Regression Project Report

Problem Description

In this project, I aimed to predict the health insurance charges based on a person's age, body mass index (BMI) and smoking status. I chose this problem as I am always interested in health related issues and it provides a chance to work with both numerical and categorical data, to handle missing data and to compare models to improve predictions.

Dataset

The dataset comes from Healthcare insurance available on Kaggle (link: https://www.kaggle.com/code/jayrdixit/healthcare-insurance). It contains 1284 records of individuals, including six features like age, sex, bmi, children, smoker, region and one target (charges).

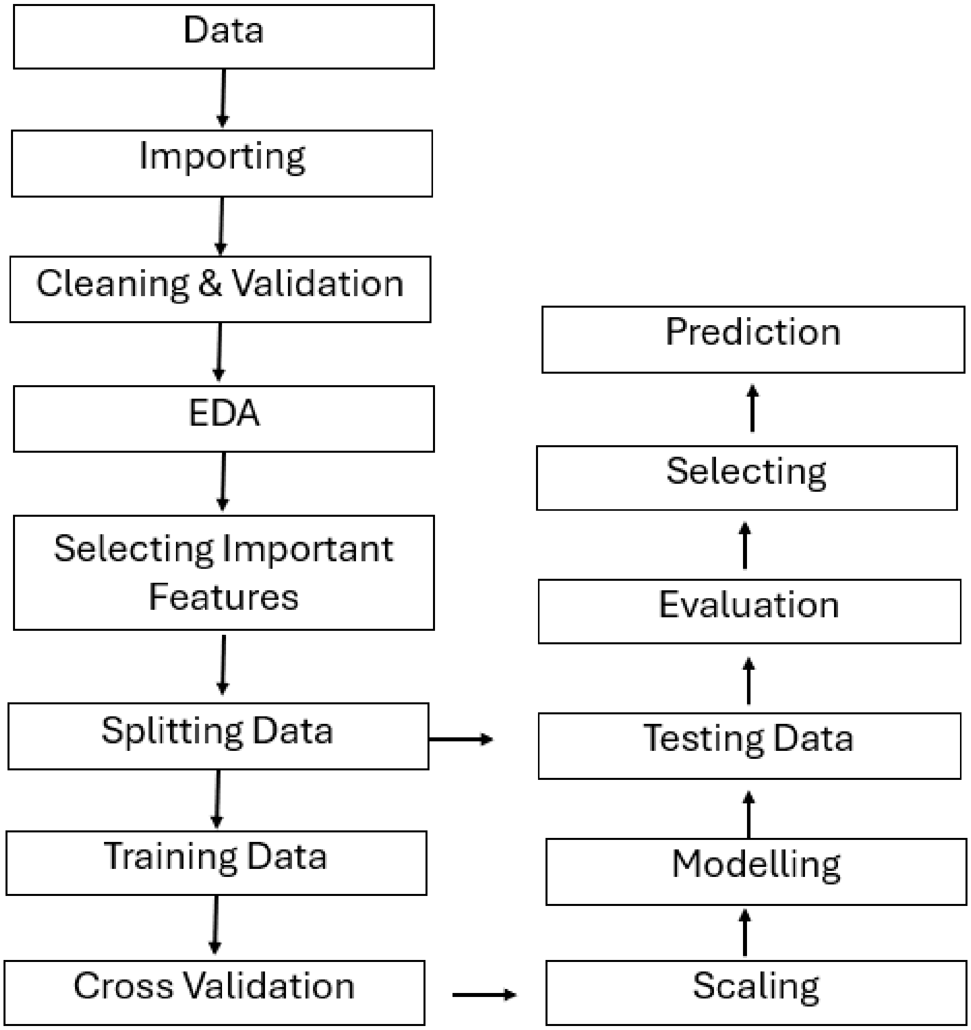


Figure 1: Workflow

Modelling Approach

In this project, I used linear regression, Ridge regression, Lasso regression and Poly regression to predict the insurance charge. The workflow is shown in Figure 1. During preprocessing, the dataset missing data but the amount is less than 5% of dataset. So, the missing data were dropped. Then, age and children features showed negative values and these values are not possible, there values were dropped again.

After cleaning and validation of data, there is only three features that can potentially affect insurance charges(Figure 2). So, I selected three key features:

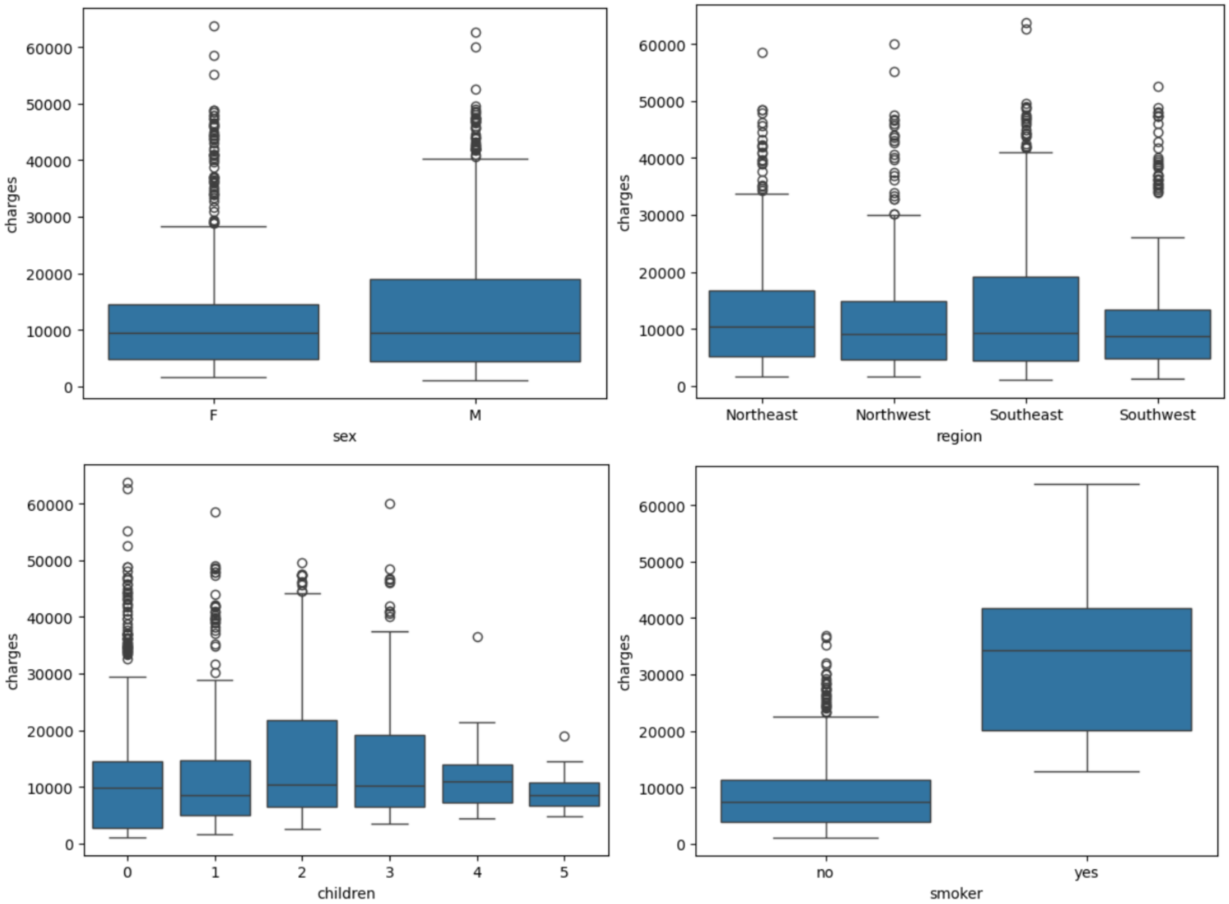
age : an individual's age (integer)

bmi: a measure of ratio of height and weight(floating)

smoker: a variable that indicates if the person smokes(category)

After selecting target and features, I first converted the categorical variable smoker into a binary variable (smoker_yes: 1 for "yes", 0 for "no"). Then I applied StandardScaler to normalize the numerical features (age and bmi). After scaling, I used PolynomialFeatures to expand the input features, allowing the model to capture nonlinear relationships

Dataset was split into training(70%) and testing(30%). Then, each model was trained on training data using pipeline that do scaling and fitting. Targe values were predicted using testing data and performance metric such as r squared, mean absolute error and mean standard error were also calculated using predicted target values.



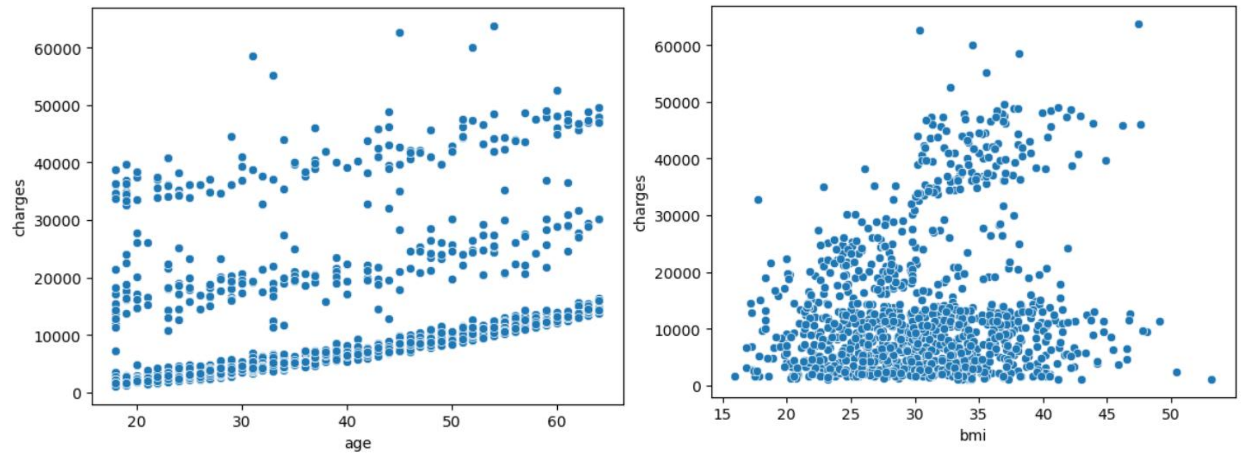


Figure 2: Effect of Features on Target

Evaluation

To evaluate model performance, I used the following metrics:

R^2 (coefficient of determination): measures how well the model explains variance in the target.

MSE (mean squared error): penalizes larger errors.

MAE (mean absolute error): gives the average magnitude of errors.

The final Polynomial Regression model (degree= 2) achieved:

$R^2 = 0.82$

MSE = 2.44e7

MAE = 2916

These scores were better than all the other models tested, indicating that Polynomial Regression captured the complexity of the relationship between features and insurance charges.

Reflection

During preprocessing, when I tried to deal with negative value of children and age, I reversed the negative value to positive value and I tried to test the models and the performance seemed not good. So, I finally dropped it.

The other challenge I faced is rename category in sex, region columns. The data in the dataset is messy and I want to collapse these category columns to simple and unified columns. It tried .cat.rename_categories and .replace method. Sometimes, the first method work but it run into error so I used .replace method. I didn't know the reason.

I also learned the importance of feature selection — reducing the input to just three meaningful variables improved interpretability without sacrificing model accuracy. Moreover, I understood how regularization methods like Ridge and Lasso perform similarly to Linear Regression when the input features are already informative and the dataset is clean.

Overall, this project strengthened my understanding of the regression modeling workflow, preprocessing pipelines, and the trade-offs between model complexity and interpretability.

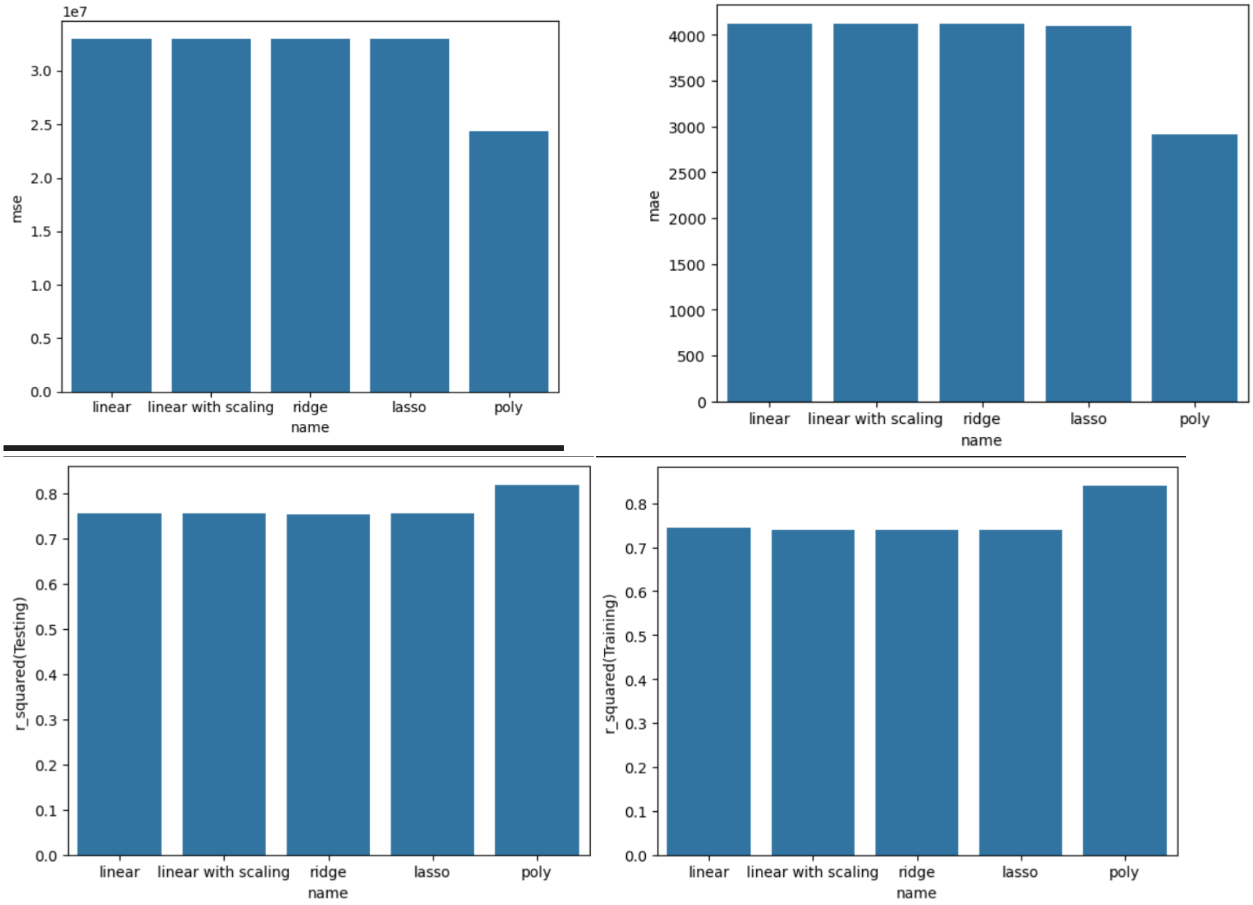


Figure 3 Performance Metrics of Different Models