Insurance Charges Prediction Report

1 Introduction

This report details the analysis and modeling of an insurance dataset aimed at predicting insurance charges based on various features such as age, sex, BMI, number of children, smoking status, and region. The dataset contains records of individuals with attributes that could influence their insurance costs.

2 Data Overview

2.1 Dataset Description

The dataset comprises several features, including:

- **age**: Age of the policyholder
- **sex**: Gender of the policyholder
- **bmi**: Body Mass Index
- **children**: Number of children/dependents covered by the policy
- **smoker**: Smoking status
- **region**: Geographical region
- **charges**: The insurance charges (target variable)

2.2 Data Exploration

- Shape: The dataset consists of n rows and m columns.
- Data Types: Each feature's data type was assessed, and categorical features were converted for analysis.
- Missing Values: No missing values were found in the dataset.

3 Exploratory Data Analysis (EDA)

3.1 Distribution of Charges

The distribution of insurance charges was visualized using a density plot, exhibiting a right-skewed distribution. A log transformation of charges was applied to normalize the distribution.

3.2 Regional Charges

The total charges were aggregated by region, revealing differences in insurance costs across different areas. Bar plots highlighted variations in charges based on sex, smoking status, and number of children.

3.3 Relationships

Linear models indicated relationships between charges and numerical variables, differentiated by smoking status. Violin plots illustrated the impact of smoking on charges, showing that smokers tend to have higher charges.

4 Modeling

4.1 Data Preparation

Categorical features were converted into numerical formats using label encoding.

4.2 Regression Models

4.2.1 Linear Regression

A linear regression model was fitted to the training data, yielding an intercept and coefficients indicating the relationship strength between features and charges. The model achieved an R^2 score on the test set.

4.2.2 Ridge and Lasso Regression

Ridge regression added a penalty term to mitigate overfitting, while Lasso regression demonstrated effective feature selection.

4.2.3 Random Forest Regressor

A Random Forest model was employed, providing robust predictions and insights into feature importance.

4.2.4 Polynomial Regression

A polynomial regression model was utilized to capture non-linear relationships, improving predictive performance.

5 Model Evaluation

The evaluation of the models was conducted using metrics:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

5.1 Performance Metrics

For the polynomial regression model:

- MAE: $x.xx$
- MSE: $x.xx$
- RMSE: $x.xx$

6 Conclusions

6.1 Key Insights

Insurance charges are influenced significantly by factors such as smoking status, age, and BMI. Regions exhibit distinct differences in average charges.

6.2 Model Effectiveness

The Random Forest model outperformed linear models, highlighting the importance of non-linear relationships.

6.3 Recommendations

Further studies could explore interactions between features or implement advanced machine learning techniques.

7 Future Work

Data enrichment and model tuning could yield better performance.