Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Abstract

Since health insurance system emerged, it has been gaining popularity among developed and developing countries. Due to health insurance system, an individual has low risk of bankruptcy to meet medical bills. Approximate estimation of health costs plays important role for different organizations such as hospitals. We propose a regression model for prediction of health insurance costs based of important features. The original data was simulated by Brett Lantz for his book *Machine Learning with R* using demographic statistics in USA Lantz (2013). Our goal is to explore the dataset and apply appropriate model to predict insurance costs based on analysis and understanding of the dataset. In our experiment we applied EDA for understanding of the dataset and regression model to predict outcome.

Keywords: EDA, Linear Regression

1. Introduction

One clinical research study published in the American journal of medicine around 2007 claimed that 62.1% of bankruptcies were caused by medical issues Himmelstein et al. (2009). Basic health insurance policy can be two types, one is public that is offered by the state and another is private that is offered by various insurance companies. These insurance plans of various companies ensures financial security, domiciliary treatment, free health check-up, tax benefits, coverage against critical illness, etc.

Health insurance companies key task is to estimate insurance cost allowance for an individual. To predict the insurance cost, it requires to collect years of data set an individual spends on medical care. The features that may effect one individual's medical costs contains include age, sex, bmi, children, smoking habit, region, charges etc. The dataset that we are exploring comprised of these mentioned features. Health insurance premiums cost's basic factors would be an individuals age, his or her gender,the persons body-mass balance, number of dependents, personal habits, residential area etc.

The objective of this project predict health insurance cost for an individual exploring the collected dataset. The key motivation for this project is to help health insurance policy makers to allocate health insurance boundaries more suited to specific individuals lifestyle for specific geographical location. This will help them to explore individuals medical insurance costs vs. required insurance plan model. Our goal is to consider these factors and estimate tentative medical insurance costs of an individual.

2. Related Work

One of the world's most pressing issues is the increasing cost of health care and many invested to predict health cost to prevent bankruptcy due to medical cost. Yet, a perfect model to get a perfect estimation cannot be ensured so far. People found data mining method to evaluate past medical records as predictor and predict future medical cost Himmelstein et al. (2009). Researchers have worked with patient-level health care expenditures in both the short and long terms and found strong temporal correlation over multiple periods Yang et al. (2018). supervised Learning Methods for Predicting Healthcare Costs using Artificial Neural Network (ANN) and the Ridge regression model using Empirical Evaluation gives approximate results but fails for time-series implementation Morid et al. (2018). Feature engineering was crucial in capturing temporal patterns in the data and reducing

the number of features with minimal loss Jödicke et al. (2019). People have worked with comparative different approaches like supervised learning, Time-Series forecasting using statistical, neural, and ensemble architectures Kaushik et al. (2020).

3. Methodology and Experiment

3.1 Dataset Description

This data was collected from the repository of the book where they used using demographic statistics from the US Census Bureau which indicates approximately reflection of real-world conditions Lantz (2013). This dataset contains the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government), policy holder's gender (it can be either male or female), the body mass index or BMI (weight (in kilograms) divided by height (in meters) squared), the number of children/dependents covered by the insurance plan, either smoker or non-smoker (it can be only two strings, either yes or no), and the beneficiary's place of residence in the US. A total of 1338 patients were enrolled in the dataset, with features reflecting patient characteristics as well as total medical costs paid to the plan for the calendar year. The structure of the csv file that contains the dataset can be described as in the table.

Apart from these features, the dataset contains another attribute representing medical charges as the label to train the dataset.

Features	**Value type**	**Value description**
Age	An integer	: A number indicating the beneficiaries age less than 64 years
Sex	A string	: male or female
BMI	A float	: weight (in kilograms) divided by height (in meters) squared
Children	An integer	: the number of children/dependents covered by the insurance plan
Smoker	A string	: yes or no
Region	A string	: four geographic regions: northeast, southeast, southwest, or northwest

3.2 Categorical Data Transformation

We identified the following categorical features in the dataset: sex, smoker, region. As these feature values are fixed and limited, we transformed them and gave them a numerical value starting from one.

3.3 Removing Outliers

There are some outliers in the dataset such as charges over 30000 which indicates noise for the dataset. We calculated inter-quartile range for (IQR) for that dataset by removing those outside the IQR. IQR fro each feature is presented below,

Features	IQR	Features	IQR
Age	24.000000	sex	1.000000
BMI	8.397500	children	2.000000
smoker	0.000000	region	1.000000
charges	11899.625365		

It removed 283 data from 1338 and the considered dataset length is 1055. We removed the outlier of the dataset. We applied standard linear regression to predict charges of an individual. Linear regression assumes a linear relationship between the input variables (X) and the output variable (y). we also split the dataset as training and testing dataset. We separated 15% dataset as the testing dataset. After training the training set, we implemented the developed model in testing dataset.

3.4 Architecture

We used linear regression model to predict the medical charges. We specially focused on regression after observing explanatory data analysis (EDA) of various features. In future, we intend to use random forest and ridge regression to get better results.

4. Discussion

In this milestone of our study we specially focused on how EDA to see how different features affect individuals charges. We found that among all the features, smoking habit of an individual has the strongest correlation with medical costs and than age, bmi of that individual correspondingly.

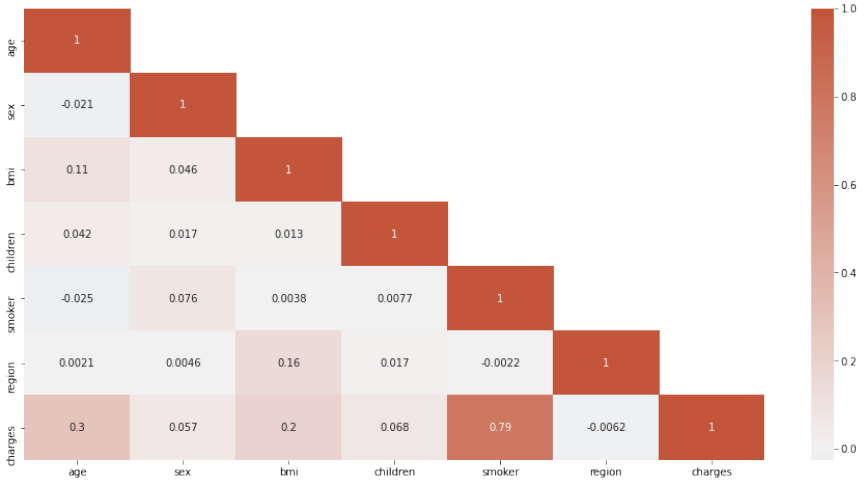

Figure 1: correlation between individuals important features and medical costs

From the understanding of the dataset and all features and categorical features relation to medical costs, we implemented linear regression and ridge regression approach to predict the medical cost for an individual. In our study both the approach produced almost similar

results. In our experiment, we got training error 6016.7425 and testing error 6206.3164 using root mean squared error method. If our dataset were a bit more enlarged and informative. training our model would be more accurate.

5. Conclusion

To predict insurance cost of an individual, smoking habits affect the cost estimation the most. Our available information size affected out final result. If we could get more data we could predict medical cost more accurately.

References

David U. Himmelstein, Deborah Thorne, Elizabeth Warren, and Steffie Woolhandler. Medical Bankruptcy in the United States, 2007: Results of a National Study. *The American Journal of Medicine*, 122(8):741–746, August 2009. ISSN 0002-9343, 1555-7162. doi: 10.1016/j.amjmed.2009.04.012. URL https://www.amjmed.com/article/S0002-9343(09)00404-5/abstract.

Annika M. Jödicke, Urs Zellweger, Ivan T. Tomka, Thomas Neuer, Ivanka Curkovic, Malgorzata Roos, Gerd A. Kullak-Ublick, Hayk Sargsyan, and Marco Egbring. Prediction of health care expenditure increase: how does pharmacotherapy contribute? *BMC Health Services Research*, 19(1):953, December 2019. ISSN 1472-6963. doi: 10.1186/s12913-019-4616-x. URL https://doi.org/10.1186/s12913-019-4616-x.

Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A. Pickett, and Varun Dutt. AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures. *Frontiers in Big Data*, 3, 2020. ISSN 2624-909X. doi: 10.3389/fdata.2020.00004. URL https://www.frontiersin.org/articles/10.3389/fdata.2020.00004/full.

Brett Lantz. *Machine Learning with R*. Packt Publishing, Birmingham, October 2013. ISBN 9781782162148.

Mohammad Amin Morid, Kensaku Kawamoto, Travis Ault, Josette Dorius, and Samir Abdelrahman. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *AMIA Annual Symposium Proceedings*, 2017:1312–1321, April 2018. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977561/.

Chengliang Yang, Chris Delcher, Elizabeth Shenkman, and Sanjay Ranka. Machine learning approaches for predicting high cost high need patient expenditures in health care. *BioMedical Engineering OnLine*, 17(1):131, November 2018. ISSN 1475-925X. doi: 10.1186/s12938-018-0568-3. URL https://doi.org/10.1186/s12938-018-0568-3.