HEALTH INSURANCE PREMIUM PREDICTION

by

Hemashruthi Ganesh
Saishrawan Vutharkar
Siddhanth Reddy

May 03, 2023

TEXAS A&M UNIVERSITY
College Station, Texas 77843-3135

DISCLAIMER

The following project is the original work of the undersigned, completed as a requirement for STAT654 – Statistical Computing with R and Python course at Texas A&M University, College station. The information and data presented in this project are based on our own research and analysis. The opinions and conclusions presented in this project are solely ours and we have made our best efforts to ensure the accuracy and reliability of the information presented, but we cannot guarantee that it is entirely error-free.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Dr. Sharmistha Guha for providing us with the opportunity to conduct this project and her invaluable guidance and support during this semester. Her insights and feedback have been instrumental in shaping our understanding of the subject matter and improving our quality of work.

TABLE OF CONTENTS

1. OBJECTIVE

The main objective of this project is to develop a health insurance price prediction model that can accurately predict the annual medical expenses of a new customer based on their demographic and lifestyle factors. We aim to achieve this by analyzing the existing customers' data, including their age, smoking habits, body mass index (BMI), and other demographic factors, and their corresponding medical expenses.

We intend to use a variety of data analysis techniques, including exploratory data analysis, feature engineering, and model building, to develop a comprehensive and accurate insurance price prediction model. By developing an accurate insurance price prediction model, this project can help health insurance companies better manage their risk and improve their profitability while providing their customers with affordable and fair insurance pricing.

The key questions we aim to answer through this project are:

1) What are the key demographic and lifestyle factors that most affect a customer's annual medical expenses?

2) What statistical analysis techniques and machine learning algorithms are best suited for developing an insurance price prediction model?

3) How can we improve the accuracy and performance of the insurance price prediction model and ensure that it is robust and reliable?

2. BACKGROUND AND MOTIVATION

Health insurance companies operate in a highly competitive market, and their profitability depends on their ability to accurately predict the medical expenses of their customers as it is essential for formulating pricing policies that ensure both the affordability of the insurance product and the profitability of the insurance company.

Insurance companies use various factors to determine the premium rates for their policies. These factors include age, gender, occupation, lifestyle habits, pre-existing medical conditions, and location. Insurance companies use these factors to assess the level of risk associated with each customer and charge premiums accordingly. Certain medical conditions are more prevalent among specific segments of the population. For example, lung cancer is more common among smokers, and heart disease is more common among people who are overweight. The companies charge higher premiums for older people as they are more likely to require medical care. Similarly, tobacco users are charged up to 50% more due to their higher risk of developing smoking-related diseases. Moreover, location is a significant factor that affects medical expenses as the cost of living, state government policies, and the availability of healthcare facilities can vary significantly across different regions.

To address these challenges, this project aims to develop an insurance price prediction model using machine learning algorithms like regression that can accurately predict the medical expenses of a new customer based on their demographic, lifestyle, and physical parameters.

3. DATASET DESCRIPTION

The dataset used in this project contains 1338 rows and 7 columns.

The independent or predictor variables in the dataset are:

1. Age: This variable represents the age of the primary beneficiary and is expressed in years.
2. Sex: This variable represents the gender of the beneficiary and can take two values: "male" or "female".
3. BMI: Body mass index is a measure of body fat based on height and weight. It is calculated as the weight in kilograms divided by the square of the height in meters.
4. Children: This variable represents the number of children covered by the health insurance policy or the number of dependents.
5. Smoker: This variable represents whether the beneficiary is a smoker or not and can take two values: "yes" or "no".
6. Region: This variable represents the beneficiary's residential area in the United States and can take one of four values: northeast, southeast, southwest, or northwest.

The dependent or target variable in the dataset is:

> Charges: This variable represents the annual medical expenses of the individual and is expressed in US dollars.

4. METHODOLOGY

The methodology section of this project report outlines the steps taken to collect, clean, preprocess, and analyze the dataset in order to develop an accurate health insurance price prediction model.

4.1 Data Collection

The dataset used in this project was obtained from an open data repository. The dataset contains information on 1338 individuals and their annual medical expenses along with various demographic and lifestyle factors.

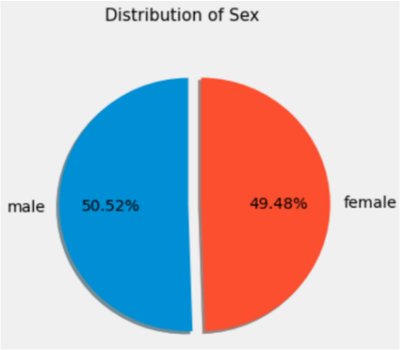
Link to the dataset: https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset

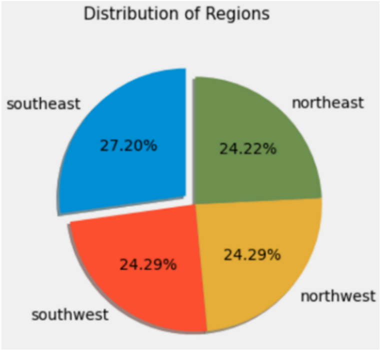
4.2 Data Cleaning

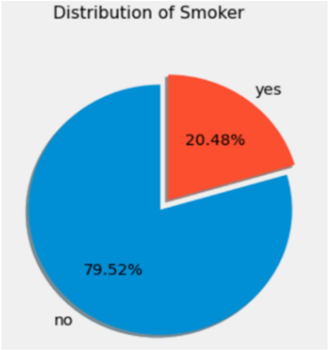
Before analyzing the dataset, we performed data cleaning to ensure that the data was free from errors and inconsistencies. This process involved removing duplicates, handling missing data, and correcting data entry errors.

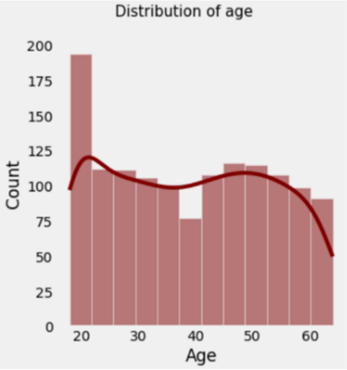
4.3 Univariate Analysis

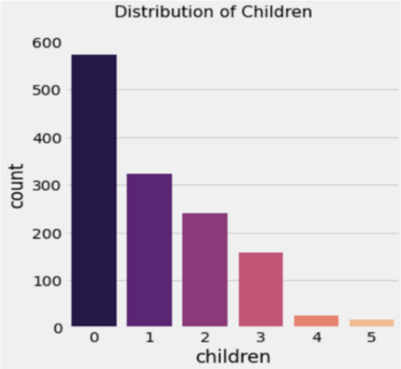
We performed univariate analysis on each independent and target variable in the dataset to understand the distribution of the data and identify any outliers or extreme values. This analysis helped us to identify potential issues with the dataset and to determine whether any transformations or scaling would be required before model building.

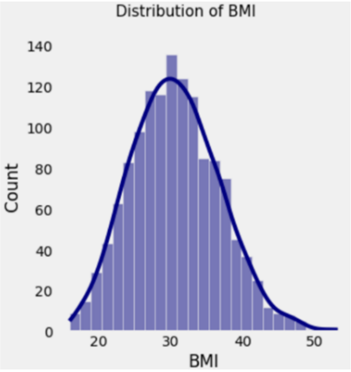


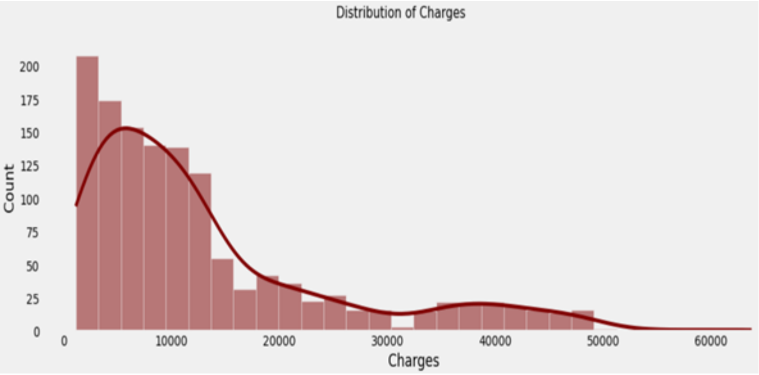












From the univariate analysis, we can deduce the following:

- The variable 'sex' shows an almost equal distribution between the two genders.

- The variable 'Region' also exhibits an equal division among the four regions - northeast, southeast, southwest, and northwest.

- About 20% of the observations correspond to smokers, and the distribution is skewed towards non-smokers.

- The variable 'age' has an almost uniform distribution, except for the age of 20, which has an unusually high number of observations, indicating that it might contain all observations under the age of 20.

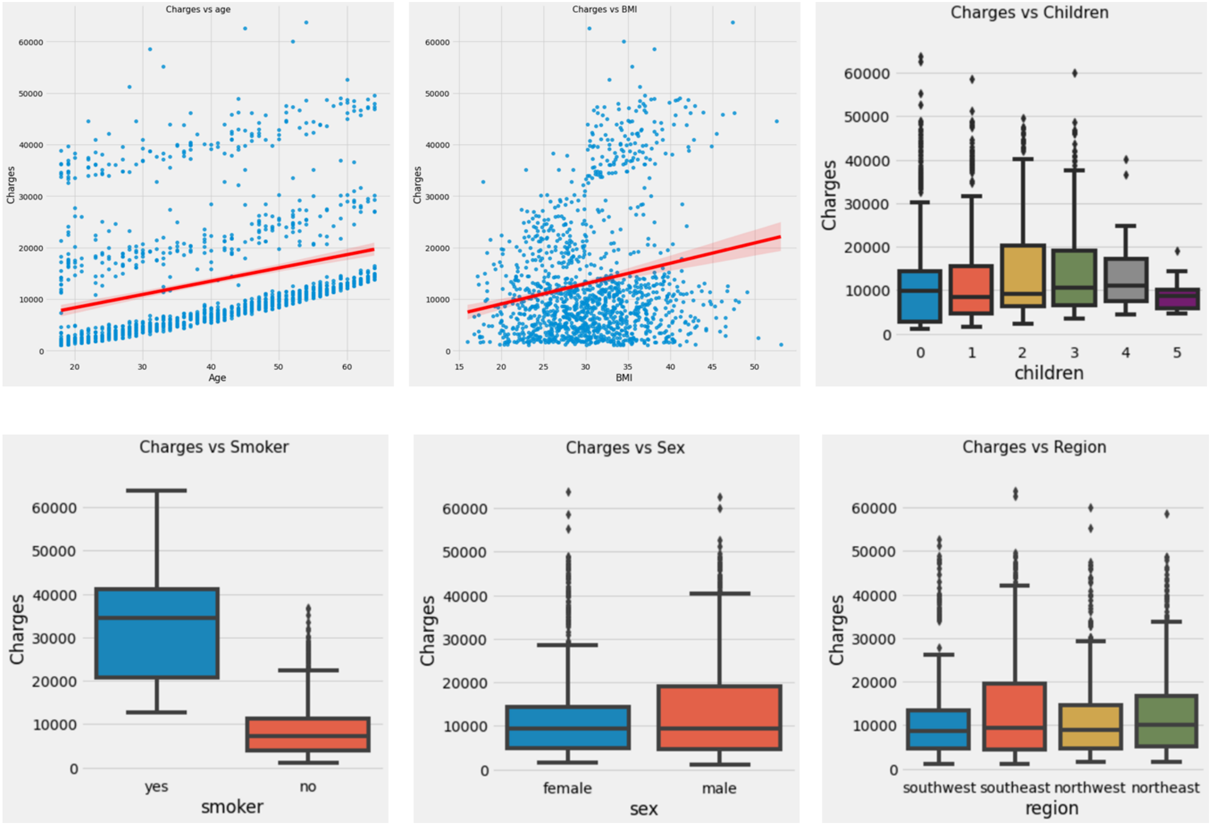
- The variable 'Children' indicates that most customers have 0-3 children. For simplicity, we can group the observations with 4 and 5 children under the 3 children category.

- The variable 'BMI' seems to follow a normal distribution, mostly centered around 30, indicating that most of the customers in the dataset are overweight or obese.

- Finally, the variable 'Charges' shows a right-skewed distribution, indicating that the majority of the observations have lower charges, while a small proportion of observations have very high charges.

4.4 Bivariate Analysis

Bivariate analysis was performed to examine the relationships between the independent variables and the dependent variable. This analysis helped us to identify any significant associations between the variables and to determine which variables were most strongly related to medical expenses.



Following are the observations made in our analysis:

- Age vs charges: We found a linear increasing relationship between age and medical charges, which is expected as medical expenses tend to increase as an individual gets older.

- BMI vs charges: We observed that higher values of medical expenses were associated with higher BMI, indicating the presence of medical conditions related to obesity.

- Children vs charges: We observed an increasing trend in medical expenses as the number of dependents increased until 2, after which it declined. This could be due to the lesser data available for customers with more dependents.

- Smoker vs charges: The medical expenses of smokers were significantly higher than those of non-smokers.

- Sex vs charges: We found that male customers tend to have higher medical expenses than female customers.

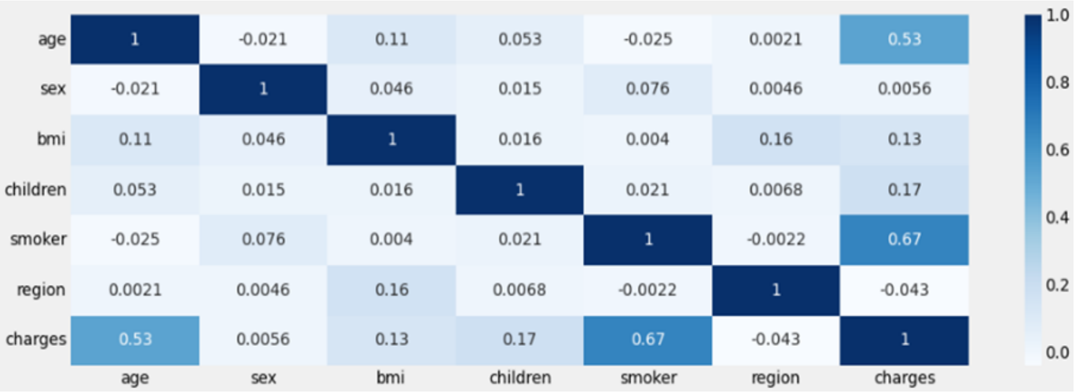
- Region vs charges: We observed that the southeast region had higher medical expenses than the other regions, but there was no significant difference among the other regions.

4.5 Correlation and Multicollinearity

Correlation Plot:

Correlation analysis was performed to examine the strength and direction of the relationships between pairs of independent variables. A correlation plot helps in identifying the correlation or bivariate relationship between two independent variables.

From the correlation plot, we observed that age and smoker have the highest correlation with the target variable (charges). We did not observe any significant correlation between the other predictor variables.

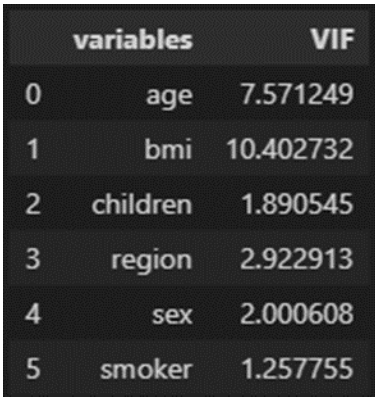


	age	sex	bmi	children	smoker	region	charges
age	1	-0.021	0.11	0.053	-0.025	0.0021	0.53
sex	-0.021	1	0.046	0.015	0.076	0.0046	0.0056
bmi	0.11	0.046	1	0.016	0.004	0.16	0.13
children	0.053	0.015	0.016	1	0.021	0.0068	0.17
smoker	-0.025	0.076	0.004	0.021	1	-0.0022	0.67
region	0.0021	0.0046	0.16	0.0068	-0.0022	1	-0.043
charges	0.53	0.0056	0.13	0.17	0.67	-0.043	1

Multicollinearity:

Multicollinearity analysis was performed to determine whether any of the independent variables were highly correlated with each other, which could lead to problems with model interpretation and accuracy.

We used VIF (Variation Inflation Factor) to check for multicollinearity among the independent variables. A VIF value of 1 indicates no multicollinearity, while a value greater than 1 indicates a correlation between the independent variable and other independent variables in the model. In general, a VIF value of less than 5 is considered acceptable, as values above 5 indicate high multicollinearity. In our analysis, we found that age and BMI have a VIF value above 5, indicating high multicollinearity between these variables.



	variables	VIF
0	age	7.571249
1	bmi	10.402732
2	children	1.890545
3	region	2.922913
4	sex	2.000608
5	smoker	1.257755

4.6 Data Preprocessing

After completing the data analysis, we preprocessed the data by performing feature engineering, transforming variables, and scaling the data as required. This step prepared the data for use in model building by ensuring that it was in the appropriate format.

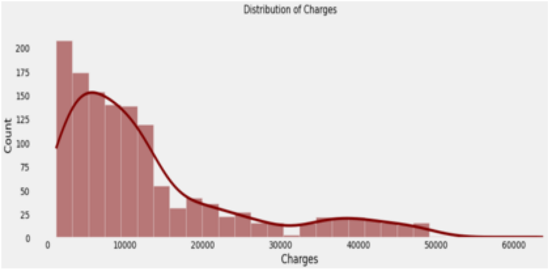
The following techniques were employed to prepare the data for model building:

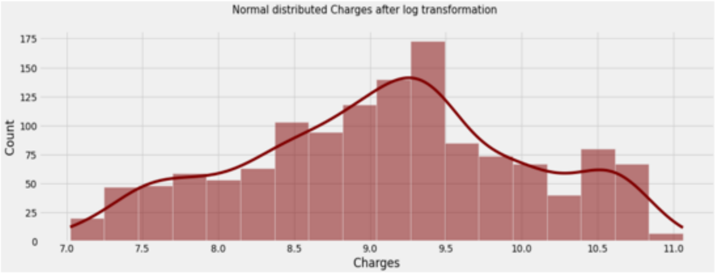
- Log transformation was used to transform the "charges" variable which had a highly skewed distribution into a more normal distribution. This is to ensure the residuals also have a normal distribution, a requirement for the assumptions of linear regression.

- Ordinal Encoder was used to transform the "region" variable which is a categorical variable into a numerical one by assigning each unique value in the dataset to a numerical value.

- One hot encoding was used to transform the "sex" and "smoker" variables into binary features. This data preprocessing technique creates a binary feature for each unique categorical value in the column.

- Standard Scaler was applied to all the features to normalize the dataset by transforming the features having different scales to have a mean of zero and a standard deviation of one.





Before Log transformation
Skewness: 1.515880
Kurtosis: 1.606299

After Log transformation
Skewness: -0.089817
Kurtosis: -0.636855

5. MODEL SELECTION

5.1 Linear Regression

Multiple linear regression models are of the form $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdot \cdot \cdot + \beta_k x_k + \epsilon$ where β_0 is the intercept and β_i is the coefficient associated with predictor xi. To obtain β_i we attempt to minimize the square loss defined as $f(b_0, b_1, . . . , b_k) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + \cdot \cdot \cdot + b_k x_{ik})]^2$ and solve for $b_0, b_1, . . . , b_k$.

5.2 Linear regression with feature selection

In multiple linear regression it is important to identify the best subset of predictors. This is useful because a smaller subset can be as good at prediction and simpler models are preferable. AIC and BIC can be used to compare models having different numbers of predictors. We can use the leaps package in R to check for the best subset of predictors for each subset size. We can fit a linear model for each of the subsets suggested and compare them using AIC and BIC to identify the best subset of predictors.

5.3 Polynomial Regression

Polynomial regression is a type of regression analysis used to model the relationship between a dependent variable and one or more independent variables by fitting a polynomial equation to the data. In polynomial regression, the relationship between the dependent variable and the independent variable(s) is represented by an nth degree polynomial equation, where n is the degree of the polynomial.

The general form of the polynomial regression equation is: $y = a + b_1x + b_2x^2 + ... + b_nx^n$ where y is the dependent variable, x is the independent variable, and a, b_1, b_2, ..., bn are coefficients that represent the relationship between the dependent variable and the independent variable(s).

Polynomial regression can be useful when the relationship between the dependent variable and the independent variable(s) is not linear but can be better approximated by a curved or nonlinear function. By using a polynomial function to model the data, polynomial regression can capture this nonlinearity and provide a better fit to the data.

5.4 Ridge Regression with hyperparameter tuning

Ridge regression is a regularization technique used to prevent overfitting in linear regression models. It adds a penalty term to the cost function that shrinks the coefficients towards zero, thereby reducing the complexity of the model. The strength of the penalty is controlled by a hyperparameter called alpha, which needs to be tuned to optimize the model performance. In ridge regression we attempt to minimize the loss function given by $f(b_0, b_1, . . . , b_k) = \sum_{i=1}^{n} [y_i - (b_0 + b_1x_{i1} + \cdot\cdot\cdot + b_kx_{ik})]^2 + \lambda \sum_{i=0}^{k} (b_i * b_i)$ and solve for $b_0, b_1, . . . , b_k$.

Ridge regression is particularly useful when dealing with multicollinearity, which is when two or more predictor variables are highly correlated with each other. In such cases, ordinary least squares (OLS) regression can give unstable and unreliable estimates of the coefficients, while Ridge regression can provide more stable and accurate estimates.

Ridge regression can be used in both simple and multiple linear regression models. The penalty term added by Ridge regression has a shrinkage effect on the coefficients, making them smaller in magnitude than the coefficients estimated by OLS. The higher the value of alpha, the greater the degree of shrinkage, and the simpler the model becomes.

Ridge regression assumes that the error terms are normally distributed with a mean of zero and constant variance. It also assumes that the predictor variables are standardized to have zero mean and unit variance.

Ridge regression can be extended to other types of regression models, such as polynomial regression and logistic regression, by adding a penalty term to the cost function.

5.5 Lasso Regression with hyperparameter tuning

Lasso regression is another regularization technique used in linear regression to avoid overfitting by penalizing large coefficients. In Lasso regression, the sum of the absolute values of the coefficients is added to the least squares objective function to be minimized. The degree of penalty is controlled by the hyperparameter alpha, which is selected through cross-validation.

Lasso regression is particularly useful when the number of features in the dataset is large and some of the features are not as relevant as others. Lasso regression can automatically perform feature selection by driving the coefficients of less important features to zero, effectively dropping them from the model.

In lasso regression we attempt to minimize the loss function given by $f(b_0, b_1, . . . , b_k) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + \cdot \cdot \cdot + b_k x_{ik})]^2 + \lambda \sum_{i=0}^{k} | b_i|$ and solve for $b_0, b_1, . . . , b_k$.

6. MODEL BUILDING AND EVALUATION

6.1 Linear models

We initially started with linear models as we observed some linear relationships from the exploratory data analysis. The model used is a multiple linear regression model with age, sex, bmi, children, smoker and region as the columns being used to predict expenses. We used the lm function in R and observed the below results:

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.07141 -0.19836 -0.04914  0.06601  2.16602 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            8.79672    0.03011 292.127  < 2e-16 ***
datad$smokeryes        1.55415    0.03027  51.336  < 2e-16 ***
datad$age              0.48577    0.01225  39.655  < 2e-16 ***
datad$bmi              0.08156    0.01278   6.382 2.41e-10 ***
datad$children         0.10183    0.01010  10.084  < 2e-16 ***
datad$sexmale         -0.07539    0.02440  -3.090 0.002043 ** 
datad$regionnorthwest -0.06377    0.03490  -1.827 0.067867 .  
datad$regionsoutheast -0.15715    0.03508  -4.480 8.09e-06 ***
datad$regionsouthwest -0.12892    0.03502  -3.681 0.000241 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4442 on 1329 degrees of freedom
Multiple R-squared:  0.768,	Adjusted R-squared:  0.7666
```

```
F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

We observe that when using a significance level of 0.05 all predictors except for region northwest are significant. The R-squared value of 0.768 indicates that our model is able to explain 76.8% of the variance in y using the linear model. The F-Statistic has a low p value and so we can conclude that the model is useful.

We fit the multiple regression model for each predictor subset suggested by leaps. We then note the R-squared, AIC and BIC values to compare the different models.

Model	R2	AIC	BIC
smoker	0.44	2794	2809
smoker + age	0.73	1779	1799
smoker + age + children	0.752	1686	1712
smoker + age + bmi+ children	0.762	1661	1692
smoker + age + bmi+ children + region	0.766	1644	1690
smoker + age + bmi+ children + sex + region	0.767	1636	1688

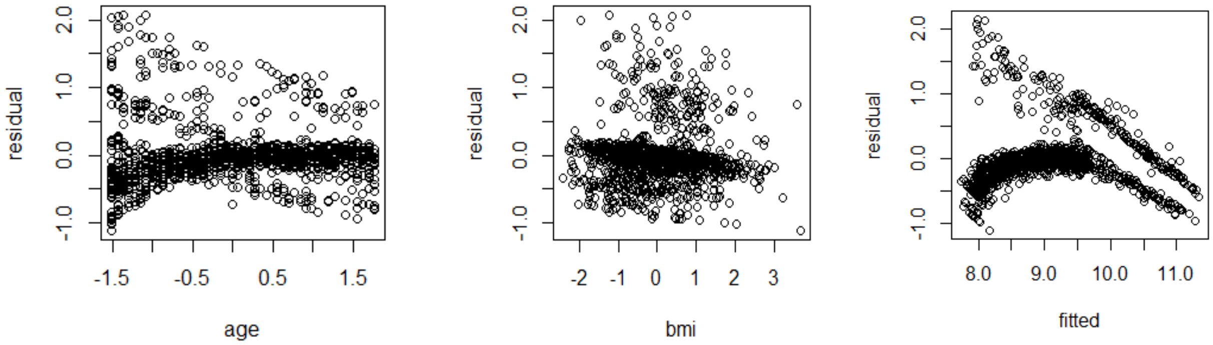
The model with features smoker + age + bmi+ children+ region seems to be the best model as it has significantly lower AIC and BIC compared to the other smaller models and the other larger model does not significantly further decrease AIC and BIC values. The model summary using the five predictors is shown below.

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.1026 -0.1971 -0.0521  0.0656  2.1506

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            8.76035    0.02781 315.051  < 2e-16 ***
datad$smokeryes        1.54712    0.03029  51.085  < 2e-16 ***
datad$age              0.48671    0.01229  39.617  < 2e-16 ***
datad$bmi              0.07971    0.01281   6.224 6.49e-10 ***
datad$children         0.10129    0.01013  10.000  < 2e-16 ***
datad$regionnorthwest -0.06333    0.03501  -1.809 0.070719 .
datad$regionsoutheast -0.15677    0.03519  -4.455 9.08e-06 ***
datad$regionsouthwest -0.12853    0.03513  -3.658 0.000264 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.4456 on 1330 degrees of freedom
Multiple R-squared:  0.7663,   Adjusted R-squared:  0.7651
F-statistic:   623 on 7 and 1330 DF,  p-value: < 2.2e-16
```

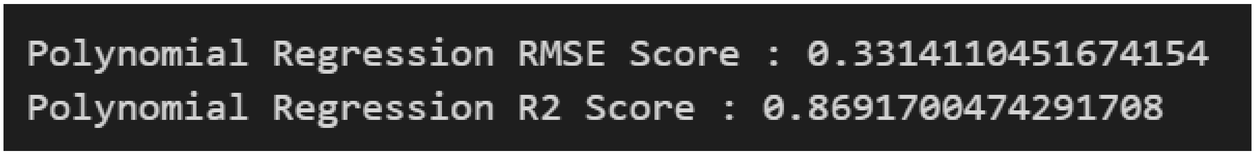
We then plot residual plots for our selected model using the five predictors to check for assumptions in the residuals.We can see from the plots of residuals against age and bmi that though there does not seem to be any non linear trend, there are a few points which have abnormally high positive residuals.



The plot of residuals against the fitted values however suggests that there is some kind of non-linear relationship as the residuals are not randomly distributed. Polynomial and other non-linear models could be better models for our data.

6.2 Polynomial Regression

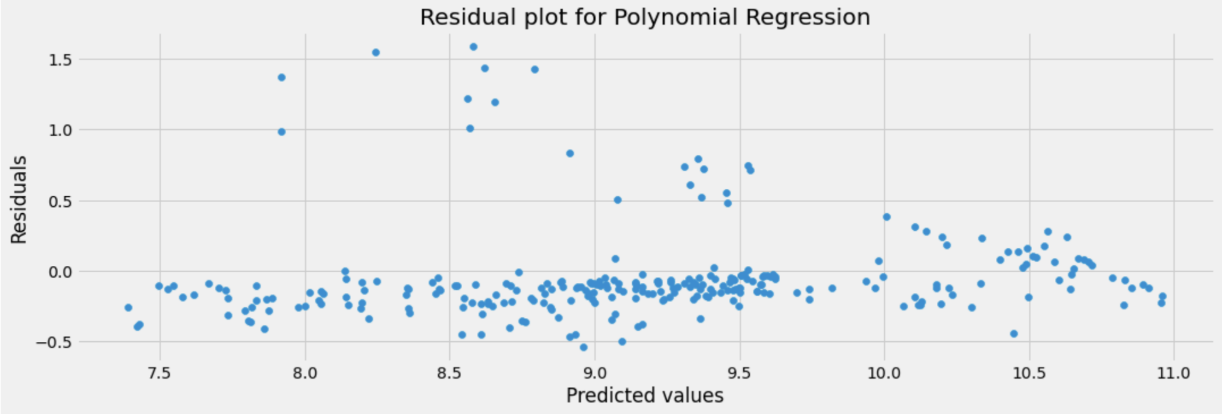
In our code, the PolynomialFeatures class from sklearn.preprocessing is used to create polynomial features up to the third degree. These polynomial features are then used to train a linear regression model using the LinearRegression class from sklearn.linear_model. Finally, the model is evaluated using the RMSE and R2 score.



```
Polynomial Regression RMSE Score : 0.3314110451674154
Polynomial Regression R2 Score : 0.8691700474291708
```

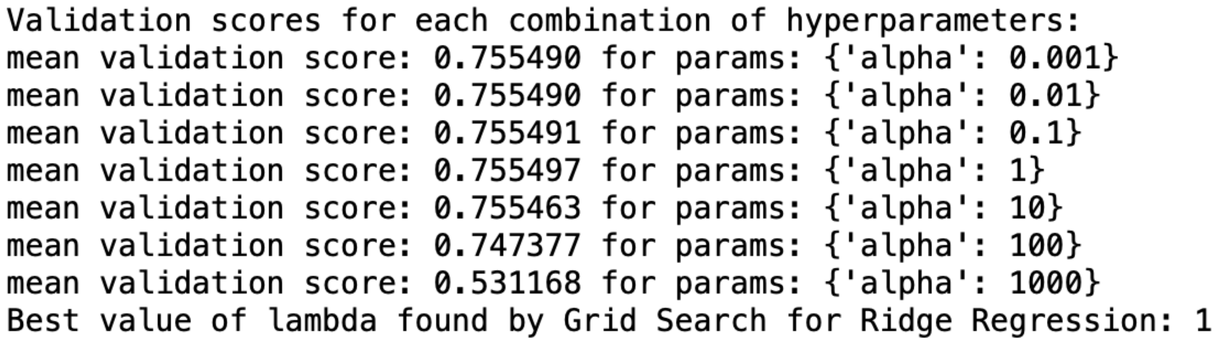
The residuals are then plotted to visualize the performance of the model. If the residuals are randomly scattered around zero, it indicates that the model is performing well. If there is a pattern or trend in the residuals, it suggests that the model is not capturing some important information in the data or that the model assumptions are not met.

In our residual plot, we see that there is not any definite pattern and they are randomly scattered around zero, which suggests that the model is capturing some important information in the data.



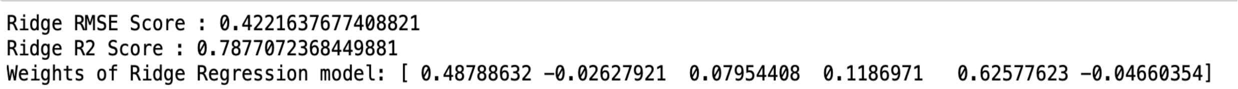
6.3 Ridge Regression with hyperparameter tuning

In the code, we perform a Grid Search cross-validation to select the optimal value of alpha, and then fit a Ridge regression model with that alpha value.



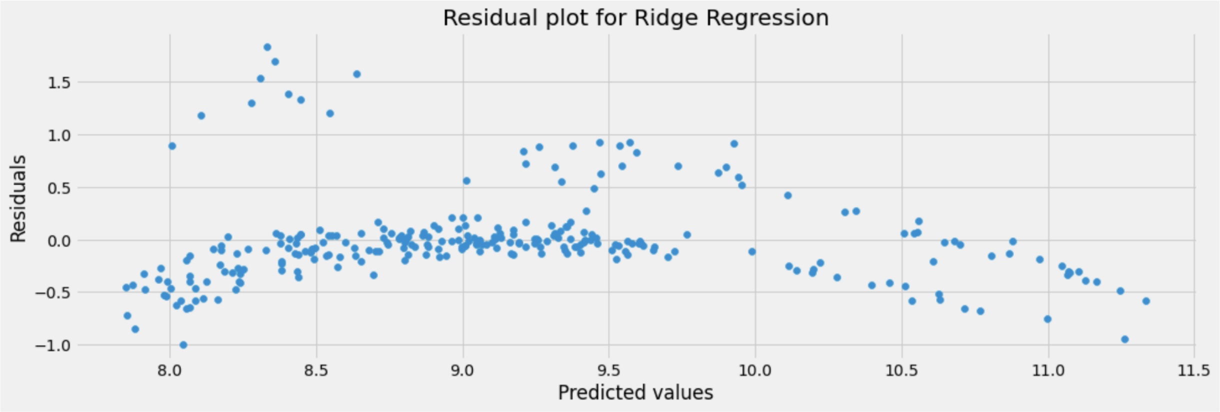
```
Validation scores for each combination of hyperparameters:
mean validation score: 0.755490 for params: {'alpha': 0.001}
mean validation score: 0.755490 for params: {'alpha': 0.01}
mean validation score: 0.755491 for params: {'alpha': 0.1}
mean validation score: 0.755497 for params: {'alpha': 1}
mean validation score: 0.755463 for params: {'alpha': 10}
mean validation score: 0.747377 for params: {'alpha': 100}
mean validation score: 0.531168 for params: {'alpha': 1000}
Best value of lambda found by Grid Search for Ridge Regression: 1
```

We also evaluate the performance of the model using the root mean squared error (RMSE) and the R-squared (R2) score.



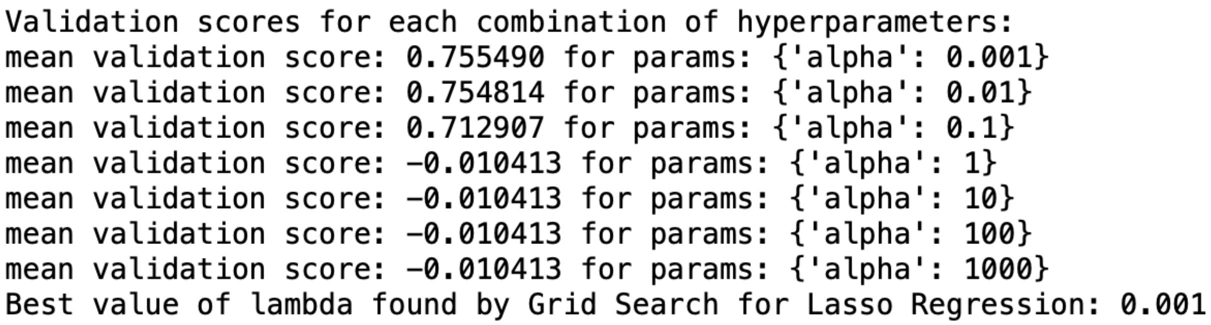
```
Ridge RMSE Score : 0.4221637677408821
Ridge R2 Score : 0.7877072368449881
Weights of Ridge Regression model: [ 0.48788632 -0.02627921  0.07954408  0.1186971   0.62577623 -0.04660354]
```

Finally, we plot the residuals to check the assumptions of the model. We see that the residuals are randomly scattered around zero, which suggests that the model is capturing some important information in the data but there is some non-linearity.



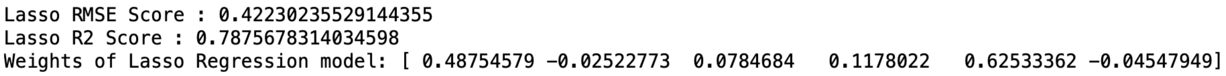
6.4 Lasso Regression with hyperparameter tuning

In the code, we perform a Grid Search cross-validation to select the optimal value of alpha, and then fit a Lasso regression model with that alpha value.



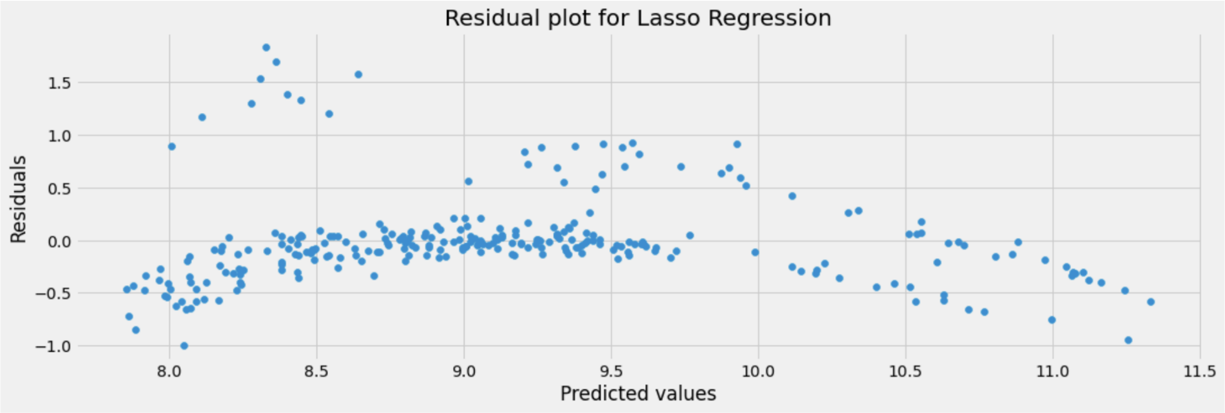
```
Validation scores for each combination of hyperparameters:
mean validation score: 0.755490 for params: {'alpha': 0.001}
mean validation score: 0.754814 for params: {'alpha': 0.01}
mean validation score: 0.712907 for params: {'alpha': 0.1}
mean validation score: -0.010413 for params: {'alpha': 1}
mean validation score: -0.010413 for params: {'alpha': 10}
mean validation score: -0.010413 for params: {'alpha': 100}
mean validation score: -0.010413 for params: {'alpha': 1000}
Best value of lambda found by Grid Search for Lasso Regression: 0.001
```

We also evaluate the performance of the model using the root mean squared error (RMSE) and the R-squared (R2) score.



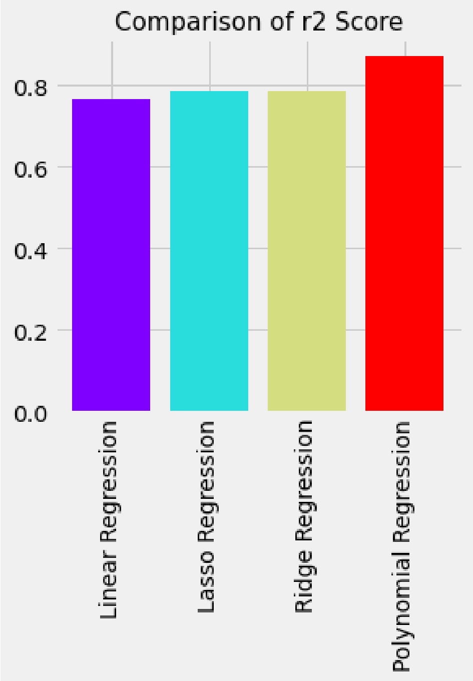
```
Lasso RMSE Score : 0.42230235529144355
Lasso R2 Score : 0.7875678314034598
Weights of Lasso Regression model: [ 0.48754579 -0.02522773  0.0784684   0.1178022   0.62533362 -0.04547949]
```

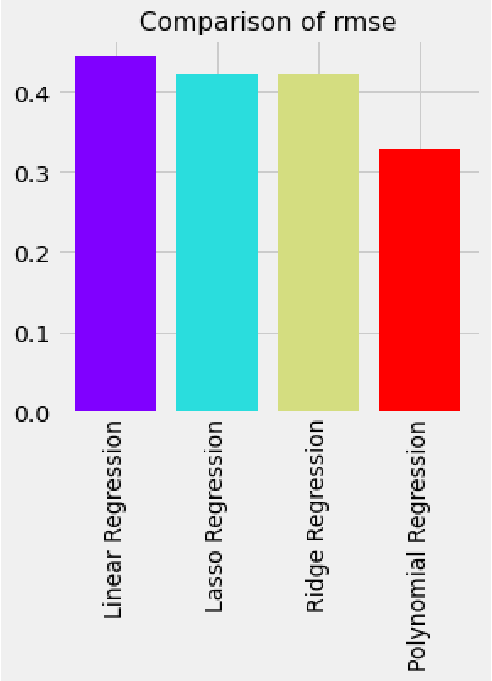
Finally, we plot the residuals to check the assumptions of the model. We see that the residuals are randomly scattered around zero, which suggests that the model is capturing some important information in the data but there is some non-linearity.



7. MODEL COMPARISON AND RESULTS

We compared all the four models and found out that the R2 score of the Polynomial Regression model is higher than the rest of the models. So we conclude that Polynomial regression is the best fit.





The rmse also further supports the conclusion that the Polynomial regression model is the best as Polynomial regression has the lowest rmse.Lasso regression and Ridge regression have similar rmse and Linear regression has the highest rmse. We can also observe from the residual plots that ridge, lasso and linear regression have similar residual plots with some non-linearity. The polynomial regression residual plot suggests that it is a better model as it has lower non-linearity in residuals. The residual plots further support our conclusion that Polynomial regression is the best fit.