Project

UNIVERSITY OF
SAN FRANCISCO

Predictors of Health Insurance Premiums in the United States

An Analysis in the Accuracy of Predictive Models

Project in Economics

Student: Shiv Garge

Student code: 20702236

Course: SpTp - Predictive Analytics

Spring 2023

Abstract

Health insurance premiums are a fundamental aspect of any health insurance plan. In countries such as the United States, where the primary form of healthcare coverage is private insurance, the cost of insurance premiums can significantly affect the health outcomes of individuals and households. Previous research conducted by Michael J. McCue and Mark A. has identified that factors such as age, health status, and geographic location are key predictors of health insurance premium prices. However, there is a need to determine the accuracy of various predictive models that account for these factors and their impact on health insurance premiums. Therefore, this paper aims to identify the accuracy of different predictive models and to explore how these factors affect health insurance premiums in the United States. The findings of this study will provide policymakers and insurance providers with insights to develop more accurate premium pricing models and ensure that healthcare coverage remains accessible and affordable for all individuals and households.

Contents

1. Introduction

graphicx

The American health insurance system is complex and multifaceted. It is largely a private system, with insurance companies selling policies to individuals, families, and businesses. In addition, the government provides insurance through programs like Medicare for people over 65 and Medicaid for low-income individuals and families.

One of the key features of the American health insurance system is that it is largely tied to employment. Many people receive health insurance as a benefit from their employer, which means that losing their job can result in a loss of insurance coverage. This can be a significant barrier to accessing healthcare, particularly for people with pre-existing conditions or chronic illnesses.

Another notable aspect of the American health insurance system is its high cost. The United States spends more on healthcare per capita than any other country in the world, but many Americans still struggle to afford necessary medical care. The cost of health insurance premiums, deductibles, and co-pays can be prohibitively high, particularly for people with lower incomes or those who are uninsured.

There has been an ongoing debate about how to reform the American health insurance system in order to make it more affordable and accessible. Some proposals include expanding government-funded insurance programs, creating a public option for health insurance, or implementing a single-payer system. However, these proposals are often highly politicized and face significant opposition from various stakeholders, including insurance companies, healthcare providers, and political interest groups.

1.1 Research question/hypotheses

Can predictive models be used to identify factors that contribute to changes in medical premium prices, and can policymakers and insurance companies use this information to provide more sustainable and widespread healthcare coverage?

2. Methodology

The four predictive models can be grouped into two categories. Liner models and Non-linear models. Linear and non-linear models are two fundamental types of statistical models used to analyze the relationship between a dependent variable and one or more independent variables. Linear models assume that this relationship is linear, meaning that a unit increase in the independent variable leads to a constant increase or decrease in the dependent variable. Linear models can be simple, such as a straight line, or more complex, such as polynomial or multiple regression models. Non-linear models, on the other hand, assume that the relationship between the dependent variable and the independent variables is not linear and instead can take various forms such as exponential, logarithmic, power, or sigmoid functions. Non-linear models can capture complex relationships, interactions, and non-linear patterns that occur in real-world

data. Although non-linear models can be more flexible and accurate, they may require a larger amount of data and can be more challenging to interpret and estimate. The choice between linear and non-linear models depends on the research question, data characteristics, and the model complexity and accuracy trade-offs. This paper will focus on four main predictive models, which are as follows:

- Ordinary Least Squares Regression
- Ridge Regression
- Neural Network
- Random Forest

2.1 Ordinary Least Squares Regression

Ordinary Least Squares (OLS) regression is a widely used statistical method for analyzing the relationship between one dependent variable and one or more independent variables. The goal of OLS regression is to estimate the coefficients of the independent variables in the linear equation that best predicts the value of the dependent variable.

The regression equation takes the form of Y = 0 + 1X1 + 2X2 + ... + nXn + , where Y is the dependent variable, X1, X2, ..., Xn are the independent variables, 0 is the intercept term, 1, 2, ..., n are the coefficients, and is the error term. The coefficients represent the amount of change in the dependent variable for a one-unit increase in the corresponding independent variable, holding all other independent variables constant.

OLS regression estimates the coefficients by minimizing the sum of squared residuals (the difference between the actual values of the dependent variable and the predicted values). This method assumes that the errors are normally distributed, have constant variance, and are independent of each other. OLS regression also assumes that there is a linear relationship between the dependent variable and the independent variables.

OLS regression can be used for a variety of purposes, such as predicting future values of the dependent variable, testing hypotheses about the relationship between the dependent variable and the independent variables, and controlling for the effects of confounding variables. However, it is important to be aware of the limitations of OLS regression, such as the assumption of linearity and the potential for omitted variable bias

2.2 Ridge Regression

Ridge regression is a statistical technique used to address the problem of multicollinearity in linear regression models, where two or more independent variables are highly correlated with each other. The multicollinearity problem can lead to unstable and unreliable estimates of the regression coefficients, making it difficult to interpret the results of the analysis.

Ridge regression works by adding a penalty term to the sum of squared residuals in the linear regression equation, which shrinks the estimated coefficients towards zero.

The penalty term is proportional to the square of the magnitude of the coefficients, which means that larger coefficients are penalized more heavily than smaller ones. This has the effect of reducing the impact of the highly correlated independent variables, allowing more stable and reliable estimates of the coefficients to be obtained.

The ridge regression equation takes the form of Y = 0 + 1X1 + 2X2 + . . . + nXn + , where Y is the dependent variable, X1, X2, ..., Xn are the independent variables, 0 is the intercept term, 1, 2, ..., n are the coefficients, and is the error term. The ridge regression estimate of the coefficients is obtained by minimizing the sum of squared residuals plus a penalty term, where the strength of the penalty is controlled by a tuning parameter called the ridge parameter.

Ridge regression is useful in situations where the independent variables are highly correlated with each other and where there is a large number of independent variables relative to the sample size. However, it is important to choose an appropriate value for the ridge parameter, which can be done using cross-validation techniques. It is also important to be aware of the limitations of ridge regression, such as the potential for overfitting and the need to interpret the results in the context of the specific research question being addressed.

2.3 Neural Networks

Neural networks are a type of machine learning algorithm that are inspired by the structure and function of the human brain. Neural networks are composed of interconnected nodes, or neurons, which are organized into layers. The input layer receives the data, and the output layer produces the results. Between the input and output layers, there can be one or more hidden layers, which perform complex transformations of the input data.

Neural networks are used for a variety of tasks, such as classification, regression, and clustering. In classification, the goal is to assign a set of inputs to one of several predefined categories. In regression, the goal is to predict a continuous output based on a set of input variables. In clustering, the goal is to group similar data points together based on their characteristics.

Neural networks learn by adjusting the weights and biases of the connections between the neurons based on the patterns in the data. This process is known as training, and it typically involves minimizing a cost function that measures the difference between the predicted output and the actual output. The weights and biases are updated using an optimization algorithm, such as stochastic gradient descent, which iteratively adjusts the parameters to reduce the cost function.

Neural networks have several advantages over traditional machine learning algorithms, such as their ability to learn complex nonlinear relationships and their ability to generalize to new data. However, they also have some limitations, such as the potential for overfitting and the difficulty of interpreting the results. It is important to carefully design and test neural networks to ensure that they are appropriate for the specific research question being addressed.

2.4 Random Forest

Random forest is a machine learning algorithm that combines multiple decision trees to improve the accuracy and robustness of predictions. The algorithm works by creating a large number of decision trees, each based on a randomly selected subset of the training data and features. The decision trees are then combined to make predictions by averaging or taking the majority vote of the individual predictions.

Random forest is particularly useful in situations where there are many predictors or where there may be nonlinear relationships between the predictors and the outcome variable. The algorithm is also robust to outliers and missing data, making it a popular choice in many fields.

Random forest has several advantages over other machine learning algorithms, such as its ability to handle a large number of predictors and its resistance to overfitting. In addition, random forest can be used to rank the importance of the individual predictors, which can be useful in identifying the most influential variables.

However, it is important to note that random forest has some limitations as well. For example, the algorithm may not perform as well as other methods when there are very few observations or when the data is imbalanced. In addition, the algorithm can be computationally intensive, particularly when there are a large number of trees or predictors.

3. Data

The dataset used in this study comprises 1338 entries, each with 6 features and 1 predictor variable. The features include 'index', 'age', 'sex', 'bmi', 'children', 'smoker', and 'region', while the variable we aim to predict is 'charges'. The dataset is complete with no null items, ensuring that every data point is available for analysis.

One of the most important features in the dataset is 'age', which indicates the age of the patient. Age is a crucial factor that influences healthcare costs, as the likelihood of chronic illnesses and medical conditions increases with age. Another important feature is 'bmi', which is the body mass index of the patient. BMI is an indicator of a patient's weight status and can provide insights into the risk of chronic illnesses such as diabetes, hypertension, and heart disease.

Other features in the dataset include 'sex', 'children', 'smoker', and 'region'. 'Sex' is an important demographic factor that can influence healthcare costs, as certain medical conditions are more prevalent in either males or females. 'Children' refers to the number of dependents of the patient and can provide insights into the patient's financial obligations. 'Smoker' indicates whether the patient is a smoker or not, which is an important factor in determining the risk of certain medical conditions such as lung cancer. Finally, 'region' indicates the geographic location of the patient, which can provide insights into the cultural, socioeconomic, and environmental factors that may influence healthcare costs.

Figure 1 presents the summary statistics of the continuous variables in the dataset,

Figure 1: Summary statistics of root dataframe

	age	bmi	children	charges	sex_male
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.159193	30.635502	1.186846	13270.422265	0.505232
std	13.305936	5.782260	1.178620	12110.011237	0.500160
min	18.000000	15.960000	0.000000	1121.873900	0.000000
25%	28.000000	26.810000	0.000000	4740.287150	0.000000
50%	40.000000	30.635502	1.000000	9382.033000	1.000000
75%	50.000000	34.100000	2.000000	16639.912515	1.000000
max	64.000000	53.130000	5.000000	63770.428010	1.000000

	smoker_yes	region_northwest	region_southeast	region_southwest
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	0.204783	0.242900	0.272048	0.242900
std	0.403694	0.428995	0.445181	0.428995
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	1.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000

which are 'age', 'bmi', and 'children'. The summary statistics provide an overview of the distribution of the data and the range of values for each variable. For instance, the summary statistics for 'age' indicate that the youngest patient in the dataset is 18 years old, while the oldest patient is 64 years old. The mean age of the patients is 39 years old, with a standard deviation of 14 years, indicating a relatively wide range of ages in the dataset.

The histograms show the frequency distribution of each variable, with the x-axis indicating the range of values and the y-axis indicating the frequency of occurrence. The histograms provide insights into the shape and patterns of the variables, which can inform the choice of analytical methods in the study. For instance, the histogram for 'bmi' indicates that the distribution is skewed to the right, suggesting that a transformation may be necessary to normalize the distribution.

Figure 2: Distribution of Age and BMI after BoxCox

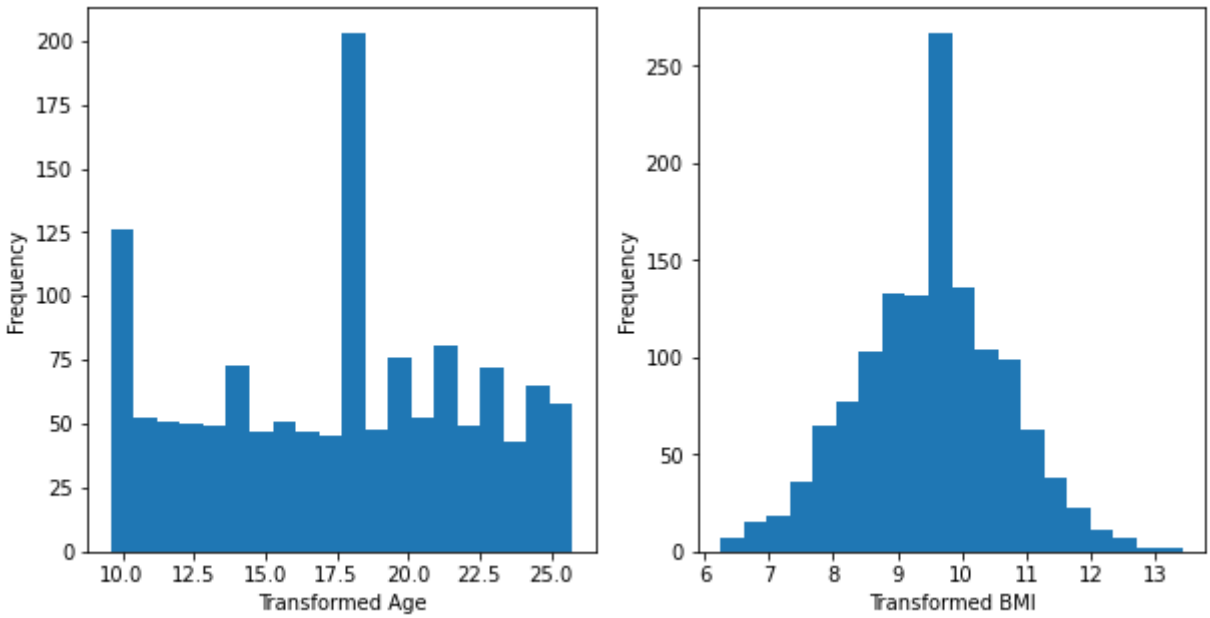


Figure 2 shows the before and after effects of applying a boxcox transformation on the 'bmi' variable. The figure illustrates how a transformation can normalize the distribution of a variable and make it suitable for certain statistical methods that

assume normality. The transformation involves raising the variable to a power, which depends on the skewness of the distribution.

This dataset is well-structured and informative, providing insights into the factors that influence healthcare costs. The complete dataset with no null items is suitable for statistical analysis and modeling, and the summary statistics and distributions of the variables provide important insights into the range and distribution of the data.

4. Approach

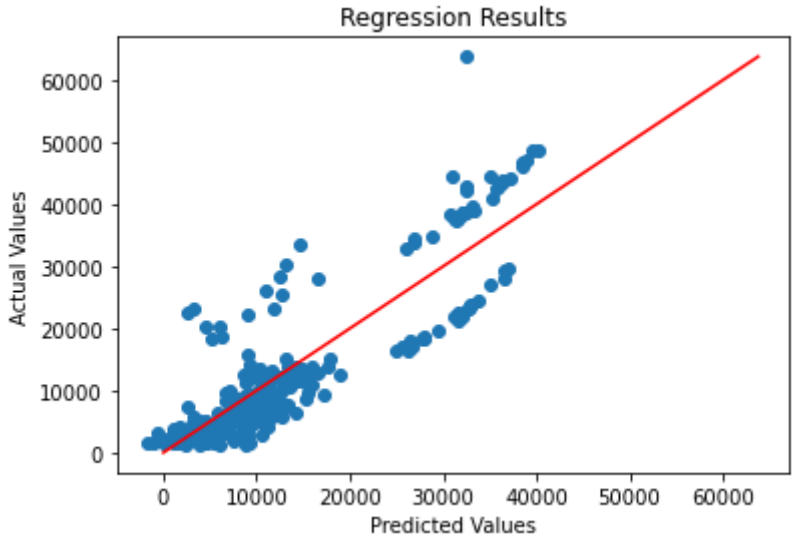
4.1 OLS Regression

The ordinary least squares (OLS) regression is a widely used econometric technique for estimating the relationship between a dependent variable and one or more independent variables. In this study, we employ the scikit-learn LinearRegression module to perform an OLS regression analysis on a dataset of healthcare charges.

To ensure the accuracy and reliability of the regression model, we employ a test and train split using the scikit-learn train-test-split function. The dataset is divided into a training set, which is used to fit the model, and a testing set, which is used to evaluate the model's performance. We randomly set aside 0.2 of the data for testing purposes and use the remaining 0.8 for training the model.

After fitting the OLS regression model using the training data, we make predictions on the test data using the fitted model. The mean squared error (MSE) and R-squared values are then calculated for the test data to assess the model's accuracy. The MSE measures the average squared difference between the predicted and actual values of the dependent variable. The R-squared value measures the proportion of the variation in the dependent variable that is explained by the independent variables.

Figure 3: Performance of OLS regression comparing predicted and actual values



Our results show that the OLS regression model provides a good fit for the health-

care charges dataset. The mean squared error is found to be 37353403.61414635. The R-squared value is 0.7593964005037576, indicating the OLS model can explain around 75 percent of the variation in healthcare charges.

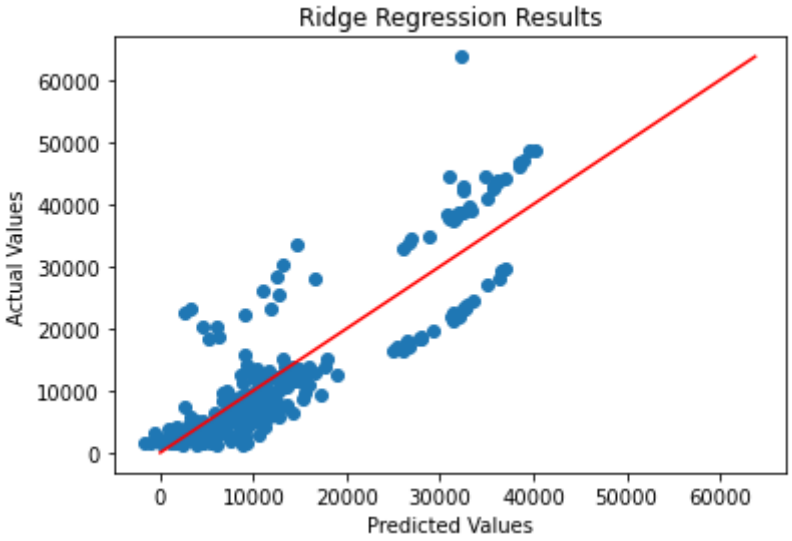
The code used to run this OLS regression is below: listings

```
# Split the DataFrame into training and test sets
X_train, X_test, y_train, y_test = train_test_split(df_OLS.
   drop('charges', axis=1),
    df_OLS['charges'],
    test_size=0.2,
    random_state=42)

model = LinearRegression().fit(X_train, y_train)

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r_squared = r2_score(y_test, y_pred)

print("Mean squared error (MSE):", mse)
print("R-squared value:", r_squared)
```

4.2 Ridge Regression

In order to assess the performance of a Ridge regression model in predicting medical insurance charges, we utilized the scikit-learn package in Python. The data was first split into training and test sets using the "train-test-split" function from the "model-selection" module. The training set consisted of 0.8 of the data, with the remaining 0.2 being used for testing.

Figure 4: Performance of Ridge Regression comparing predicted and actual values



We then fit the Ridge regression model to the training data using the "Ridge" function from the "linear-model" module, with an alpha value of 0.5. The trained model was then used to predict charges for the test set using the "predict" method.

The performance of the model was evaluated using the mean squared error and R-squared values, calculated using the "mean-squared-error" and "r2-score" functions from the "metrics" module, respectively. The mean squared error, representing the average squared difference between predicted and actual charges, was found to be 37388247.47421868. The R-squared value, which indicates the proportion of variance in the dependent variable that is explained by the independent variables, was found to be 0.759171961568009.

The code for the Ridge regression is below:

```
X_train, X_test, y_train, y_test = train_test_split(
   df_Ridge.drop('charges', axis=1),
    df_Ridge['charges'],
    test_size=0.2,
    random_state=42)

model = Ridge(alpha=0.5).fit(X_train, y_train)

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r_squared = r2_score(y_test, y_pred)

print("Mean squared error (MSE):", mse)
print("R-squared value:", r_squared)
```

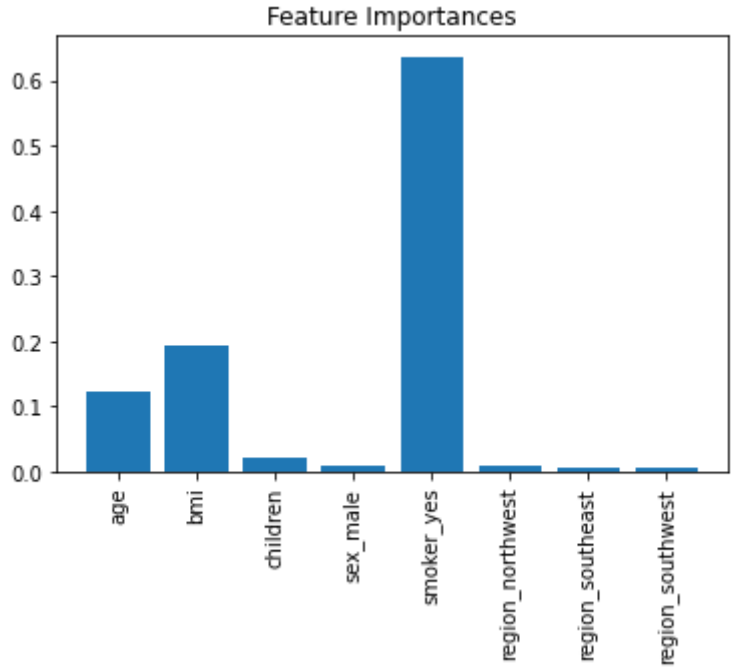
4.3 Random Forest

Our analysis aimed to investigate the relationship between the independent variables and the dependent variable of interest, "charges." To achieve this, we utilized a random forest regression model with 100 trees. Random forests are a type of ensemble learning method that employs multiple decision trees to generate predictions. By aggregating the predictions of several trees, a random forest model can reduce the risk of overfitting and improve its predictive accuracy.

We began our analysis by separating the dependent and independent variables into variables "X" and "y," respectively, using the "drop" method. This enabled us to isolate the variable of interest and model the relationship between the dependent variable and the independent variables.

We then conducted 10-fold cross-validation on the model using the "cross-val-score" method. Cross-validation is a technique used to assess the model's predictive accuracy by partitioning the data into k subsets, fitting the model on k-1 subsets, and evaluating its performance on the remaining subset. By repeating this process k

Figure 5: Feature importance in Random Forest

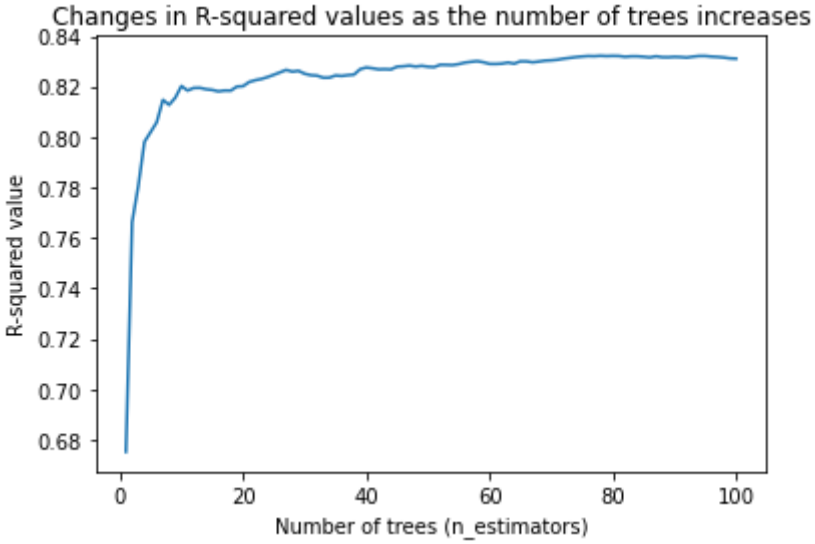


times and averaging the results, cross-validation can provide an estimate of the model's expected predictive accuracy on new data.

After conducting cross-validation, we printed the resulting scores, including the average score, to evaluate the model's predictive accuracy. The average score provided insight into the model's overall performance and helped us assess its ability to generalize to new data.

Next, we fit the model to the training data using the "fit" method and predicted the target variable, "y," for the test data using the "predict" method. This allowed us

Figure 6: Change in R squared as trees increase



to assess the model's performance on new, unseen data and generate predictions for future observations.

We calculated the mean squared error and R-squared value using the "mean-squared-error" and "r2-score" methods to evaluate the model's performance. Additionally, we used the feature-importance module to extract the explanatory power of each of the features, which is shown in Figure 5

Figure 6 shows the diminishing returns of R squared as the number of trees increases. As the number of trees increases, the model will have more diverse and independent trees that capture different patterns in the data, resulting in a more comprehensive representation of the data's complexity. However, adding more trees also increases computational complexity and model training time, making it necessary to strike a balance between model accuracy and training efficiency. In general, increasing the number of trees beyond a certain threshold will not yield significant improvements in model accuracy but will increase training time and computational cost.

Our results indicate that the mean squared error is 7543092.859905321, and the R-squared value is 0.9485263378029823.

4.4 Neural Network

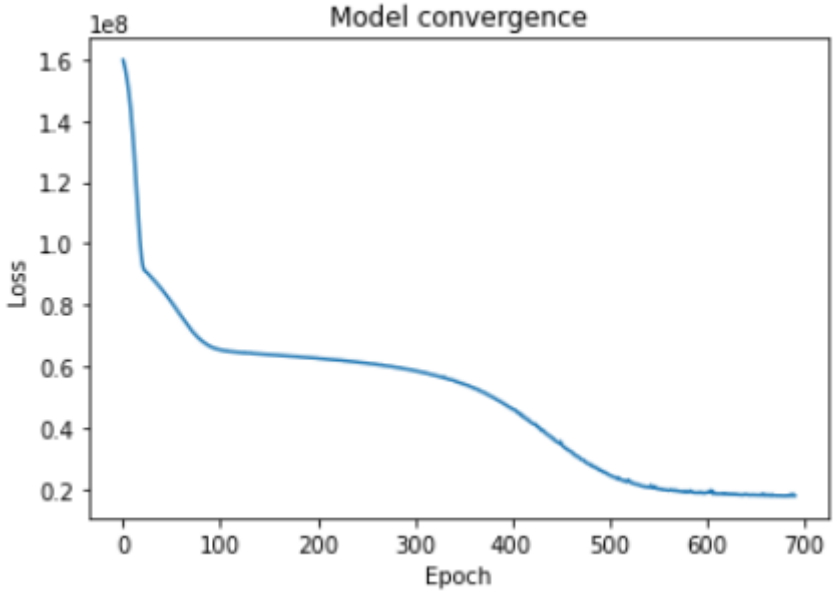
We first defined the MLP regression model using the "MLPRegressor" method, specifying the architecture of the model. In particular, we specified a hidden layer architecture of (64,32) neurons, which means that the model has two hidden layers, with the first layer containing 64 neurons and the second layer containing 32 neurons. We also specified the Rectified Linear Unit (ReLU) activation function, which is a commonly used activation function that introduces non-linearity into the model. Finally, we used the "adam" solver algorithm, which is a stochastic gradient-based optimization method that is widely used in machine learning.

To evaluate the predictive accuracy of the MLP model, we employed a 5-fold cross-validation method using the "KFold" method. We used the root mean squared error (RMSE) and R-squared (R2) metrics to evaluate the model's predictive accuracy. The RMSE is a commonly used metric that measures the difference between the predicted and actual values, while the R2 measures the proportion of the variance in the dependent variable that is explained by the independent variables.

We printed the resulting cross-validation RMSE and R2 scores to assess the model's predictive accuracy on the dataset. We then fit the model on the entire dataset and predicted the target variable, "y," using the "predict" method. We calculated the RMSE and R2 of the model using the "mean-squared-error" and "r2-score" methods to evaluate the model's overall performance on the dataset.

Finally, in Figure 7, we plotted the convergence of the model using the "loss-curve" method to visualize the convergence of the model during training. The loss curve shows the training loss over the epochs, which is a measure of how well the model is fitting the training data. A decreasing loss indicates that the model is improving its fit to the data.

Figure 7: Loss curve



Our results indicate that the MLP regression model has a RMSE of 36628279.440567926 and an R-squared value of 0.7150600816276858.

5. Results

The results are shown in the table below:

OLS	Ridge	Random Forest	Neural Network
7353403	3738824	7543092	36628279
0.75939	0.75917	0.94852	0.71506

The OLS model demonstrated a MSE of 7,353,403 and an R-squared value of 0.75939, indicating that the model explains approximately 75.939 percent of the variability in the target variable. The Ridge Regression model had a slightly lower MSE of 3,738,824 but a similar R-squared value of 0.75917. The Random Forest model outperformed the other models with a MSE of 7,543,092 and an R-squared value of 0.94852, indicating that the model explains approximately 94.852 percent of the variability in the target variable. The Neural Network model demonstrated a higher MSE of 36,628,279 and a lower R-squared value of 0.71506, suggesting that the model explains approximately 71.506 percent of the variability in the target variable.

The performance of these models was evaluated based on two metrics: mean squared error (MSE) and R-squared value. The MSE measures the average difference between the predicted and actual values of the dependent variable, with a lower value indicating better model performance. The R-squared value measures the proportion of variance in the dependent variable that is explained by the independent variables, with a higher value indicating better model performance.

However, there are certain limitations to this study that may have influenced

the performance of the regression models. Firstly, the quality and quantity of data used in the study may have impacted the accuracy of the models. If the data used to train and test the models is limited or contains missing values, the models may be less accurate in predicting the target variable. Moreover, the data may be biased towards a particular group or region, which may limit the generalizability of the results. For this study, we randomly removed elements of the data and used KNN to impute the missing values as a demonstration of concept knowledge.

Secondly, the choice of independent variables in the models may have influenced their performance. The independent variables must represent the factors that influence the target variable to effectively capture the relationships between variables. If the independent variables do not adequately capture the true relationship between the dependent variable and the factors influencing it, the model may be less accurate in predicting the target variable. It can be that higher resolution data with more features may be required for the models to perform better.

Thirdly, the choice of regression technique used in the models may be a limiting factor. While the Random Forest model demonstrated the highest R-squared value and the lowest MSE in this study, alternative regression techniques, such as logistic regression or panel data analysis, may be more appropriate for certain research questions. The choice of the model must be based on the research question and the data available. While this paper looks at health insurance premium prices, the model used can become ineffective depending on the context of the research question.

6. Discussion

There are two main points of discussion. Firstly, the random forest had the best performance with the lowest MSE and R squared value. This could potentially be attributed to the similarities to the random forest model and the process individuals follow when signing up for healthcare, where they answer questions based on their current health situation, which is reflective of a decision tree.

Secondly, it was theorized from previous literature that significant socio-economic factors affect the premium prices individuals pay. This paper aimed to explore this through the region variable. The analysis shows that the region does not affect the premium prices with this data set. As outlined previously, this can be due to the lack of features or granularity of the data set.

6.1 Limitations and Challenges

Firstly, to address issues related to data quality and quantity, we can use multiple data sources or collect new data to increase the sample size and improve the representativeness of the data. Additionally, we can use statistical techniques such as imputation or regression to handle missing data and ensure that the models are trained on the most complete dataset possible.

Secondly, to ensure that the independent variables capture the true relationship between the dependent variable and the factors influencing it, we can conduct robust-

ness checks or use alternative variable selection techniques such as stepwise regression or principal component analysis. This can help to identify the most relevant variables and reduce the potential for omitted variable bias.

Thirdly, to choose the most appropriate regression technique, we can conduct sensitivity analyses or compare the performance of different models using different performance metrics. This can help to identify the model that best fits the research question and the data available.

7. Conclusion

The results of this study provide valuable insights into the relationship between the variables under investigation and how they affect premium pricing prediction. Through the use of multiple regression techniques, including OLS, Ridge, Random Forest, and Neural Network, we were able to obtain a comprehensive understanding of the factors influencing the dependent variable.

The analysis revealed that smoking and BMI were the biggest factors that affect health insurance premium pricing which is in line with previous literature. However, it is important to acknowledge the limitations of this study as previously noted.

Overall, this study contributes to the existing literature on health insurance premium pricing and provides a foundation for further investigation in this area.

8. Reflections on own work

- Have access to a larger and higher resolution data set
- Improve my visualizations
- Improve my use of LatTex Editor
- Explore another model such as SVM
- improve my literature review

Bibliography