Assignment: Data Analytics Process and Interpretation

IS4116 - Business Intelligence Systems

By
H.H.S.Fernando - Index No: 20020376

Degree of Bachelor of Science Honours in Information Systems

University of Colombo School of Computing
35, Reid Avenue, Colombo 07,
Sri Lanka

February 2025

1 Introduction

Healthcare costs are a significant concern for individuals and insurance companies. Understanding the factors that influence medical insurance costs can help insurance providers develop fair pricing strategies .This report presents an analysis of a dataset containing medical insurance records to identify key patterns in insurance charges and develop a predictive model to estimate insurance costs based on customer profiles.

1.1 Objective

- Explore the dataset to understand patterns and factors affecting medical insurance costs.
- Apply statistical methods to determine relationships between variables.
- Build a predictive model to estimate insurance costs and provide actionable insights for insurance companies and policyholders.

1.2 Business Question

Medical insurance companies need accurate cost estimation to offer competitive and fair premiums. Understanding how factors like age, smoking habits, and BMI impact costs enables companies to make data-driven decisions, optimize pricing models, and reduce financial risks.

1.3 Overview of the Dataset

The Medical Insurance Cost Prediction dataset provides information about individuals and their medical insurance charges. he dataset is sourced from Kaggle and contains 1,338 records with 7 features

1.3.1 Columns & Description

Column Name	**Data Type**	**Description**
age	Integer	Age of the individual (18-64 years)
sex	Categorical	Gender of the individual (male, female)
bmi	Float	Measure of body fat based on height and weight
children	Integer	Number of dependents (children) covered by insurance
smoker	Categorical	Smoking status (yes = smoker, no = non-smoker)
region	Categorical	Geographic region of residence (northeast, northwest, southeast, southwest)
charges	Categorical	Target Variable - Medical insurance cost (in USD)

Table 1: Column Descriptions

2 Methodology

2.1 Data Preprocessing & Cleaning

To ensure high-quality data for analysis and modeling, the following preprocessing steps were performed:

1. Handling Missing Values - The dataset was checked for missing values.Since no missing values were found, no imputation was necessary.
2. Checking for Outliers - Boxplots were generated using seaborn.boxplot() to detect outliers in numerical features (e.g., bmi, charges).
3. Encoding Categorical Variables - The dataset contained three categorical variables (sex, smoker, region), which were converted into numerical values to be used in machine learning models: One-hot encoding was applied to region, Label encoding was used for sex

2.2 Exploratory Data Analysis (EDA)

Before applying machine learning models, EDA was conducted to understand the dataset's structure and characteristics:

1. Summary Statistics: Measures like mean, median, and standard deviation were analyzed to understand the distribution of features.
2. Correlation Analysis: Pearson correlation coefficients were computed to identify relationships between features and the target variable.
3. Outlier Detection: Boxplots and histograms were used to detect extreme values that might impact model training.
4. Feature Importance: Initial assessments using Random Forest and Gradient Boosting models helped determine the most significant predictors.

The insights gained through EDA guided feature selection for predictive modeling.

2.3 Statistical Analysis

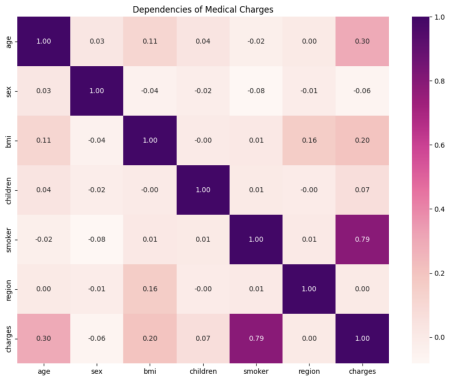
To predict insurance charge levels, 4 models (Linear Regression, Decision Tree Regressor , Random Forest Regressor ,Gradient Boosting Regressor (XGBoost))were trained and evaluated. The dataset was divided into training sets (80%) and testing sets (20%) to evaluate the model performance. Feature importance analysis was also performed to identify the most influential factors that affect insurance charges.

3 Results

The dataset was examined for patterns, distributions, and relationships among variables affecting insurance charges.

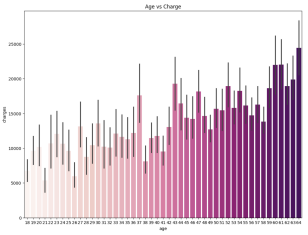
3.1 Correlation Analysis

A correlation analysis was performed to determine relationships between variables and insurance charges. The key findings include are Smoking status has the strongest correlation, BMI and Age is also positively correlates but not as strongly as smoking , Gender has weak correlation with Number of Children as a weak correlation to charges

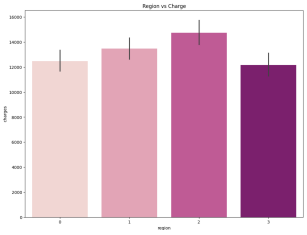


3.2 Insurance Charges vs. Risk Factors

The following figures illustrates relationship between risk factors and insurance charges using a barplots and scatterplots . Key observations include:
- Smoking is the biggest factor affecting medical insurance charges .
- Age is positively correlated with higher charges, BMI influences medical expenses
- Region does not significantly impact charges , Gender does not have a significant impact

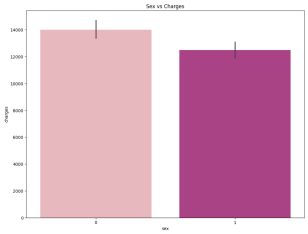


(a) Age vs Charges

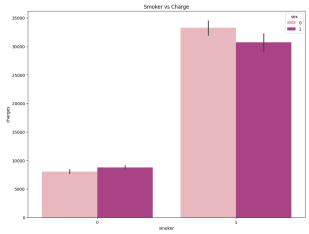


(b) Region vs Charges

(c) BMI vs Charges



(d) Sex vs Charges



(e) Smoker vs Charges

3.3 Statistical Methods - Models

Multiple machine learning models were applied and evaluated based on performance metrics

Model	R^2	RMSE	MAE	Note
Linear Regression	0.74	6319.27	4160.25	Improved accuracy by capturing non-linear relationships
Random Forest Regressor	0.95	2446.07	695.44	struggled with complex relationships,sensitive to outliers
Gradient Boosting Regressor (GBR)	0.96	6319.27	4160.25	reducing errors by optimizing weak learners iteratively.
XGBoost)	0.96	2504.11	694.18	Efficiently handled categorical features.

The XGBoost was considered the best model and the predicted vs actual charges were visualized giving the following

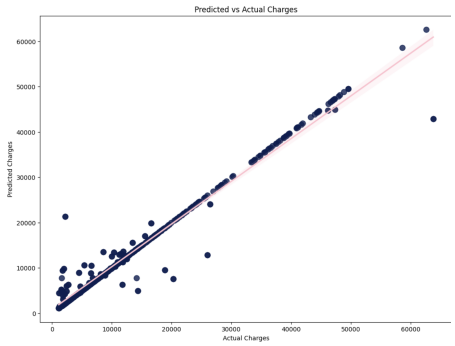
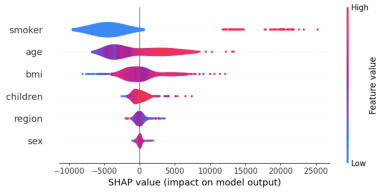


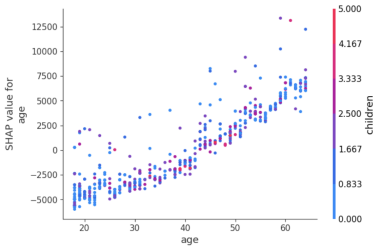
Figure 1: Predicted vs actual

The spread around the regression line is minimal, meaning the model has low variance and captures trends well.

Feature importance analysis confirmed smoking, BMI, and age as the strongest predictors.



(a) SHAP Analysis



(b) SHAP for Age

SHAP force plot for the random person explaining which factors influence the prediction the most

Figure 2: SHAP for random person

3.4 Business Implications

Based on the findings, the following business strategies can be considered:

- Risk-Based Premium Adjustments: Implement tiered pricing models where smokers and individuals with high BMI pay higher premiums.
- Preventive Health Programs: Offer incentives for maintaining a healthy BMI and quitting smoking, potentially reducing long-term claims.
- Regional Pricing Strategies: Adjust premiums based on regional risk factors and demographic trends.
- Customizable Insurance Plans: Provide personalized insurance plans based on individual health metrics and lifestyle choices.
- Gender-Specific Plans: Consider gender-based pricing models if further data supports a statistically significant difference in medical costs.
- Health Monitoring Incentives: Introduce wellness programs that reward policyholders for maintaining healthy lifestyles, such as regular health check-ups and fitness activities.

4 Conclusion

This analysis highlights the significant impact of smoking, BMI, age, gender, and regional factors on insurance charges. Business strategies should focus on risk-based pricing, preventive health programs, and regional adjustments to optimize profitability while maintaining competitive insurance offerings. By leveraging data-driven insights, insurance providers can create more tailored and efficient pricing models that align with customer health profiles and behaviors.

5 Github Repository

https://github.com/shamin2021/BIS_Medical_Insuarance_Cost_Analysis.git