Comparative Study of Regression Models and Deep Learning Models for Insurance Cost Prediction

Aditya Shinde(✉) and Purva Raut

Dwarkadas J. Sanghvi College of Engineering, Mumbai, India
adityashinde989@gmail.com, purvapraut@gmail.com

Abstract. In the finance world, Insurance is a product that reduces or eliminate the cost of loss caused due to different risks. There are various factors associated that affect the insurance charges. These factors contribute in formulating the insurance policies. Using machine learning, we can predict insurance charges by developing a model that generalize the entire process. Machine Learning in the Insurance industry would enable seamless formulation of insurance policies with better performance and will time. This study presents how insurance charges can be predicted using various regression models. Also, comparison between the performances of models like Multiple Linear Regression, Support Vector Machine, Random Forest Regressor, XGBoost and Deep Neural Networks is done. This paper provides the most optimal solution using Deep Neural Networks with a root mean square error RMSE value of 0.0695 and model accuracy of 87.95.

Keywords: Regression · Insurance · Random Forest Regressor · XGBoost · Deep Neural Networks

1 Introduction

We live in a world full of uncertainties and risks. Individuals, families, businesses, assets and properties are at various types of risks. The levels of these risks may vary. These risks include the risk of loss of life, health, assets or properties. Of these, life and health are the most valuable aspects of individual's lives. However, it is not always possible to prevent risks from occurring, therefore the financial world has developed various products that protects individuals and businesses against such losses by compensating them with financial resources. Thus, Insurance is a type of financial product that reduces or eliminate the cost of loss caused due to different risks [1, 2].

Knowing the importance of Insurance in people's lives, it becomes imperative for the insurance policy providers to be accurate enough to estimate or calculate the amount covered under the policy and the amount or the insurance charges one must pay for the same. These amounts are estimated using various parameters. Each and every factor is important. If in case any factor is missed out while calculating the amounts, the overall policy changes. Thus, it becomes important to perform this tasks with a high precision. As human errors are about to happen, the insurance companies use people who have expertise in this field. Also, they use various softwares for calculating the

A. Abraham et al. (Eds.): ISDA 2018, AISC 940, pp. 1102–1111, 2020.
https://doi.org/10.1007/978-3-030-16657-1_103

insurance charges. This is where Machine Learning is helpful. Using machine learning (ML), the efforts or process of formulating the policy can be generalized. These ML models can learn on their own. The model is trained on the past insurance data. Then, the factors required for calculating the charges can be given as an input to the model, the model can correctly predict the insurance charges for the policy. Thus, this reduces human efforts as well as time and increases the throughput of the company. With ML, the accuracies can be increased.

Under Machine Learning, supervised learning is used to predict the output or the target variable(s) using the various inputs. In supervised learning, the output class or labels are known beforehand. Regression is one of the supervised learning techniques used to predict a dependent variable based on the values of one or more independent variables. In regression, there is only one dependent variable or target variable or class or label. In this paper, our aim is predict insurance charges. The value of insurance charges depend on various factors. Thus, the insurance charges are continuous values. To satisfy our needs, Regression is the best available option. Regression can be of three types i.e. Linear Regression, Multiple Linear Regression and non-linear Regression. In this study, we use Multiple Linear Regression because we have more than one independent variable used to calculate the dependent variable. The Medical Cost Personal Dataset is used for the study [3]. First, the dataset is pre-processed. Then, it is trained on regression models such as Multiple Linear Regression, Support Vector Machine, Random Forest Regressor and XGBoost. Finally, the dataset is trained on a Deep learning model such as Deep Neural Networks. It is found that the Deep Neural Networks gives the best accuracy of 87.95 which can be further increased as size of data increases.

The main motivation behind carrying out the research is to provide a novel way to forecast insurance charges. Also, using Deep learning models, it is observed that the accuracy of the model increases as amount of data increases. Thus, to handle such huge data and correctly predict the insurance charges, deep learning models such as deep neural networks can be used to overcome the issue with machine learning regression models which reach a constant level of accuracy after certain quantity of data is fed. In this paper, it is presented that how Deep Learning models can change the current use of the regression models and provide a way for getting more accuracies.

2 Dataset

2.1 Dataset Description

Medical Cost Personal Dataset from Kaggle.com is used for this study. The dataset has attributes such as age, body mass index (bmi), number of children, region where the person resides (southwest, southeast, northwest, northeast), sex and smoker (whether a person is a smoker or not). The charges attribute is the target variable, which we are predicting using different regression models and a deep learning model [3].

2.2 Data Pre-processing

The dataset has 7 relevant attributes (age, bmi, children, region, sex, smoker, charges). Each attribute has some contribution in predicting the insurance costs i.e. the charges, which is our target variable. In this step, the data is analyzed and modified into an appropriate format so that the data can be efficiently applied to the machine learning algorithms. First, the data is cleaned by identified all the unknown values. The unknown numeric values are replaced by the mean of the corresponding feature and the unknown categorical values are replaced by the value which appears the most in the corresponding feature. Next, the target variable (charges) is analyzed. It is observed that the distribution of insurance charges is skewed towards the right. Figure 1. Thus, feature scaling is performed. Feature scaling is a method used to standardized the range of the features in the dataset. It is also called as Data Normalization [4]. The objective functions of various machine learning models will not work efficiently without normalization because the ranges of values in a dataset can vary widely. Also, it may happen that features have different units, for example, 5 kg and 5000 g. If data is not normalized, the machine learning algorithm neglects the units and only take magnitudes in consideration. Although the value of both the features is same, but due to their magnitudes, the algorithm treats them differently. Thus, the data is rescaled using normalization so that all the data is of the same scale. Normalization also ensures that the algorithm does not commit small errors on low values and large errors for high values. By normalizing the data, the mean of the entire data becomes zero and each feature has a unit variance. To normalize the insurance charges, the log transformation are applied (Fig. 1).

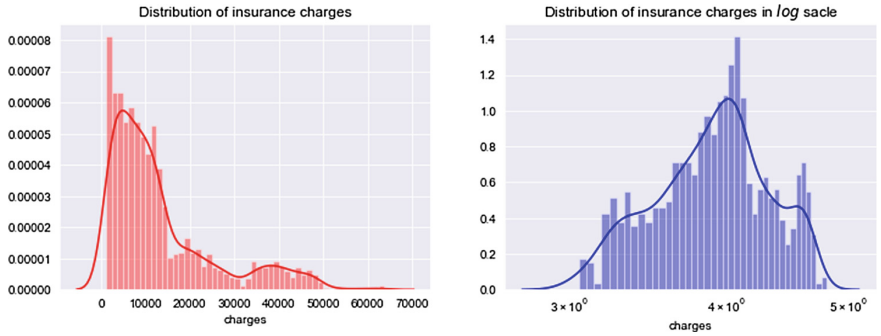

Fig. 1. Right skewed (left), normalized data (right)

Finally, the categorical values are converted into numeric or binary values by using label encoding and one-hot encoding. The machine learning models only work on numeric data. They cannot interpret categorical data. For using categorical data in prediction, one-hot encoding is performed [5]. New column is created for every categorical value and 1 is assigned to it if that particular row has that value otherwise 0. After performing one-hot ending, now we have 12 attributes of interest. After this step, the data is pre-processed and ready to be applied to various models.

3 Related Works

3.1 Regression

Regression is a statistical measure that is used to determine the relationship between one dependent variable and one or more changing variables also known as independent variables. The dependent variable is usually denoted by Y. The independent variables are denoted by x1, x2 and so on. The value of Y is determined by the values of independent variables. Regression is used in various domains such as finance, investing, machine learning and many other disciplines. Regression is of two basic types, Linear Regression and Multiple Linear Regression. There are also non-linear regression methods for more complex data analysis. In linear regression, one independent variable is used to predict the dependent variable Y. In multiple linear regression, two or more independent variables are used to predict the dependent Variable Y. The equation of linear regression is by Eq. 1

$$Y = a + bX + u \tag{1}$$

where Y is the dependent variable (the variable is to be predict), X is the independent variable (the variable used to be predict Y), a is the intercept, b is the slope and u is the regression residual.

In this study, regression is used to predict the insurance charges which is our target variable based on the features values. Multiple Linear Regression is used as we have more than one feature or independent variables to predict our outcome i.e. the insurance charges.

3.2 Regression Models

Multiple Linear Regression. In Multiple Linear Regression models, multiple independent variables (or predictors) are used to determine the relation between independent variables and the dependent variable Y. Each independent variable is assigned a weight depending upon its contribution in predicting the dependent variable. These weights are called as regression coefficients. A regression coefficient measures the change in the dependent variable Y when an independent variable changes by a unit, while other independent variables are constant [6].

A multiple linear regression model with n independent variable X1, X2, X3, …, Xn and the outcome Y can be written as in Eq. 2.

$$Y = \alpha_0 X_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots\ldots + \alpha_n X_n + u \tag{2}$$

where u is the regression residual and α is the regression coefficients (weights) assigned to each parameter or the independent variables.

Random Forest Regressor. A decision tree is a tool that uses a tree-like graph of decisions and their possible outcomes. It is a tree-like structure in which each internal node specifies a condition based on which correct path is chosen. Each leaf node of the

tree represents a label or a value. The decision tree model has a drawback that it overfits the data. Overfitting - a model is said to overfitting the data when it exactly or closely predicts a value corresponding to a particular set of data and may therefore fail to fit additional data or correctly predict future observations. Random forests are an ensemble learning method for regression that consists of multiple decision trees. Many decision trees together form a random forest. In regression, a random forest produces a mean prediction value of individual trees as an output. Random forests overcome the problem of overfitting, which is a drawback of decision trees [6]. A random forest regressor model can be expresses as in Eq. 3

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots\ldots\ldots + f_n(x) \tag{3}$$

where g is the final model which is the sum of all individual model. Each individual model $f(x)$ is a decision tree.

Each tree is formed by branching the tree at a splitting attribute. The attribute with minimum standard deviation is chosen as a splitting attribute. Standard deviation is a measure that shows the amount of variation or dispersion between the values of a feature. Using Random forests, it is easy to see what features contributes to the regression and their relative significance based on whether a leaf node i.e. the output is higher or lower in the tree.

XGBoost. XGBoost is the improved version of Gradient Boosting. XGBoost stands for eXtreme Gradient Boosting. Boosting algorithms consists of repetitively learning weaker models and adding them to the final strong models. The fundamental principle of Boosting is to convert weak models (those which make errors in prediction) into a strong models (those with low error rate). XGBoost is a technique in which new models are built on top of each other such that the new models rectify the errors made by the existing models. The algorithm assigns larger weight to the wrongly predicted data [7].

Support Vector Machine. Support Vector Machine (SVM) can be also used for regression problems. Support Vector Regression (SVR) uses the same logic as the SVM used for classification, with a few minute differences. It is very difficult to predict the information, which has infinite possibilities because output of regression is a real number. In regression using SVM, a margin of tolerance is used. However, it is difficult to implement SVM for regression as compared to SVM for classification. Therefore, the algorithm is more complicated SVM gives high accuracy in case of classification as compared to regression.

3.3 Deep Learning

Deep Learning is a part of machine learning. Deep learning models such as deep neural networks, recurrent neural networks and deep belief networks have their applications in field of natural language processing, computer vision, speech recognition, bio-informatics, etc. Deep learning mainly consists of algorithms which are based on the structure and functions of the human brain. These algorithms produce outputs

comparable and sometimes superior to humans. Using the analogy of a brain, It is the most revolutionary advances in the field of Artificial Intelligence (AI) [8]. The reason deep learning algorithms is used because it performs better than other learning algorithms. These algorithms work efficiently with huge amount of data. Deep learning algorithms can be used for regression problems as well as for classification problems. They can be used in supervised learning as well as unsupervised learning methods. The performance of deep learning models can be increased by feeding more and more data to the model. This is quiet different from other learning techniques that reach a plateau in performance after a certain amount of data is feed into the model as shown in Fig. 2.

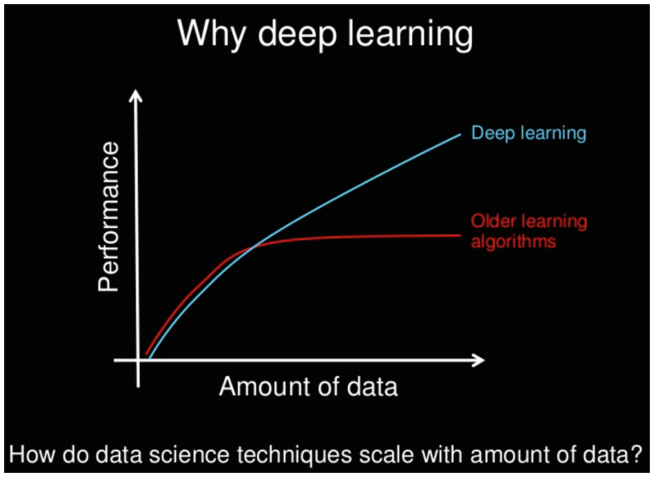

Fig. 2. Performance comparison between Deep Learning models and other learning algorithms [9].

3.4 Deep Neural Networks

A neural network is a system of hardware and software patterned after the operation of neurons in the human brain. Neural Network can be thought as a system which can compute operations as our human brain does. Deep learning is capable of establishing correlations between present events and future events. A neural network has a thousand of units (analogous to neutrons in the brain) which are arranged in a series of layers. A layer is made of many nodes. A node is like a neuron in a brain where computations take place. A node takes input from the data and combines with a set of coefficients or weights. These weights assign significance to a input while performing a task. The higher the weight, more influence one unit has on another. A deep neural network has one input layer, two or more hidden layers and an output layer. Unlike other deep learning models such as artificial neural networks which has only one hidden layer. So, a deep neural networks consists of many layers which leads to a deep structure. A neural networks with more than 3 layers qualifies as a “deep” network (Fig. 3).

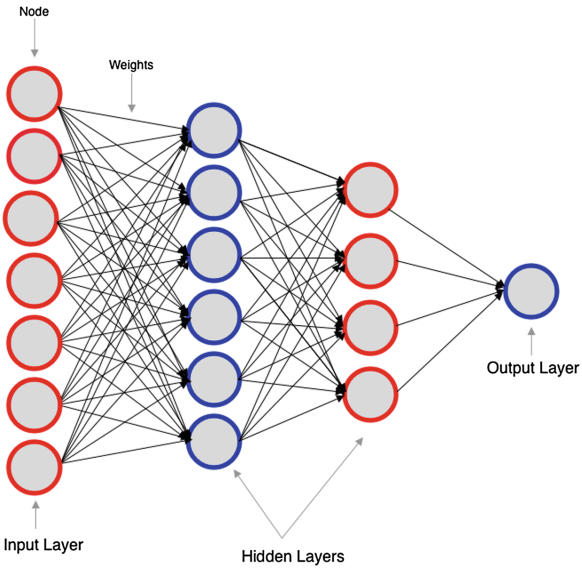

Fig. 3. Deep Neural Networks

4 Implementation

For this study, Medical Cost Personal Dataset from Kaggle.com. The study aims to measure the insurance charges based on parameters such as age, body mass index (bmi), number of children, region where the person lives, sex and smoker (whether a person is a smoker or not). These attributes contribute towards the prediction of insurance charges which is our target variable. Various regression models are used to measure the insurance charges. The dataset is divided into two parts. One part for training the model and the other part for evaluating or testing the model. The dataset can be divided into 80% training data and 20% testing data or 75% training data and 25% testing data or 60% training data and 40% testing data. In this study, the data set is divided into 80% training data and 20% testing data. Each model is trained on the training data and it is evaluated on the testing data [10–13].

For this study, Jupyter Notebook with python 3.0 is used for implementing the models. For regression models, scikit-learn libraries are used. Scikit-learn is a free machine learning library and it is an extension to SciPy. For Deep Neural Networks, Keras Library for Deep Learning is used. For regression we are using root mean squared error (RMSE) and r^2 score (R squared) as measure of comparison. The RMSE is calculated as the square root of the difference between the predicted values and the actual values. The RMSE must be low or 0 for an accurate prediction because there will less difference between the predicted and the actual values. The formula for RMSE is given by Eq. 4.

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y} - y)^2} \tag{4}$$

where N = total number of observations, $\hat{y}$ = predicted values of insurance charges, y = actual values of insurance charges.

The R-squared is also known as coefficient of determination. It is the proportion of variance in the dependent variable that is predicted from the independent variables. The formula for R-squared is given by Eq. 5.

$$R{-}squared = \frac{Explainedvariance}{Totalvariance} \tag{5}$$

The value of R-squared should be high for a better performance of a model. A higher values shows that the model deviates less from actual values in predicting the values. A R-squared score of 1 means it is a perfect fit.

5 Results

Regression models used in the study are Multiple Linear regression, Support Vector regression, Random Forest Regressor and XGBoost. These all models are trained using the training data and evaluated on the testing data. The root mean square error (RMSE) and R-squared (R^2 score) are calculated for each model. The multiple linear regression model predicts insurance charges with an RMSE value of 0.0925 and model accuracy of 78.35. With Support vector regression, RMSE value of 0.0965 and model accuracy of 76.44. Random Forest regressor gives a RMSE value of 0.0737 and model accuracy of 86.26. With XGBoost, the model provides a RMSE value of 0.0694 and model accuracy of 87.83. Deep Learning models such as Deep Neural Networks predicts insurance charges with a RMSE value of 0.0695 and model accuracy of 87.95. These results are tabulated in Table 1.

Table 1. Results of all trained models evaluated on testing data

Algorithm used	Root mean squared error (RMSE)	R-squared (model accuracy)
Multiple Linear Regression	0.0925	78.35
Support Vector Regression	0.0965	76.44
Random Forest Regressor	0.0737	86.26
XGBoost	0.0694	87.83
Deep Neural Networks	0.0695	87.95

6 Conclusion

This study performs regression using various regression models and deep neural networks on the medical cost personal dataset from Kaggle.com to predict the insurance charges based on certain attributes. The results as shown in the Sect. 5 depicts that deep neural networks provides the best performance and the most optimal prediction with a RMSE value of 0.0695 and accuracy of 87.95. Hence, it is proved that deep learning models can be used in the prediction of insurance charges with better results as compared to other regression models. Predicting insurance cost based on given parameters would enable the insurance policies providers to attract more customers to them and save their time in formulating the policies for each and every individual. Using machine learning, these human efforts of policy making can be reduced to great extent, as these machine learning models can calculate the costs in small amount of time, while on the other hand, a human employee would take considerable time for the same task. This would help the companies in increasing their throughput. Also, the Deep Neural Networks can handle huge amount of data. Hence, these networks can be scaled upto a large scale and can be used in large companies. It is also found that deep neural networks show increased accuracy when more and more data is feed into the networks. This is opposite to the other regression models which after certain amount of data, reach a plateau of performance beyond which the model accuracy cannot be increased. Thus, the most optimal solution is found out using Deep neural Networks whose applications can be used in forecasting insurance charges for insurance policy companies.

7 Future Scope

The dataset used in the study consists of 1338 rows. As the accuracy of deep neural networks increases as the size of the data increases, the size of the dataset can be increased so that the overall accuracy increases. Also, the most significant features among the available attributes that actually contribute to the final outcome so that unnecessary attributes are removed and time will be saved which would in turn result in increased performance. This task can be accomplished by taking into consideration every attribute and calculating the R-squared score which signifies how much each attribute is significant for the prediction.

References

1. HDFC. https://www.hdfcergo.com/blogs/general-insurance/importance-of-insurance.html
2. Gupta, S., Tripathi, P.: An emerging trend of big data analytics with health insurance in India. In: 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH), Noida, pp. 64–69 (2016)
3. Kaggle Medical Cost Personal Datasets. Kaggle Inc. https://www.kaggle.com/mirichoi0218/insurance
4. Feature Scaling. https://en.wikipedia.org/wiki/Feature_scaling

5. One Hot Encoding. https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f
6. Kayri, M., Kayri, I., Gencoglu, M.T.: The performance comparison of multiple linear regression, random forest and artificial neural network by using photovoltaic and atmospheric data. In: 2017 14th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, pp. 1–4 (2017)
7. Gumus, M., Kiran, M.S.: Crude oil price forecasting using XGBoost. In: 2017 International Conference on Computer Science and Engineering (UBMK), Antalya (2017)
8. Ian, G., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
9. https://machinelearningmastery.com/what-is-deep-learning/
10. Kansara, D., Singh, R., Sanghvi, D., Kanani, P.: Improving accuracy of real estate valuation using stacked regression. Int. J. Eng. Dev. Res. (IJEDR) **6**(3), 571–577 (2018)
11. Yerpude, P., Gudur, V.: Predictive modelling of crime dataset using data mining. Int. J. Data Min. Knowl. Manag. Process (IJDKP) **7**(4) (2017)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**(Oct), 2825–2830 (2011)
13. Grosan, C., Abraham, A.: Intelligent Systems: A Modern Approach, Intelligent Systems Reference Library Series. Springer, Cham (2011)