International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 03 Issue: 08 | August -2019 ISSN: 2590-1892

DOCTOR BILL PREDICTION USING PYTHON FOR DATASCIENCE AND MACHINE LEARNING TECHNIQUES

1: UMA B,2: SNEHA C D, 3: SHWETHA R, 4: I MSHIVAPRASAD, 5: PAVITHRA R,

CO - AUTHOR - MAHENDRA KUMAR

A1 – A5 M.C.A. Post GraduateScholler D.S.C.E
CO-AUTHOR - AssistantProf. Dept of MCA,D.S.C.E

Abstract –

Predictive Analytics is an increasingly important tool in a medical field since modern machine learning methods can use large amounts of available data to predict the individual outcomes for a doctor bill. Machine Learning predictions can help hospital providers to determine doctor bill of a particular patient.

In this work, we will develop a medical price prediction system using machine learning algorithms which will aid in steering patients to cost effective providers and thereby curb health spending. The policymakers can also use the tool to better understand which providers are relatively expensive and take punitive actions if necessary. The prediction of the medical price will be done using implementing Random Forest Regression algorithm in machine learning. Additionally, we plan to include the experiments on the same data with other machine learning models such as Gradient Boosted Trees and Linear Regression and compare results. The findings from these experiments will also be included(*ref [1]).

Keywords-**Logistic Regression, Doctor, bill, Prediction, machine learning**

1. INTRODUCTION

The data sets contain the patient details and bill details of all patients and predicting the doctor bill of a particular patient on the basis of patient record. The outcome that we have to predict is doctor bill. The record contains patient name, room number, Date, medicine bills, room bills, food bills and other bills.

Machine Learning predictions can help hospital providers to determine doctor bill of a particular patient regarding to their name, room number, and bills. The doctor bill is considered as target variable and other variables are considered as a independent variable on which we can predict the value of a doctor bill of a particular patient. (*ref[3])Since the source variables contain various data types, such as dates, numeric, text, they needed to be pre-processed before they could be used for modelling. Numeric variables were kept in their numeric format. All categorical variables were enumerated, i.e., replaced by integers for the purpose of making the whole feature matrix numeric and conserving computing memory. For some of the categorical variables, binary features were also extracted by generating a separate column for each of the categories, as a category indicator. Some additional features were also generated using specific calculations(*ref[3]).

2. BODY OF PAPER

The project depends on accuracy of information. The undertaking is meant to build up an AI model dependent on information given (*ref[2])by the World Health Organization to decide the specialist charge forecast for various nations in years. The information offers a time allotment from 2000 to 2015. Here we select the proper calculation/model that is important for the investigation reason, we have chosen the accompanying models for preparing the dataset. The yield calculations have been utilized to test in the event that they can keep up their exactness in anticipating the future for information (*ref[2])they haven't been prepared. Four algorithms have been used(*ref [2]):

- Logistic Regression
- Logistic Regression with polynomial features
- Decision Tree
- Random Forest

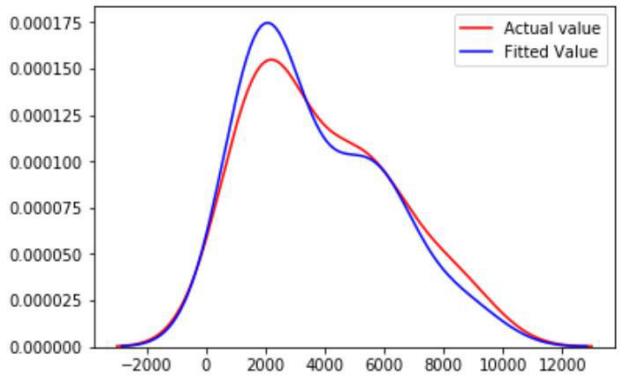
LOGISTIC REGRESSION

A few calculations will be given it a shot. First the old-style calculated relapse. (*ref[2])The model is fitted first on the prepared information the R square is 0.92 (*ref[2])on the preparation information. Later its R square which is coefficient of precision assurance is kept an eye on the testing information. The score is 87 % in the cycle of composing. We additionally compute the MAE, the modulus between the anticipated and the genuine incentive at 2.32 and the MSE (the equivalent just put to the intensity of 2) at 9.8(*ref[2]).

The objective of relapse investigation is to demonstrate the normal estimation of a needy variable y as far as the estimation of an autonomous variable (or vector of free factors) x. In straightforward strategic relapse, the model.

The underneath distplot figure production line shows a mix of factual portrayals of numerical information, for example, the connection between genuine qualities and anticipated qualities utilizing Logistic Regression

Fig 1–Output of Logistic Regression Model



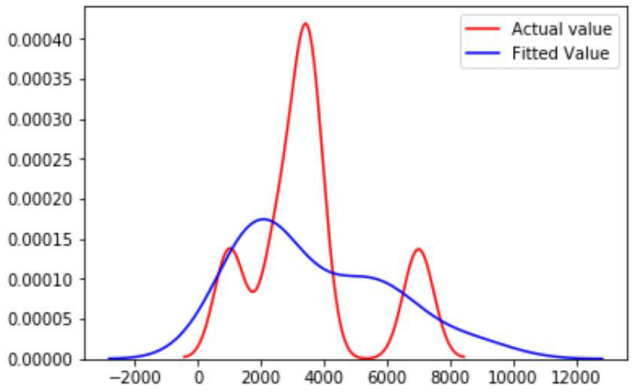
LOGISTIC REGRESSION WITH POLYNOMIAL FEATURES

The Logistic (*ref[2])Regression is being tried on the preparation information with the new (*ref[2])Polynomial Features. The Polynomial Features capacity has been utilized to get the collaborations of the information factors just to the intensity of 2.

In this venture, we will examine the utilization of Logistic Regression to anticipate the protection guarantee.

Presently we will import pandas to peruse our information from a CSV document and control it for further use. We will likewise utilize numpy to change over out information into a configuration reasonable to bolster our characterization model. We'll utilize seaborn and matplotlib for perceptions. We will at that point import Logistic Regression

Fig 2 – Output of Logistic Regression Model with Polynomial Features

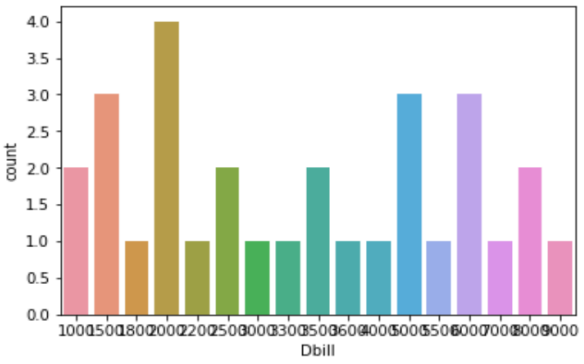


DECISION TREE

In a standard characterization tree, the idea is to part the dataset subject to homogeneity of data. In a backslide tree the idea is this: since the target variable does not have classes, we fit a backslide model to the target variable using all of the self-governing components. By then for each self-governing variable, the data is part at a couple of part centers. At each part point, the "botch" between the foreseen worth and the certifiable characteristics is squared to get an "Entire of Squared Errors (SSE)". The split point botches over the elements are taken a gander at and the variable/point yielding the least SSE is picked as the root center/split point. This strategy is recursively continued and (*ref[2])cross Validation has been performed. The R square on the arrangement data is 1 inferring that the estimation has taken in the data by heart, with the cross endorsement the figure diminishes to 77% and using the test date we get 80%.(*ref[2])After performing grid search with least models split in the range some place in the scope of (*ref[2])2 and 10 we get the best split of 3. The R square on the arrangement data is 98%, the computations has nearly slanted the data by hearth. On the test data we get R square of 81 %, the MAE is 2.71 and MSE is 16.59.

Choice trees are useful imagining help for breaking down a progression of anticipated results for a particular model. In that limit, it is normally used as an upgrade (or even choice rather than) backslide assessment in choosing how a movement of legitimate elements will influence the dependent variable.

Fig 1 – Output of Decision Model



RANDOM FOREST

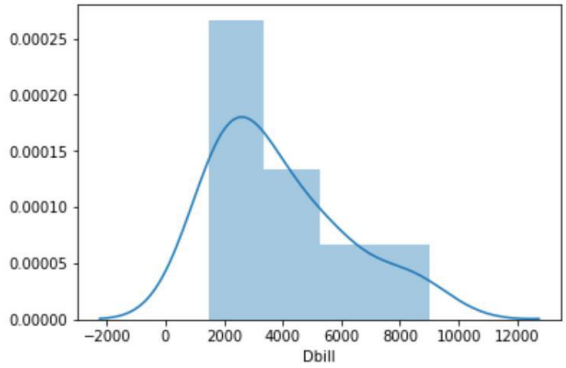
A Random Forest is a troupe system equipped for performing both relapse and order errands with the utilization of numerous choice trees and a method called Bootstrap Aggregation, usually known as packing. The essential thought behind this is to join different choice trees in deciding the last yield as opposed to depending on individual choice trees. The calculation (*ref[2])has learned 98% on the preparation information without cross approval and 88% with, the worth is 92 % on the test information.

$$\text{Random Forest Prediction s} = \frac{1}{K}\sum_{K=1}^{K} K^{th}\ \text{tree response}$$

The after effect of each tree depends upon a great deal of foreseen characteristics picked uninhibitedly with overriding and with a comparative appointment for all of the trees in the model, which is a subset of the pointer estimations of the main enlightening record. The perfect size of the subset of marker elements is given by log2 M+1, where M is the amount of wellsprings of information. Self-assertive Forest system describes an edge work that gauges how much the typical number of rulings for the correct class outperforms the ordinary ruling for some different class present in the poor variable. This measure outfits us not simply with a supportive strategy for making estimates, yet furthermore with a technique for accomplice an assurance measure with those desires. For relapse issues, Random Forests are shaped by

developing basic trees, each fit for creating a numerical reaction value.

Fig 1 – Output of Random Forest Mode



3. CONCLUSIONS

In this project, we have successfully implemented Regression Trees and Random Forest Regression algorithms from scratch to predict the medical prices from the input dataset. We also compared the results of Regression Trees, Random Forest Regression, Gradient Boosted Regression Trees and Linear Regression for the same dataset. From the results, we cannot conclude which model performed the best because the model performance can vary depending upon the configuration tried while testing. Hence, the model performing best for some configurations can give unsatisfactory results for some other configurations. Overall for the test configuration parameters, the order of performance of each model from the best to worst is Gradient Boosted Regression Trees, Random Forest Regression, Regression Trees and Linear Regression. The average medical payments predicted by Gradient Boosted Regression Trees, Random Forest Regression and Regression Trees are close to the actual values of payments(*ref[1]).

REFERENCES

[1]scholarworks.sjsu.edu

[2] www.datasciencesociety.net

[3] Yang Xie, Günter Schreier, Michael Hoy, Ying Liu, Sandra Neubauer, David C.W. Chang, Stephen J. Redmond, Nigel H. Lovell. "Analyzing health insurance claims on different timescales to predict days in hospital", Journal of Biomedical Informatics, 2016

[4] A. Tike and S. Tavarageri. (2017). A Medical Price Prediction System using Hierarchical Decision Trees. In: IEEE Big Data Conference 2017. IEEE.

[5] Morbidity rates (sickness) in America will continue to increase compared to the developed world

[6] Preventive, Medical, Surgical and Palliative Outcomes will continue to be the only 4 healthcare industry products manufactured and produced by physicians with their patients.

[7] J. Cubanski and T. Neuman,"The Facts on Medicare Spending and Financing," The Henry J. Kaiser Family Foundation, Menlo Park, CA, 2017.

[8] The Henry J. Kaiser Family Foundation. (2017). Total Number of Medicare Beneficiaries.