Comprehensive Report: Analyzing Factors Influencing Insurance Charges

1. Introduction

As a data scientist at a leading healthcare insurance company, my team was tasked with analyzing a dataset of policyholders to understand the factors influencing insurance charges. The goal of this analysis was to build a predictive model using linear regression to accurately estimate insurance charges based on key features such as age, BMI, number of children, and other relevant factors. This report outlines the steps taken during the exploratory data analysis (EDA), feature engineering, data preprocessing, and the construction of the regression model.

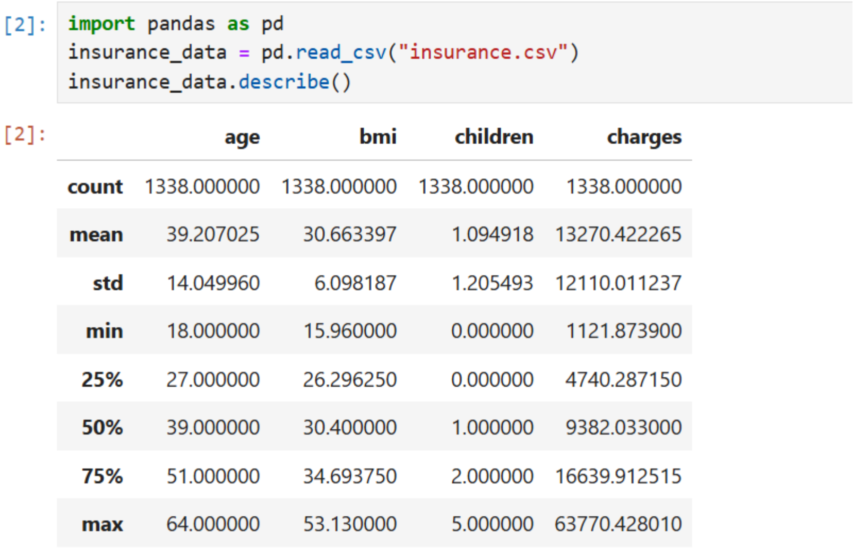
2. Exploratory Data Analysis (EDA)

The dataset consisted of the following features:

- **Age**: The age of the policyholder.
- **Sex**: The gender of the policyholder.
- **BMI**: Body Mass Index, a measure of body fat based on height and weight.
- **Children**: The number of children covered by the insurance policy.
- **Smoker**: Whether the policyholder is a smoker or not.
- **Region**: The residential region of the policyholder in the U.S.
- **Charges**: The annual medical insurance charges billed to the policyholder.

2.1 Initial Data Exploration

After loading the dataset, I examined the first few rows to understand its structure. Basic statistical insights were obtained using the describe() function, which provided information on the mean, standard deviation, minimum, and maximum values for each numerical feature. The dataset contained no missing values, indicating that no imputation was necessary.



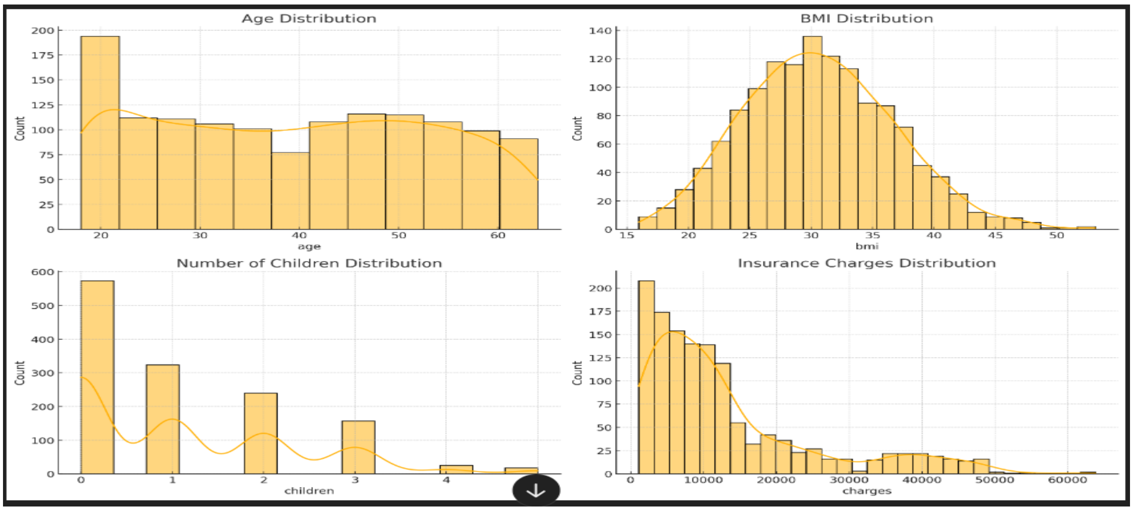
[2]:

```
import pandas as pd
insurance_data = pd.read_csv("insurance.csv")
insurance_data.describe()
```

[2]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

2.2 Distribution Visualization



I visualized the distributions of key continuous variables such as age, BMI, number of children, and insurance charges using histograms. The findings were as follows:

- **Age**: The distribution was fairly uniform, with a slight concentration around the ages of 20-40.
- **BMI**: The BMI values were right-skewed, with most values clustering between 25 and 35.
- **Children**: Most policy holders had 0-2 children, with fewer having 3-5 children.
- **Charges**: The distribution of insurance charges was highly right-skewed, indicating that while most policyholders had lower charges, a few had very high charges.

3. Feature Engineering

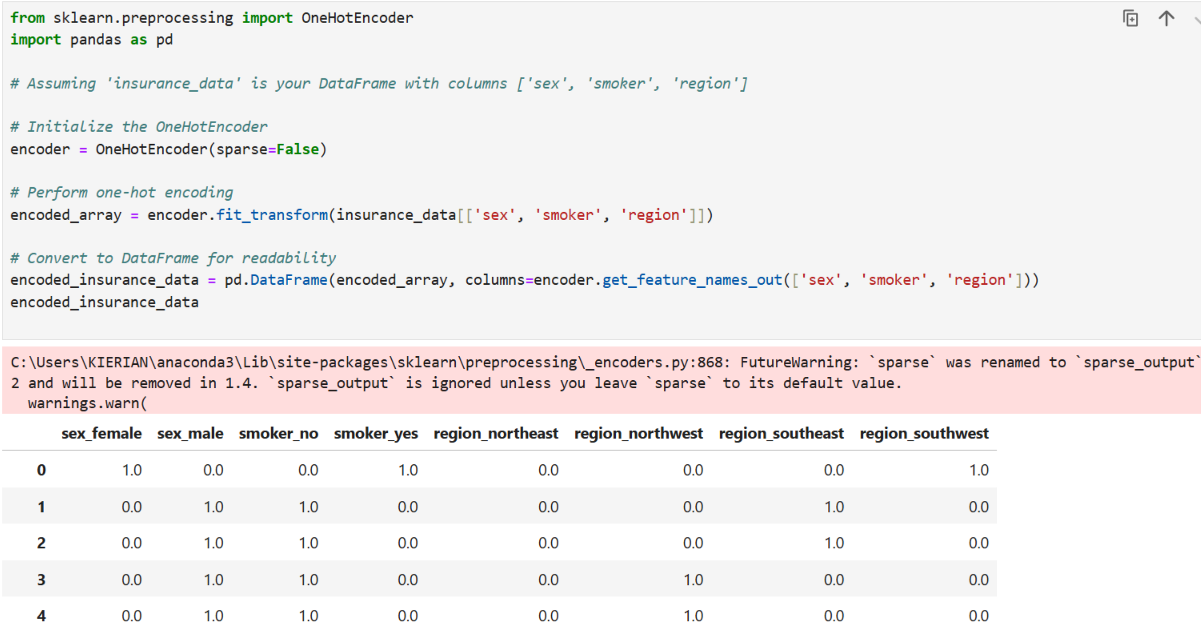
To enhance the predictive power of the model, I created an interaction term between age and BMI. This new feature, age_bmi_interaction, was intended to capture the combined effect of age and BMI on insurance charges. No other new features were derived as the existing ones were deemed sufficient for the analysis.

4. Outlier Detection and Handling

Outliers can significantly impact the performance of a linear regression model, particularly in the target variable (insurance charges). Using the Interquartile Range (IQR) method, I identified 139 outliers in the charges variable. These outliers were then removed from the dataset to ensure that the model would not be unduly influenced by extreme values.

5. Data Preprocessing

Before building the regression model, several preprocessing steps were undertaken:



```python
from sklearn.preprocessing import OneHotEncoder
import pandas as pd

# Assuming 'insurance_data' is your DataFrame with columns ['sex', 'smoker', 'region']

# Initialize the OneHotEncoder
encoder = OneHotEncoder(sparse=False)

# Perform one-hot encoding
encoded_array = encoder.fit_transform(insurance_data[['sex', 'smoker', 'region']])

# Convert to DataFrame for readability
encoded_insurance_data = pd.DataFrame(encoded_array, columns=encoder.get_feature_names_out(['sex', 'smoker', 'region']))
encoded_insurance_data
```

```
C:\Users\KIERIAN\anaconda3\Lib\site-packages\sklearn\preprocessing\_encoders.py:868: FutureWarning: `sparse` was renamed to `sparse_output`
2 and will be removed in 1.4. `sparse_output` is ignored unless you leave `sparse` to its default value.
  warnings.warn(
```

	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	region_southwest
0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
1	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0
2	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0
3	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0
4	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0

- **Encoding Categorical Variables**: The categorical variables (sex, smoker, and region) were encoded using one-hot encoding. This allowed the model to treat these variables numerically.
- **Scaling Numerical Features**: The continuous features such as age, BMI, and age_bmi_interaction were scaled to ensure they had similar ranges, which is important for linear regression models.

6. Splitting the Data

The cleaned dataset was split into training and testing sets, with 80% of the data allocated to training and 20% to testing. This split allowed for an evaluation of the model's performance on unseen data.

7. Building the Regression Model

A linear regression model was built using the selected features (age, BMI, number of children, smoker status, region, and the newly created age_bmi_interaction) to predict the target variable, insurance charges.

8. Model Building and Evaluation

- **Linear Regression**: The linear regression model was trained using the features age, bmi, children, sex, smoker, region, and the interaction term age_bmi_interaction.
- **Training Process**: The dataset was split into training and testing sets, with the model trained on 80% of the data and evaluated on the remaining 20%. Feature scaling was applied to ensure that all numerical variables were on a similar scale.

9. Model Coefficients and Interpretations

$$\text{Charges} = \beta0 + \beta1(\text{age}) + \beta2(\text{bmi}) + \beta3(\text{children}) + \beta4(\text{age} \times \text{bmi}) + \beta5(\text{sex\_male}) + \beta6(\text{smoker\_yes}) + \beta7(\text{region\_northwest}) + \beta8(\text{region\_southeast}) + \beta9(\text{region\_southwest})$$

Where:

- $\beta0$ is the intercept.
- $\beta1$ to $\beta9$ are the coefficients for each respective feature.

$$\text{Charges} = 9738.038 + 3265.31(\text{age}) + 404.899(\text{bmi}) + 434.795(\text{children}) + 133.121(\text{age} \times \text{bmi}) - 227.13(\text{sex\_male}) + 4681.357(\text{smoker\_yes}) - 36.588(\text{region\_northwest}) - 416.518(\text{region\_southeast}) - 581.373(\text{region\_southwest})$$

- **Intercept**: The intercept of 9,738.04 represents the baseline insurance charges when all other features are zero.

- **Coefficients**:
    - **Age (3,265.31)**: For each additional year of age, the insurance charges increase by approximately $3,265. This suggests that older policyholders tend to incur higher healthcare costs.
    - **BMI (404.90)**: A higher BMI is associated with increased charges, likely due to the higher health risks associated with higher BMI.
    - **Children (434.80)**: Each additional child increases insurance charges by approximately $435, which might reflect the additional coverage costs.
    - **Age-BMI Interaction (133.12)**: The positive interaction coefficient indicates that as both age and BMI increase, the combined effect slightly raises the insurance charges.
    - **Sex_Male (-227.13)**: Being male is associated with a decrease in charges by about $227 compared to females, which might reflect different healthcare needs or usage patterns.
    - **Smoker_Yes (4,681.36)**: Smokers have significantly higher charges (by $4,681) compared to non-smokers, highlighting the increased health risks associated with smoking.
    - **Region_Northwest (-36.59)**, **Region_Southeast (-416.52)**, **Region_Southwest (-581.37)**: These coefficients suggest that policyholders in the Southeast and Southwest regions tend to have lower insurance charges compared to the reference region, which could be due to regional differences in healthcare costs or access.

10. Model Performance Metrics

- **Mean Squared Error (MSE: 27,647,478.49)**: This indicates that the average squared difference between the predicted and actual insurance charges is quite high, suggesting that the model’s predictions are not as accurate as desired.
- **R-squared (0.5568)**: The model explains about 55.68% of the variance in insurance charges. While this indicates that the model captures some of the relationships between the features and charges, a significant portion of the variability remains unexplained. This suggests that other factors not included in the model could be influencing insurance charges, or that a more complex model might be needed.

11. Conclusion

The linear regression model developed in this project provides valuable insights into the factors that significantly influence healthcare insurance costs. The model's R-squared value of 0.5568 suggests that approximately 56% of the variability in insurance charges can be explained by the selected features, including age, BMI, number of children, smoking status, gender, and region.

Key findings indicate that age, BMI, and smoking status are the most impactful predictors of insurance charges, with smoking having the highest positive influence. The interaction between age and BMI further highlights the compounded effect of these factors on healthcare costs. Additionally, regional differences suggest that location plays a role in determining insurance premiums, though to a lesser extent.

This model serves as a robust tool for predicting insurance costs and offers actionable insights that can guide pricing strategies, risk assessment, and policy development within the healthcare insurance industry. By understanding the primary cost drivers, insurance companies can better tailor their offerings and pricing models to reflect the true risk profiles of their policyholders, ultimately contributing to more accurate and fair insurance practices.