

HealthCare Capstone Project

PG Program in Data
Science and Business
Analytics

By:

*Sangram
Keshari Patro*



Presentation Overview

1

Business Problem

Understanding the core challenge and project scope.

2

Data Report

Overview of data collection, structure, and key variables.

3

Exploratory Data Analysis

Univariate and bivariate insights from the data.

4

Modeling Building Approach

Building different models and find the best one.

5

Key Insights & Recommendations

Actionable strategies derived from analysis.

Business Problem Understanding

Problem Statement

Estimate individual insurance costs based on health, lifestyle, and demographic variables.

Need of Project

- To assist insurance companies in fairly pricing insurance policies.
- To minimize risks involved.
- To ensure affordability for customers based on their risk profile.

Business Opportunity

- Optimize pricing
- Reduce claim payouts and improve profitability
- **Customer Benefit:** Encourage preventive care and promote healthy lifestyle.

This project focuses on building a robust predictive model for insurance cost estimation.

Data Report: Collection & Overview

Data Collection

- Timeframe: Rolling 1-year period.
- Frequency: Snapshot-based per applicant.
- Methodology: Internal databases, health records, self-reported, wearables.

Variable	Business Definition
applicant id	Applicant unique ID
years of insurance with us	Years with current insurance company
regular checkup last year	Health checkups in last year
adventure sports	Participates in adventure sports
Occupation	Customer's occupation
visited doctor last 1 year	Doctor visits in last year
cholesterol level	Current cholesterol level
daily avg steps	Average daily steps
age	Customer's age
heart disease history	Past heart disease history

Data Overview

- 25,000 rows, 24 columns.
- Numeric and categorical data.
- No duplicate rows.
- 'Year last admitted' removed (47% nulls).
- 'BMI' imputed using KNN.
- Columns renamed for clarity.
- Insurance cost is the target variable.

Variable	Business Definition
bmi	Body Mass Index
smoking status	Smoking status
Year last admitted	Last hospital admission year
Location	Hospital location
weight	Weight in kg
covered by any other company	Other insurance coverage
Alcohol	Alcohol consumption status
exercise	Regular exercise status
weight change in last one year	Weight change in past year
fat percentage	Body fat percentage

Exploratory Data Analysis: Univariate



Years of Insurance

- Slight left skew
- Most customers 2-6 years.
- Highest tenure: 7-8 years.



Weight

- Symmetric
- Most 64-78 kg.
- Range: 52-96 kg.



Checkups Last Year

- Highly right-skewed,
- 60%+ no checkup (Most 0-1)
- Few outliers with 4-5 checkups. (Range: 0-5)



Insurance Cost

- Slight right skew
- Most (Rs. 16,024-Rs. 37,020)
- Range (Rs. 2,468-Rs. 67,870).



Fat percentage

- Slight left skew
- Most 21-36 %.
- Range: 11 - 42 %



BMI

- Right-skewed
- Most 26.1-35.6 .
- Range: 12.3-100.6



Covered by other company

- Almost 70% of the users don't have insurance cover by other company. **Target them for insurance.**



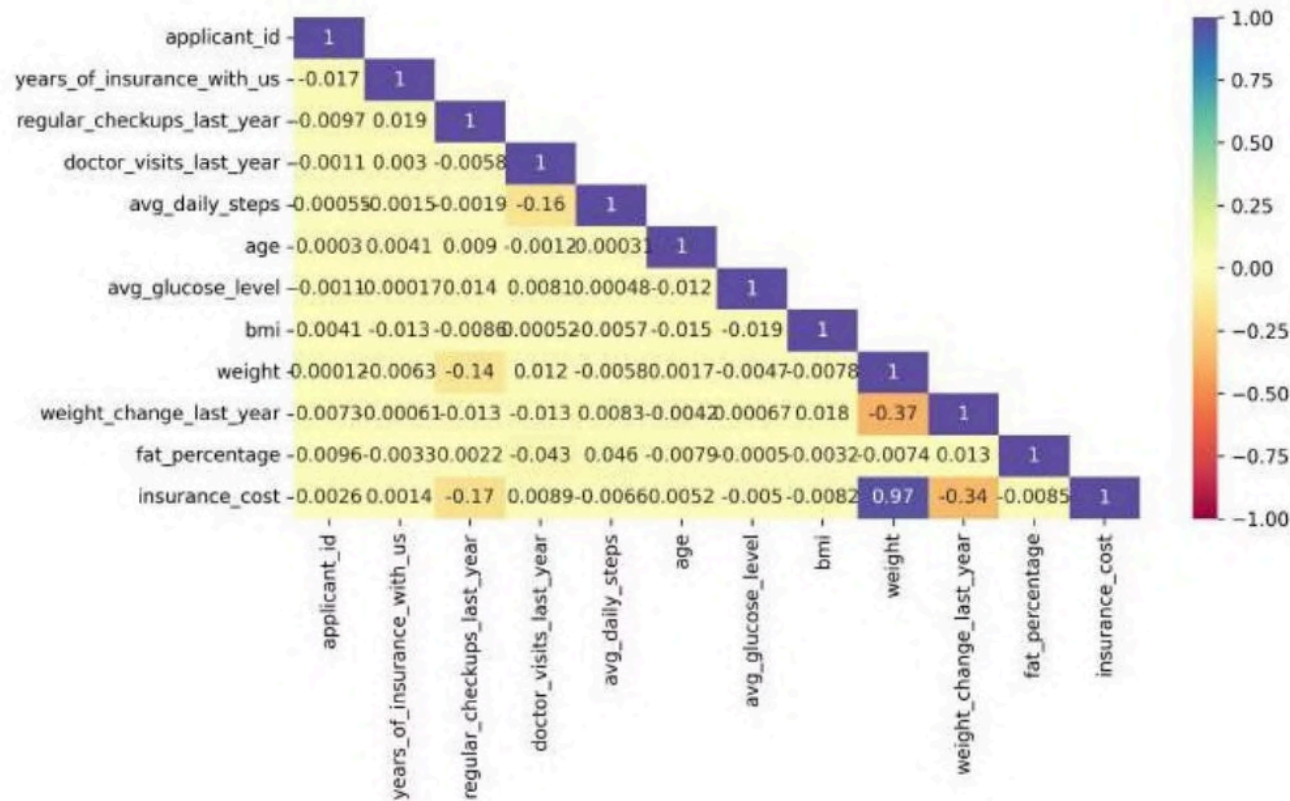
Doctor visits last year

- Right-skewed
- Most 2-4 visits.
- A few high frequency outliers (up to 12 visits).

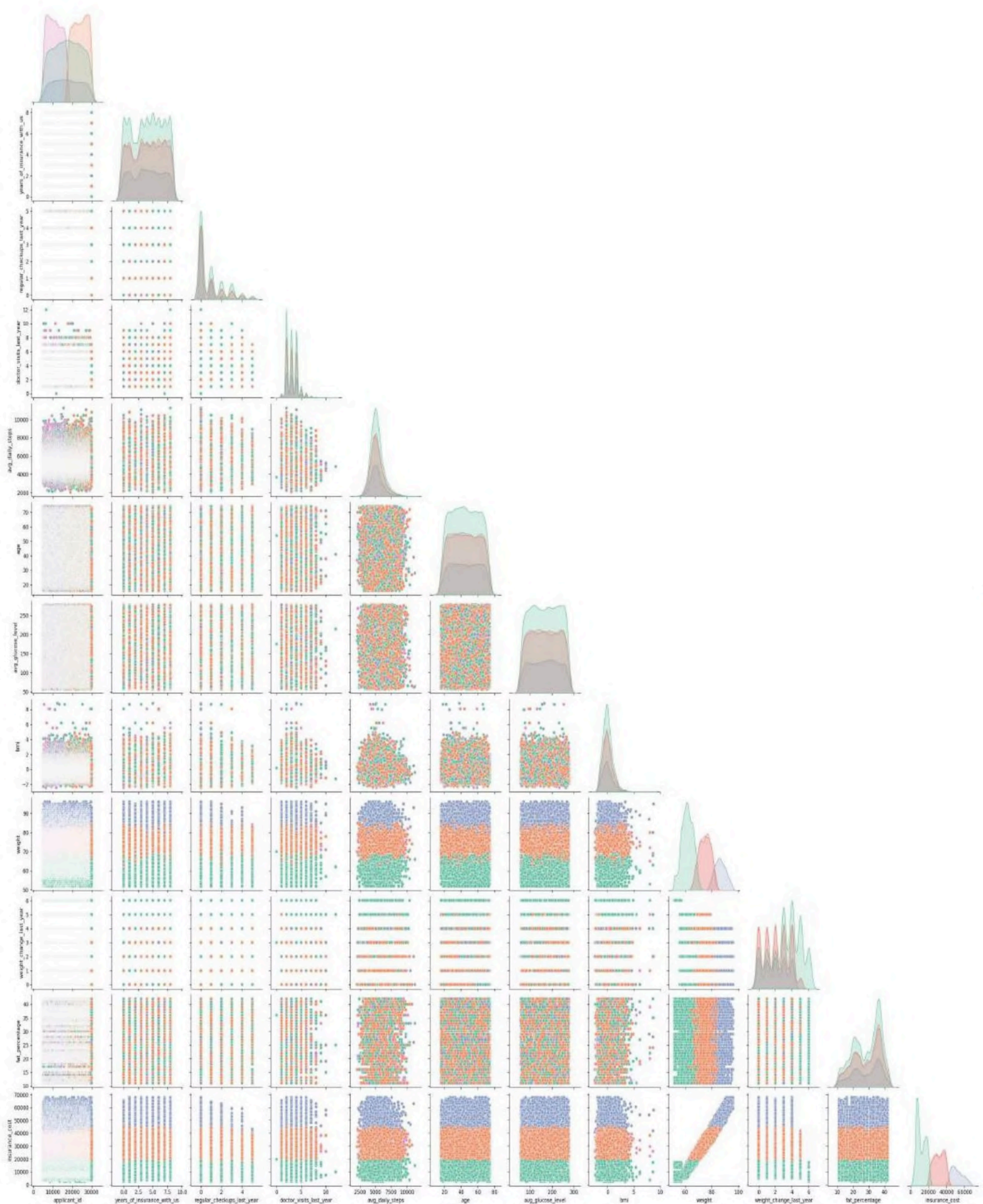
Exploratory Data Analysis: Bivariate

Numerical Variables

- Weight: Strong positive correlation (0.97) with insurance cost.
- BMI: Weak positive correlation with insurance cost.
- Weight change: Negative correlation (-0.34) with insurance cost.
- Other variables: Weak or negligible correlations.



Multivariate Relationships (K-Prototype Clustering)



Insights from Cluster Profiling

1

Group 0: Health-Conscious (8950)

Moderate cost (Avg: 11,909), highest checkup rate, active lifestyle, low disease history.

2

Group 1: High-Risk (6413)

Highest cost (Avg: 31,691), slightly lower checkups, higher BMI, more disease history.

3

Group 2: Chronically Ill (3273)

Second highest cost (Avg: 51,726), lowest checkups, very high weight, high fat percentage.

4

Group 3: Low Engagement (6364)

Lowest cost (Avg: 31,357), moderate checkups, moderate BMI/steps/fat, moderate chronic disease.

Actionable Insights & Recommendations

Risk-Based Pricing

Prioritize weight and fat percentage.
Higher premiums for Groups 1 & 2,
discounts for Group 0.

Preventive Health

Target Groups 2 & 3 with wellness
programs, screenings, and disease
management.

Product Personalization

Tailored plans for
students/professionals. Discounts for
healthy activities. Wearable-
integrated plans.

Data Preprocessing Pipeline

The objective was to preprocess the insurance dataset for predicting insurance cost, including imputation, encoding, scaling, and splitting.

Key Steps

- Numerical Features: Scaled using StandardScaler.
- Ordinal Features: Encoded with OrdinalEncoder.
- Nominal Features: One-Hot Encoded.
- Missing Values: Handled using KNNImputer on BMI.

Dataset Split

Dataset	Ratio	Samples
Train	70%	17500
Validation	15%	3750
Test	15%	3750

Feature Engineering

To capture deeper relationships and improve model performance, new features were created based on interactions and transformations.

Checkups in the last year & Weight Change Interaction

Multiplied regular checkups with weight change to reveal hidden non-linear relationships.

Log Transformation of Weight

Applied log transformation to handle skewness and reduce impact of extreme weight values.

Squared Term of Weight

Created a squared version to learn quadratic effects of weight on insurance cost.

Weight & Checkups in the last year Interaction

Identifies if heavier individuals with frequent checkups influence costs differently.

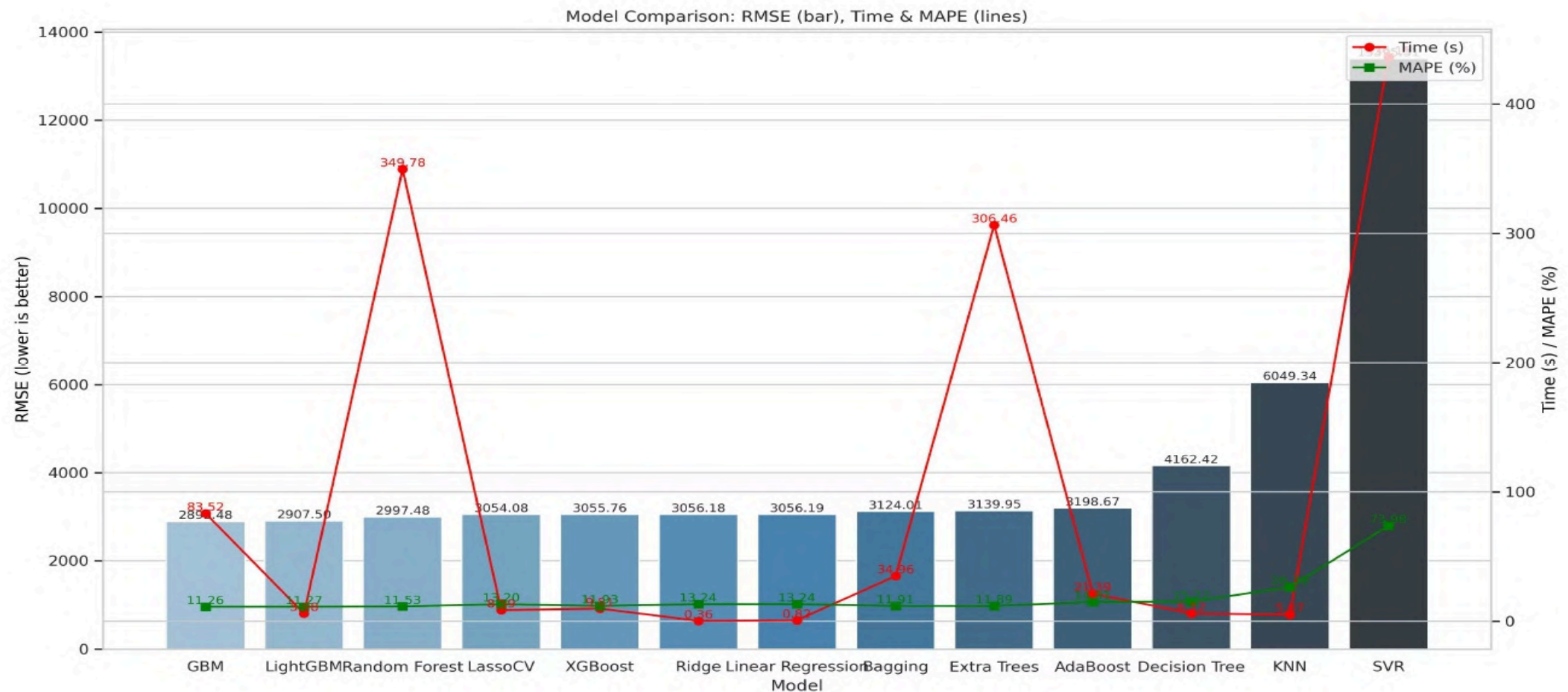
Weight and Weight change Interaction

It captures whether individuals with high body weight and significant weight change influence insurance costs differently.

These engineered features were added to new copies of the datasets.

Modeling Approach & Why

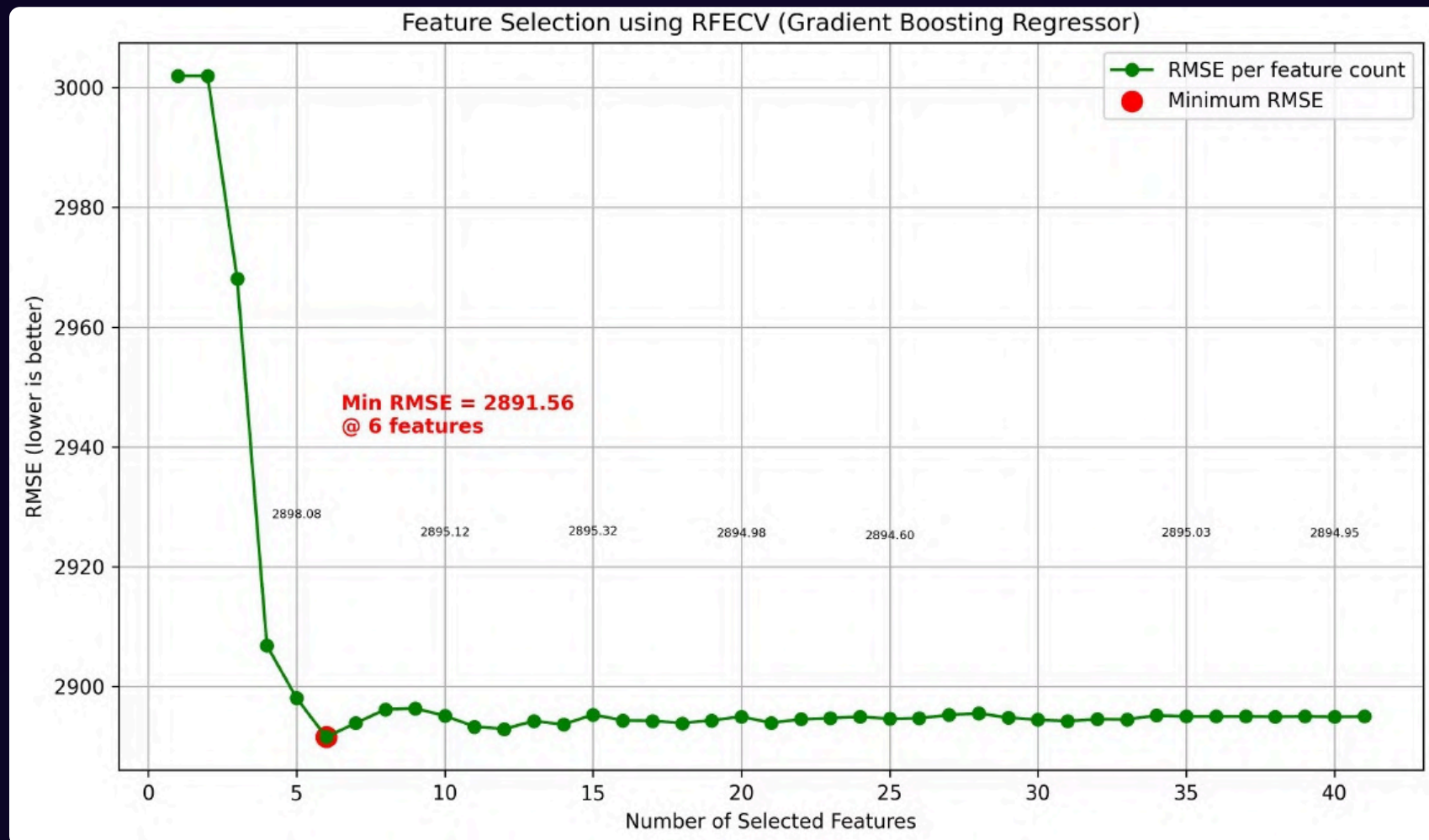
- Various parametric and non-parametric models were built and evaluated using cross-validation to identify top performers.
- Top performing models (GBM, LightGBM, Random Forest, Lasso, Ridge, Linear Regression, XGBoost) were selected for further tuning based on their RMSE and MAPE scores.



Feature Elimination: RFECV Method

Recursive Feature Elimination with Cross-Validation (RFECV) was used to select the optimal subset of features, minimizing RMSE.

This method systematically removes features and evaluates model performance, ensuring only the most impactful features are retained.



Model Performance Comparison (Overall)

A comprehensive comparison of the final tuned models on both training and validation datasets.

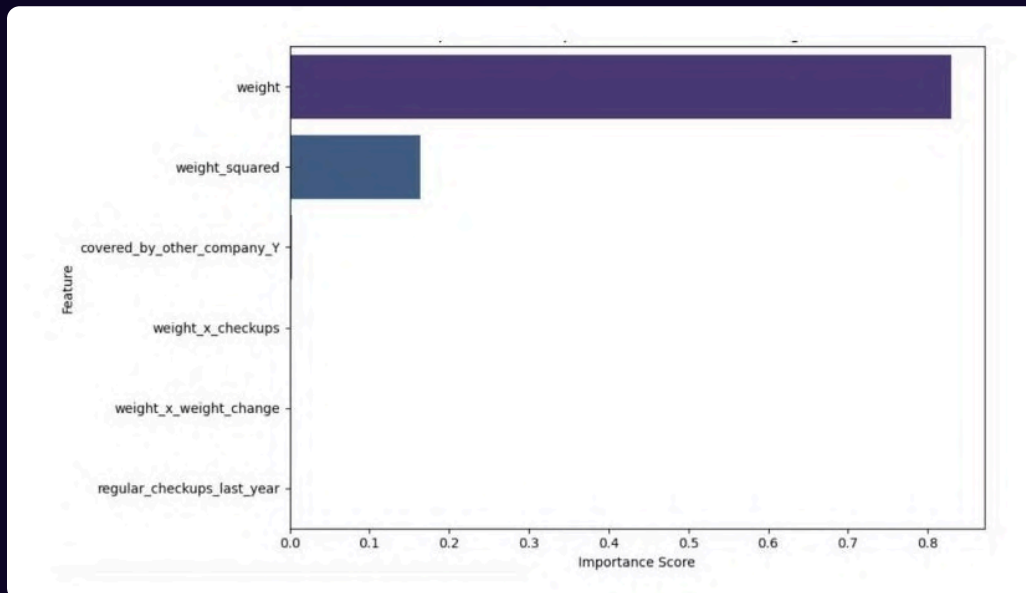
Table 8: Training and Validation Performance of Final Models								
Model	Training Performance				Validation Performance			
	MAPE (%)	RMSE	R^2	Adj. R^2	MAPE (%)	RMSE	R^2	Adj. R^2
Final LGBM Model	11.2114	2876.67	0.9559	0.9558	11.2786	2877.34	0.9544	0.9540
Final Ridge Model	12.5881	3013.66	0.9516	0.9515	12.3600	2982.66	0.9510	0.9507
Final GBM Model	11.0879	2852.82	0.9566	0.9566	11.2342	2875.45	0.9544	0.9544
Final XGBoost Model	11.2958	2884.37	0.9556	0.9556	11.3270	2878.16	0.9543	0.9543
Final Lasso Model	12.5963	3014.48	0.9515	0.9515	12.3638	2982.10	0.9510	0.9508
Final Random Forest Model	12.6027	3015.47	0.9515	0.9515	12.3596	2982.98	0.9509	0.9509
Final Linear Regression Model	10.3245	2695.01	0.9613	0.9612	11.2621	2891.05	0.9539	0.9536

The Gradient Boosting Model (GBM) is identified as the best overall model based on its superior RMSE, MAPE, R^2 , and Adjusted R^2 values.

Insights from Analysis: SHAP & PDP

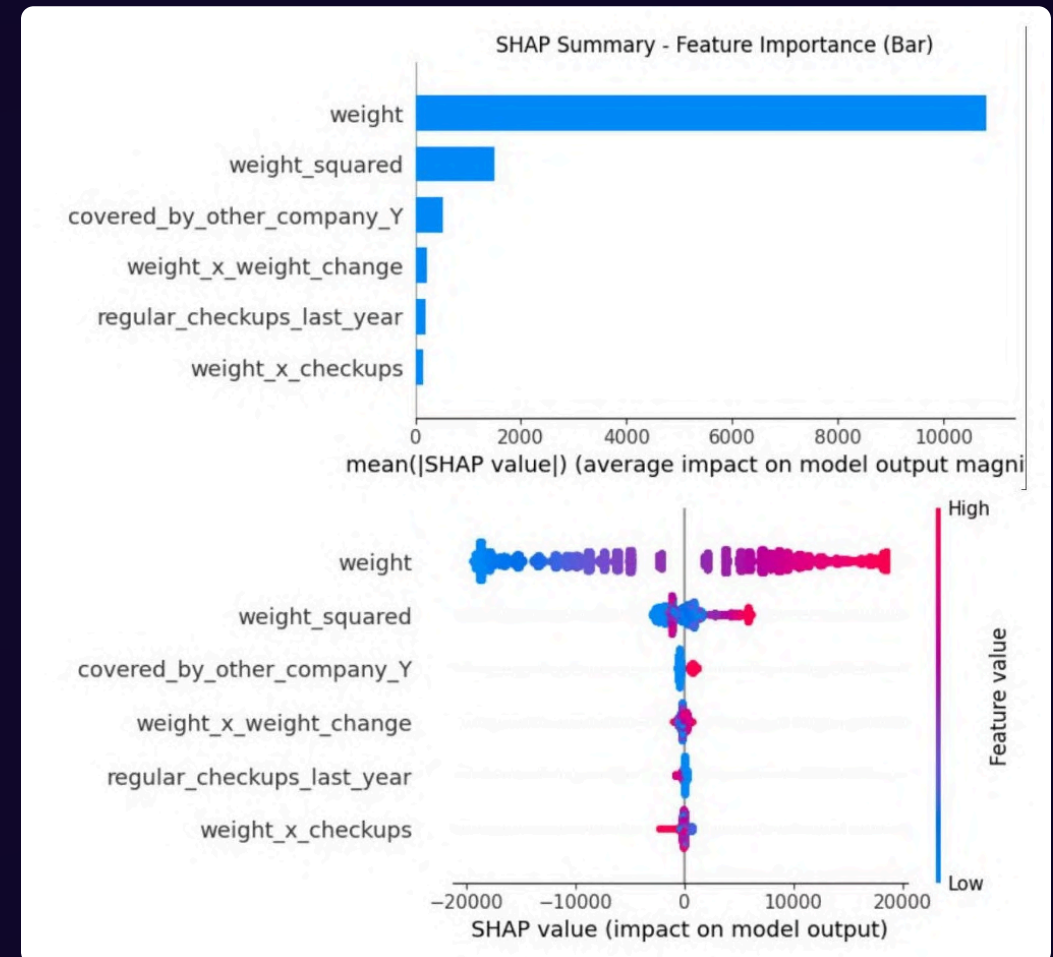
SHAP and Partial Dependence Plots (PDP) provide deep insights into feature impact and relationships.

SHAP Bar Plot



Weight is the most impactful feature, followed by weight squared. Other features like 'covered by other company Y' and 'regular checkups' also contribute.

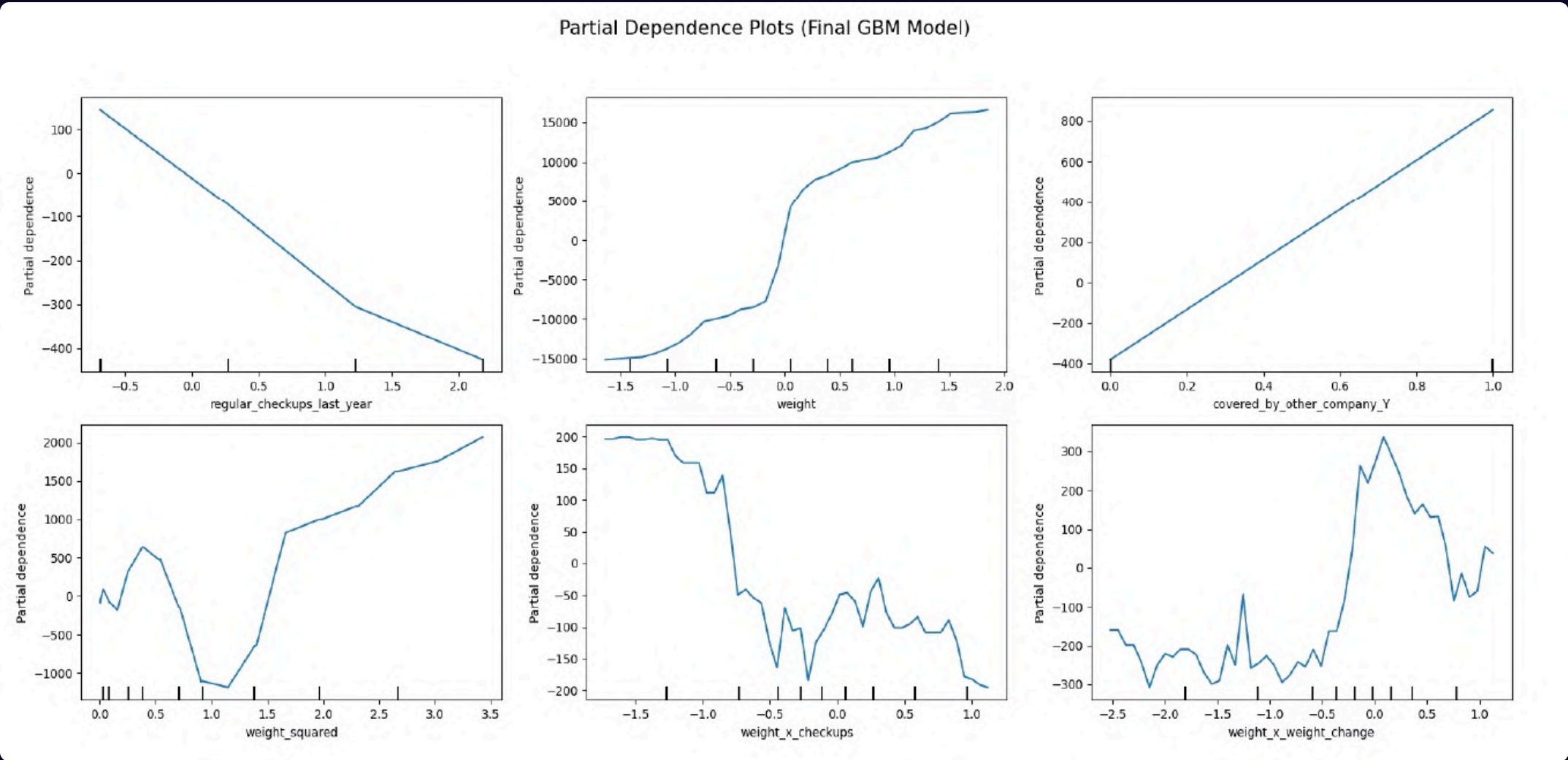
SHAP Beeswarm Plot




Higher weights increase insurance costs. Regular checkups slightly reduce costs. Complex interactions exist for weight changes.


Insights from Analysis: PDP

Partial Dependence Plots reveal the marginal effect of each feature on insurance cost.




- 


Regular Checkups

More frequent checkups lead to lower predicted insurance costs, promoting preventive care.
- 


Weight

As weight increases, predicted insurance cost rises steeply, indicating higher risk.
- 


Covered by Other Company

Individuals with other coverage tend to have higher insurance costs, possibly due to complex health needs.
- 

Weight Squared

Exhibits a non-linear, upward trend, capturing escalating risk with increasing body mass.
- 

Weight & Checkups Interaction

It Suggests that higher weight combined with more checkups may slightly reduce predicted costs.
- 

Weight & Weight change Interaction

It reflects varying risk levels associated with rapid or unstable weight changes.

Business Recommendations

For Insurance Companies

- Personalized Premiums: Tailor premiums based on weight and health checkup history.
- Early Risk Detection: Flag individuals with rapid weight changes for follow-up.
- Incentivize Checkups: Reward regular preventive health checkups to lower risk.
- Cross-insurance Awareness: Design policies for applicants with existing coverage.
- Product Personalization: Offer premium discounts to policyholders who engage in healthy activities such as regular physical exercise and maintaining optimal cholesterol.
- Launch wearable-integrated plans that track steps, workout frequency, and health milestones to incentivize healthy living.
- Charge higher premiums to individuals in Groups 1 and 2, who exhibit higher health risks, while offering discounts to healthier Group 0 members.
- Incorporate recent weight change as a pricing factor, rewarding individuals who maintain or reduce their weight.

For Consumers

- Maintain Healthy Weight: Significantly reduce premiums by maintaining a healthy BMI.
- Routine Health Checkups: Lower predicted costs through regular medical checkups.
- Monitor Weight Stability: Avoid drastic weight changes; stable health is preferred.
- Maximize Preventive Behavior: Engage in healthy lifestyles to reduce chronic risks.

Strategic Implications

- Policyholder Behavior Change: As customers become aware that their lifestyle choices (e.g., regular checkups, stable weight) affect premiums, they are more likely to adopt healthier behaviors.
- Data-Driven Product Innovation: Insurers can introduce dynamic or usage-based insurance plans (e.g., rewarding healthy steps recorded via wearables), tailored for tech-savvy, health-conscious individuals.