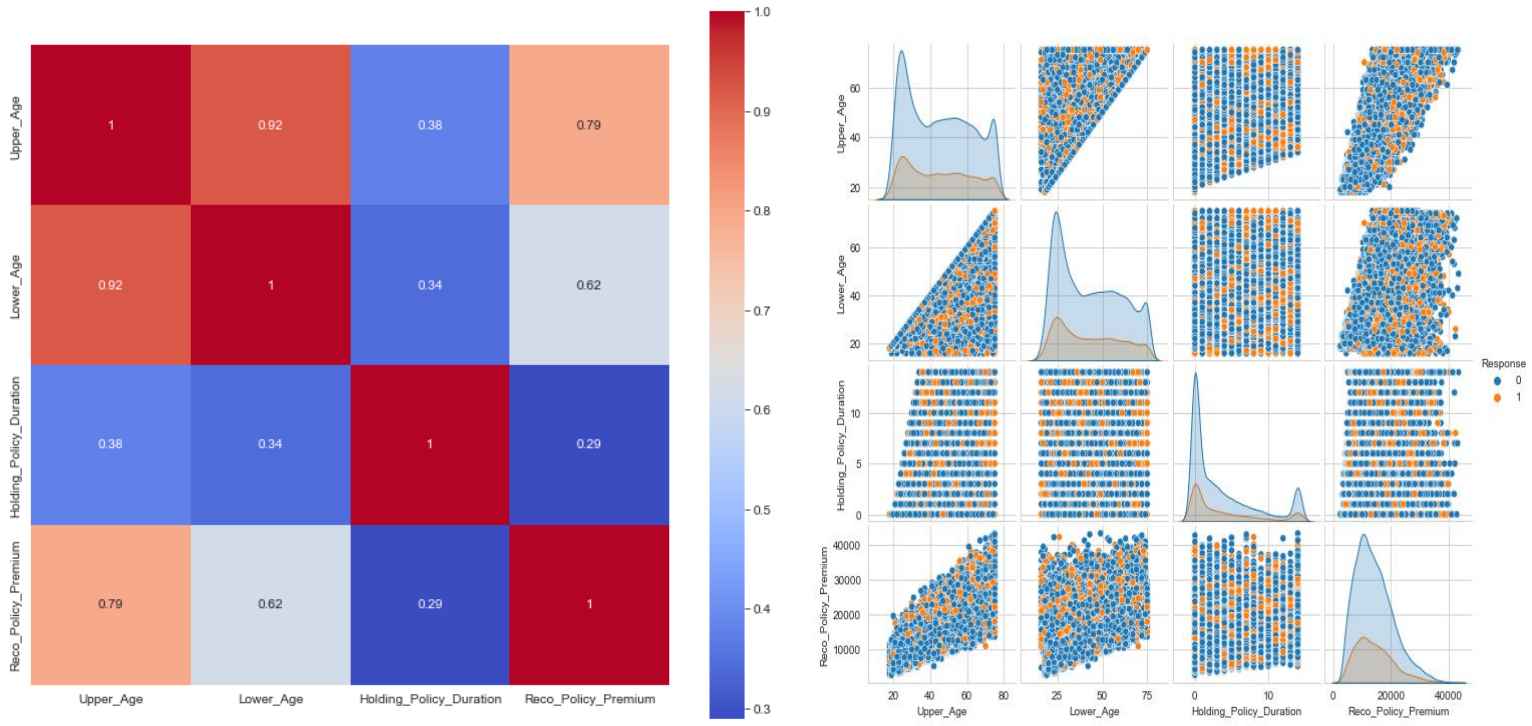
- **Insight:** Longer policy durations tend to correlate with higher response rates.

  Details:

  - **Duration Ranges:** Shows the spread of policy durations for responders vs. non-responders.
  - **Box Plot:** Highlights the median, quartiles, and outliers in holding policy durations.

7. Correlation Heatmap

- **Description:** This heatmap shows the correlation between different numerical feature



- **Insight:** Identifies which features are most strongly correlated with each other and with the response variable.

  Details:

  - **Color Intensity:** Represents the strength of correlations (stronger correlations shown in darker colors).
  - **Feature Pairs:** Highlights important relationships, such as the correlation between age and policy premium.

8. Response Rate by Health Indicator

- **Description:** This grouped bar plot shows response rates for different health indicators.
- **Insight:** Certain health indicators significantly affect response rates, providing a potential area for targeted marketing.

  Details:

  - **Health Indicators:** Comparison of positive response rates across different health conditions.
  - **Grouped Bars:** Shows the proportion of responders and non-responders within each health indicator category.

Modeling Approach

Model Selection

- **Logistic Regression:** Simple and interpretable model.
- **K-Nearest Neighbors (KNN):** Non-parametric method.
- **Support Vector Machine (SVM):** Effective in high-dimensional spaces.
- **Random Forest:** Ensemble method that handles overfitting well.

- **Gradient Boosting Classifier:** Boosting method that combines weak learners.

Performance Evaluation

- **Metrics:** Accuracy, ROC-AUC scores were used to evaluate models.
- **Results:**
  - **Gradient Boosting:** Highest ROC-AUC score of 0.5975.
  - **Random Forest:** ROC-AUC score of 0.5715.
  - **Logistic Regression:** ROC-AUC score of 0.5575.
  - **KNN:** ROC-AUC score of 0.5448.
  - **SVM:** ROC-AUC score of 0.5110.

Model Tuning

- **Grid Search:** Hyperparameter tuning was conducted for Gradient Boosting and Random Forest to optimize performance.
  - **Gradient Boosting Parameter Grid:**
    - `learning_rate`: [0.01, 0.1, 0.05]
    - `n_estimators`: [100, 200, 300]
    - `max_depth`: [3, 5, 7]
    - `subsample`: [0.8, 0.9, 1.0]
  - **Random Forest Parameter Grid:**
    - `n_estimators`: [100, 200, 300]
    - `max_depth`: [None, 10, 20, 30]
    - `min_samples_split`: [2, 5, 10]
    - `min_samples_leaf`: [1, 2, 4]

Key Findings and Insights

Summary of Modeling Results

- **Gradient Boosting:** Best performance with an AUC-ROC score of 0.5975, indicating a strong predictive capability.
- **Random Forest:** Good performance but showed signs of overfitting with an AUC-ROC score of 0.5715.
- **Logistic Regression, KNN, SVM:** Lower performance compared to ensemble methods.

Insights

- **Targeted Marketing:** Identified customer segments with higher positive response rates.
- **Policy Adjustments:** Suggested modifications in policy offerings to align with customer profiles.
- **Health Data Utilization:** Emphasized the importance of health indicators in predicting customer behavior.

Recommendations

1. **Targeted Marketing:** Utilize the predictive model to focus marketing efforts on customer segments with higher response probabilities, enhancing the efficiency of marketing campaigns.
2. **Policy Optimization:** Adjust insurance policy details and premiums based on customer profiles and predicted responsiveness to increase engagement and satisfaction.
3. **Health Indicator Utilization:** Incorporate health indicators more prominently in marketing strategies to personalize insurance offers, leveraging the strong correlation between health conditions and customer responses.

Ideas for Further Research

- **Feature Engineering:** Explore additional features and transformations that could improve model accuracy, such as interaction terms or polynomial features.
- **Advanced Models:** Investigate the use of deep learning models for potentially higher predictive performance.
- **Customer Feedback:** Incorporate customer feedback data to refine models and improve prediction accuracy.
- **Time-Series Analysis:** Analyze temporal trends in customer responses to adjust strategies dynamically.