Introduction

The aim of this project is to develop a predictive model for medical charges billed by an insurance company using linear regression. The dataset employed in this analysis, titled "insurance," encompasses various attributes including age, gender, body mass index (BMI), number of children, smoking status, geographic region, and medical charges.

Task Description:

To accomplish the objective of predicting medical charges, the following tasks were undertaken:

a. Summary Statistics: Summary statistics for the variable "charges" were computed to gain insights into the distribution and central tendency of medical charges within the dataset.

b. Regional Distribution: A tabular representation displaying the count of individuals within each geographic region was generated to understand the demographic distribution.

c. Visualization: A scatterplot matrix was employed to visualize the relationships among all features in the dataset. This facilitated the exploration of potential correlations and patterns between variables.

d. Model Training: Linear regression models were trained on the dataset to establish a predictive relationship between the input features and medical charges.

e. Model Evaluation: The performance of each regression model was evaluated using various metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score.

f. Model Improvement: Strategies were employed to enhance model performance, including the addition of nonlinear relationships and interaction effects. Specifically, nonlinear relationships were incorporated by considering age as the sole input variable, while interaction effects between smoking status and obesity (BMI > 30) were introduced to capture potential synergistic impacts on medical charges.

Evaluation Criteria:

The evaluation of the project encompasses several criteria:

- Data Understanding and Exploration
- Code Quality and Documentation
- Visualization and Feature Relationships
- Model Performance and Evaluation

These evaluation criteria serve as benchmarks for assessing the quality and efficacy of the predictive models developed throughout the project.

Data Exploration

Missing Values:
There are no missing values in the dataset across any of the columns.

Data Types:
The dataset consists of a combination of integer, float, and object data types.

Dataset Information:
The dataset contains 1338 entries and 7 columns. The columns include age, sex, BMI, number of children, smoking status, region, and medical charges.

Dataset Summary Statistics:

- Age: The age variable ranges from 18 to 64 years, with a mean of approximately 39 years.
- BMI: Body Mass Index ranges from 15.96 to 53.13 kg/m2, with a mean of approximately 30.66 kg/m2.
- Children: The number of children ranges from 0 to 5, with a mean of approximately 1.09.
- Charges: Medical charges range from $1121.87 to $63770.43, with a mean of approximately $13270.42. The distribution of charges exhibits a positive skew, with a median charge of $9382.03.

Observations:

<u>Observation 1:</u>

Age Distribution: The age variable shows a nearly uniform distribution, indicating a relatively even representation across different age groups.
BMI Distribution: The BMI variable exhibits a distribution close to normal, with a slight positive skew indicating a few outliers with higher BMI values.
Number of Children: The distribution of children follows a pattern resembling a log-normal distribution, with a decrease in observations as the number of children increases.
Charges Distribution: The charges variable displays a log-normal distribution, indicating a higher concentration of individuals with lower charges and a smaller number of individuals with higher charges.

Observation 2:

1. Sex Distribution: The dataset includes 676 male and 662 female individuals, suggesting a balanced distribution between the two sexes.
2. Smoking Habits: Among the individuals, 1064 are non-smokers, while 274 are smokers, indicating a larger proportion of non-smokers.
3. Regional Distribution: Individuals are distributed across four regions: southeast, southwest, northwest, and northeast, with a relatively balanced representation across regions.

Outlier Analysis:

Outliers primarily consist of individuals with relatively higher medical charges, often associated with attributes such as older age, higher BMI, and smoker status. Most outliers are from the southeast region. However, addressing outliers may lead to a loss of valuable information and adversely affect model performance.

This comprehensive exploration of the dataset provides insights into the distribution of variables and helps in understanding the characteristics of the data, which is crucial for further analysis and model development.

Data Exploration

Summary Statistics of Charges:

The summary statistics of the variable 'charges' in the dataset are as follows:

1. Count: There are 1338 entries in the dataset, indicating the number of observations we have.
2. Mean: The average charge is approximately $13,270.42.
3. Standard Deviation (Std): The standard deviation of charges is approximately $12,110.01, indicating the extent of variability or dispersion in the charges.
4. Minimum (Min): The minimum charge recorded is $1,121.87, representing the lowest charge in the dataset.
5. 25th Percentile (25%): 25% of the charges fall below $4,740.29, indicating the first quartile of the distribution.
6. Median (50%): The median charge, also known as the 50th percentile, is $9,382.03. This represents the middle value in the dataset when arranged in ascending order.
7. 75th Percentile (75%): 75% of the charges fall below $16,639.91, indicating the third quartile of the distribution.
8. Maximum (Max): The maximum charge recorded is $63,770.43, representing the highest charge in the dataset.

Number of People in Each Region:
A table showing the count of individuals in each region is presented below:

Region	Number of People
southeast	364
southwest	325
northwest	325
northeast	324

Visualize Relationship Among Features:

A scatter plot matrix was created to visualize the relationship among all features, categorized by relevant variables such as gender, smoking status, and region. The analysis of the scatter plots identified patterns and insights regarding the impact of different factors on medical charges.

Insights from the scatter plots include:

1. Gender Impact on Charges: No significant disparity based on gender was observed in the distribution of charges.
2. Smoking Habits Influence: Smokers generally exhibited higher charges compared to non-smokers.

3. Regional Disparities: No clear separation or pattern was observed, indicating that regional differences may not have a significant influence on medical charges.
4. Age as a Contributing Factor: Older individuals tended to incur higher medical charges, with age being a significant contributing factor.
5. Parental Status Impact: Individuals without children tended to have higher charges compared to those with children.

Noteworthy Finding: Individuals classified as obese and smokers exhibited the highest charges, indicating the significance of the intersection between obesity and smoking in influencing medical charges.
These insights provide valuable information for understanding the factors influencing medical charges in the dataset.

Model Development and Evaluation

Model Training and Evaluation

1. Simple Linear Regression with All Columns:

Performance Metrics:

- Mean Absolute Error (MAE): 4181.19
- Mean Squared Error (MSE): 33596915.85
- Root Mean Squared Error (RMSE): 5796.28
- R-squared (R2 score): 0.78

Performance Analysis:
The model showed moderate performance. It predicted medical charges based on basic linear relationships between all available variables. However, it might have oversimplified the complexities present in the data, leading to inaccuracies.

Linear Regression with Nonlinear Relationship (Age as the Only Input):

Performance Metrics:

- MAE: 9189.48
- MSE: 136815004.01
- RMSE: 11696.79
- R2 score: 0.12

Performance Analysis:

Unfortunately, this model didn't perform well. It assumed a linear relationship between age and medical charges, which didn't accurately capture the true dynamics. Medical charges often don't increase linearly with age.

Model with Interaction Feature between Smoker and Obesity:

Performance Metrics:

- MAE: 4181.19
- MSE: 33596915.85
- RMSE: 5796.28
- R2 score: 0.78

Performance Analysis:
This model showed some improvement over the simple linear regression. By considering the interaction between smoking and obesity, it captured a bit more of the variability in charges.

Model with Interaction Effects and Polynomial Features:

Performance Metrics:

- MAE: 2729.49
- MSE: 20713528.93
- RMSE: 4551.21
- R2 score: 0.87

Performance Analysis:
This model performed the best among the ones tested. By incorporating interaction effects and polynomial features, it captured more of the complex relationships in the data, resulting in more accurate predictions.

Model Improvement Strategies

Simple Linear Regression with All Columns: Consider incorporating more complex features or interaction effects to capture the nuances in the data better.

Linear Regression with Nonlinear Relationship: Explore additional features or nonlinear transformations of existing features to capture the nonlinear relationships present in the data.

Model with Interaction Feature between Smoker and Obesity: Further investigate the interaction between smoker and obesity and consider additional features or transformations to improve model performance.

Model with Interaction Effects and Polynomial Features: Continue exploring interaction effects and polynomial features to capture the complex relationships and improve model accuracy further.

Conclusion

In this project, we effectively tackled the task of predicting medical charges for an insurance company using linear regression. We began by thoroughly exploring the dataset, calculating summary statistics, and visualizing relationships among features. Through analysis, we identified patterns such as the impact of smoking and obesity on charges.

Next, we trained multiple regression models, starting with a simple linear regression and progressively incorporating nonlinear relationships and interaction effects to improve predictive accuracy. Evaluation metrics such as MAE, MSE, RMSE, and R2 score were used to assess model performance.

Ultimately, we successfully achieved the task goals outlined in the project brief, demonstrating a systematic approach to data exploration, model development, and evaluation. Our documentation provides clear insights into the factors influencing medical charges and the steps taken to develop accurate predictive models for the insurance company.