CSE3505 – Essentials of Data Analytics

PROJECT REPORT

Medical Insurance Cost Prediction

BY

19BCE1442 **ANKAN ROY**

19BCE1522 **MOHIT KAUSHIK**

B. Tech Computer Science and Engineering

Submitted to

SIVAKUMAR R.

School of Computer Science and Engineering

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABSTRACT

A health insurance company can only make money if it collects more than it spends on the medical care of its beneficiaries. On the other hand, even though some conditions are more prevalent for certain segments of the population, medical costs are difficult to predict since most money comes from rare conditions of the patients. The objective of this article is to accurately predict insurance costs based on people's data, including age, Body Mass Index, smoking or not, etc. Additionally, we will also determine what the most important variable influencing insurance costs is. These estimates could be used to create actuarial tables that set the price of yearly premiums higher or lower according to the expected treatment costs. This is a regression problem. A hierarchy perception structure was suggested to choose significant words, health checks, and diagnoses for training phase informative data representations, because various words, diagnoses, and previous health care have varying significance for expense calculation. In this system model, linear regression analysis, naive Bayes classifier, and random forest algorithms were compared using a business analytic method that applied statistical and machine-learning approaches. According to the results of our forecasting method, linear regression has the maximum accuracy of 97.89 percent in forecasting overall healthcare costs. In terms of financial statistics, our methodology provides a predictive method.

1. Introduction

The incidence of overweight and obesity has increased significantly in most countries in recent decades. Excess weight is associated with an increased incidence of many chronic diseases, including vascular disease, respiratory disease, osteoarthritis, some cancer, type 2 diabetes, and premature death. There is consistent evidence that an increased BMI is associated with higher health costs, and these costs are expected to increase as obesity. Modeling uses machine-learning methods, in which the machine learns from the data and uses it to forecast new data. The most commonly predictive analytic model used is regression. The proposed model for accurate prediction of future outputs has applications in banking, economics, e-commerce, sports, business, entertainment, etc. A method used to forecast healthcare costs for BMI is based on several factors. Multiple linear regression is one of the statistical techniques for estimating the relationship among the dependent (target) and independent variables. The regression method is commonly used to develop a system based on a number of factors to predict the cost.

The regression analysis is performed to determine the relationship among two or more variables with cause-effect relationships and to make predictions for the topic using the relationships. If regression used one independent variable, then it is known as univariate regression analysis, or else if it used more than two independent variables then it is known as multivariate regression analysis. Linear regression involves initially uploading the data and then analysing the data. Subsequently, the data are cut, and then, the data are trained and separated to create the model. At last, it will evaluate the accuracy. The main aim of regression is to develop an efficient technique for predicting dependent properties from a set of characteristic variables. A regression problem is the actual or continuous value of the output variables, that is, area, salary, and weight. Regression can be defined as a statistical method used in applications such as predicting the healthcare costs. Regression is used to predict the relationship among the dependent variable and set of independent variables. There are various types of regression techniques available namely simple linear regression, multiple linear regression, polynomial regression, support vector regression, and random forest regression.

Fast-growing healthcare costs have become a significant challenge in several developed countries. Existing evidence suggests that healthcare costs have accumulated among a large number of BMI. Even though experiments have attempted to develop accurate models for predicting healthcare costs for BMI, their effectiveness is excellent due to the lack of detailed clinical information in the data used to create complex intervals and prognostic models. ad to improvements in the prognostic model.

1.1 Objective and goal of the project

Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Various factors influence the cost of insurance.

The objective of this project is to accurately predict insurance costs based on people's data, including age, Body Mass Index, smoking or not, etc.

1.2 Problem Statement

Everyone's life revolves around their health. Good health is essential to all aspects of our lives. Health refers to a person's ability to cope up with the environment on a physical, emotional, mental, and social level. Because of the quick speed of our lives, we are adopting many habits that are harming our health. One spends a lot of money to be healthy by participating in physical activities or having frequent health check-ups to avoid being unfit and get rid of health disorders. When we become ill we tend to spend a lot of money, resulting in a lot of medical expenses.

So, an application can be made which can make people understand the factors which are making them unfit, and creating a lot of medical expenses, and it could identify and estimate medical expense if someone has such factors.

2. Literature Survey

Some of the recent literature that describes the various mechanism of estimating the costs of physical healthcare is summarized below. It is critical to implement tailored treatment programs for high hazard patients of readmission in an attempt to prevent readmissions and lower healthcare costs. This necessitates recognizing high individuals at the time of hospital release. Using specialist characteristics and situational integration of medical knowledge provides a cost-sensitive implementation of the long short-term memory neural net. Using both machine-derived and professional characteristics, including frequent patterns, and resolving the issue of class imbalances, this research focuses on important parts of an EHR-driven forecasting system in a single framework. We assess each element's impact on forecasting effectiveness and price benefits. In at least 2 evaluating criteria, it shows that the technique with all critical features outperforms the simplified approaches in terms of discriminating capability. Researchers also propose a basic economic assessment to predict annual income if high-risk patients are provided tailored therapies.

Patients with heart failure (HF) require precise hazard classification to implement tailored therapies focused on enhancing their efficiency of living. To assess the economic benefit of complementing claim-based forecasting analytics with electronic medical record -derived data and to contrast machine-learning techniques to conventional logistic regression in forecasting critical results in patients with HF, healthcare patients with HF from 2 healthcare professional systems in Massachusetts, Boston, were included in predictive research with a one-year follow-up duration. "Providers" comprise therapists, various medical professionals, clinicians, and their organization including the network. Logistic regression, gradient boosted modeling, regression trees, random forests, least absolute shrinkage, classification, and selection operation regression were used to predict all-cause morbidity, top cost decile, HF hospitalization, gradient

boosted modelling, and home days loss larger than 25%. Information from network 1 was used to educate all algorithms, which were then evaluated in network 2. The area under high accuracy curves (AUPRCs) and overall value estimations from decision curves were obtained after choosing the best effective modeling strategy depending on the Brier score, calibration, and discrimination.

The goal of this study was to evaluate the effectiveness of machine-learning methodologies for predicting healthcare expenses connected with spinal fusion in aspects of gains or losses and to use these techniques to investigate the major features connected with spinal fusion medical costs.

Because of the ageing populations and enhanced therapy of fundamental conditions, cardiac arrest is among the most complicated chronic disorders with a higher incidence. The incidence is projected to gradually climb, reaching 3% of the population in Western countries [24]. It is the leading reason for hospitalizations in people aged 65 and above, leading to substantial expenses and a significant societal effect. In the therapy of HF, the present "one-size-fits-all" strategy does not produce the optimal results for all patients. These facts pose a serious danger to the proper treatment of heart failure patients. It will take an unconventional method from a unique perspective on health care. We offer a unique forecasting, preventive, and personalized healthcare strategy, in which patients are actually in charge of their care, aided by a user-friendly online form that employs artificial intelligence (AI). This technique study outlines the demands in HF care, as well as the necessary paradigm shift and the factors necessary to make it happen. A digital physician is being developed through an exciting combination of medical and high-tech partners from patient coaching, serious gaming, North-West Europe, artificial intelligence, and combining state-of-the-art HF health care. The findings are intended to improve and customize self-care, in which patients conduct routine care chores without the intervention of healthcare experts, allowing them to focus on more difficult problems. This innovative approach to health care will lower prices per patient while increasing results, ensuring the long-term viability of top-tier HF health care.

DRG codes are useful for price tracking and allocation of resources since healthcare operators obtain predetermined levels of compensation for certain treatments under diagnosis-related group (DRG) payments. Coding, on the other hand, is usually done after the fact, after the patient has been discharged. They want to use normal medical text to forecast DRGs and DRG-based case mix index (CMI) at initial inpatient admission to forecast

hospital costs in an acute context. Without manual coding, a deep learning-based natural language processing (NLP) method is tested to forecast cost-reflecting weights and per-episode DRGs on 2 cohorts (paid by All Patient Refined (APR) DRG or Medicare Severity (MS) DRG). In fivefold cross-validation trials on the first day of ICU admission, it attained macro-averaged area under the receiver operating characteristic curve (AUC) scores of 0•871 (SD 0•011) on MS-DRG and 0•884 (0•003) on APR-DRG. When applied to hypothetical patient populations to predict average cost-reflecting weights, the algorithm improved over time, yielding absolute CMI errors of 12•79 (2•31%) and 2•40 (1•07%) on the first day, correspondingly. Because the system can adjust to changes in admission time and cohort size while requiring no additional manual coding, it has the potential to aid in cost estimation for active patients and enable improved functional outcome in hospitals.

3. Requirements Specification

3.1 Hardware Requirements

No hardware component required for running this project

3.2 Software Requirements

Software Requirements deal with defining software resource requirements and prerequisites that need to be installed on a computer to provide optimal functioning of an application. These requirements or pre-requisites are generally not included in the software installation package and need to be installed separately before the software is installed. The software requirements that are required for this project are:

- R Studio
- Required Libraries

4. Implementation

4.1 Dataset Description and Exploratory Data Analysis

The medical insurance charges dataset was selected. It consists of 7 attributes and 1338 vectors. The task is to predict individual payments for health insurance.

Data preprocessing for mentioned dataset consists of the following stages:

- Missing data imputation
- Data transformation

In the missing data imputation stage MICE algorithm [30] is used. For data transformation stage one-hot encoding is used for binary (sex, smoker) and categorical (region) variables.
Finally, the dataset consists of 11 features, namely:

- As we can see, we got these features:

`age`: age of the primary beneficiary

`sex` : insurance contractor gender, female, male

`bmi`: Body Mass Index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9

`children`: number of children covered by health insurance, number of dependents

`smoker`: smoking or not

`region`: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

`charges`: individual medical costs billed by health insurance

The next step is feature selection. To do this, Pierson coefficient is used . A significant correlation between features is absent. However, smokers correlated with the target variable charges. For non-smoker patients, the correlation between bmi and charges is not clear.

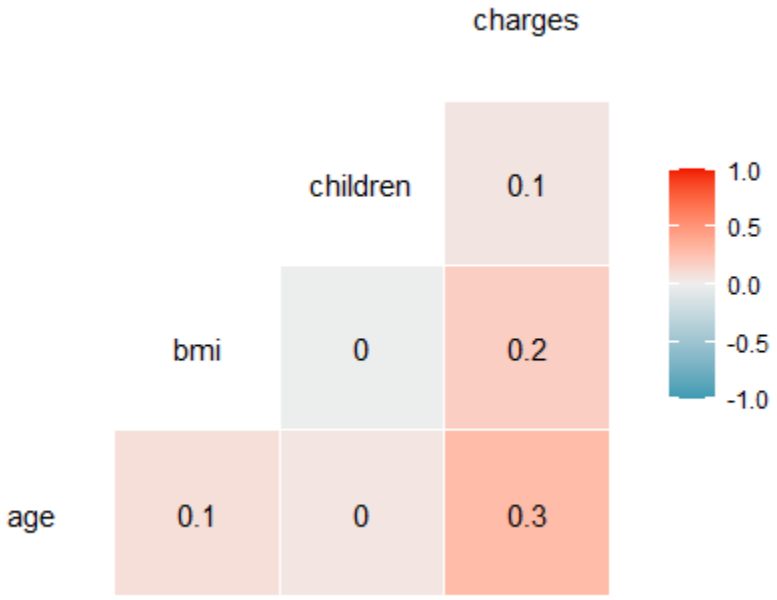
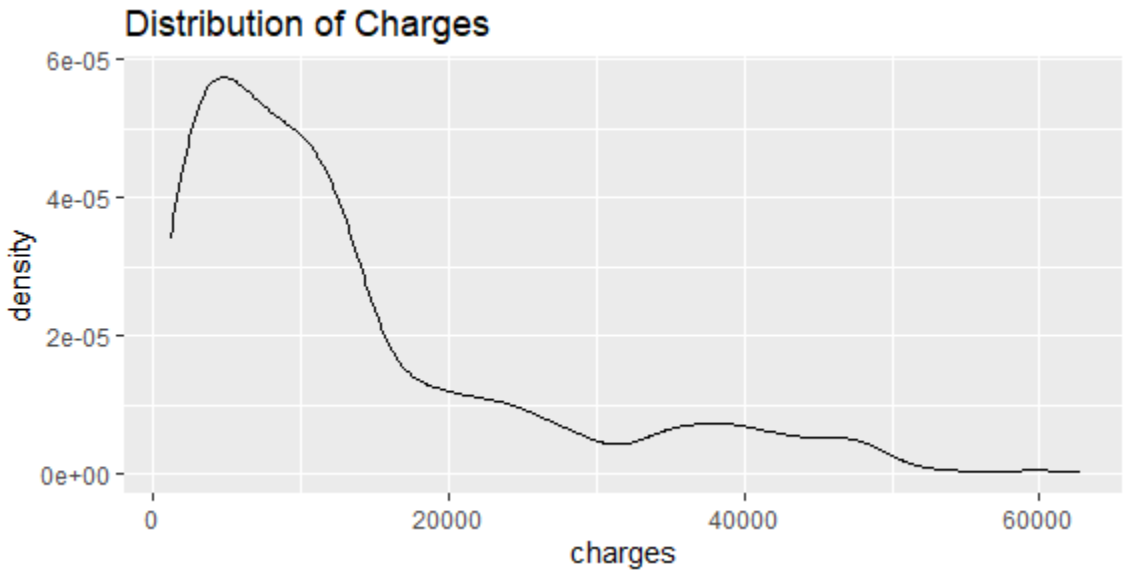


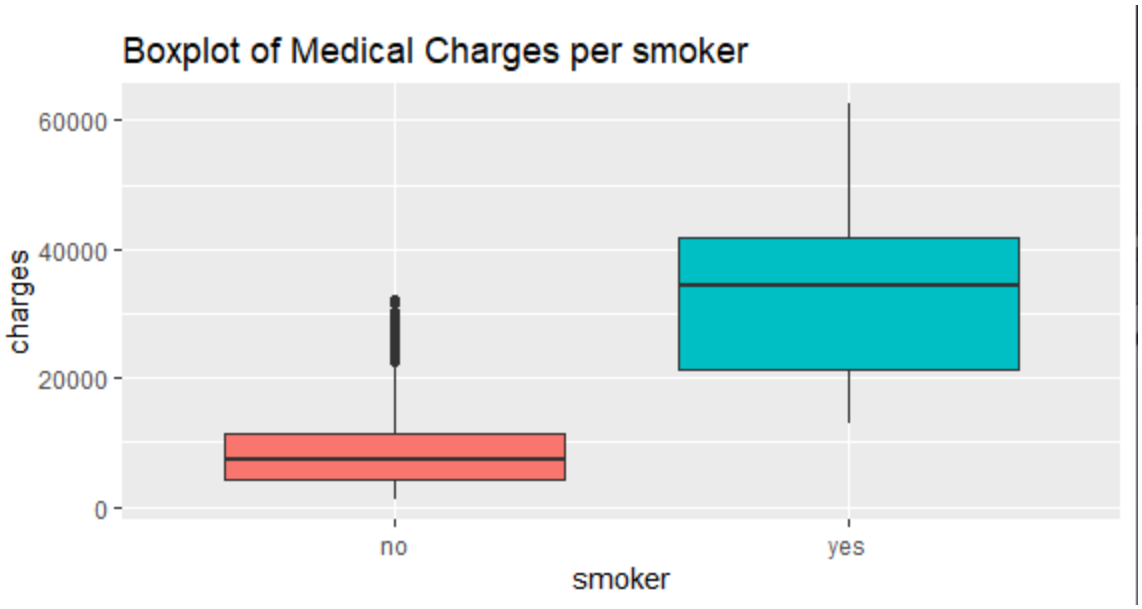
Fig. Correlation Matrix

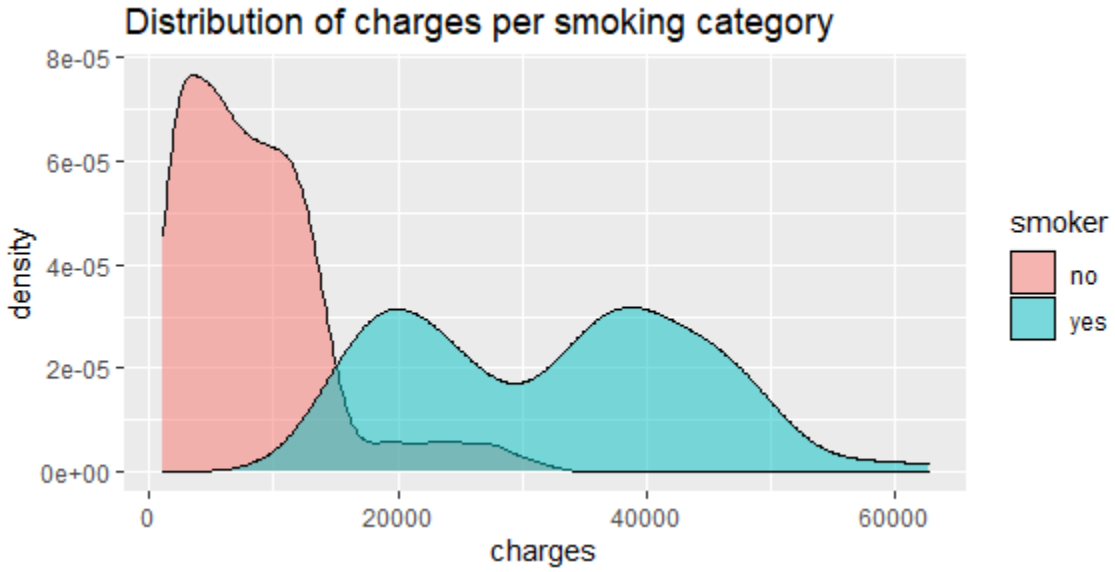
4.2 Data Exploration and Visualization

```
> summary(train)
      age            sex                 bmi           children        smoker
 Min.   :18.00   Length:955         Min.   :16.82   Min.   :0.000   Length:955
 1st Qu.:27.00   Class :character   1st Qu.:26.12   1st Qu.:0.000   Class :character
 Median :40.00   Mode  :character   Median :30.50   Median :1.000   Mode  :character
 Mean   :39.68                      Mean   :30.70   Mean   :1.074
 3rd Qu.:52.00                      3rd Qu.:34.96   3rd Qu.:2.000
 Max.   :64.00                      Max.   :53.13   Max.   :5.000
    region             charges
 Length:955         Min.   : 1132
 Class :character   1st Qu.: 4868
 Mode  :character   Median : 9550
                    Mean   :13413
                    3rd Qu.:16841
                    Max.   :62593
> |
```

In terms of categorical features, the dataset has a similar number of people for each category, except for smoker. We have more non-smokers than smokers, which makes sense. The charges itself varies greatly from around $1,000 to $64,000.



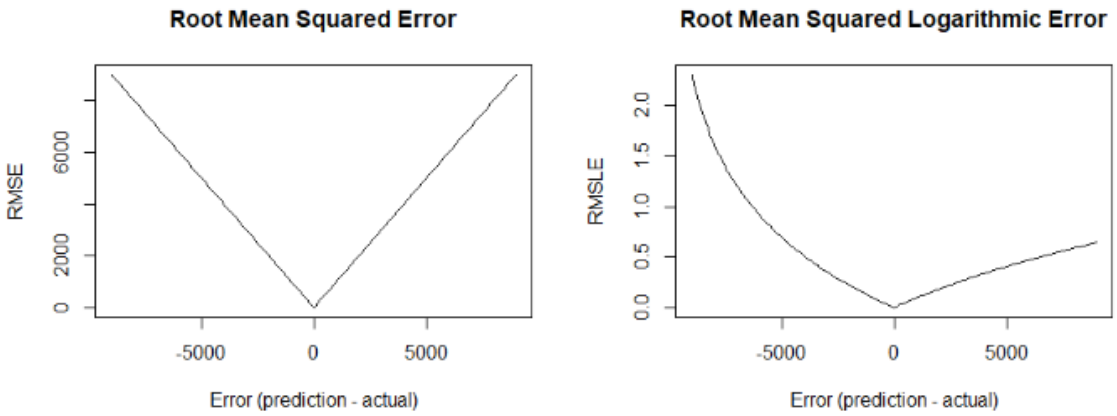




4.3 Metrics and Validation Strategy

We will use Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Root Mean Squared Logarithmic Error (RMSLE) as our metrics. These three metrics can be used depends on the business

point of view. To see what we mean, consider the true value of one observation of charges be $10,000. Assume the model predictions are exactly the same as true values, except for this particular observation which the model predicts as *x*. We will vary *x* from $1,000 to $19,000 and see the resulted error.



As we can see, RMSLE incurs a larger penalty for the underestimation of the actual variable than the overestimation. Also, RMSLE metric only considers the relative error between the predicted and the actual value, and the scale of the error is not significant. On the other hand, RMSE value increases in magnitude if the scale of error increases. This means RMSLE should be more useful than RMSE when underestimation is undesirable.

MAE and RMSE are indifferent to the direction of errors. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful than MAE when large errors are particularly undesirable.

Knowing the metrics, we validate the performance of our model simply by applying new data `test.csv` to it and see the metrics score. We don't do k-fold cross validation since the data is small.

4.4 Modeling

We will build and train Linear Regression model for this problem. For starters, let's use all available features in the model.

Linear Regression

Linear Regression will be implemented with automatic feature selection using *backward elimination*. Starting from using all features, the *backward elimination* process will iteratively discard some and evaluate the model until it finds one with the lowest Akaike Information Criterion (AIC). Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models based on information loss. Lower AIC means better model. We'll use `step()` function to apply *backward elimination* in a greedy manner.

```
> step(temp)
Start:  AIC=23868.38
data1_smoker$charges ~ data1_smoker$smoker

                      Df  Sum of Sq        RSS   AIC
<none>                                7.4554e+10 23868
- data1_smoker$smoker  1 1.2152e+11 1.9607e+11 25160

Call:
lm(formula = data1_smoker$charges ~ data1_smoker$smoker, data = train)

Coefficients:
        (Intercept)  data1_smoker$smoker
              -15182                23616

> summary(temp)

Call:
lm(formula = data1_smoker$charges ~ data1_smoker$smoker, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-19221  -5042   -919   3705  31720

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -15181.7      643.0  -23.61   <2e-16 ***
data1_smoker$smoker  23616.0      506.1   46.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7470 on 1336 degrees of freedom
Multiple R-squared:  0.6198,    Adjusted R-squared:  0.6195
F-statistic:  2178 on 1 and 1336 DF,  p-value: < 2.2e-16

> summary(temp)$r.squared
[1] 0.6197648
>
```

Polynomial Regression

We can improve our model by *feature engineering*, specifically, by making new features that capture the interactions between existing features. This is called polynomial regression. The idea is to generate a new feature matrix consisting of all polynomial combinations of the features with degrees less than or equal to the specified degree. For example, if an input sample is two-

dimensional and of the form [a, b], the degree-2 polynomial features are [1, a, b, a^2, ab, b^2]. We will use degree 2.

We don't want charges to be included in the process of generating the polynomial combinations, so we take out charges from train and test and save them as y_train and y_test, respectively.From EDA we know that sex and region have no correlation with charges. We can drop them. Also, since polynomial combinations don't make sense to categorical features, we mutate smoker as numeric.

We can see that our new datasets train_poly and test_poly now have 16 columns:

- (Intercept) is a column consists of constant 1, this is the constant term in the polynomial.
- age , bmi , children , smoker are the original features.
- age² , bmi² , children² , smoker² are the square of the original features.
- age x bmi, age x children , age x smoker , bmi x children , bmi x smoker , children x smoker are six interactions between pairs of four features.
- charges is the target feature.

Now, we are ready to make the model. As usual, we start with all features and work our way down using *backward elimination*.

Let’s see the summary of our new Regression model.

```
> summary(lm_all)

Call:
lm(formula = charges ~ age + bmi + children + smoker, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12120.9  -2826.9   -971.4   1509.4  29374.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11797.53    1116.52  -10.57  < 2e-16 ***
age            252.33      13.91   18.14  < 2e-16 ***
bmi            317.39      31.93    9.94  < 2e-16 ***
children       465.19     162.64    2.86  0.00433 **
smokeryes    24024.90     483.65   49.67  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6046 on 950 degrees of freedom
Multiple R-squared:  0.7529,    Adjusted R-squared:  0.7518
F-statistic: 723.5 on 4 and 950 DF,  p-value: < 2.2e-16

> summary(lm_all)$r.squared
[1] 0.7528564
```

Conclusion

As it shown, we can make the conclusion that adding more features to the linear Regression increases the accuracy of the given model. The measures the average magnitude of the errors in a set of forecasts is 10270.2207578 and the root mean square deviation value is 14517.5488946.

FUTURE SCOPE

- The constructed algorithm can be made even more advanced by removing the outliers and doing more data cleaning.

- This could also be converted into an application for vast use instead of running as
a computer software.

- A better algorithm for prediction can be used to further reduce the error in prediction.

REFERENCES

1. L. Hu, L. Li, J. Ji, and M. Sanderson, "Identifying and understanding determinants of high healthcare costs for breast cancer: a quantile regression machine learning approach," *BMC Health Services Research*, vol. 20, no. 1, pp. 1066–1110, 2020.View at: Publisher Site | Google Scholar
2. M. A. Aefa, M. Mahmoud, and M. M. Nassar, "Parameter estimation for a mixture of inverse chen and inverse compound Rayleigh distribution based on type-I hybrid censoring scheme," *Journal of Statistics Applications & Probability*, vol. 10, no. 3, pp. 647–663, 2021.View at: Google Scholar
3. W. A. Afifi and A. H. El-Bagoury, "Optimal multiplicative generalized linear search plan for a discrete randomly located target," *Information Sciences Letters*, vol. 10, no. 1, pp. 153–158, 2021.View at: Google Scholar
4. R. A. Ganaie, V. Rajagopalan, and S. Aldulaimi, "The weighted power shanker distribution with characterizations and applications of real life time data," *Journal of Statistics Applications & Probability*, vol. 10, no. 1, pp. 245–265, 2021.View at: Google Scholar
5. M. H. Abu-Moussa, A. M. Abd-Elfattah, and E. H. Hafez, "Estimation of stress-strength parameter for Rayleigh distribution based on progressive type-II

censoring," *Information Sciences Letters*, vol. 10, no. 1, pp. 101–110, 2021.View at: Google Scholar

6. S. Sana and M. Faizan, "Bayesian estimation using lindley's approximation and prediction of generalized exponential distribution based on lower record values," *Journal of Statistics Applications & Probability*, vol. 10, no. 1, pp. 61–75, 2021.View at: Google Scholar

7. K. Sahu and R. K. Srivastava, "Needs and importance of reliability prediction: an industrial perspective," *Information Sciences Letters*, vol. 9, no. 1, pp. 33–37, 2020.View at: Google Scholar

8. A. A. Soliman, Al-W. A. Farghal, and G..A. Abd-Elmougod, "Statistical inference under copula approach of accelerated dependent generalized inverted exponential failure time with progressive hybrid censoring scheme," *Applied Mathematics & Information Sciences*, vol. 15, no. 6, pp. 687–699, 2021.View at: Google Scholar

9. S. Kent, J Green, G Reeves et al., "Hospital costs in relation to body-mass index in 1·1 million women in England: a prospective cohort study," *The Lancet Public Health*, vol. 2, no. 5, pp. e214–e222, 2017.View at: Publisher Site | Google Scholar

10. V. S. Kadam, S. Kanhere, and S. Mahindrakar, "Regression techniques in machine learning &applications: a review," *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, no. 10, pp. 826–830, 2020.View at: Publisher Site | Google Scholar

11. B. Panay, N. Baloian, J. Pino, S. Peñafiel, H. Sanson, and N. Bersano, "Predicting health care costs using evidence regression," *Proceedings*, vol. 31, p. 74, 2019, https://www.mdpi.com/2504-3900/31/1/74.View at: Publisher Site | Google Scholar

12. B. J. Moore, S. White, R. Washington, N. Coenen, and A. Elixhauser, "Identifying increased risk of readmission and in-hospital mortality using hospital administrative data," *Medical Care*, vol. 55, no. 7, pp. 698–705, 2017.View at: Publisher Site | Google Scholar

13. R. S. Suidan, W. He, C. C. Sun et al., "Impact of body mass index and operative approach on surgical morbidity and costs in women with endometrial carcinoma and hyperplasia," *Gynecologic Oncology*, vol. 145, no. 1, pp. 55–60, 2017.View at: Publisher Site | Google Scholar

14. H. J. Kan, H. Kharrazi, H.-Y. Chang, D. Bodycombe, K. Lemke, and J. P. Weiner, "Exploring the use of machine learning for risk adjustment: a comparison of standard and penalized linear regression models in predicting health care costs in older adults," *PloS one*, vol. 14, no. 3, Article ID e0213258, 2019.View at: Publisher Site | Google Scholar

15. S. Kent, S. A. Jebb, A. Gray et al., "Body mass index and use and costs of primary care services among women aged 55-79 years in England: a cohort and linked data study," *International Journal of Obesity*, vol. 43, no. 9, pp. 1839–1848, 2019.View at: Publisher Site | Google Scholar

16. J. A. Irvin, A. A Kondrich, M Ko et al., "Incorporating machine learning and social determinants of health indicators into prospective risk adjustment for health plan payments," *BMC Public Health*, vol. 20, no. 1, pp. 608–610, 2020.View at: Publisher Site | Google Scholar

17. S. Kent, F. Fusco, A. Gray, S. A. Jebb, B. J. Cairns, and B. Mihaylova, "Body mass index and healthcare costs: a systematic literature review of individual participant data studies," *Obesity Reviews*, vol. 18, no. 8, pp. 869–879, 2017.View at: Publisher Site | Google Scholar

18. A. Ashfaq, A. Sant'Anna, M. Lingman, and S. Nowaczyk, "Readmission prediction using deep learning on electronic health records," *Journal of Biomedical Informatics*, vol. 97, Article ID 103256, 2019.View at: Publisher Site | Google Scholar