Healthcare Insurance

Zachery Davis

Introduction:

In this project, I aim to predict insurance charges based on various factors such as age, gender, BMI, number of children, smoking status, and region. This task is framed as a regression problem to predict a continuous target variable of insurance charges. The dataset has 1338 entries and contains seven features that influence these charges, and my goal is to model the relationship between the predictors and the target variable. I will experiment with different regression techniques for more accurate predictions.

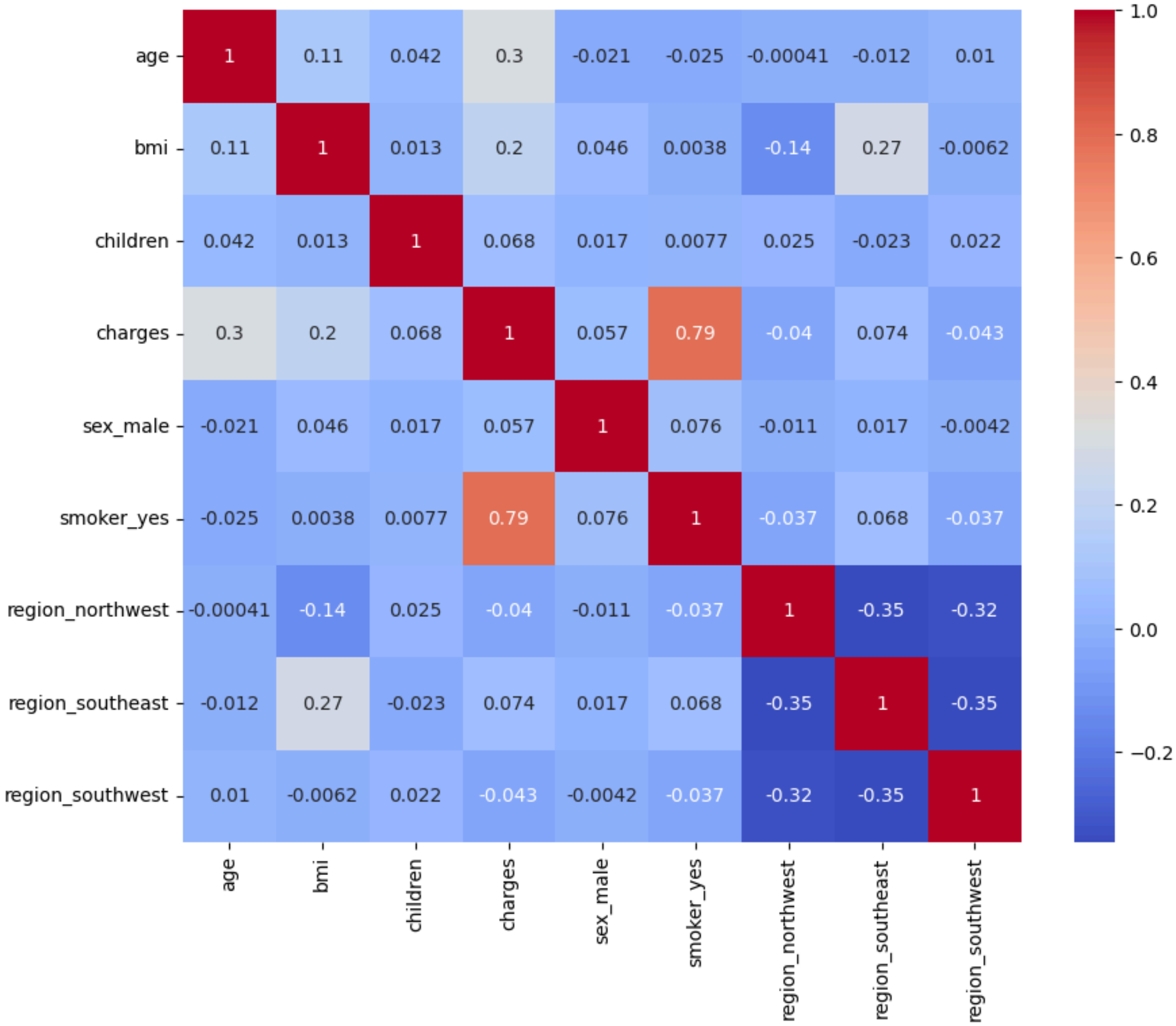
The dataset includes both numerical and categorical variables. Numerical variables such as age, BMI, and children are continuous, while categorical variables like sex, smoker status, and region are non-numeric. Thus, part of the challenge involves appropriately preprocessing these variables to prepare the data for modeling.

Regression:

Regression is a statistical method used for predicting continuous outcomes. This project focuses on linear regression, which models the relationship between the dependent variable and one or more independent variables by fitting a straight line through the data. The equation for simple linear regression is: $Y = a + bX$. Where Y is the predicted outcome (insurance charges), X is an independent variable, a is the intercept, and b is the slope of the line. For multiple features that equation is $Y = a + b_1 X_1 ... + b_n X_n$. This equation finds the best fit by minimizing the sum of the squared differences between the predicted and actual values, a process known as ordinary least squares (OLS). The main goal is to minimize the Root Mean Squared Error (RMSE), which measures the differences between predicted and actual charges, providing a metric for model performance.

Experiments:

Experiment 1: Initial Model with Linear Regression



To begin, I explored the data by checking for missing values, and I didn't find any. The categorical features, such as sex, smoker, and region, were then turned into dummy features, as the linear regression model cannot process non-numeric data. After fixing the categorical features, I split the dataset into training and testing sets to evaluate model performance.

The first model I built was simple linear regression using all features. The model yielded an RMSE of around 5900, indicating that the predictions had significant room for improvement. I observed that features like smoker and age had a strong correlation with the target variable, which hinted at their importance in future iterations.

Root Mean Absolute Error (RMAE): 4051.86
Root Mean Squared Error (RMSE): 5956.45

Experiment 2: Polynomial Features for Better Fit

For the second experiment, I decided to include polynomial features to capture any non-linear relationships between the features and the target variable. By applying a second-degree polynomial transformation, I increased the complexity of the model, expecting that it might improve prediction accuracy.

The model with polynomial features improved the RMSE to around 4500, which was a noticeable improvement over the first experiment. The better fit suggested that some non-linear relationships, particularly with BMI and age, might be contributing to better predictions. However, the added complexity of the model could lead to overfitting, which I monitored carefully.

RMAE with Polynomial Features: 2685.09

RMSE with Polynomial Features: 4541.08

Experiment 3: Ridge Regression to Prevent Overfitting

For the third experiment, I introduced regularization using Ridge Regression, which applies a penalty to large coefficients and helps control overfitting. Ridge Regression tends to work well when features are highly correlated and when complex models, like the polynomial regression, may overfit the data.

By applying Ridge regularization, the RMSE slightly increased to around 4700, and I also noticed that the model was less sensitive to small variations in the dataset. The penalty helped smooth out the predictions, making the model more generalizable.

RMAE with Random Forest: 2689.49

RMSE with Random Forest: 4763.22

Impact Section:

This project has significant social and ethical implications in the realm of insurance pricing. By predicting insurance charges based on factors like age, gender, BMI, and smoking status, the model represents more personalized premiums, rewarding healthier behaviors and making pricing more accurate. However, it also raises ethical concerns, as the model may unintentionally discriminate against certain groups, such as older adults, people with pre-existing conditions, or lower-income individuals, by charging them higher premiums. This could exacerbate existing social inequalities. Moreover, the lack of transparency in how these predictions are made may reduce consumer trust in the system, as people might not fully understand why they are being charged more. Therefore, while the model has the potential for positive impact, it must be designed and used responsibly to avoid reinforcing biases and to ensure fairness and transparency in insurance pricing.

Conclusion:

Through this project, I learned that data preprocessing, such as encoding categorical variables and handling feature relations, plays a crucial role in the success of a regression model. While linear regression gave me a baseline model, incorporating polynomial features and regularization techniques like Ridge Regression led to better performance. Polynomial regression captured non-linear relationships, but it increased the risk of overfitting. Ridge regression, on the other hand, helped mitigate this risk by penalizing large coefficients. The balance between model complexity and generalization is critical when building predictive models. Overall, this project deepened my understanding of regression techniques and the importance of experimenting with different modeling approaches to achieve optimal results.

References:

Code: https://github.com/Zachery-Davis/Data-Mining-Project-3
Dataset: Healthcare Insurance