CAPSTONE PROJECT REPORT

By,

Ann Raichel John

PG Program in DSBA, Great Learning

INTRODUCTION & PROBLEM STATEMENT

- Healthcare is an important domain to determine an individual's life. Money is an important factor in Healthcare sector where the Insurance companies plays a major role.

- The company provides a support in financial base to each individual by helping in hospital or health bills. If an individual is unaware of his health and routine check-up's/ follow-ups may affect his/her life in hospital which may be a risk factor to an insurance company.

- In order to optimize the insurance-cost, analyzing various parameters of an individual may help to reduce the risk of financial company.

- The **objective** of this project is **to develop a predictive model** that helps:
  - Understand the key factors influencing insurance cost.
  - Predict and optimize insurance premiums for individuals based on health and lifestyle indicators.

Data Set Information

Table 1: Data information

Dataset	25000 rows and 24 columns	
Target variable	Insurance cost, continuous variable	
ML model	Supervised ML	
Data Nature	Regression analysis	
Data cleaning	Missing / Null values	Treated with median for bmi & year_last_admitted
	▪ No duplicate values ▪ Spelling errors corrected	
Feature engineering	Columns created – bmi_category and years_since_admission	
Encoding	New column - cholesterol_level_encoded	

Dataset Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   applicant_id                    25000 non-null  int64  
 1   years_of_insurance_with_us      25000 non-null  int64  
 2   regular_checkup_lasy_year       25000 non-null  int64  
 3   adventure_sports                25000 non-null  int64  
 4   Occupation                      25000 non-null  object 
 5   visited_doctor_last_1_year      25000 non-null  int64  
 6   cholesterol_level               25000 non-null  object 
 7   daily_avg_steps                 25000 non-null  int64  
 8   age                             25000 non-null  int64  
 9   heart_decs_history              25000 non-null  int64  
 10  other_major_decs_history        25000 non-null  int64  
 11  Gender                          25000 non-null  object 
 12  avg_glucose_level               25000 non-null  int64  
 13  bmi                             24010 non-null  float64
 14  smoking_status                  25000 non-null  object 
 15  Year_last_admitted              13119 non-null  float64
 16  Location                        25000 non-null  object 
 17  weight                          25000 non-null  int64  
 18  covered_by_any_other_company    25000 non-null  object 
 19  Alcohol                         25000 non-null  object 
 20  exercise                        25000 non-null  object 
 21  weight_change_in_last_one_year  25000 non-null  int64  
 22  fat_percentage                  25000 non-null  int64  
 23  insurance_cost                  25000 non-null  int64  
dtypes: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

Figure 1: Data information on datatypes.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
applicant_id	25000.0	NaN	NaN	NaN	17499.5	7217.022701	5000.0	11249.75	17499.5	23749.25	29999.0
years_of_insurance_with_us	25000.0	NaN	NaN	NaN	4.08904	2.606612	0.0	2.0	4.0	6.0	8.0
regular_checkup_lasy_year	25000.0	NaN	NaN	NaN	0.77368	1.199449	0.0	0.0	0.0	1.0	5.0
adventure_sports	25000.0	NaN	NaN	NaN	0.08172	0.273943	0.0	0.0	0.0	0.0	1.0
Occupation	25000	3	Student	10169	NaN	NaN	NaN	NaN	NaN	NaN	NaN
visited_doctor_last_1_year	25000.0	NaN	NaN	NaN	3.1042	1.141663	0.0	2.0	3.0	4.0	12.0
cholesterol_level	25000	5	150 to 175	8763	NaN	NaN	NaN	NaN	NaN	NaN	NaN
daily_avg_steps	25000.0	NaN	NaN	NaN	5215.88932	1053.179748	2034.0	4543.0	5089.0	5730.0	11255.0
age	25000.0	NaN	NaN	NaN	44.91832	16.107492	16.0	31.0	45.0	59.0	74.0
heart_decs_history	25000.0	NaN	NaN	NaN	0.05464	0.227281	0.0	0.0	0.0	0.0	1.0
other_major_decs_history	25000.0	NaN	NaN	NaN	0.09816	0.297537	0.0	0.0	0.0	0.0	1.0
Gender	25000	2	Male	16422	NaN	NaN	NaN	NaN	NaN	NaN	NaN
avg_glucose_level	25000.0	NaN	NaN	NaN	167.53	62.729712	57.0	113.0	168.0	222.0	277.0
bmi	24010.0	NaN	NaN	NaN	31.393328	7.876535	12.3	26.1	30.5	35.6	100.6
smoking_status	25000	4	never smoked	9249	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Year_last_admitted	13119.0	NaN	NaN	NaN	2003.892217	7.581521	1990.0	1997.0	2004.0	2010.0	2018.0
Location	25000	15	Bangalore	1742	NaN	NaN	NaN	NaN	NaN	NaN	NaN
weight	25000.0	NaN	NaN	NaN	71.61048	9.325183	52.0	64.0	72.0	78.0	96.0
covered_by_any_other_company	25000	2	N	17418	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Alcohol	25000	3	Rare	13752	NaN	NaN	NaN	NaN	NaN	NaN	NaN
exercise	25000	3	Moderate	14638	NaN	NaN	NaN	NaN	NaN	NaN	NaN
weight_change_in_last_one_year	25000.0	NaN	NaN	NaN	2.51796	1.690335	0.0	1.0	3.0	4.0	6.0
fat_percentage	25000.0	NaN	NaN	NaN	28.81228	8.632382	11.0	21.0	31.0	36.0	42.0
insurance_cost	25000.0	NaN	NaN	NaN	27147.40768	14323.691832	2468.0	16042.0	27148.0	37020.0	67870.0

Figure 2: Data summary

EXPLORATORY DATA ANALYSIS

- Performed **Univariate analysis** for all features including target variable to learn distribution of all variables (Figure 3 & 4)

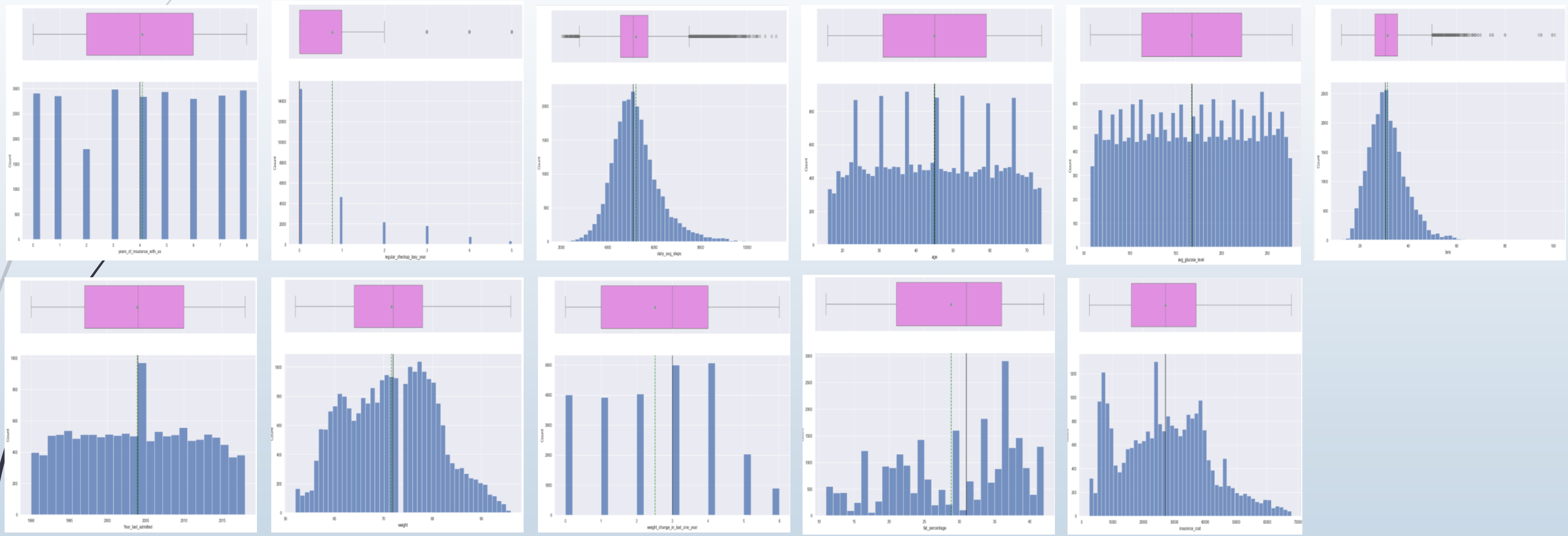


Figure 3: EDA 1

EXPLORATORY DATA ANALYSIS

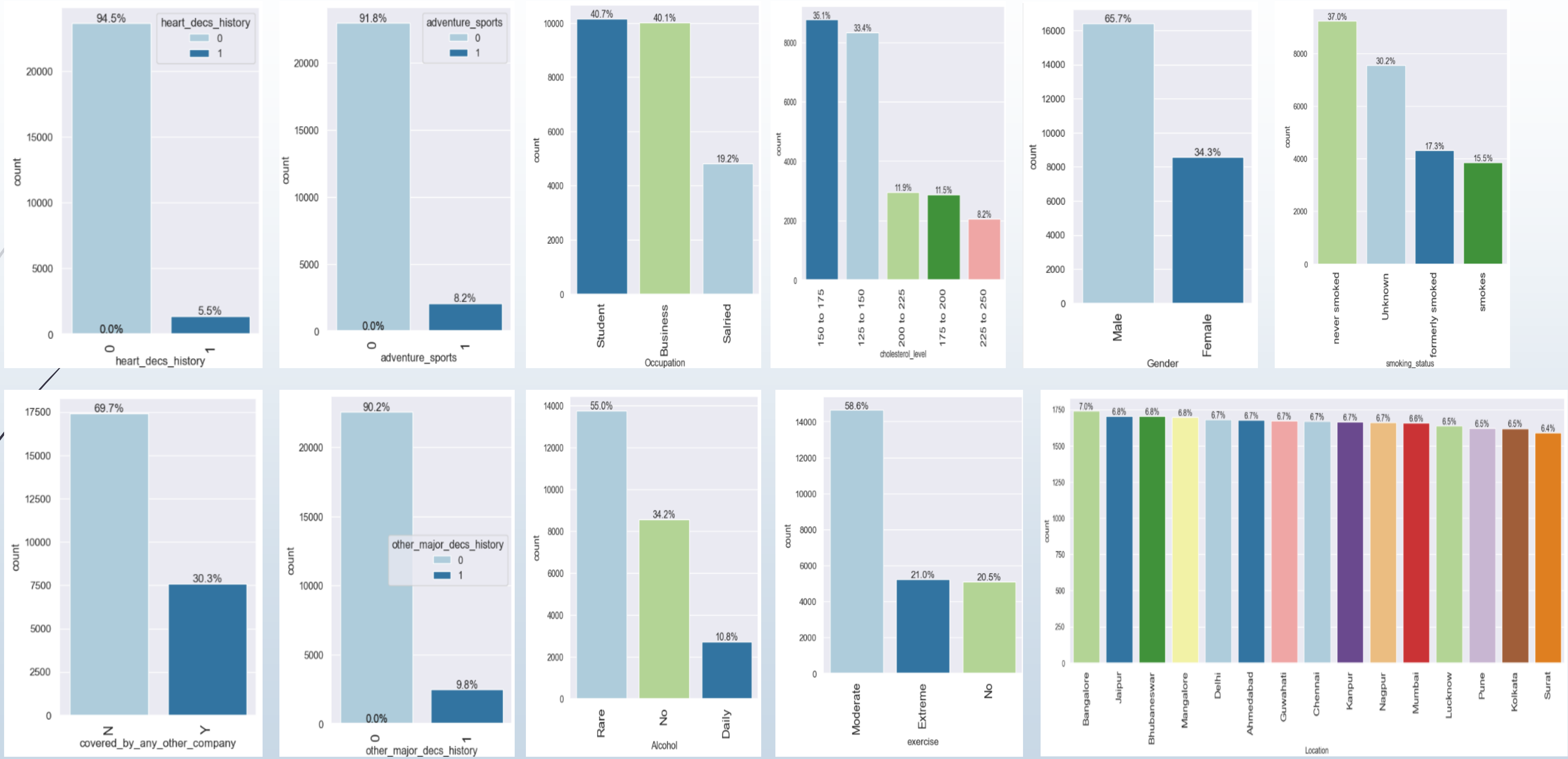


Figure 4: EDA 2

BIVARIATE ANALYSIS

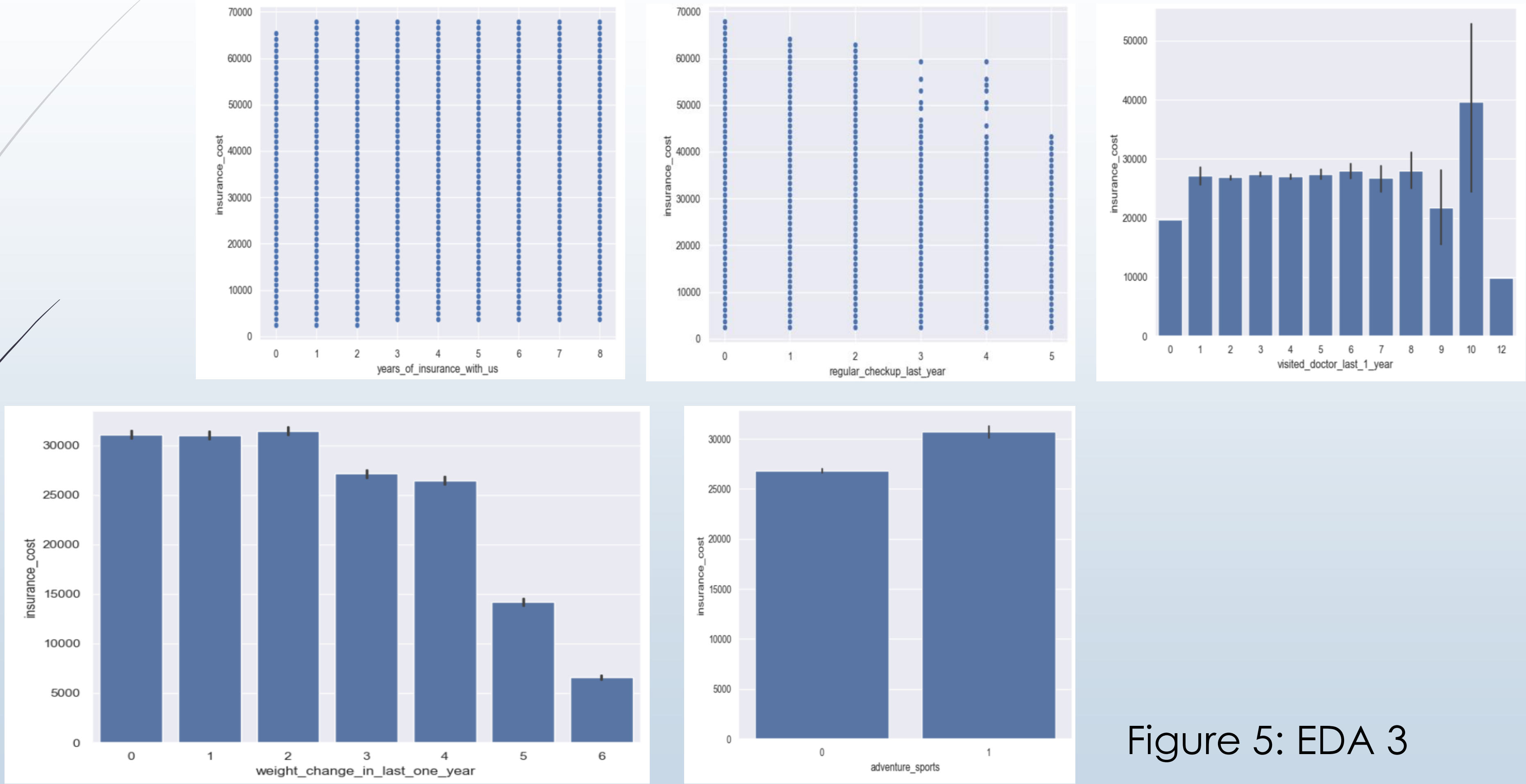


Figure 5: EDA 3

BIVARIATE ANALYSIS

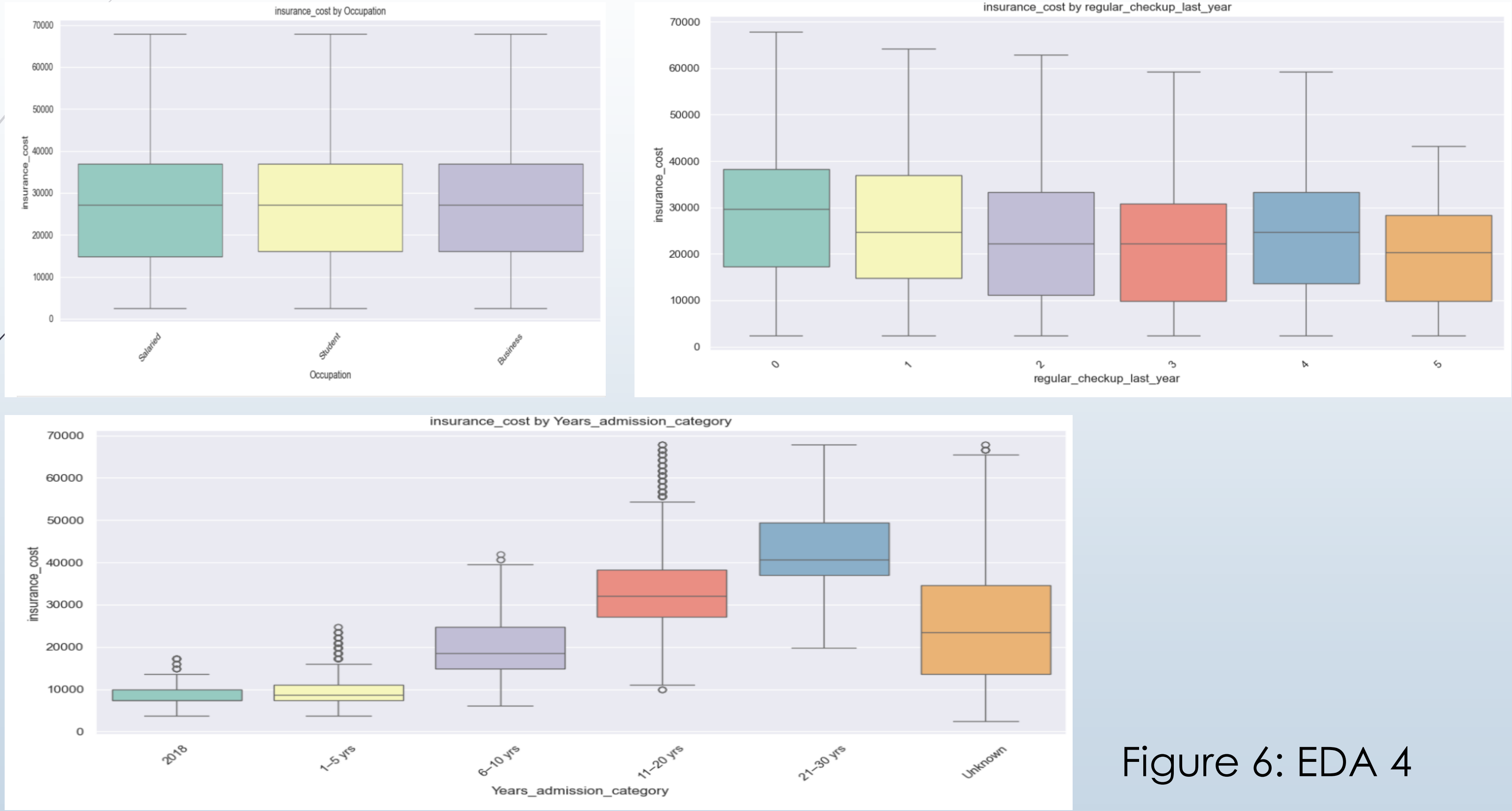


Figure 6: EDA 4

INSURANCE COST – FEATURE RELATIONSHIPS

Table 2: Feature relationship summary

Weight	Strong positive correlation – higher weight → higher cost
Weight Change	Negative correlation – less activity linked to higher cost
Years of Insurance with Us	No correlation
Avg Glucose Level	No clear relationship
Daily Avg Steps	No significant relationship
Age	No significant relationship
Heart Disease / Other Major Illness	No correlation observed
Years of Admission	Longer history (11–20 yrs) → higher cost
Regular Check-ups	More check-ups → slightly lower cost
Adventure Sports	Slightly higher cost among active participants
Cholesterol Level (200–225)	Mild increase in cost observed
Covered by Other Company	Associated with higher insurance cost
Occupation, Gender, Smoking, etc.	No meaningful relationship

DATA PREPARATION AND MODEL BUILDING

Table 3: Data Preparation & Model building summary

Model	Linear Regression analysis
Data preparation	
• Dropped column **applicant_id** for model building	
Outlier detection and treatment	• Used IQR method and treated outliers
Splitted the data into Train and test Number of rows in train data = 17500 Number of rows in train data = 7500	• Defined X for independent variables and Y for target variable • Created dummy variables, added constant • Splitted data in 70:30 ratio
Linear regression model building	
Model performance check	• RMSE • MAE • MSE • MAPE

MODEL BUILDING –LINEAR REGRESSION

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 1.118e+04
Date:                Sat, 28 Jun 2025   Prob (F-statistic):               0.00
Time:                        20:57:12   Log-Likelihood:            -1.6693e+05
No. Observations:               17500   AIC:                         3.339e+05
Df Residuals:                   17472   BIC:                         3.341e+05
Df Model:                          27
Covariance Type:            nonrobust
====================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                        -8.238e+04    423.399   -194.557      0.000   -8.32e+04   -8.15e+04
years_of_insurance_with_us     -25.8752     10.159     -2.547      0.011     -45.789      -5.962
regular_checkup_last_year     -646.0151     28.180    -22.925      0.000    -701.251    -590.780
adventure_sports              1.914e-11   1.88e-11      1.018      0.309   -1.77e-11     5.6e-11
visited_doctor_last_1_year     -34.2981     23.261     -1.475      0.140     -79.891      11.295
daily_avg_steps                 -0.0300      0.027     -1.107      0.268      -0.083       0.023
age                              2.7156      1.579      1.719      0.086      -0.380       5.811
heart_decs_history           -1.683e-11   1.57e-12    -10.691      0.000   -1.99e-11   -1.37e-11
other_major_decs_history     -3.553e-12   7.59e-14    -46.824      0.000    -3.7e-12    -3.4e-12
avg_glucose_level                0.3948      0.406      0.973      0.331      -0.400       1.190
weight                        1462.1779      3.601    405.998      0.000    1455.119    1469.237
weight_change_in_last_one_year 174.0329     16.252     10.708      0.000     142.177     205.889
fat_percentage                  -0.3840      3.099     -0.124      0.901      -6.458       5.690
cholesterol_level_encoded       33.9094     24.154      1.404      0.160     -13.435      81.253
Years_since_admission          308.7663     22.869     13.502      0.000     263.941     353.591
Occupation_Salaried             85.6449     74.331      1.152      0.249     -60.052     231.342
```

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	3361.429138	2700.7052	0.945276	0.945179	15.13952

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	3322.303772	2690.187342	0.945379	0.945152	14.923196

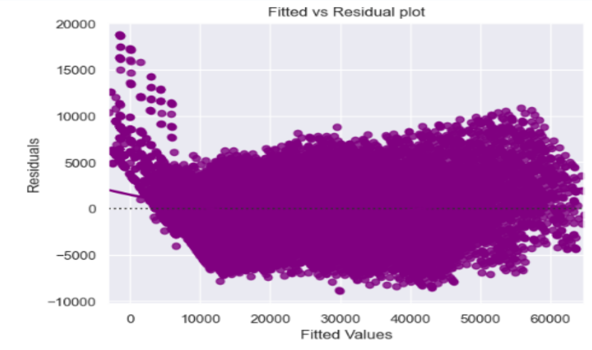
Figure 7: LR Model building

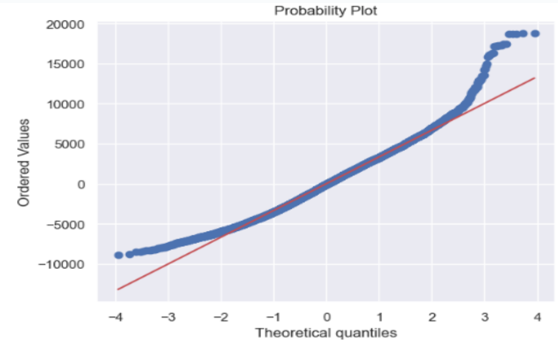
- Model was well-fitted and generalized well to unseen data.
- Values shows that there was no signs of overfitting or underfitting.
- Seems like features were strong, and target modeling was effective.

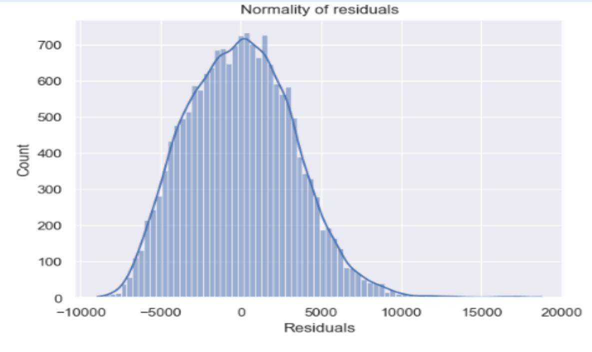
LINEAR REGRESSION ASSUMPTIONS

	feature	VIF
0	const	277.200557
1	years_of_insurance_with_us	1.084652
2	regular_checkup_last_year	1.031342
3	adventure_sports	NaN
4	visited_doctor_last_1_year	1.038667
5	daily_avg_steps	1.061165
6	age	1.000899
7	heart_decs_history	NaN
8	other_major_decs_history	NaN
9	avg_glucose_level	1.001896
10	weight	1.754490
11	weight_change_in_last_one_year	1.172145
12	fat_percentage	1.104885
13	cholesterol_level_encoded	1.431172
14	Years_since_admission	1.559064
15	Occupation_Salaried	1.324811
16	Occupation_Student	1.658478
17	Gender_Male	1.301879
18	smoking_status_formerly smoked	1.478197
19	smoking_status_never smoked	1.587139
20	smoking_status_smokes	1.416420
21	covered_by_any_other_company_Y	1.082248
22	Alcohol_No	2.766150
23	Alcohol_Rare	2.764576
24	exercise_Moderate	1.587139
25	exercise_No	1.598475
26	bmi_category_Normal	8.299270
27	bmi_category_Overweight	11.144043
28	bmi_category_Obesity I	12.167261
29	bmi_category_Obesity II	7.656289
30	bmi_category_Obesity III	6.843999

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 5.031e+04
Date:                Sat, 28 Jun 2025   Prob (F-statistic):               0.00
Time:                        21:00:51   Log-Likelihood:            -1.6694e+05
No. Observations:               17500   AIC:                         3.339e+05
Df Residuals:                   17493   BIC:                         3.340e+05
Df Model:                           6
Covariance Type:            nonrobust
==================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                      -8.221e+04    297.540   -276.292      0.000   -8.28e+04   -8.16e+04
years_of_insurance_with_us   -26.3056     10.153     -2.591      0.010     -46.206      -6.405
regular_checkup_last_year   -645.8177     28.163    -22.931      0.000    -701.021    -590.615
adventure_sports            1.496e-10   5.46e-13    274.045      0.000    1.49e-10    1.51e-10
heart_decs_history         -4.999e-11   1.83e-13   -272.815      0.000   -5.03e-11   -4.96e-11
weight                      1461.9696      3.598    406.291      0.000    1454.917    1469.023
weight_change_in_last_one_year 173.7986   16.243     10.700      0.000     141.962     205.636
Years_since_admission        309.3255     22.853     13.535      0.000     264.531     354.119
covered_by_any_other_company_Y 1210.2474  57.377     21.093      0.000    1097.783    1322.712
==============================================================================
Omnibus:                      568.852   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              688.185
Skew:                           0.395   Prob(JB):                    3.65e-150
Kurtosis:                       3.566   Cond. No.                     2.87e+18
==============================================================================
```








```
ShapiroResult(statistic=0.9881370009861654, pvalue=2.0678052801854352e-35)
```

```
[('F statistic', 1.0021299907290038), ('p-value', 0.4604222008267006)]
```

VIF score check - Multicollinearity

No feature has p-value greater than 0.05

Figure 8: Linear Regression Assumptions

LINEAR REGRESSION ASSUMPTIONS

Table 4: Linear Regression Assumptions - summary

Linear regression assumptions	
1. Multicollinearity	• Checked for VIF Score, symmetrically dropped columns bmi_category I and II to attain ideal VIF score • Rechecked the model performance, much effect not observed(R2 – 0.945) • Checked p-value of independent variables, removed all variables having p-value>0.05 • Rechecked the performance, no change in R2
2. Linearity and Independence	• No pattern is observed. Hence, the assumptions of linearity and independence are satisfied.
3. Normality	• Since p-value < 0.05, the residuals are not normal as per the Shapiro-Wilk test. • The residuals are not normal, as an approximation, we can accept this distribution as close to being normal. So, the assumption is satisfied.

4. Homoscedascity	• p-value > 0.05, i.e 0.483. we can say that the residuals are homoscedastic. So, this assumption is satisfied.

Test performance comparison:

	Linear Regression (initial)	Linear Regression (final)
RMSE	3331.873445	3333.699033
MAE	2697.706834	2700.113530
R-squared	0.945064	0.945004
Adj. R-squared	0.944732	0.944960
MAPE	14.997179	15.011629

- The train and test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting.
- Hence, we can conclude the model olsmodel4(final) is good for prediction as well as inference purposes.

Final Model - Summary

With our linear regression model, we have been able to capture ~94 of the variation in our data.

The model indicates that the most significant factors affecting the insurance cost are the following:

- regular checkup last year
- weight
- weight change in last 1 year
- year since admission
- insurance covered by other company

FINAL MODEL SUMMARY

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 5.031e+04
Date:                Sat, 28 Jun 2025   Prob (F-statistic):               0.00
Time:                        21:03:36   Log-Likelihood:            -1.6694e+05
No. Observations:               17500   AIC:                         3.339e+05
Df Residuals:                   17493   BIC:                         3.340e+05
Df Model:                           6
Covariance Type:            nonrobust
====================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                       -8.221e+04    297.540   -276.292      0.000   -8.28e+04   -8.16e+04
years_of_insurance_with_us    -26.3056     10.153     -2.591      0.010     -46.206      -6.405
regular_checkup_last_year    -645.8177     28.163    -22.931      0.000    -701.021    -590.615
adventure_sports             1.496e-10   5.46e-13    274.045      0.000    1.49e-10    1.51e-10
heart_decs_history          -4.999e-11   1.83e-13   -272.815      0.000   -5.03e-11   -4.96e-11
weight                       1461.9696      3.598    406.291      0.000    1454.917    1469.023
weight_change_in_last_one_year 173.7986    16.243     10.700      0.000     141.962     205.636
Years_since_admission         309.3255     22.853     13.535      0.000     264.531     354.119
covered_by_any_other_company_Y 1210.2474    57.377     21.093      0.000    1097.783    1322.712
==============================================================================
Omnibus:                      568.852   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              688.185
Skew:                           0.395   Prob(JB):                    3.65e-150
Kurtosis:                       3.566   Cond. No.                     2.87e+18
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.16e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Training Performance

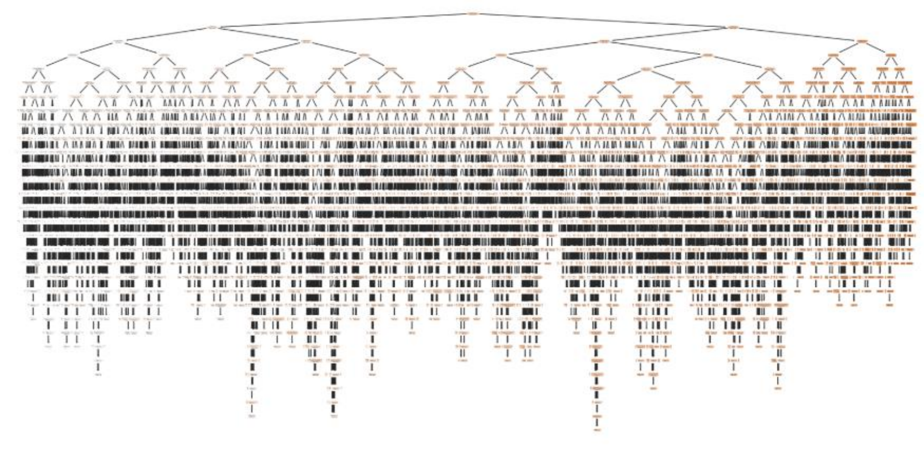
	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	3363.102908	2702.192815	0.945222	0.945194	15.152233

Test Performance

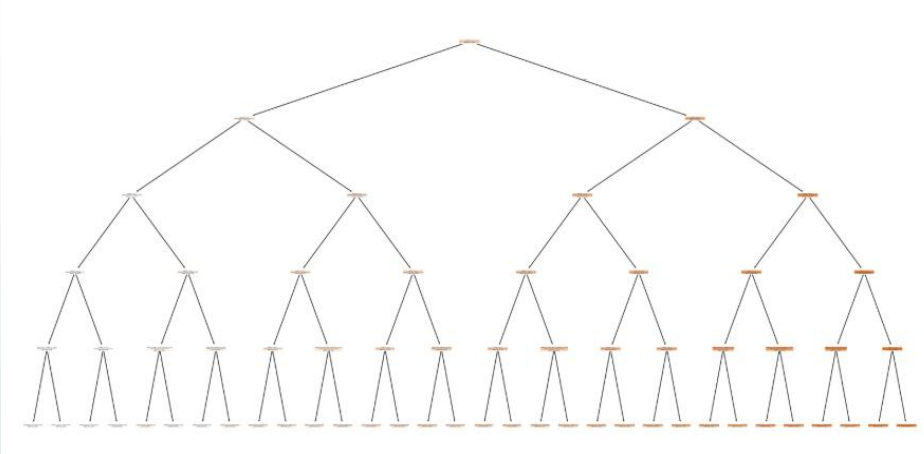
	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	3321.75315	2689.953861	0.945397	0.945332	14.918289

Figure 9: Final Model

ENSEMBLING TECHNIQUES – DECISION TREE REGRESSOR



```
DecisionTreeRegressor
DecisionTreeRegressor(random_state=42)
```

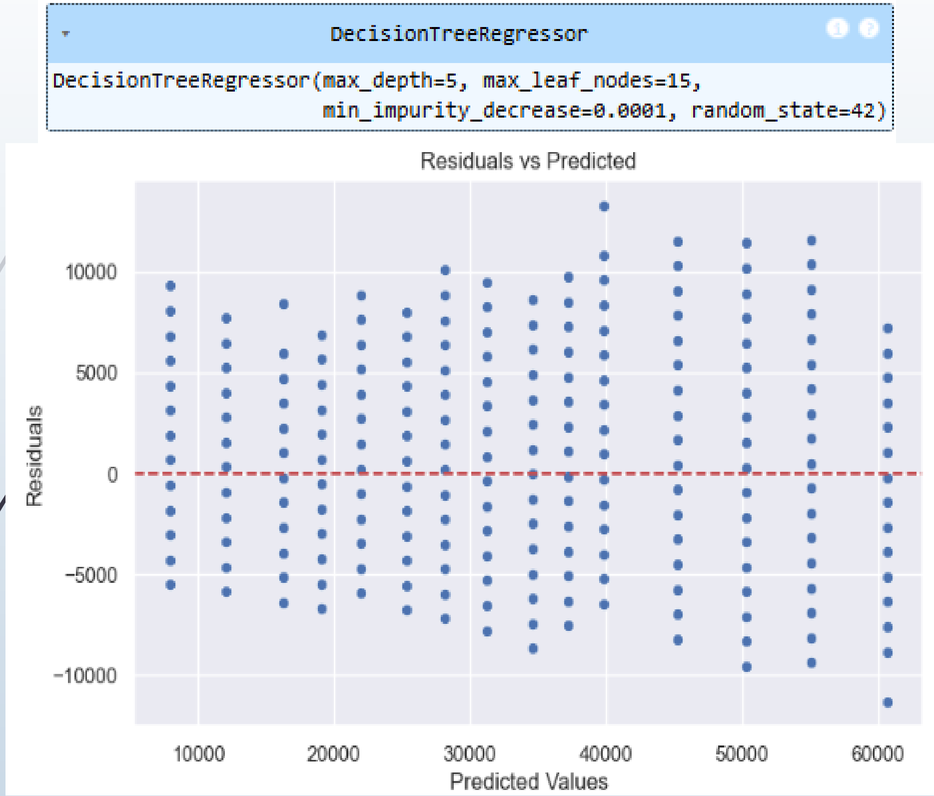


```
DecisionTreeRegressor
DecisionTreeRegressor(max_depth=5, min_samples_split=10, random_state=42)
```

Figure 10: Decision Tree Regressor

TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.95 MAE: 2532.861 MSE: 9894087.4787	R^2 Score: 0.95 MAE: 2525.739 MSE: 9938145.3981

DECISION TREE REGRESSOR – HYPERTUNING



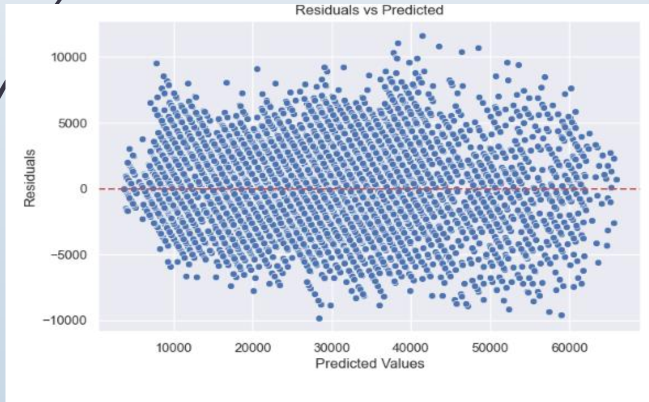
TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.947 MAE: 2650.34 RMSE: 3290.81	R^2 Score: 0.947 MAE: 2627.35 RMSE: 3277.87

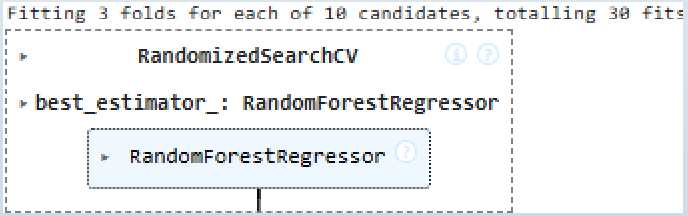
Figure 11: Decision Tree Regressor Hypertuned performance

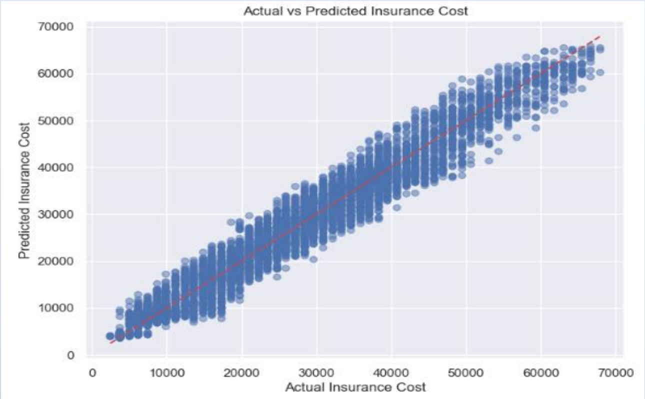
RANDOM FOREST REGRESSOR & HYPER-PARAMETER TUNING

TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.99 MAE: 905.589 RMSE: 1148.10	R^2 Score: 0.95 MAE: 2406.79 RMSE: 3039.03

Figure 12: Random Forest regressor

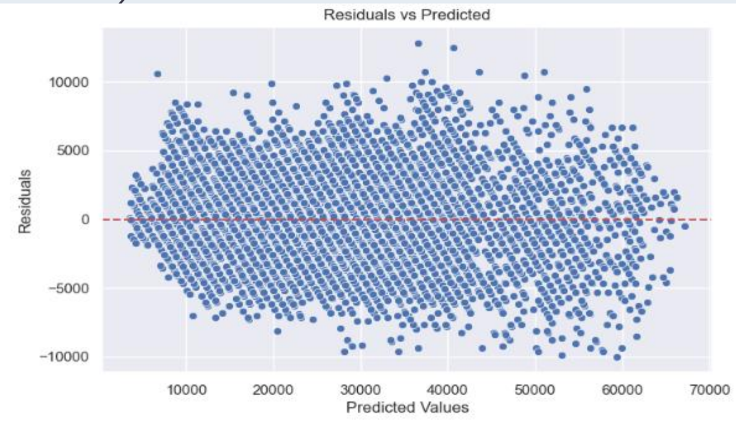






BAGGING REGRESSOR AND HYPER-PARAMETER TUNING

TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.99 MAE: 1005.04 RMSE: 1362.76	R^2 Score: 0.95 MAE: 2520.25 RMSE: 3196.08



TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.98 MAE: 1207.44 RMSE: 1603.85	R^2 Score: 0.95 MAE: 2511.22 RMSE: 3174.72

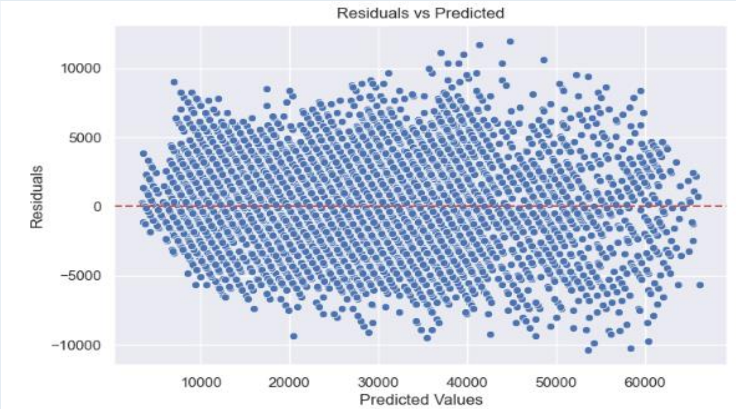
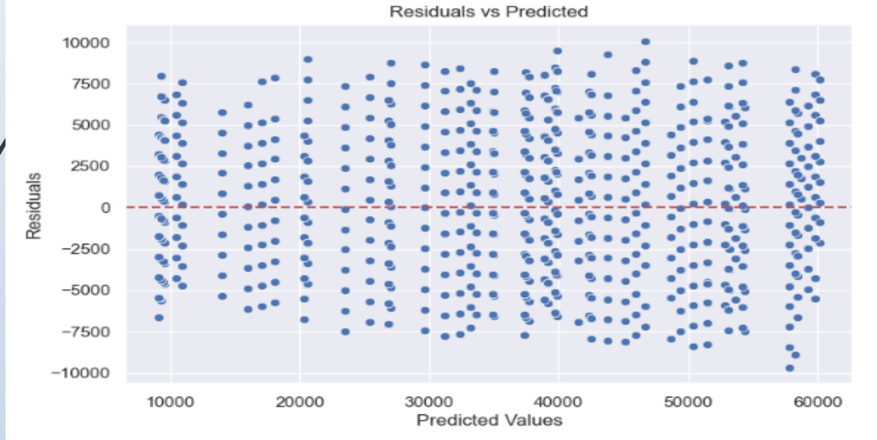


Figure 13: Bagging Regressor

BOOSTING REGRESSOR & HYPER-PARAMETER TUNING

TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.99 MAE: 905.59 RMSE: 1148.10	R^2 Score: 0.95 MAE: 2406.79 RMSE: 3039.03

TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.95 MAE: 2309.87 RMSE: 2875.20	R^2 Score: 0.95 MAE: 2356.45 RMSE: 2937.92



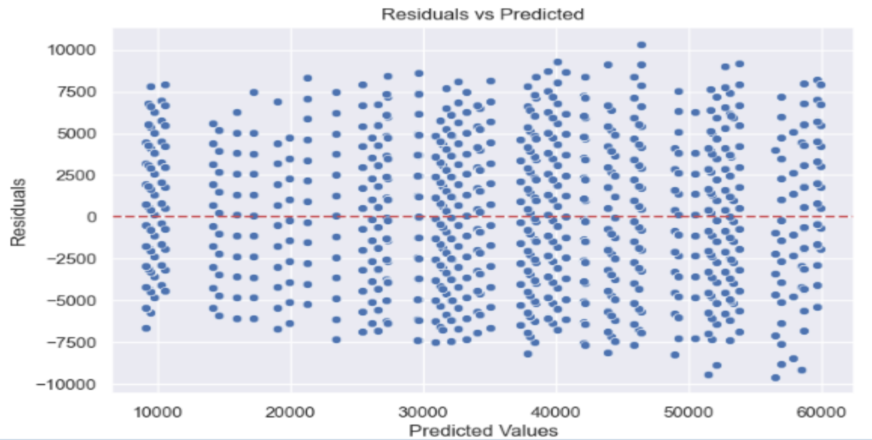
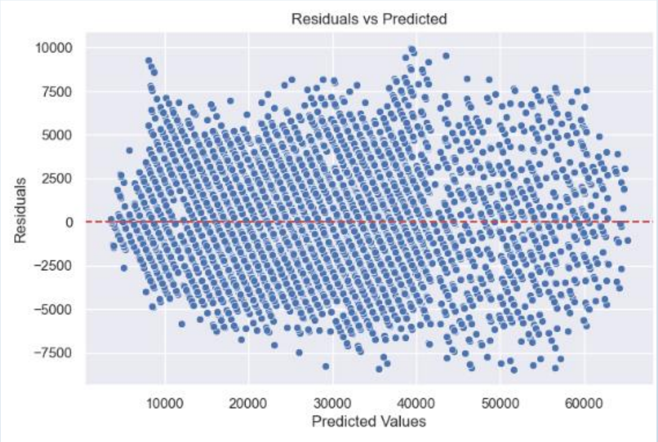


Figure 14: Boosting Regressor

GRADIENT BOOSTING REGRESSOR &HYPER TUNING

TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.95 MAE: 2371.29 RMSE: 2947.64	R^2 Score: 0.95 MAE: 2384.08 RMSE: 2964.58

TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.95 MAE: 2309.87 RMSE: 2875.20	R^2 Score: 0.95 MAE: 2356.45 RMSE: 2937.92



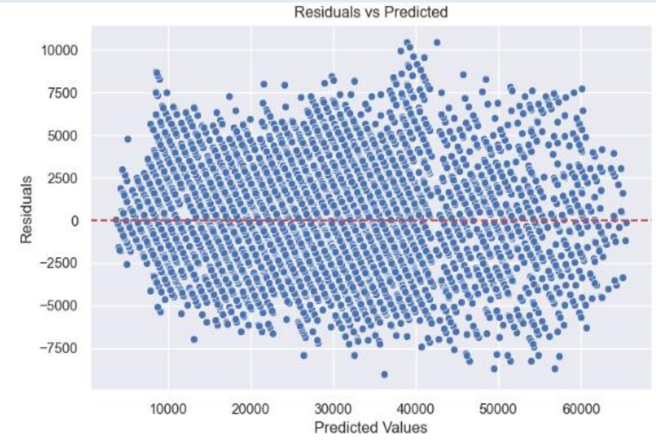
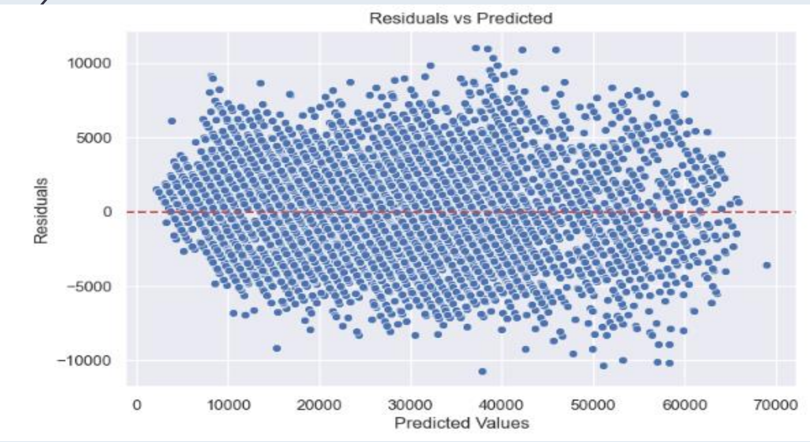


Figure 15: Gradient Boosting Regressor

XGBOOST REGRESSOR & HYPERPARAMETER TUNING

TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.97 MAE: 1661.54 RMSE: 2117.23	R^2 Score: 0.95 MAE: 2446.71 RMSE: 3072.006

TRAIN PERFORMANCE	TEST PERFORMANCE
R^2 Score: 0.95 MAE: 2309.87 RMSE: 2875.20	R^2 Score: 0.95 MAE: 2356.87 RMSE: 2937.92



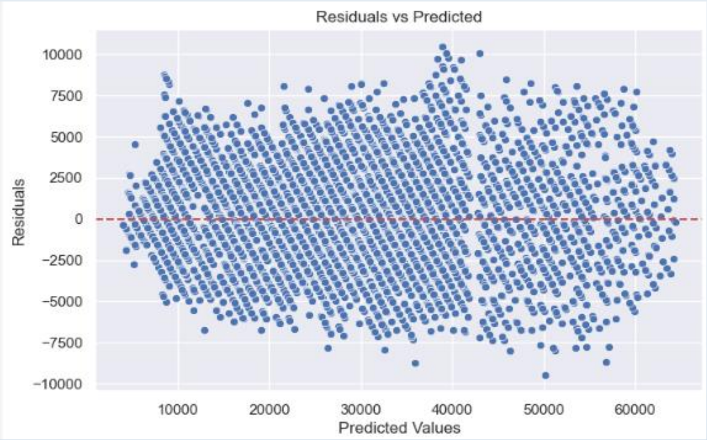


Figure 16: XGBooster regressor

ENSEMBLING TECHNIQUES

- Checked various ensemble techniques on the train and test set
- Carried out Hypertuning for each model
- Since, the model is regressive analysis, we have used MAE/MSE and R^2 as the metrics to check their performance.
- Following techniques were performed

Training performance comparison:

	Decision Tree Estimator	Random Forest Tuned	Bagging Estimator Tuned	Adabosst Classifier Tuned	Gradient Boost Classifier Tuned	XGBoost Classifier Tuned
RMSE	3290.813375	1148.104899	2875.208119	3261.278989	2814.967617	2875.208119
MAE	2650.349398	905.589586	2309.877448	2689.273159	2257.882674	2309.877448
R-squared	0.947322	0.993588	0.959787	0.948263	0.961455	0.959787
MAPE	13.765495	4.345967	11.286928	15.319856	10.920485	11.286928

Testing performance comparison:

	Decision Tree Estimator	Random Forest Tuned	Bagging Estimator Tuned	Adabosst Classifier Tuned	Gradient Boost Classifier Tuned	XGBoost Classifier Tuned
RMSE	3277.874081	3039.037581	2937.926204	3234.294017	2938.188465	2937.926204
MAE	2627.359074	2406.795245	2356.454856	2660.576024	2352.719948	2356.454856
R-squared	0.947378	0.954767	0.957727	0.948768	0.957720	0.957727
MAPE	13.701020	11.564470	11.547602	15.244621	11.411264	11.547602

Figure 17: Ensembling techniques performance's comparison

INSIGHTS

- Random Forest had the best training performance but a slightly higher RMSE on test data.
- Bagging Regressor and XGBoost showed a better balance between training and testing performance, making them great choices if generalization is the priority.
- The feature importance is given below: Dominant predictor is weight—strongly associated with insurance_cost or health status.
- Other important features are insurance covered by other company, years since admission, regular check up last year

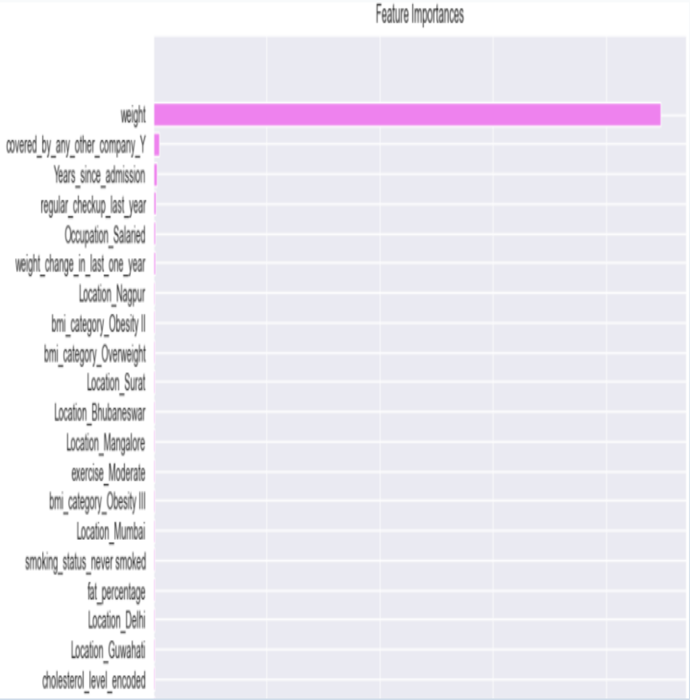


Figure 18: Feature importance

CONCLUSION

- In this project, a regression model is built to predict insurance costs based on health and lifestyle factors.
- After trying multiple models, **XGBooster Regressor** gave the best results.
- It was found that weight is the most important factor influencing insurance costs, while other features had very little impact.
- This insight can help insurance companies focus more on weight-related health risks when planning policies.