IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTING POSSIBLE PROSPECTS TO BUY INSURANCE USING DATA ANALYTICS

[1] M. Ramya, [2]A. Sankeerthana, [3]S. Harshitha, [4]Dr. Sunil Bhutada, [5]Dr.Y. Rohita
[1,2,3] B. Tech Student, [4]Professor, [5]Assistant Professor
Department of Information Technology
Sreenidhi Institute of Science and Technology, Hyderabad, India

Abstract **:** Health insurance is insurance that covers the whole or a part of the risk of a person incurring medical expenses, spreading the risk over numerous persons. By estimating the overall risk of health care and health system expenses over the risk pool. Most health insurance plans encompass each inpatient and outpatient covers the grounds that injuries and extended infection while overseas may be extraordinarily highly-priced and very stressful. In this paper, we are Predicting the Insurance Claim of each user. This observes the cost of healthcare for a sample of the population given smoking habits, age, sex, body mass index, and region. The data features observations of statistical information and regression analysis of the dataset. The goal is to predict the best estimator of insurance charges using machine learning algorithms. Machine Learning algorithms for Regression analysis and Random Forest Regression analysis are used and Data Visualization is also performed to support Analysis.

Index Terms : Insurance, Health System, Smoking Habits, Age, Sex, Body Mass Index, Random Forest Regression.

1. INTRODUCTION

1.1 INTRODUCTION TO THE PROJECT

Health insurance is a type of insurance product that specifically the fitness fees or care of the coverage individuals in the event that they fall sick or have an or have an accident. Broadly speaking, there are two types of treatments offered by insurance companies, namely inpatient (in-patient treatment) and outpatient (out-patient treatment). Where In-patient treatment means that an overnight stay in hospital is required. This is generally protected through each non-public medical insurance policy. All medical insurance overseas plans cowl in affected person hospital therapy that calls for a live in a clinic, and the affected person can pick any clinic and any surgeon. Out-patient plans cover conditions that do not require a patient to be admitted to a hospital. Most health insurance plans consist of each inpatient and outpatient cover considering the fact that injuries and extended infection while overseas may be extraordinarily steeply-priced and really stressful.

In this paper we're Predicting the Insurance Claim with the aid of using every user, Machine Learning algorithms for Regression analysis and Random Forest Regression analysis are used and Data Visualization is also performed to support Analysis.

This paper is composed of different sections. It starts with a literature survey in the context of the present work. We will begin via way of means of explaining approximately hassle statement. Then we will move on to the process flow. Further, we communicate approximately methodology. Then we present results that we had obtained by performance evaluation. Lastly about the conclusion and future scope.

1.2 OVERVIEW OF PROJECT:

It is a tough task to determine the premiums for their insurable customers. While the health care law in the different countries have several rules for the companies to follow to determine the premiums to their customers, it's really up to the companies on what factors they want to hold more weight age to. This observes the cost of healthcare for a sample of the population given smoking habits, age, sex, body mass index (BMI), and region. The data features observations of statistical information and regression analysis of the dataset. The goal is to predict the best estimator of insurance charges using random forest regression.

1.3 EXISTING SYSTEM APPROACH:

Health-Care insurance prediction using linear regression analysis:

Linear regression is used to predict data. Linear regression is used for modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted x. There are advances in this field but the limitations remain the same. Simple Linear Regression is the one where only one explanatory variable is used.

Disadvantages:

1. Considers only two columns of the dataset for analysis.
2. The open value and close value are considered.
3. But the accuracy given is not satisfactory.

1.4 PROPOSED SYSTEM APPROACH:

Health-Care insurance prediction using random forest regression technique:

Random forest is a Supervised Learning algorithm that uses ensemble learning method for classification and regression.

Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. A random forest is a meta-estimator (i.e. it combines the result of multiple prediction)which aggregates many decision trees, with some helpful modifications.

The number of features that can be split at each node is limited to some percentage of the total (which is known as the hyper parameter). This ensures that the ensemble model does not rely too heavily on any individual feature, and makes fair use of all potentially predictive features.

Each tree draws a random sample from the original data set when generating its splits, adding a further element of randomness that prevents over fitting.

Advantages:

1. Advantages: It is one of the maximum correct studying algorithms available. For many records sets, it produces a exceedingly correct classifier.
2. It runs efficaciously on huge databases.
3. It can cope with hundreds of enter variables without variable deletion.
4. It offers estimates of what variables are crucial with inside the classification.
5. It generates an inner independent estimate of the generalization blunders because the woodland constructing progresses.
6. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

2. LITERATURE SURVEY

Data visualization is the graphical representation of information and data. By the use of visible factors like charts, graphs, and maps, statistics visualization equipment offer an on hand manner to look and recognize trends, outliers, and styles in statistics. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions. Our eyes are drawn to colours and patterns. We can speedy discover purple from blue, rectangular from a circle. Our tradition is visual, which include the whole thing from artwork and classified ads to TV and movies.

Data visualization is every other shape of visible artwork that grabs our hobby and maintains our eyes at the message. When we see a chart, we quickly see trends and outliers.

If we are able to see something, we internalize it quickly. It's storytelling with a purpose. If you've ever stared at a huge spreadsheet of facts and couldn't see a trend, you know the way a good deal extra powerful a visualization can be.

As the "Age of Big Data" kicks into high gear, visualization is an increasingly key tool to make sense of the trillions of rows of data generated every day. Data visualization facilitates to inform memories with the aid of using curating information right into a shapeless difficult to understand, highlighting the developments and outliers. A desirable visualization tells a story, putting off the noise from statistics and highlighting beneficial information.

However, it's not simply as easy as just dressing up a graph to make it look better or slapping on the "info" part of an info graphic. Effective information visualization is a sensitive balancing act among shape and function. The plainest graph will be too uninteresting to trap any word or it make inform a powerful Point; the maximum beautiful visualization ought to thoroughly fail at conveying the proper message or it may communicate volume. The statistics and the visuals want to paintings together, and there's an artwork to combining high-quality evaluation with high-quality storytelling.

There's an entire choice of visualization strategies to offer information in powerful and exciting ways. Common general types of data visualization:

- Charts
- Tables
- Graphs

- Maps
- Info graphics
- Dashboards

5.PROBLEM STATEMENT

There is a need to identify the health insurance for preventing the financial loss of the patient. The manual way of predicting health insurance has become difficult. Thus, there has been a dire necessity that compelled researchers to work on predicting health insurance.

Health insurance has been treated as important for avoiding financial losses. As opposed to treating age as a regression problem, Although, we expect age, sex, BMI, region to be number as output. We treated them as a categorical problem and trained the machine.

4.SYSTEM ARCHITECTURE

This architecture will give a demonstration of how the project is carried out.

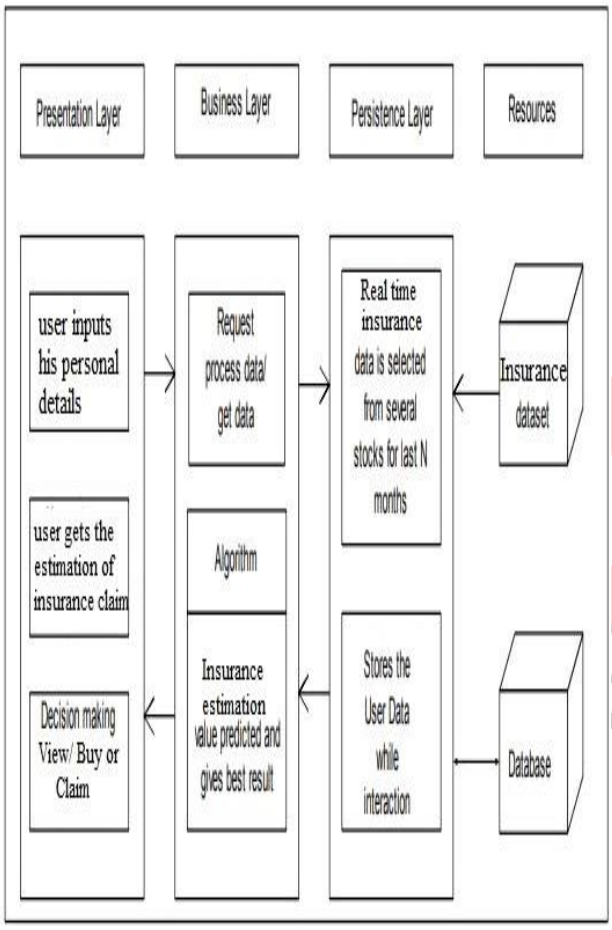

Fig 4.1 system architecture

5. METHODOLOGY

5.1 REGRESSION

In statistics, regression evaluation is a statistical system for estimating the relationships amongst variables. It consists of many strategies for modeling and studying numerous variables while the focus Is on the connection among a established variable and one or extra unbiased variables. More specifically, regression evaluation facilitates one recognize how the everyday price of the based variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression evaluation estimates the conditional expectation of the established variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile or other location parameter of the conditional distribution of the dependent variable given the independent variables.

5.2 RANDOM FOREST

REGRESSION

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

5.3 Parameters

n_estimators: integer, optional (default=10) The number of trees in the forest.

max_depth : integer or None, optional (default=None)
The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

n_jobs : int or None, optional (default=None)
The number of jobs to run in parallel. fit, predict, decision_path, and apply are all parallelized over the trees. None means 1 unless in a joblib.parallel_backend context. -1 means using all processors.

random_state : int, RandomState instance or None, optional (default=None)
Controls both the randomness of the bootstrapping of the samples used when building trees (if bootstrap=True) and the sampling of the features to consider when looking for the best split at each node **(if max_features<n_features)**

6.DATA VISUALIZATION

Visualizing the data to see gives an insight into our data set.

SEX HISTOGRAM

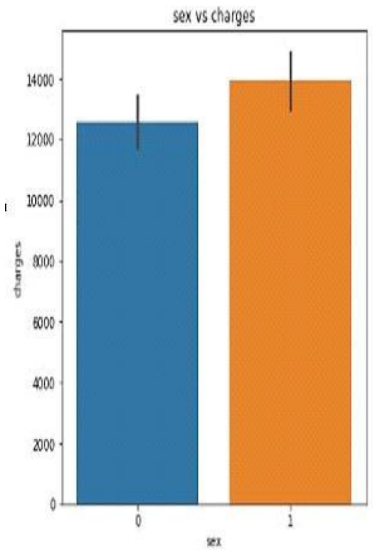

Fig 6.1 The categorical observations of sex, 0 being female and 1 being male.

CHILDREN HISTOGRAM

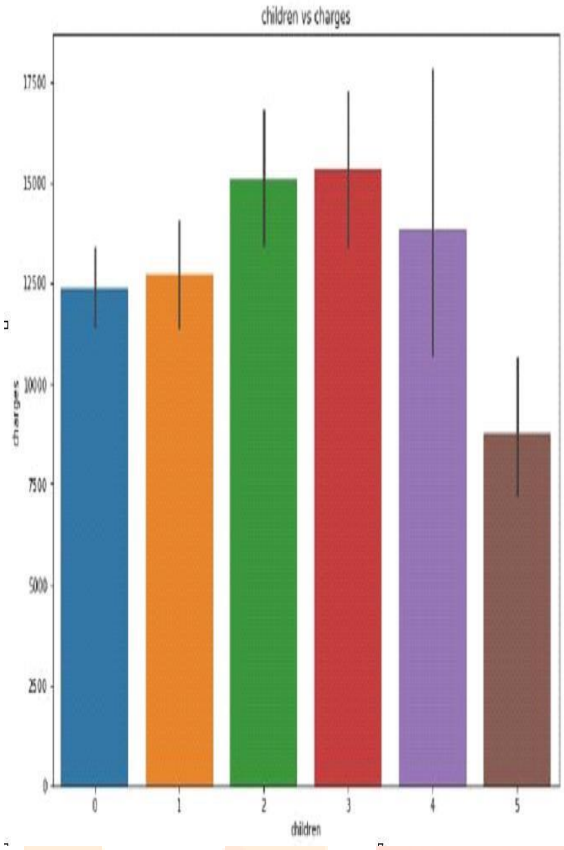

Fig 6.2 The children are plotted in the histogram based on the frequency count.

REGION HISTOGRAM

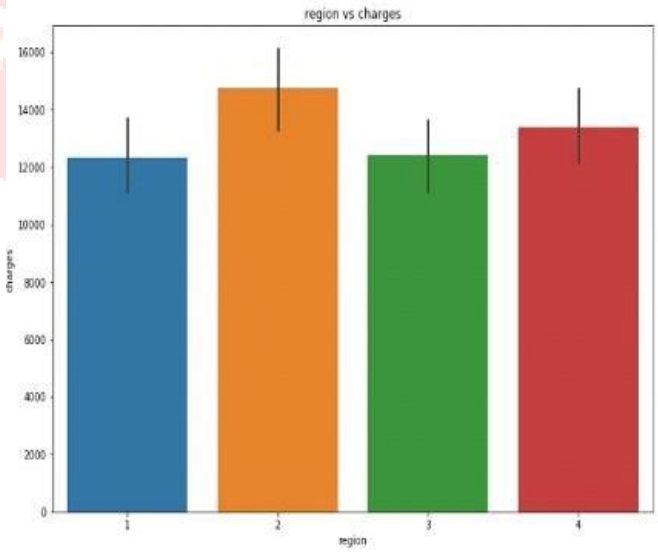

Fig 6.3 The region are plotted in the histogram based on the frequency count.

SMOKER HISTOGRAM

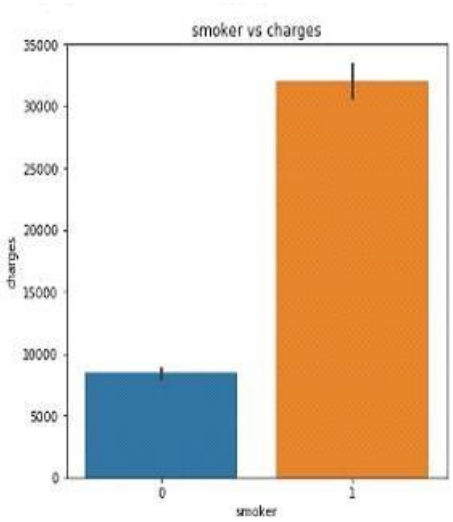

Fig 6.4 The smoker are plotted in the histogram based on the frequency count.

7.TECHNOLOGY

There are plenty of Python scientific packages for data visualization, machine learning, natural language processing, complex data analysis, and more. All of those elements make Python a extraordinary device for medical computing and a strong opportunity for industrial programs inclusive of MatLab. The most popular libraries and tools for data science are:

PANDAS

A library for data manipulation and analysis. The library affords information systems and operations for manipulating numerical tables and time series.

NUMPY

The fundamental package for scientific computing with Python, adding support for large, multi-dimensional arrays and matrices, alongside a huge library of high-stage mathematical capabilities to perform on those arrays

SCIPY

A library used by scientists, analysts, and engineers doing scientific computing and technical computing.

SCIKIT-LEARN

Scikit-learn is a machine learning library. It functions diverse classification, regression and clustering algorithms consisting of assist vector machines, random forests, gradient boosting, k-means, and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

SCIKIT-IMAGE

Scikit-Image is an image processing library. It consists of algorithms for segmentation, geometric transformations, shadeation area manipulation, analysis, filtering, morphology, feature detection, and more.

GRAPH-TOOL

Graph-tool is a module for the manipulation and statistical analysis of graphs.

8.PERFORMANCE EVALUATION

Accuracy
It tells the number of accurate predictions made to the number of total predictions.
Accuracy is equal to the Number of Correct predictions to a Total number of predictions made.
The model has an accuracy of 0.86 for health insurance prediction **Mean Absolute Error (MAE)**

$$MAE = \frac{\Sigma|y - \hat{y}|}{N}$$

where y is the actual value ŷ is the predicted value and |y−y^| is the absolute value of the difference between the actual and predicted value. N is the number of sample points.

Root Mean Square Error (RSME)

$$RMSE = \sqrt{\frac{\Sigma(y - \hat{y})^2}{N}}$$

Another evaluation metric for regression is the root mean square error (RMSE). Its calculation is very similar to MAE, but instead of taking the absolute value to get rid of the sign on the individual errors, we square the error (because the square of a negative number is positive).

9. CONCLUSION

This paper focuses on machine learning techniques to do the analysis of the insurance claims efficiently and compare their performances using different metrics and predict the estimation of the insurance claim. The aim is to apply data science analytics in the insurance it is the same as in the other industriesto improve the business and optimize its marketing strategies, reduce costs, and enhance its income.

The application of statistics in insurance has a long history. It is a fact that insurance companies are actively using data science analytics. With the help of various models and modern technologies, it is easy for the insurance industry to move extremely fast making its way into various fields of the business.

10. FUTURE ENHANCEMENT

Many research directions might be considered in future work. Improving the accuracy of the predictive models is one of them. Accuracy can be improved by considering an entirely different aspect i.e., with a type of disease or what kind of accident they had. As the future scope of the Insurance industry is limitless, the demand for its data analysis will be ever-increasing. By changing only the training data, the proposed system can be used for any kind of Insurance Industries in other countries. With few alterations, the system can be used for various purposes such as predicting prices of commodities like gold, predicting the fuel consumption of a vehicle as well as monitoring the health of a patient.

References:

.Belhadji, E., G. Dionne, and F. Tarkhani**,** ―A Model for the Detection of Insurance Fraud, Geneva Papers on Risk and insurance Theory‖, 25: 517-538, May 2012.

. Crocker, K. J., and S. Tennyson,‖ Insurance Fraud and optimal Claims Settlement Strategies: An empirical investigation of Liability Insurance Settlements‖ The Journal of Law and Economics, 45(2), April 2010.

.Kajiamuller, ―The Identification of Insurance Fraud – an empirical Analysis Working papers on Risk Management andInsurance‖ no: 137, June 2013. **.** S. B. Kotsiantis, ―Supervised Machine Learning: A Review of Classification Techniques,‖Informaticavol 31, pp 249-268, May 2011.

. Sivarajah U, Kamal M, Irani Z, Weerakkody V (2017) Critical analysis of big data challenges and analytical methods. J BusRes 70:263–286

. Mishr K (2016) Fundamentals of life insurance theories and applications. In: 2nd ed, Delhi: .PHI Learning Pvt Ltd

. The Kaggle Website. [Online]. https://www.kaggle.com/c/prudential-lifeinsurance-assessment/data/-The Accenture website https://www.accenture.com