IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 8, April 2023

Health Insurance Cost Prediction and Analysis

Jeyakumar D[1], Bharath Teja S[2], Lokesh Reddy Y[3], Ruthik P[4]

Head of the Department, Computer Science and Engineering[1]
Students, Computer Science and Engineering[2,3,4]
Dhanalakshmi College of Engineering, Chennai, India

Abstract: *The Indian authorities spends 1.5 percent annually GDP on public health, which is significantly lower than GDP other countries. On the other hand, global public health spending hand, has almost doubled in line with inflation in the last two decade, reaching $8.5 trillion in 2019, or 9.8%globallyGDP. Transnational multi-private sectors provide around 60%. comprehensive medical treatment and 70% outpatient care, Which price sufferers astronomically excessive fees. Insurance data has increased dramatically in recent times decades and carriers now have access to it. Health insurance system is exploring predictive modelling to support its business operations and services. Computer algorithms and the machine Learning (ML) is used to study and analyse past insurance data and predict new output values based on customer trends behaviour, insurance contracts and data-driven business decisions, and support in formulating new schemes. In addition to it Machine Learning found huge and potential application in insurance industry. It develops real-time insurance costs a price prediction system called ML Health Insurance Prediction System using ML algorithms to help insurance companies on the market easily and quickly determination of premium values and thus limit health expenses. The proposed model includes a demonstration of Random Forest Regression to predict insurance costs and assess model results. In proposed model, achieved a forest regression model better results with R-squared value of 0.80 compared to all other models.*

Keywords: Insurance

I. INTRODUCTION

Public health is an integral part of the society, of the country and of the world we live in and is an important matter of concern. Human lives and public health may be endangered by natural calamities, global epidemic and pandemics, global crisis of medical aids, etc. which increases the vulnerability of public health, unforeseen circumstances at any point of their lifetime. Individuals, families, companies, and properties are uncovered and uninsured to diverse hazard forms, natural calamities, and the likelihood can shift. These perils include the possibility of mortality, health, and property disaster or resource depletion. People's lives revolve around two main elements: life and prosperity. However, keeping a safe and sound distance from unforeseen events is impossible. The Government of India spends 1.5% of GDP on public health, the lowest level globally [1]. In the last two decades, universal health care has been almost doubled, surpassing US $ 8.5 trillion in 2019, or 9.8% of global GDP [2]. The multi-private sector having state-of-the-art facilities provides almost 60% of total hospitalizations and 70% outpatient services [3]. Thus, Health insurance is becoming an essential commodity for every individual because of the increasing cost of quality healthcare combined with higher life expectancy and widespread to alleviate these types ofproblems, the world of fund has developed a style of equipment to guard human beings and agencies from such unseen catastrophic situations, through the utilization of monetary capital to repay and compensate them. In this manner, insurance is an arrangement that diminishes or evacuates misfortune fees delivered via way of means of exceptional dangers.

Today, data has evolved drastically since the recent past decade and insurance carriers have access to it. The health insurance system is exploring ways to use predictive modelling to boost their business operations and services. Computer algorithms and Machine Learning (ML) are used to study and analyse the historical insurance data and predict new output values primarily based totally on developments in patron behaviour, insurance policies and data-driven business decisions, and supports in formulation of new schemes. Besides, most insurance companies use conventional databases to store their data which is primarily structured data. Moreover, merely 10- 15 percent of the total data available is processed for gaining insights. Thus, transformation of the data is necessary to gain valuable

DOI: 10.48175/568

insights that may be very crucial for the growth of such companies. The main advantage of ML is that it can be effectively applied to a massive volume of structured, semi-structured, or unstructured datasets. The ML model can be used across multiple value chains to understand the weightage of risk involved, claims made and customer behaviour with greater predictive accuracy. ML applications in the health insurance sector include various tasks such as understanding risk tolerance and premium Leakage main to inaccurately pricing of premiums, loss deterrence, claims handling, expense management, litigation, and fraud identification.

When it comes to the value of health insurance in people's lives, it is vital for insurance firms to be as explicit as possible for measuring the sum secured by this approach and the protection charges which must be paid for it. The calculation of health insurance charges in the traditional process are a hefty task for the insurance companies. The intervention of humans in this process may sometime produce faulty or inaccurate results. Additionally, as the data increases manual calculations becomes lethargic and time consuming. Again, in such scenarios the implementation of ML models can be beneficial for such companies. Therefore, ML may generalize the exertion or strategy to define such an approach. These models can perform self-learning to predict the cost of insurance using past insurance data of the companies. The model inputs are the main parameters that are utilized to calculate the installments made. This enables the algorithm to precisely estimate the disbursement of insurance coverage. In this way, the correctness can be progressed with ML. The objective of the proposed model is to perform rapid estimation and prediction of coverage expenses at a hospital incurred by a patient, using ML models upon the Kaggle dataset. Thus, this paper develops a real-time insurance cost price prediction system named ML Health Insurance Prediction System using Machine Learning algorithms which will aid the insurance companies in the market for easy and rapid determination of values of premiums and thereby curb down health expenditure. The proposed model incorporates and demonstrates Random Forest Regression to predict the insurance costs and compares the models based on their results.

II. LITERATURE SURVEY

There are several studies on drug price estimation have been published in the field of health in different contexts [4], [5], [6], [7], [8]. Machine learning has many probabilistic assumptions, but its performance depends on choosing a near-exact algorithm for the required trouble area and the following appropriate procedures for building, training, and deploying the model.

Moran et al. [4] "used complex linear regression method that was used to predict the cost of intensive care unit (ICU) using patient profile data, DRG (diagnosis-related groups), length of time spent in health facility and different capabilities like capabilities.'

Sushmita et al. [5] "Proposed a version primarily based totally on a person's medical and past spending history to make predictions about him/her future health costs. Quarterly projected spending on future 3, 6, 9 and 12 months were estimated with usage Model.

Lahiri et al. [6] "used linear regression to evaluate forecast costs. Used a type of algorithm predict whether an individual's medical expenses increase next year, considering medical previous year's expenses."

Gregori et al. [7] used the logit a model An OLS method to study multivariate health care modelling cost data.

III. METHODOLOGY

The regression techniques used are statistical methods that establish an association between the target or dependent variable and a set of independent or predictors variables. It presumes that both the target and the predictor variables have numerical values and there is some sort of relation among the two. The models we are implementations in our problem are described below,

Model Selection

1) Simple Linear Regression: In simple linear regression [16] is the target variable (Y) dependent on one independent variable (X) and the model specifies a linear the connection among the 2 variables. The equation of a straight line is given by:

$$Y = a + bX \qquad (1)$$

where "a" and "b" are called model parameters as regression coefficients, "a" is the Y-intercept value which the line forms when X equals zero and "b" is the slope that is, change in Y with change in X. More the "b" value means that a minor change in X will cause a significant one change in Y and vice versa. The value of "a" and "b" can be found by the method of ordinary Least Square Method.

In this model, there can be only predicted values not always exact. There will always be some difference hence we include an error term to the original equation (1) that results for the difference and thus help in making better predictions.

$$Y = a + bX + \varepsilon$$

Assumptions in Linear Regression

- The sample size of data should exceed the number of available parameters.
- Only over a restricted range of data the regression can be valid.
- Error term is normally distributed. This also means that the mean of the error has expected value of 0.

2) Multiple Linear Regression: Similar to simple linear regression, multiple regression [17] is a statistical procedure that examines the degree of association between a set of unbiased variables and a structured variable. In multiple linear regression and the value of dependent variable is now calculated depending on the values of the predictor variables. It is assumed that there is no dependency among the predictor variables. Suppose if the target value is dependent on ‘n’ independent variables then the regressor fits the regression line in a N dimensional space.

The regressor line equation is now modified into

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots.. + b_nX_n + \quad \varepsilon$$

where “a” is the Y-intercept value and < b_1, b_2, b_3, …., bn > are the regression coefficients associated with the n independent variables and is the error term.

3) Polynomial Regression: Polynomial Regression [18] is another special case of linear regression. In Linear regression the model tries to fit a straight regression line among the established and unbiased variable. In scenarios where there doesn’t exist linear relationship between the target and predictor variable then instead of a straight line a curve is being outfitted towards the 2 variables. This is accomplished by fitting a polynomial equation of degree n on the non-linear data which establishes a curvilinear relationship among the dependent and independent variables. In polynomial regression the assumption that the independent variables must be unbiased of every different isn't mandatory. The equation of the line thus reduces to:

$$Y = a + b_1X^2 + b_2X^{\,2} + b_3X^{\,3} + \ldots. + b_nX^{\,n} + \varepsilon \qquad (4)$$

The following are some of the benefits of applying polynomial regression:

- Polynomial Regression offers the best estimate of the courting among the established and unbiased variable.
- Higher degree polynomial generally provides a good fit on the dataset. Polynomial Regression essentially suits an extensive variety of curves of varying degree to the dataset.

Drawbacks of applying Polynomial Regression:

- These are too sensitive to the existence of outliers in the dataset, as outliers cause the model's variance to rise. When the model comes across an unknown data item, it underperforms.

4) Ridge Regression: Ridge regression [19] is a standard model tuning process used to analyse the data suffering from multicollinearity. This is a way to approximate the coefficients of the regression model when the independent variables are firmly related or there exists an association between them. Ridge Regression's main objective is to take the dataset and fit a new line into it in a way that does not overfit the model. For this purpose, ridge regression adds an insignificant amount of bias that determines the fitting of the line into the data. We obtain a substantial reduction in variance, which leads to an increase in the accuracy value by the addition of bias. The least Square determines the values of the parameters for the equation (1), which diminishes the sum of squared residuals. But in contrast the Ridge

DOI: 10.48175/568

Regression regulates the value for parameters that results in minimization of the sum of squared residuals along with an additional term λ*b2. Ridge Regression performs L2 regularization.

$$Loss_{Ridge} = \Sigma\ (y_i - y^*)^2 + \lambda b^2 \qquad (5)$$

where y* = a + bX is the predicted value.

In this ridge regression method, the coefficients are penalized by a value lambda, this acts as a control parameter, which determines how severe is the penalty and how much significance should be given to Xi. The higher the values of the lambda the bigger is the penalty and therefore the magnitude of coefficients is reduced.

When the slope (b) of the line is steep then the target variable(Y) is very sensitive to relatively small change in the predictor variable variable(X). In ridge regression by the addition of the lambda value the sensitivity decreases. If lambda is zero then ridge regression reduces to linear regression and when lambda increases gradually the slope the line decreases asymptotically. To know which value of lambda is to choose we try different values of lambda and use cross validation to determine which one result in the lowest variance.

5) Lasso Regression: Least Absolute Shrinkage and Selection Operator (LASSO) [20] is appropriate to ridge regression. It performs L1 regularization. The Lasso regression process is usually used in machine learning for the selection of the significant subset of variables. The prediction accuracy of this model is higher when compared to other model interpretations. Like ridge regression lasso regression also results in a line with little amount of bias added to it which thereby decreases the variance of the model.

$$\mathrm{Loss}_{\mathrm{Lasso}} = \Sigma\ (y_i - y^*)^2 + \lambda|b|$$

The major difference between lasso and ridge regression is, ridge regression decreases the slope asymptotically close to zero but lasso regression can reduce the slope all the way down to zero, thereby eliminating the useless parameters from the line equation that do not have any significant role for predicting the value of the target variable.

Lasso regression usually works better under conditions where some predictors have high coefficients, and the rest have low coefficients. Ridge regression performs better when the result is a function of many predictors, all of which have coefficients of approximately the same size.

IV. PROPOSED METHODOLOGY

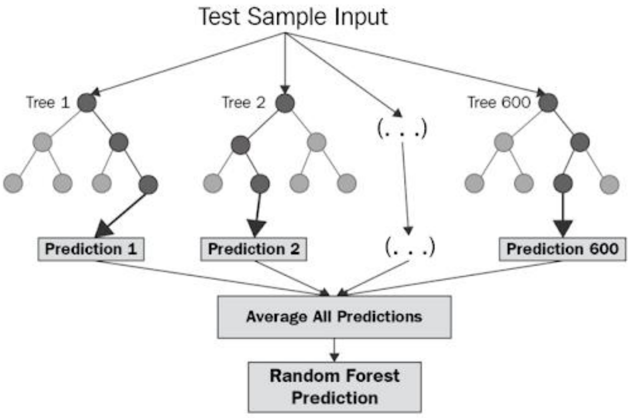
4.1 Random Forest Regression

Random forest is a supervised getting to know set of rules that makes use of an ensemble getting to know approach for class and regression. Random forest is a bagging approach and now no longer a boosting approach. The trees in random forests run in parallel, which means isn't any interplay among those trees while it constructing the trees. Random forest operates with the aid of using building a mess of choice trees at education time and outputting the elegance that's the mode of the classes (classification) or imply prediction (regression) of the individual trees.The same data set that was used for the Decision Tree Regression is utilized in this where we have one-independent patient health condition and insurance Revenue which we have to predict the expected insurance for the patient save primarily based totally on the data given by the patient.

$$\frac{1}{N}\sum_{i=1}^{N}|y_i - \mu|$$

y_i is label for an instance
N is the number of instances and
μ is the mean given by $1/N\ \Sigma_{i=1}^{N} y_i$

Every choice tree has excessive variance, however while we integrate, they all collectively in parallel then the ensuing variance is low as every choice Tree receives flawlessly educated on that unique pattern data,and consequently the output doesn't depend upon one choice tree however on a couple of choice trees.In the case of a class problem, the very last output is taken with the aid of using the use of the bulk balloting classifier.In the case of a regression problem, the very last output is the imply of all of the outputs.



4.2 Description of the Dataset

From the Kaggle site [21] we obtained our dataset for developing the ML Health Insurance Prediction System (MLHIPS). The data set obtained contains seven attributes or features and 1338 rows; out of the seven attributes or features three of them contains categorical values and the rest contain numerical values.The first component is referred to as training data, while the second is referred to as test data. The more data that is supplied to the version at some point of its schooling period, the greater correct the model will be when making predictions on unseen data. Data is typically split at a ratio of 80:20 for testing and training purposes.

Training datasets are used to build models as predictors of health insurance costs, and test sets are used to evaluate regression models. Table.1 displays the dataset's description. The dataset contained missing values in certain fields. After reviewing the distributions, it was decided to replace the missing variables with new attributes, implying that the data is missing. [9]. This is only possible if the data is lost completely at random; thus, the lacking records of data mechanism, which determines the best method to data processing, must first be developed. [10] [11]. The multilevel structure and hidden dependencies are available in medical data [12]. It is vital to figure out these hidden patterns and use various fundamental analysis techniques present in a combined fashion. This is the reason why in the medical data analysis, many researchers use different ensemble Machine Learning models. In order to address the hassle of fee prediction, researchers have also used hierarchical regression analysis. Many of them have used different ensemble learning techniques such as Random Forests, Adaboost, GBM, and XGBM. In paper [13], to forecast the modulus of elasticity of the collective recycled concrete, an ensemble of Random Forests (RF) and Support Vector Machines (SVM) is utilized. This classical ensemble has a much higher level of precision. To take advantage of the heterogeneity of distinct sets of meta-features, an ensemble of K-Nearest Neighbour (KNN) classifiers was created for recommendation purposes is analyzed in [14].

For the diabetic retinopathy dataset, researchers used an ensemble-based machine learning model that included ID3, Random Forests, Adaboost, Logistic Regression and KNN, [15].

4.3 Data Pre-processing:

The dataset contains seven variables, as shown in the table above. While calculating the cost of the Charges of a customer which is our target variable the values of the rest six of the variables are taken into consideration. In this phase, the data is reviewed, properly reconstructed, and properly applied to machine learning algorithms. The dataset will first check for missing values. The dataset was found containing missing values in the bmi and conditions columns. As regression models accept only numerical data, the categorical columns in our case the gender, smoker containing categorical columns were converted into numerical values using label encoding. Then the updated dataset was partitioned into training and testing dataset. And the model was trained using the training dataset.

DOI: 10.48175/568

TABLE I. DATA DESCRIPTION

Name	Description
Age	Patient's Age
BMI	Body mass index of the patient
Number of kids	Number of kids of the patient
Gender	Male / Female
Smoker	Whether the patient is smoker or not.
Condition	Cancer Accident HIV/AIDS Diabetes/Heart Ailments
Charges	Medical fee the patient has to pay

V. RESULTS and OUTPUT

The regression model's performance is evaluated on the basis of the following metrics

R^2_Score

Root Mean Square Error (RMSE)

R^2_Score: R-Squared is a good measure to evaluate the model fitness. The R-squared value lies between 0 to 1 (0% to 100%). Large value represents a better fit.

$R^2 = 1 – SSE/SST(7)$

TABLE II. CALCULATED RESULTS

Regressor	R2_Score	RMSE	Accuracy (in %)
Simple Linear Regression	0.62	7523.98	62.86
Multiple Linear Regression	0.75	7523.98	75.86
Polynomial Regression	0.80	5100.53	80.97
Ridge	0.75	6070.80	75.82
Lasso Regression	0.75	6066.31	75.86
Random Forest Regression	**0.86**	**1000**	**86.07**

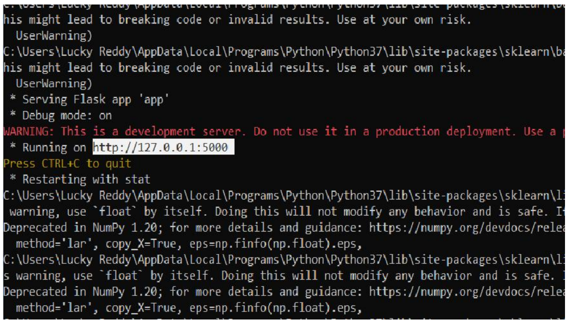
The outcomes of the above discussed models after testing the models on the test dataset were noted.

Table.2 displays the R^2, RMSE and Accuracy metrices of the models discussed so far.

The output projects how the model is performed using python as the programming language to predict results. By using software, we are able to produce the results.

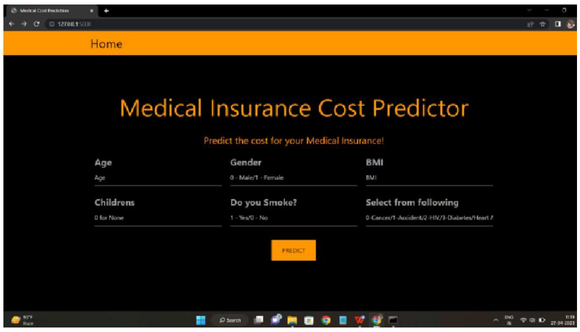
When a person needs to know the insurance covered according to the health condition, their conditions are taken into consideration to predict and produce the result.

Activating the server link using Flask



```
his might lead to breaking code or invalid results. Use at your own risk.
  UserWarning)
C:\Users\Lucky Reddy\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\ba
his might lead to breaking code or invalid results. Use at your own risk.
  UserWarning)
 * Serving Flask app 'app'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with stat
C:\Users\Lucky Reddy\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\li
 warning, use `float` by itself. Doing this will not modify any behavior and is safe. I
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/relea
  method='lar', copy_X=True, eps=np.finfo(np.float).eps,
C:\Users\Lucky Reddy\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\li
s warning, use `float` by itself. Doing this will not modify any behavior and is safe. I
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/relea
  method='lar', copy_X=True, eps=np.finfo(np.float).eps,
```

Opening the link to access predictor



Medical Insurance Cost Predictor

Predict the cost for your Medical Insurance!

Age: Age

Gender: 0 - Male/1 - Female

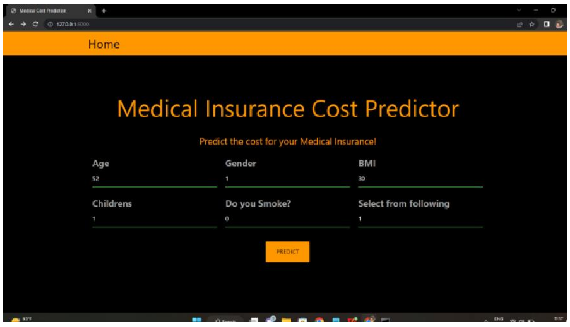
BMI: BMI

Childrens: 0 for None

Do you Smoke?: 1 - Yes/0 - No

Select from following: 0-Cancer/1-Accident/2-HIV/3-Diabetes/Heart A

Providing data



Medical Insurance Cost Predictor

Predict the cost for your Medical Insurance!

Age: 52

Gender: 1

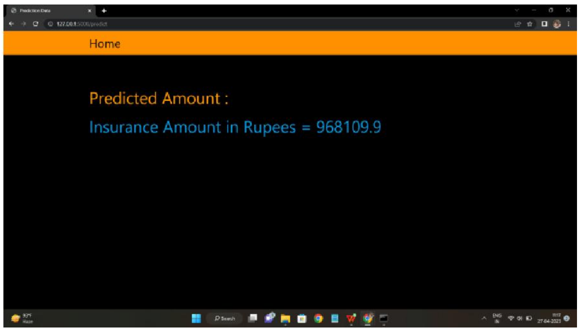
BMI: 30

Childrens: 1

Do you Smoke?: 0

Select from following: 1

Predicted Result



Predicted Amount :

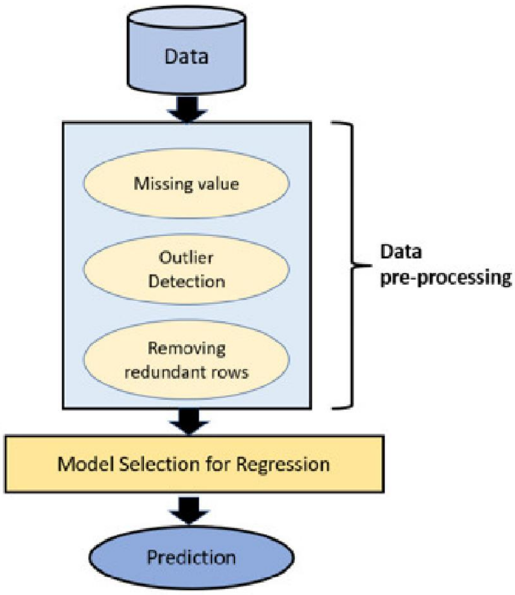
Insurance Amount in Rupees = 968109.9

VI. OBSERVATIONS

From the preceding calculations, Polynomial Regression outperforms other models for the proposed MLHIPS giving an accuracy of 80.97%. The Polynomial Regression in contrast to other models fits a curve to the dataset which increases the variance of the model thereby reducing the residual error. With an RMSE of 5100.53 and R2 value of 0.80, the model achieves superior results when using Polynomial Regression. If there exists a nonlinear relationship between the target and set of predictor variables then polynomial regression eventually produces better results. Besides, Polynomial Regression Ridge and Lasso Regression have achieved an accurate value of 75.82% and 75.86%, respectively. The 6 independent parameters have a considerable correlation amongst them as a result of which lasso regression and ridge

DOI: 10.48175/568

regression produces similar results. The Multiple linear regression model has also achieved an accuracy of 75.86%. whereas the simple linear regression had an accuracy of 62.86% which was the lowest among all.



In the above scenario, the dataset given is partitioned into 80-20 ratio for the purpose of training and testing.In the dataset of 70:30 ratio, the accuracy of the polynomial regression decreases from previous value of 80.97% to 80.54% which is negligible, and similarly the rest of the models have also produced a drop in the accuracy values. Further it was noted that on increasing the degree of the polynomial regression from n=2 to n=3 the accuracy has increased from previous value of 80.97% to 83.62%. But later, further increasing the degree to 4 and higher values there was a drop in accuracy from 83.62% to 68.06% for degree n=4 and 51.98% for degree n=5. Thus, degree n=3 the polynomial regression gives us a good accuracy in predicting the charges. A working of the MLHIPS model is shown in Fig.1.

VII. FUTURE WORK

In this paper some of the traditional regression models are discussed for our proposed problem statement, moving forward some of the other techniques like Support Vector Machine (SVM), XGBoost, Decision Tree (CART), Forest Classifier and Stochastic Gradient Boosting needs to be addressed as the future work.

Genetic Algorithm or the Gradient Descent Algorithm maybe applied on top of model evaluation as they are few of the optimization techniques. We can also apply some feature selection techniques to our dataset before we train our model to gain a good accuracy value as some of the features may be omitted while predicting the charges. Besides a model to perform well a good balanced dataset with a greater number of observations is required which will reduce the variability of the model so in the future if, we get more data than the model can be trained well.

VIII. CONCLUSION

Calculation of health insurance charges in the traditional process are a hefty task for the insurance companies. Following human intervention in the process may sometime produce faulty or inaccurate results and also when the data increases the time taken for calculation by human's increases. In scenarios like these the implementation of Machine Learning models can be very beneficial to the company. In this paper, several Machine Learning regression models are used to predict the cost of health insurance based on specific attribute values present in the dataset. The results obtained are summarized in Table II. With an RMSE of 5100.53, an R2 of 0.80, and an accuracy of 80.97 percent, Polynomial Regression is the most efficient. Based on the model's configuration parameters which are tuned during the training phase, on the basis of performance the different proposed models are arranged accordingly starting from Polynomial Regression, Lasso and Ridge Regression, Multiple Linear Regression, and Simple Linear Regression. These models can be incorporated by the companies for calculating the charges in a fast and reliant manner thereby saving the time and cost of the company. These models in the later stage of life cycle can be deployed onto the cloud platforms when the data increases integrated with high end computing resources for faster processing of real time data in short span of time interval.

REFERENCES

[1]. "National Health Accounts," National Health Systems Resource Centre. [Online]. Available:https://nhsrcindia.org/national-health-accounts records care: an introductory review," International Journal for Quality in Health Care, vol. 23, no. 3, pp. 331–341, 2011.

[2]. Bertsimas, M. V. Bjarnad´ottir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, "Algorithmic prediction of health-care costs," Operations Research, vol. 56, no. 6, pp. 1382–1392, 2008.

[3]. Stucki, O. "Predicting the customer churn with machine learning methods: case: private insurance customer data" Master's dissertation, LUT University, Lappeenranta, Finland, 2019.

[4]. Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338L.

[5]. H. Demirtas, "Flexible Imputation of Missing Data", J. Stat. Soft., vol. 85, no. 4, pp. 1–5, Jul. 2018. Available: DOI: 10.18637/jss.v085.b04 .

[6]. H. Goldstein, W. Browne and J. Rasbash, "Multilevel modelling of medical data," Statistics in Medicine, John Wiley and Sons, vol. 21, no. 21, pp. 3291–3315, 2002.

[7]. T. Han, A. Siddique, K. Khayat, J. Huang and A. Kumar, "An ensemble machine learning approach for prediction and optimization of modulus of elasticity of recycled aggregate concrete," Construction and Building Materials, vol. 244, pp. 118–271, 2020.

[8]. X. Zhu, C. Ying, J. Wang, J. Li, X. Lai et al., "Ensemble of ML-kNN for classification algorithm recommendation," Knowledge-Based Systems, vol. 106, pp. 933, 2021.

[9]. G. Reddy, S. Bhattacharya, S. Ramakrishnan, C. L. Chowdhary, S. Hakak et al., "An ensemble-based machine learning model for diabetic retinopathy classification," in 2020 Int. Conf. on Emerging Trends in Information Technology and Engineering, IC-ETITE, VIT Vellore, IEEE, pp. 1–6, 2020.

[10]. Douglas C Montgomery, Elizabeth A Peck and G Geoffrey Vining, "Introduction to linear regression analysis", John Wiley & Sons, vol. 821, 2012.

[11]. Tian Jinyu, Zhao Xin et al., "Apply multiple linear regression model to predict the audit opinion," in 2009 ISECS International Colloquium on Computing, Communication, Control, and Management, IEEE, pp.1–6, 2009.

[12]. Ostertagova et al., "Modelling using Polynomial Regression", "Procedia Engineering", vol. 48, pp. 500-506, 2012.

[13]. Donald W. Marquardt, Ronald D. Sneet al., "Ridge Regression in Practice", "The American Statistician", vol. 29, pp – 3-20, 2012.

[14]. V. Roth, "The generalised LASSO"," IEEE Transactions on Neural Networks", vol. 15, pp – 16 28, 2004.