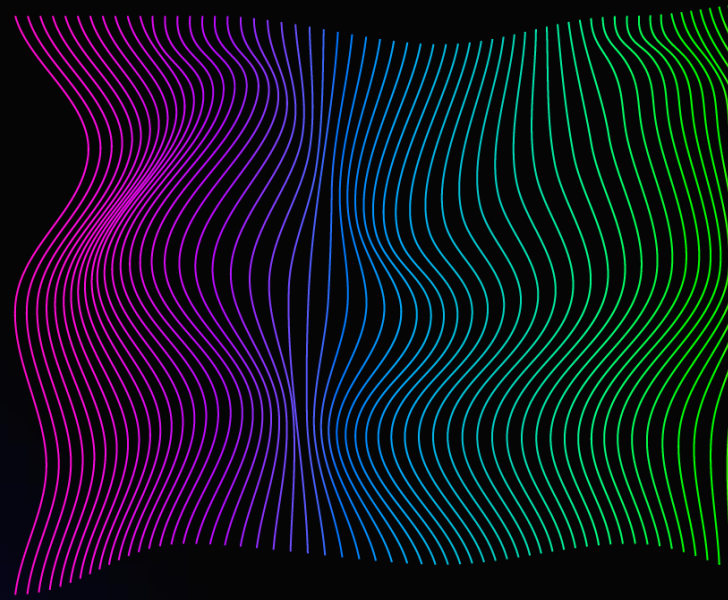
Medical Insurance Cost

A Project By Konstantinos Soufleros

TABLE OF CONTENTS

OUR DATASET

The Medical Cost Personal Datasets is a public dataset used in the book "Machine Learning with R" by Brett Lantz. It contains information about medical insurance charges for individuals based on various factors. The original dataset is available on Kaggle Medical Cost Personal Datasets.



INTRODUCTION

Medical insurance costs are a critical aspect of the healthcare industry, influencing both insurance companies and individuals. Understanding the factors that contribute to these costs can aid in accurate cost estimation, risk assessment, and decision-making in the insurance domain. In this analysis, we explore a dataset that contains information about medical insurance charges for individuals and aim to develop a predictive model for estimating insurance costs.

01

About The Project

WHAT WE WORKED ON

Project

- Exploring relationships between age, gender, BMI, children, smoking status, region, and charges.
- Predicting medical insurance costs using machine learning models.

Analysis

- Conducting data preprocessing, exploratory data analysis, and statistical analysis.
- Implementing regression algorithms (Linear, Ridge, Lasso, ElasticNet, Polynomial, SVR, Random Forest, Gradient Boosting) and evaluating their performance.

Impact

- Providing accurate cost estimations for insurance companies and individuals.
- Enhancing cost estimation, risk assessment, and decision-making in the insurance industry.

ABOUT THE PROJECT

Throughout this project, we followed a structured approach, including data exploration and preprocessing, univariate, bivariate, and multivariate analyses, as well as implementing and evaluating multiple regression algorithms. The goal was to develop a robust predictive model that accurately estimates medical insurance charges based on the available dataset.

TRANSFORM IDEAS INTO A PROJECT

1st step
Define Project Objectives

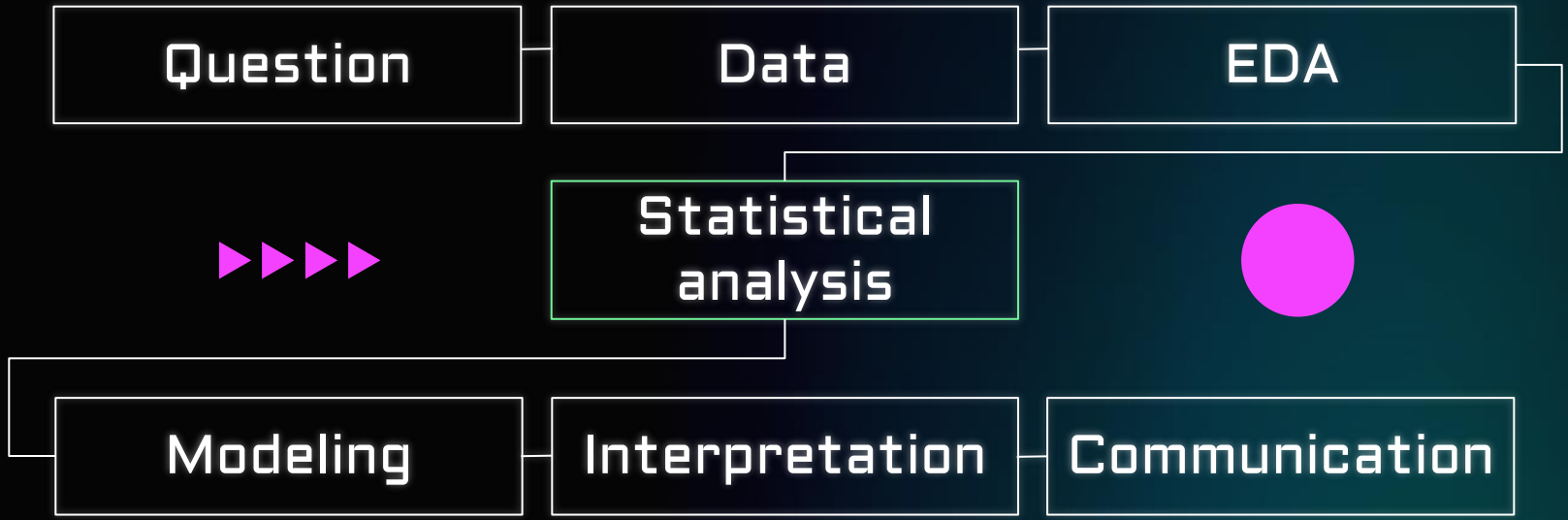
2nd step
Data Collection and Preprocessing

3rd step
Exploratory Data Analysis (EDA)

4th step
Machine Learning Modeling and Evaluation

Submit

PROJECT ARCHITECTURE

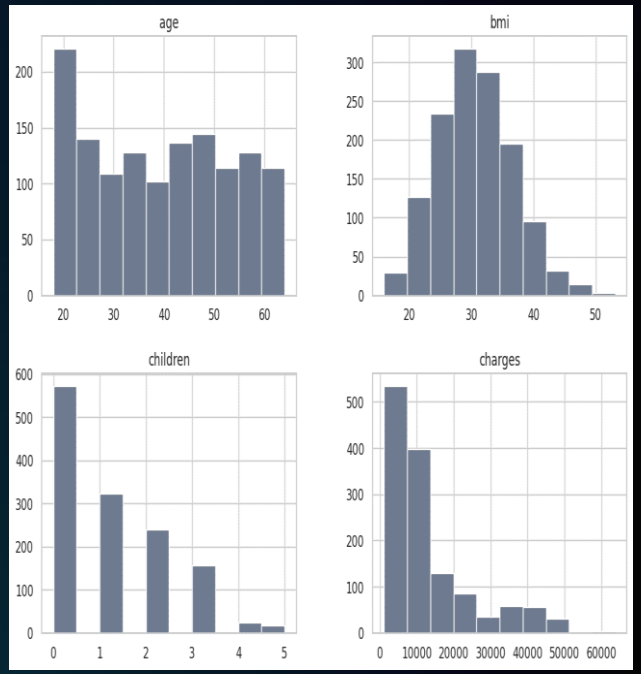


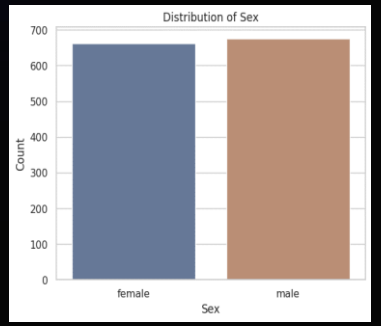
“What factors influence medical insurance costs and how accurately can they be predicted?”

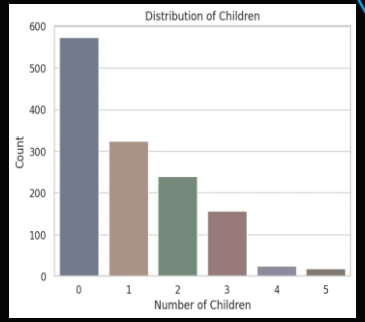
Exploratory Data Analysis

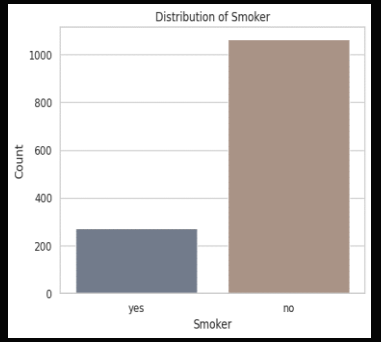
02

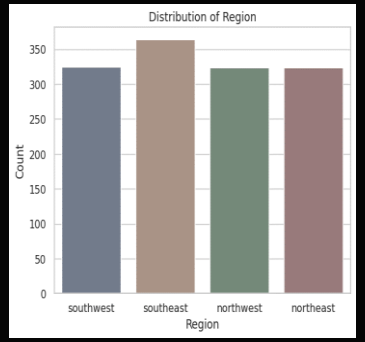
DISTRIBUTIONS





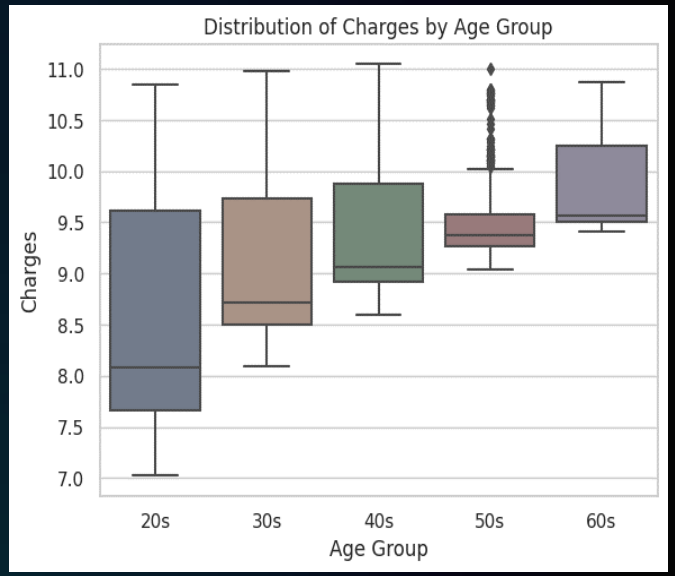


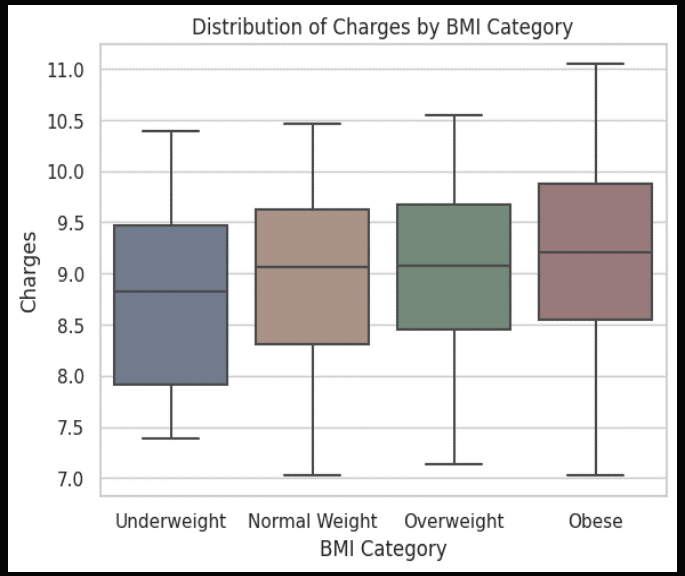




Our target variable ‘charges’ is ‘right skewed’ so we log it.

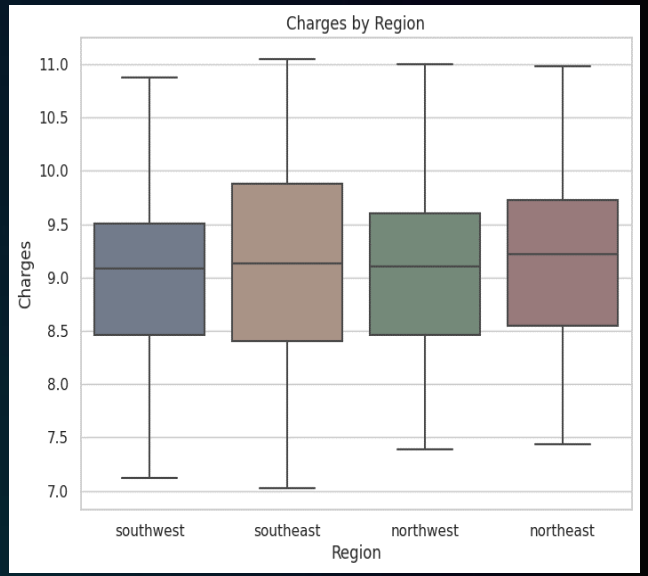
DISTRIBUTIONS

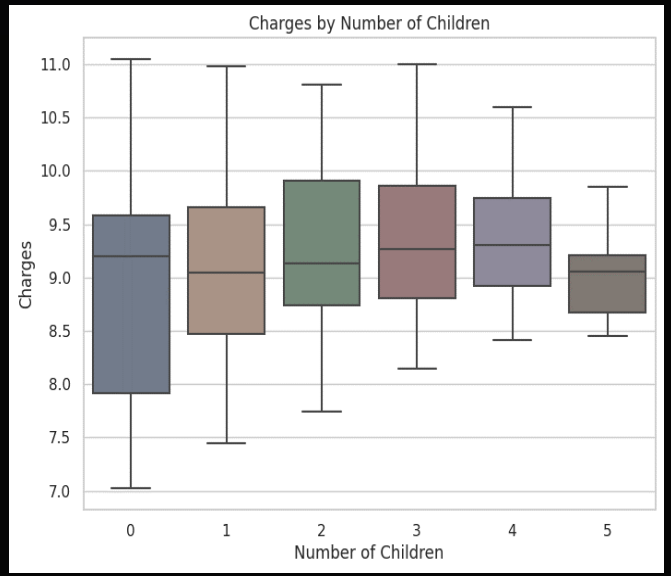




'charges' increase as 'age' increases. Same is happening in 'bmi'.

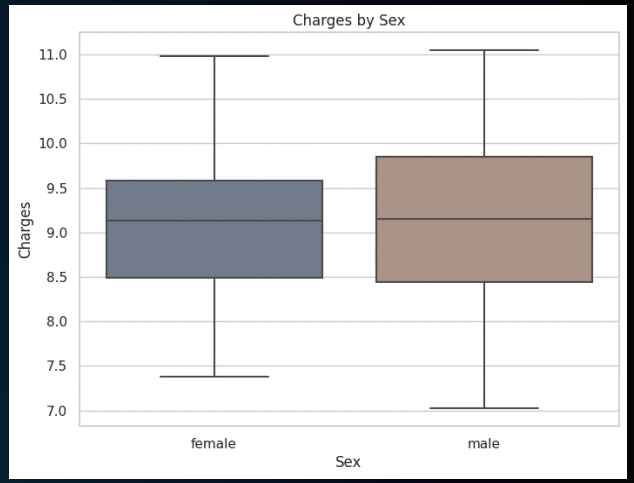
DISTRIBUTIONS

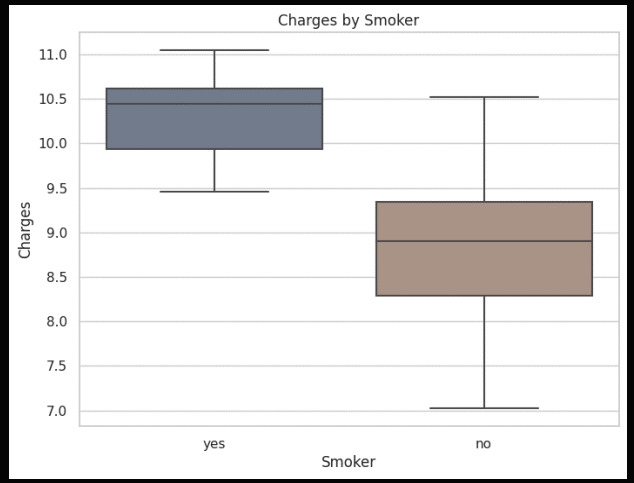




‘charges’ by ‘region’ don’t differ significantly . ‘charges’ by ‘children’ increase until the second child.

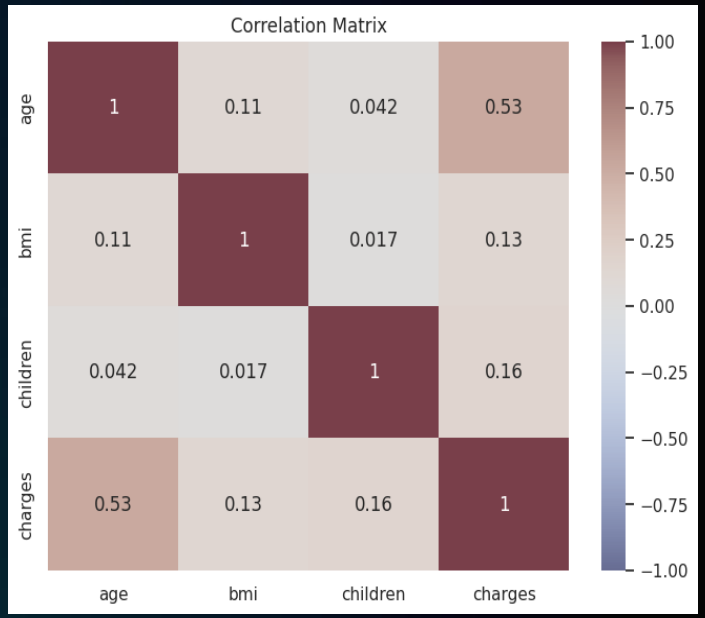
DISTRIBUTIONS

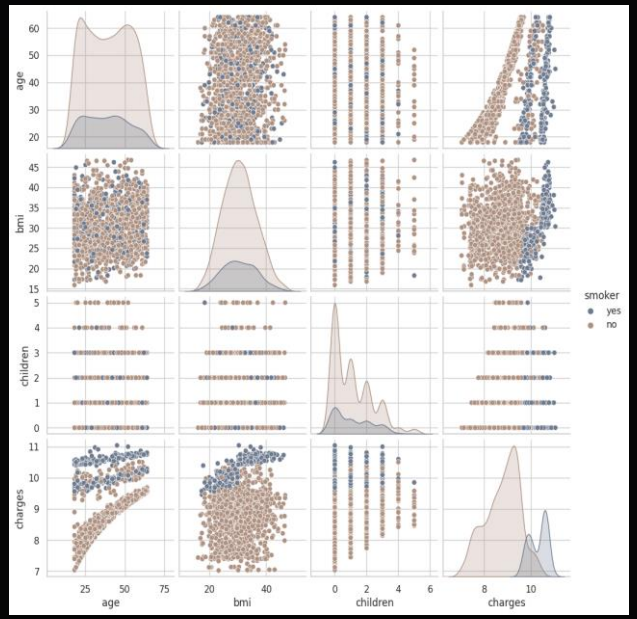




‘charges’ by smoking habits differ significantly. ‘sex’ though is not a influential factor.

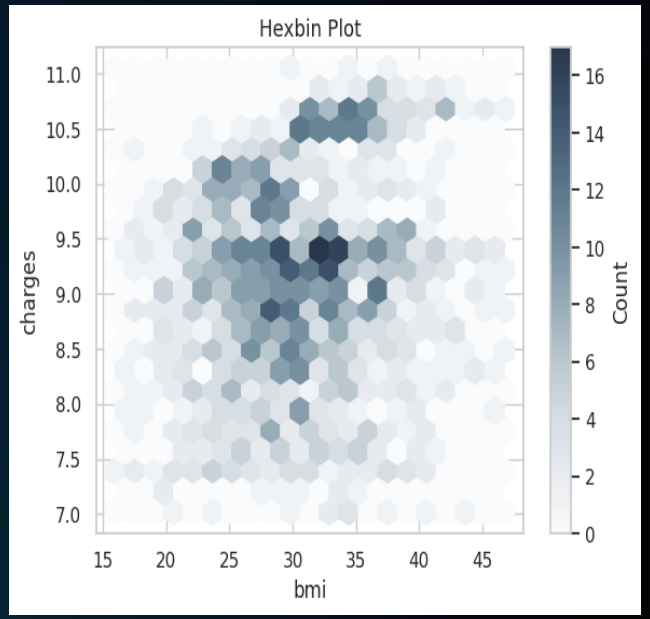
CORRELATIONS

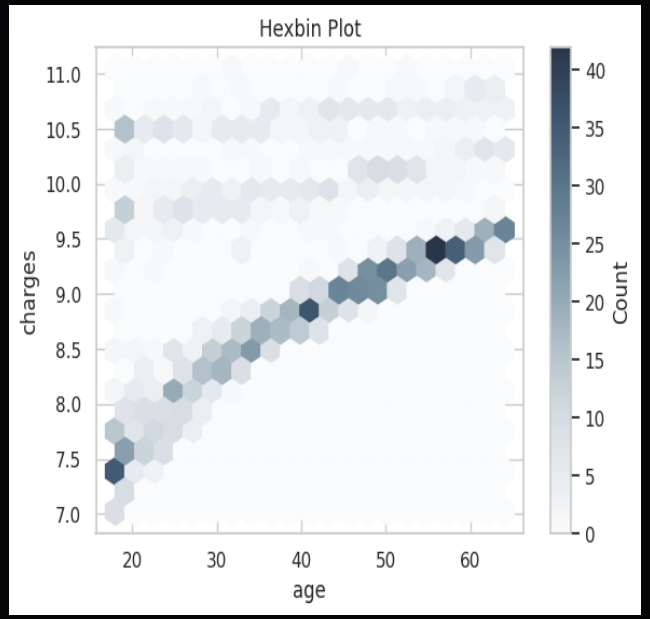




Initial examination shows correlation between ‘age’ and ‘charges’. ‘smoker’ correlates to higher ‘charges’.

CORRELATIONS





'bmi' density plot shows density in charges between 30-35 bmi. 'age' shows gradual increase as people get older.

Statistical Analysis

03

STATISTICAL TESTS

‘charges’ by ‘age_group’:

ANOVA test: p-value:9.16e-88
F-statistic:120.56

Highly significant

‘charges’ by ‘sex’

Independent t-test: p-Value: 0.679
T-Statistic: 0.413

Not significant

‘charges’ by ‘bmi_category’

Independent t-test: p-Value: 0.0003
T-Statistic: 3.62

Significant

STATISTICAL TESTS

‘charges’ by ‘smoker’:

Independent t-test: p-Value: 1.74e-169
T-Statistic: 32.31

Highly significant

‘charges’ by ‘region’

ANOVA test: p-Value: 0.222
F-Statistic: 1.46

Not significant

‘charges’ by ‘children’

ANOVA test: p-Value: 1.39e-08
F-Statistic: 9.16

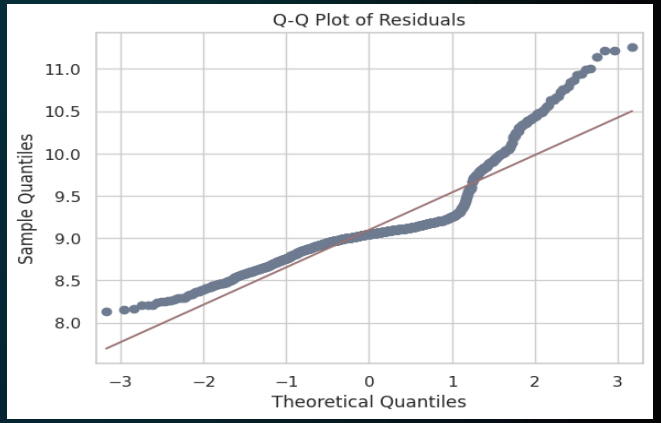
Significant

OLS ANALYSIS

	Feature	Coefficient	P-value
x1	sex_male	-0.036256	0.885553
x2	smoker_yes	0.623507	0.013604
x3	region_northwest	-0.027135	0.929988
x4	region_southeast	-0.069605	0.828275
x5	region_southwest	-0.056433	0.855546
x6	age	0.483751	0.056037
x7	bmi	0.082685	0.753160
x8	children	0.121991	0.627349

| R-squared (uncentered)| 0.008
| Adj. R-squared (uncentered) | 0.002
| F-statistic | 1.277
| Prob (F-statistic) | 0.251

The model has low explanatory power, only smoking appears to be a significant predictor of higher charges. The low R-squared value (0.008) suggests non-linear relationships. The Q-Q plot also supports this.



Regression Algorithms

04

OUR MODELS

Linear Regression

MSE: 0.20
R^2 Score: 0.74
Train Score: 0.77
Test Score: 0.74

Ridge Regression

MSE: 0.20
R^2 Score: 0.74
Train Score: 0.77
Test Score: 0.74

Lasso Regression

MSE: 0.77
R^2 Score: -0.0003
Train Score: 0.00
Test Score: -0.0003

ElasticNet Regression

MSE: 0.69
R^2 Score: 0.11
Train Score: 0.11
Test Score: 0.11

Polynomial Regression

MSE: 0.17
R^2 Score: 0.78
Train Score: 0.85
Test Score: 0.78

Support Vector Regression

MSE: 0.17
R^2 Score: 0.77
Train Score: 0.85
Test Score: 0.77

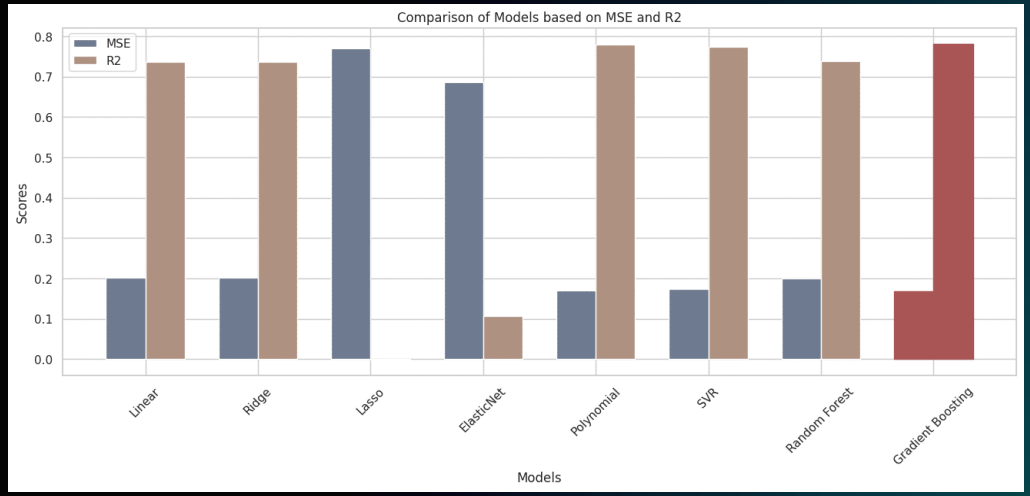
Random Forest Regression

MSE: 0.20
R^2 Score: 0.74
Train Score: 0.97
Test Score: 0.74

Gradient Boosting Regression

MSE: 0.17
R^2 Score: 0.78
Train Score: 0.89
Test Score: 0.78

The Mean Squared Error (MSE) measures the average squared difference between the predicted and actual values, with lower values indicating better performance. The R-squared score quantifies the proportion of variance in the target variable explained by the model, with higher values indicating a better fit to the data.



INTERPRETATIONS

Models Performance

The regression models performed moderately well in predicting medical insurance costs.
Train and test scores indicated reasonable performance, with a tendency for slight overfitting in some models.

Poor Performance

Lasso and ElasticNet regression models showed relatively poor performance compared to other models.
These models may have struggled to capture the complex relationships within the data due to their regularization techniques. Although, this changed later in hyperparameter tuning.

INTERPRETATIONS

Polynomial Model

The Polynomial Regression model showed promising performance in comparison to the Linear Regression model.
The polynomial model's ability to capture higher-order interactions and nonlinear effects contributed to its improved performance.

Best Performance

The Gradient Boosting Regressor demonstrated the best performance among all the models.
It achieved the lowest MSE and highest R^2 score, indicating a strong ability to predict medical insurance costs accurately.
The Gradient Boosting model effectively captured the underlying patterns and nonlinear relationships within the dataset.

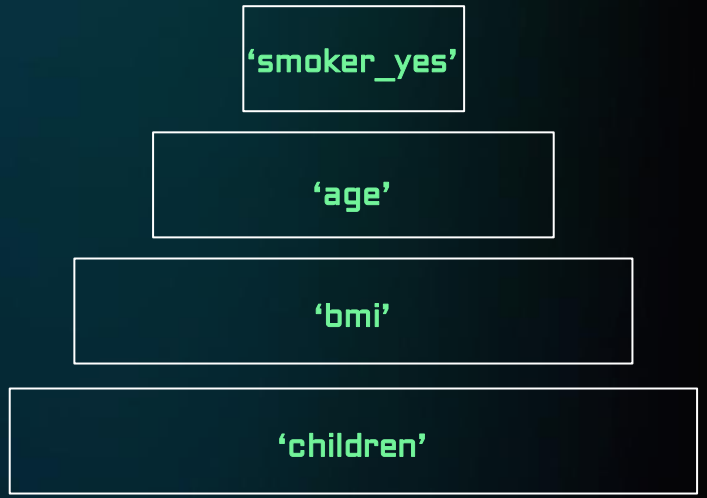
HYPERPARAMETER TUNING

Linear	Ridge	Lasso	ElasticNet
MSE: 0.20 R^2 Score: 0.74	MSE: 0.20 R^2 Score: 0.74	MSE: 0.21 R^2 Score: 0.73	MSE: 0.20 R^2 Score: 0.75
Polynomial	**SVR**	**Random Forest**	**Gradient Boosting**
MSE: 0.17 R^2 Score: 0.78	MSE: 0.17 R^2 Score: 0.77	MSE: 0.20 R^2 Score: 0.73	MSE: 0.17 R^2 Score: 0.78

FEATURE IMPORTANCE

Gradient Boosting Model

The feature importance values indicate the relative contribution of each feature in predicting the target variable.
Features with higher importance values have a stronger influence on the model's predictions.
In this model, the top four features with the highest importance are "smoker_yes," "age," "bmi," and "children" suggesting that these factors play a significant role in determining the charges.



Key Findings

05

KEY FINDINGS

AGE

Age is positively correlated with charges, indicating that as age increases, medical charges tend to increase as well.
The 50's age group shows higher charges compared to other age groups, suggesting that age may be a significant factor in determining healthcare costs.

SEX

There is no statistically significant difference in charges between males and females.
Gender alone does not appear to be a significant predictor of healthcare costs in the dataset.

BMI

Higher BMI values are associated with higher medical charges, indicating that BMI is an important factor in determining healthcare costs. There are variations in charges across different BMI categories, with the 'Obese' category having higher charges compared to the 'Normal Weight' category.

KEY FINDINGS

CHILDREN

The number of children has a slight influence on medical charges, with higher charges observed for individuals with 2 or 3 children.
There is a gradual decrease in charges as the number of children increases beyond 3, with lower charges for individuals with 5 children.

SMOKER

Smokers tend to have significantly higher healthcare charges compared to non-smokers. Smoking behavior is an important predictor of medical expenses, with smokers experiencing higher charges on average.

REGION

There are no significant differences in charges based on the region. The geographical location does not appear to be a strong predictor of healthcare costs.

STRENGTHS AND LIMITATIONS

Feature importance analysis provided insights into the relative importance of different variables.

The models relied on the available dataset and might not capture all possible factors influencing healthcare charges.

The Gradient Boosting Regression model demonstrated high predictive performance, capturing non-linear relationships.

The models' performance may vary when applied to different datasets or time periods.

FUTURE STEPS

Model

Fine-tune the Gradient Boosting Regression model.

Feature engineering

Create new features or transform existing ones to capture additional information.

External data integration

Incorporate external datasets that provide additional information.

Model ensemble and stacking

Experiment with combining multiple regression models

06 Conclusion

In conclusion, our analysis of medical insurance costs has provided valuable insights into the factors influencing insurance charges and the performance of various regression models.

- Predictive Power: Through the implementation of different regression algorithms, we demonstrated the ability to accurately predict medical insurance costs based on features such as age, BMI, smoking status, region, and gender. Our best-performing model, Gradient Boosting Regressor, achieved an impressive R^2 score of 0.78, indicating its high predictive power.
- Feature Importance: Our analysis highlighted the significant influence of smoking status, age, and BMI on insurance charges. These factors should be carefully considered when estimating healthcare expenses and setting insurance premiums.
- Practical Implications: The insights gained from this analysis can assist insurance companies in accurately estimating insurance costs, managing risk, and providing fair pricing to policyholders. Individuals can also benefit from a better understanding of the factors influencing their insurance charges, enabling them to make informed decisions regarding their healthcare coverage.

THANKS

Do you have any questions?

soufleros.kostas@gmail.com
+381 61 297 9469

https://www.linkedin.com/in/konstantinos-soufleros/

https://github.com/kostas696