Predicting Health Insurance Costs by using Regression Models

Vinayak Vatsalya J
201IT266
Information Technology
National Institute of Technology
Karnataka Surathkal, India 575025
vinayakvatsalyaj.201it266@nitk.edu.in

Jain Samyak Pankajkumar
201IT125
Information Technology
National Institute of Technology
Karnataka Surathkal, India 575025
jainsamp.201it125@nitk.edu.in

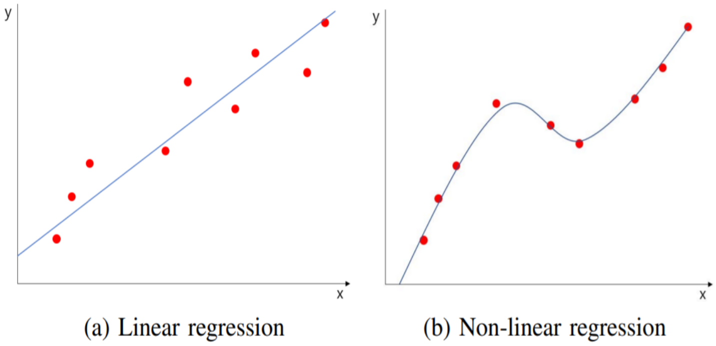
Rakshith Jain
201IT147
Information Technology
National Institute of Technology
Karnataka Surathkal, India 575025
rakshithchajjed.201it147@nitk.edu.in

***Abstract*—Insurance is a policy that tries to decrease costs incurred by accident or illness.This can be influenced by various factors. We use different regression techniques to predict the cost of insurance and compare the results.We compare the results from Linear regression, Lasso Regression, Ridge Regression, Polynomial Regression, Random Forest Regression and Decision Tree Regression.**

Code in Github - group 40

I. Introduction

Insurance is very important as it ensures the financial stability to face any type of problem in life, and this is why insurance is a very important part of financial planning. A general insurance company offers insurance policies to secure health, travel, motor vehicle, and home. Costs for insurance are therefore vital. It is therefore critical that these costs are predicted with high accuracy. Machine learning is beneficial here. With the help of machine learning models we can generalize the effort or method to predict these costs.The model is trained on insurance data from the past. The objective is to forecast insurance charges as accurately as possible.



(a) Linear regression (b) Non-linear regression

Fig. 1. The difference between linear and non-linear regression.

Regression can be classified into two types: linear and non-linear regression analysis. As shown in Fig 1, the linear regression analysis can only learn a linear function from a set of features. In many cases, however, it is important to learn a nonlinear function , so the linear regression analysis often shows lower performance than the non-linear regression analysis. The non-linear regression analysis often performs much better than the linear regression analysis on many data because it can learn a non-linear feature of a variable that the linear regression analysis cannot.

II. Literature Survey

In this section, research efforts in the field of regression (particularly for the case of claim prediction) are discussed. The most recent work is by Mohamed hanafy, Omar.M.A.Mahmoud titled “ Predict Health Insurance Cost by using Machine Learning and DNN Regression Models”. This study used various machine learning regression models and DNN models to forecast charges of health insurance based on specific attributes ,on medical cost personal data set from Kaggle.com. Different algorithms like SVM, k-NN , Random Forest Regressor, XGBoost , Stochastic Gradient descent etc were implemented and compared. The inference was Stochastic Gradient Boosting offers the best efficiency, with an RMSE value of 0.380189, an MAE value of 0.17448, and an accuracy of 85.82Jessica Pesantez-Narvaez suggested, ”Predicting motor insurance claims using telematics data” in 2019. This research compared the performance of logistic regression and XGBoost techniques to forecast the presence of accident claims by a small number and results showed that because of its interpretability and strong predictability, logistic regression is an effective model than XGBoost. Another paper “Predicting Hospital Length of Stay using Regression models” by Combes,Kadri and Chabane predicts for what duration a patient has to stay based on 8 features like No of patients in hospital, CAC, Echo-scan etc.Two Linear regression datamining models were designed and the best of the two was chosen. The best model was only 28.24Oskar Sucki published a paper , the purpose of whose research is to study the prediction of churn. Random forests were considered to be the best model (74 percent accuracies). In the work of G.Kowshalya, M. Nandhini. , three classifiers have been developed in this study to predict and estimate fraudulent claims and a percentage of premiums for the various customers based upon their personal and financial data. For classification, the algorithms Random Forest, J48, and Naïve Bayes are chosen. The findings show that Random Forest exceeds the remaining techniques depending on the synthetic dataset. This paper therefore does not cover insurance claim forecasts, but rather focuses on false claims Keeping in mind these works, we are evaluating the performance of the Polynomial regression clubbed with regularization techniques

III. Methodology

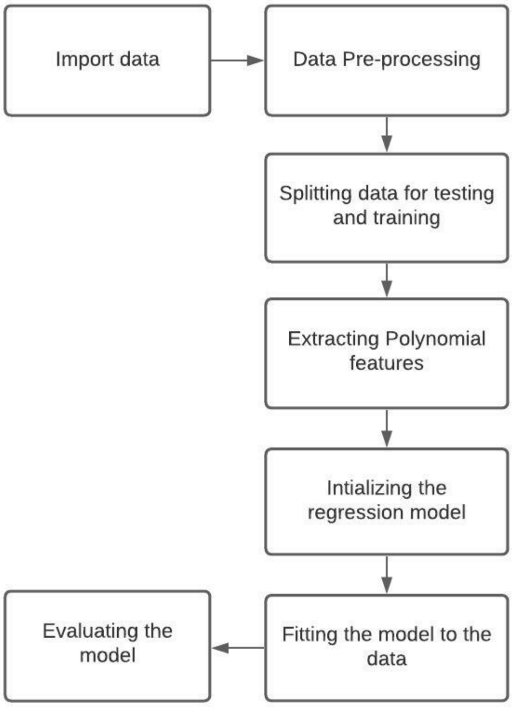

Import data
Data Pre-processing
Splitting data for testing and training
Extracting Polynomial features
Intializing the regression model
Fitting the model to the data
Evaluating the model

Fig. 2. Overview of the methodology

The methodology involves first of extracting data into the dataframe. We then perform various data pre processing techniques. Then we split the data for training and testing. We then implement the various different regression models and evaluate the results.(fig 2)

IV. Dataset

To train the models we have used data from a dataset on Kaggle (3).The dataset contains seven different columns. We extract different features from these columns for our model.The data is split into two sets for training and testing respectively. The training set makes about 80 percent of the dataset, with 20 percent kept for testing and validating the model. The following table gives a simple overview of the dataset.

V. Data pre-processing

The dataset is made up of seven variables (X), as shown in table 1, every one of these features can help us in estimating the cost of the insurance (y , which is the target variable here). In this stage we apply various pre-processing techniques to get the best model possible. To begin, the unknown numerical values (nan) are replaced with the mean.

TABLE I
Dataset overview

Feature name	Description
age	age of primary beneficiary
sex	gender
bmi	Body mass index
smoker	if the beneficiary smokes
children	number of children of beneficiary
region	region in which the beneficiary stays
charges(Target value)	costs billed

TABLE II
Categorical Features

Feature name	Encoded data
region	values are mapped from 0 to 3
sex	values are mapped as 0 or 1
smoker	values are mapped as 0 or 1

We have three categorical features -smoker, region and sex. We encode this into numerical data with the help of label encoders (table ii).

We also standardize the features by removing the mean and scaling it to unit variance column wise(equation 1). Further , we use different techniques to identify the importance of different features and perform feature selection.

$$\hat{X} = \frac{\hat{X} - mean}{std} \tag{1}$$

A. Exploring the dataset

We can get useful information by analysing the different features in the dataset.

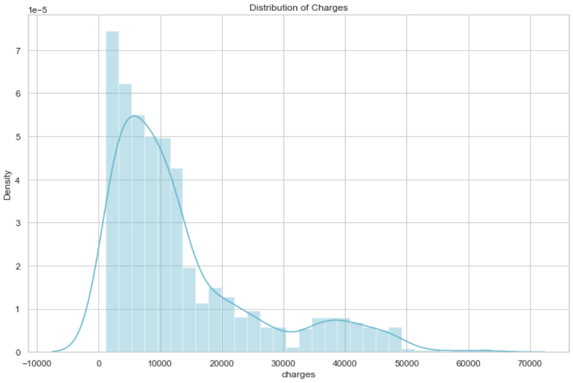


Fig. 3. Distribution of charges

By looking at the distribution of health insurance charges(fig. 3) we can see that it is right skewed. In regression, having skewed data can result in sub optimal models.We can normalize this data using natural logarithm transformation. The distribution after normalization is show in (fig 4).

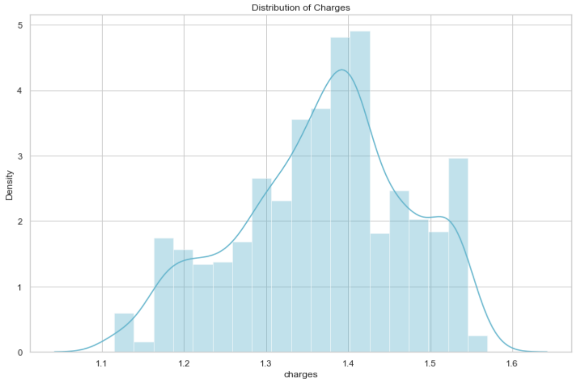


Fig. 4. Natural logarithm of the distribution of charges

Now we can analyse the other features in the dataset and the correlation between them.

Fig 5 clearly shows us the impact of smoking on the costs billed. Generally smokers have higher charges compared to non-smokers. We can see the same applies to both gender.

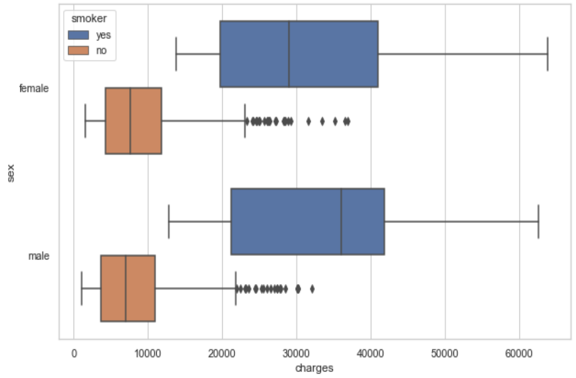


Fig. 5. Distribution of charges for smokers vs non-smokers on basis of gender

From the scatter plot(fig 6) we can see that as BMI increases smoking makes it more likely that charges are high.

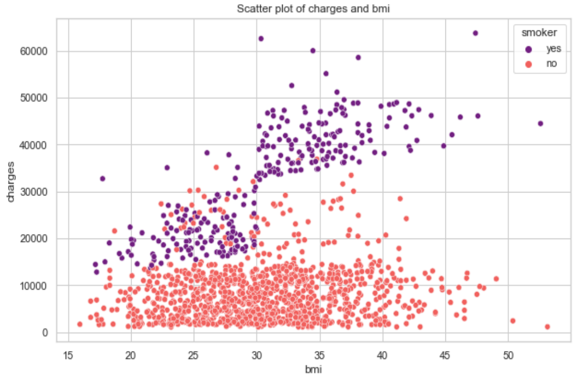


Fig. 6. Distribution of charges for smokers vs non-smokers on basis of gender

We can see that the factors which influence the cost most are age, smoking and BMI.

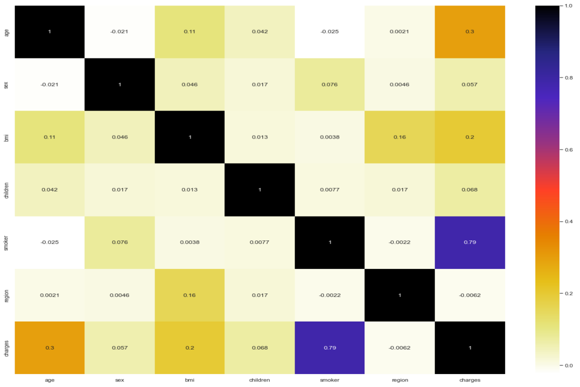


Fig. 7. Correlation matrix

The correlation matrix (fig 7) shows about the linear relationship between each other.

VI. Model

A. Linear regression

A multi-linear regression model was used to train the data in the training set, where hypothesis is given as ,

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2.. \tag{2}$$

Listing 1. h(x) in python

```
def hyp(theta ,X):
    return np.dot(X,theta)
```

and the cost function we used was the mean squared error function ,

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{3}$$

Listing 2. h(x) in python

```
def cost_function(theta ,X,y):
    hx = hyp(theta ,X) - y
    return float((np.dot(hx.T,hx))/(2*
    (X.shape[0])))
```

We trained this model using gradient descent with range or learning rates.In this step , we update the parameters theta simultaneously.

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \tag{4}$$

B. Lasso Regression

Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds “absolute value of magnitude” of coefficient as penalty term to the loss function. In this model (also known as Lasso regularization) , we used a modified version of cost function and gradient upgrade step from linear regression. Lasso shrinks the less important feature’s coefficient to zero thus, removing some feature altogether. So, this works well for feature selection.

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \sum_{i=1}^{m}(|\theta|) \tag{5}$$

C. Ridge Regression

Ridge regression adds “squared magnitude” of coefficient as penalty term to the loss function. Here the highlighted part represents L2 regularization element.This technique works very well to avoid over-fitting issue. The modified cost function - (equation 6).

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \sum_{i=1}^{m}(\theta^2) \tag{6}$$

D. Polynomial regression

Polynomial Regression is a form of regression analysis in which the relationship between the independent variables and dependent variables are modeled in the nth degree polynomial.

E. Gradient Descent

We First initialize our theta value to a matrix of zeros.

Listing 3. Initializing theta

```
theta = np.zeros(
    [X_train_scaled.shape[1],1])
```

Then we find the gradient and update the theta values over the epochs.

Listing 4. Updating the parameters after calculating the gradient

```
epochs = 20
for i in range(epochs):
    theta = theta - (alpha/m)*(np.dot(
    X_train_scaled.T ,hyp(theta ,
    X_train_scaled) - y_train))
```

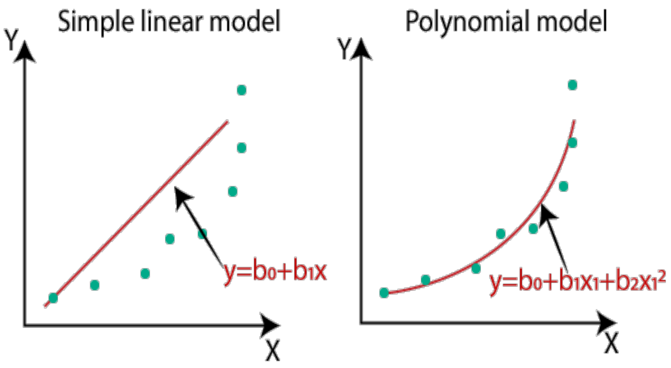


Fig. 8. Comparison between linear and polynomial models

To use this model , first we extracted polynomial features from the dataset. We used a degree of 2 and extracted more than 30 features. The cost function now is,

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1 x_2 \ldots + \theta_n x_1^n \tag{7}$$

We have also used decision tree regressor and random forest regressor for comparing the results.

VII. Results

We implemented the four different models and used three different metrics to evaluate them. The metrics we used mean squared error (MSE), mean absolute error (MAE) and root mean squared error (RMSE).

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \tag{8}$$

The Mean Absolute Error (MAE) is the difference between the original and forecast values obtained by averaging the absolute difference over the data set.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right) \tag{9}$$

The RMSE of the disparity between the expected values and the real values is determined as the square root. For an accurate forecast, the RMSE must be low so there would be less variance among the expected values and the real values

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2} \tag{10}$$

TABLE III
Results

Model	MSE	MAE	RMSE
Linear Regression	0.0334	0.1321	0.1892
Ridge Regression	0.0334	0.1146	0.1827
Lasso Regression	0.0572	0.1732	0.2392
Polynomial Regression	0.0217	0.0927	0.1475
Decision Tree Regression	0.0504	0.1114	0.2246
Random Forest Regression	0.0223	0.0854	0.1493

From the above table we can see that the polynomial regression model fits best and has the best metrics compared to the other models.

VIII. Conclusion

This project uses various regression techniques to forecast the insurance based on specific attributes, on medical insurance cost. The findings can be summarized in the table III , it shows us that the polynomial regression model fit the best with the MSE value of 0.0217 and RMSE value of 0.1475.

IX. Future Work

This dataset was confined to just the regions in the United States, in the future as we get more data in other regions , we can create more region specific models to predict the health insurance costs. As more data is accumulated , we can try larger neural networks which can perform well on large data.

REFERENCES

[1] Predict Health Insurance Cost by using Machine Learning and DNN Regression Models by Mohamed hanafy, Omar M. A. Mahmoud, in International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-10 Issue-3, January 2021.

[2] Predicting Length of Stay and Discharge Destination for Surgical Patients: A Cohort Study.Fabrizio Bert 1,2, Omar Kakaa 1,† Alessio Corradi 1,* , Annamaria Mascaro Stefano Roggero Daniela Corsi 2,Antonio Scarmozzino 2 and Roberta Siliquini 1,2

[3] Kaggle Medical Cost Personal Datasets. Kaggle Inc. https://www.kaggle.com/mirichoi0218/insurance

[4] Catherine Combes, Farid Kadri, Sond'es Chaabane. PREDICTING HOSPITAL LENGTH OF STAY USING REGRESSION MODELS: APPLICATION TO EMERGENCY DEPARTMENT. 10'eme Conf´erence Francophone de Mod´elisation, Optimisation et SimulationMOSIM'14, Nov 2014, Nancy, France. 2014. ¡hal-01081557¿

[5] Machine learning study guides tailored to CS 229 by Afshine Amidi and Shervine Amidi in the CS229 Course by Andrew NG.

[6] An introduction to seaborn¶ by seaborn.pydata.org