International Journal for Modern Trends in Science and Technology, 6(8S): 156-159, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTSTCIET30
Available online at: http://www.ijmtst.com/ncracse2020.html

Use of Predictive Modeling in Healthcare

Dr.K.R.R.Mohana Rao[1] | Dr.K.Kiran Kumar[2] | G.Ramachandra Rao[3] | P.Venkata Siva[3]

[1]Professor, Department of CSE, Chalapathi Institute of Engineering and Technology, Guntur, Andhra Pradesh. India.
[2]Professor & HOD, Department of CSE, Chalapathi Institute of Engineering and Technology, Guntur, AP, India.
[3]Assistant Professor, Department of CSE, Chalapathi Institute of Engineering and Technology, Guntur, AP, India.

To Cite this Article
Dr.K.R.R.Mohana Rao, Dr.K.Kiran Kumar, G. Ramachandra Rao and Mr.P.Venkata Siva, "Use of Predictive Modeling in Healthcare", *International Journal for Modern Trends in Science and Technology*, 6(8S): 156-159, 2020.

Article Info
Received on 16-July-2020, Revised on 15-August-2020, Accepted on 25-August-2020, Published on 28-August-2020.

ABSTRACT

The project 'Use of Predictive Modeling in Healthcare' focuses on classification of cancer stage into Malignant or Benign and finding a linear relationship between insurance charges and beneficiary's profile to predict factors that affect hospital costs since we know how costly healthcare in United States is.We used two datasets. First dataset is a collection of data that classifies cancer into malignant and benign and second one is an insurance dataset that contains hospital charges and beneficiary's profile data. We applied Logistic Regression model on Cancer Dataset and Linear Regression model on hospital charges dataset.

KEYWORDS: *Predictive Modeling, Health care, regression Model.*

I. INTRODUCTION

Machine learning in the field of healthcare can influence positively when done correctly. It can help in detecting diseases and also save lives if acted upon soon. So, the ability to be able to predict is a boon to healthcare. This is why we wanted to do a project with Healthcare in our mind.

II. PROCEDURE FOR PAPER SUBMISSION

Research Questions:
We searched for datasets that have data related to cancer patients. The reason why we searched for cancer datasets is because it is one of the most catastrophic diseases and prediction of cancer in early stages will be helpful. One of the problems we faced was getting access to the data because most of the medical datasets have restricted or limited access. Also, we wanted to even consider the financial aspect to the hospital treatment. So, we searched for datasets that have hospital bills or insurance charges details. We wanted to play around with such a dataset and see the patterns to see as to what kind of a patient(patient here means beneficiary) profile gets charged more by the health insurance companies.

III. METHOD

We selected two different datasets for our project, and we are calling it as ProjectA and ProjectB. For the **ProjectA**, we selected a dataset that contains details of Breast Cancer stages – the stages being 'Malignant' and 'Benign'. The dataset contains of 32 columns and 569 rows.

Attribute/ Column Information:

1. ID number
2. Diagnosis (M = malignant, B = benign)

3-32. are divided into three parts first is Mean (of all cells), Standard Error(of all cells) and Worst(worst cell) and each contain 10 parameter

(radius, texture, area, perimeter, smoothness, compactness,concavity, concave points, symmetry and fractal dimension) each.

Diagnosis is a categorical column and remaining all are numerical columns. We did not use ID in our model as it is merely just a unique row id and is of no importance.

The model that we chose to classify the stages of cancer is 'Logistic Regression'. The reason why we chose this type of classifier is because we needed only classification between two classes(binary) and logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary.

For the **ProjectB**, we chose a dataset that gives insurance beneficiary details. It has 1338 rows and 7 columns.

Attribute / Column Information:
1. age: age of primary beneficiary
2. sex: gender - female, male
3. 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
4. children: Number of children covered by health insurance/ Number of dependents
5. smoker: tobacco/ nicotine user
6. region: the beneficiary's residential area in the US - northeast, southeast, southwest, northwest.
7. charges: Individual medical costs billed by health insurance

Out of these 7 columns, 4 are numerical columns and 3 are categorical columns. We converted these 3 categorical columns into binary columns and chose the Linear Regression model to run on the

data. We chose Linear Regression model because we wanted to see to what extent there is linear relationship between health insurance charges and beneficiary profile.

IV. RESULTS

Project A – Breast Cancer Stage Prediction Using Logistic Regression:

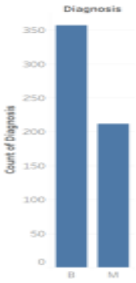

Fig 1

Initially, we did a bit of data visualization to see which stage had more occurrences in the dataset as seen in Fig 1.
The dataset had three parts – Mean, Standard error and Worst cell set of parameters, but we decided to use the mean set of features for our model because considering error or worst cell would not give a complete picture of the dataset. The next step that we did was to see the correlation between the selected features by creating a heatmap.

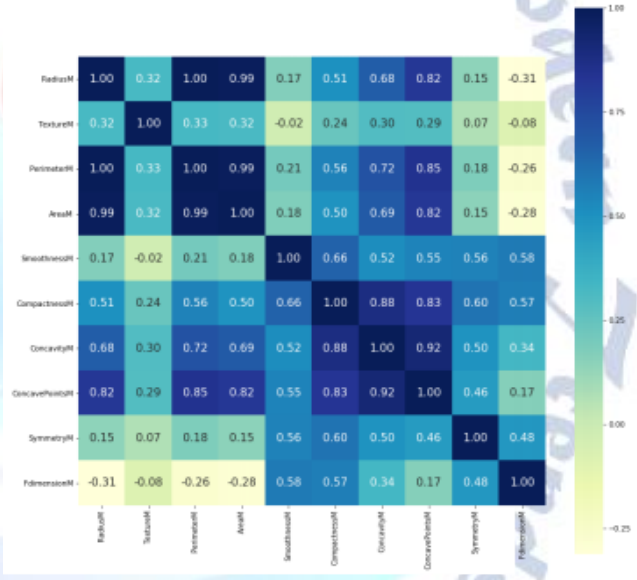

Fig 2

The outcome of this heatmap was that RadiusM, PerimeterM and AreaM have high correlation in general. So, we experimented with these features and they almost gave same accuracy with their inclusion. So, we considered one out of these 3. We selected RadiusM as it gave the highest accuracy.

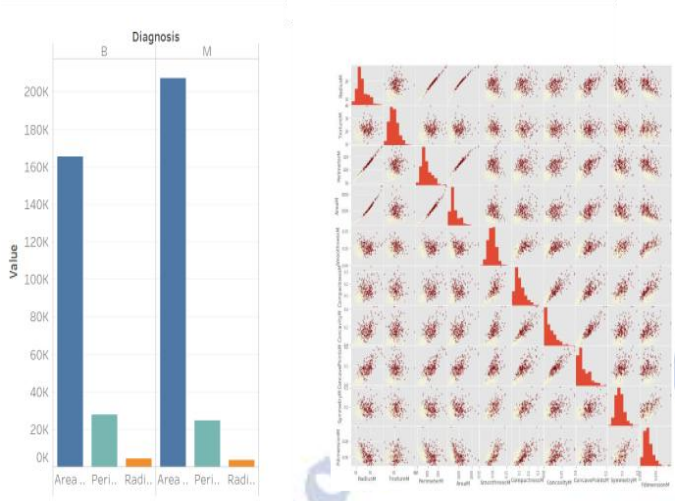

Fig 3, 4

We also tried visualization on the dataset using Tableau. We saw a positive correlation between radius, area and perimeter as seen in the above figure(Fig 3). CompactnessM, ConcavityM and ConcavepointM are also highly correlated so we used one among the three.

Once the features were selected, we did data Analysis for the selected features to understand better as to which features can be used for prediction. We plotted a scatter plot for the all features for both Diagnosis Categories(M & B – Malignant and Benign). And from it, we tried to find which can easily be used for differentiating between two categories.

In the figure(Fig 4), Color maroon represents Malignant type and Beige color represents Benign type. Radius, Area and Perimeter have a strong linear relationship as expected.

Next step was to train our model. We chose the split percent between test/ train as 30/70. We built our logistic regression model on the train data and tested it on the test data. The accuracy that we got was 90.158%.

ProjectB – Cost Analysis for Hospital Admission Using Linear Regression:

The first step that we did was to check the correlation(Fig 5) between 4 numerical columns (age, bmi, children, charges).

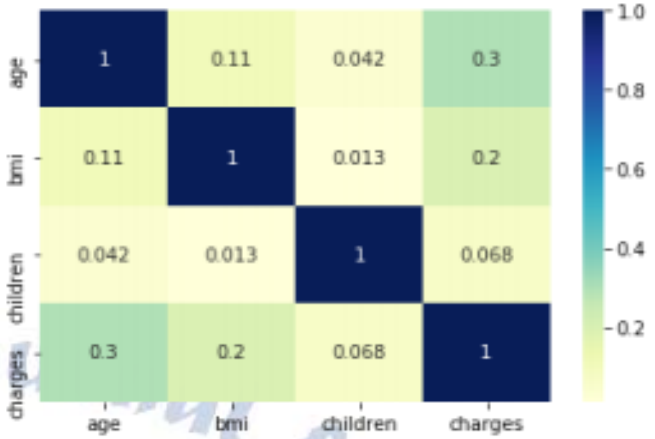

Fig 5

As part of data pre-processing, we converted the categorical columns (sex, region and smoker) into binary columns (binary vectors). So, sex was transformed into two binary fields 'female' and 'male'. Smoker was transformed into 'non_smoker' and 'smoker_new'. Region was transformed into 'northeast', 'northwest', 'southeast' and 'southwest'.

Then we assigned X and Y variables. Y included the 'charges' field and remaining fields were assigned to X since we want to analyze the relationship between 'charges' and all the other columns. Then we split the data into train and test with a test/train percent of 25/75. We built a Linear Regression model on the train data. We then found the intercept and coefficient values of the model. We plotted a scatter plot for predictions and also found the Root Mean Squared Error.

V. CONCLUSION

In project A, The research problem that we were addressing to was to predict which stage of cancer the profile has. The accuracy of the model came to 90.158%.

When we were selecting our features for the Logistic Regression Model, we came across fields that were highly correlated (positively). We checked accuracies with these fields(radius, perimeter and area) and noticed that using them together did not make much difference to the model's accuracy. So, we concluded that these are redundant fields and that using of redundant fields will not improve accuracy and in fact might increase complexity in some cases.

For Project B, The research problem that we wanted to address was to see how much a patient is charged by health insurance companies based on the beneficiary's profile. We wanted to see the linear relationships between charges (predictor) and other variables that affect charges.

One interesting fact that we noticed from the heatmap in the Fig 5 was that children category has the lowest correlation with 'charges' (lowest correlation of 0.068). We initially assumed that having children/ dependents might increase charges by insurance companies, but that column had no major correlation with charges.

Another interesting fact was that the smoker category has the highest effect on charges as seen in Fig 6. This means that being a smoker, will increase health insurance charges. So, smoking not only harms one's health but also hurts financially.

REFERENCES

[1] Ken Terry, "Futuristic Clinical Decision upport Tool Catches On," InformationWeek Healthcare, Jan. 27, 2012, accessed at http://www.informationweek.com/healthcare/clinical-systems/futuristic-clinical-decision-support-too/232500603.

[2] Ben-Chetri E, Chen-Shuali C, Zimran E, Munter G, Nesher G. "A simplified scoring tool for prediction of readmission in elderly patients hospitalized in internal medicine departments." Isr Med Assoc J. 2012 Dec;14(12):752-6.

[3] Hasan O, Meltzer DO, Shaykevich SA, Bell CM, Kaboli PJ, Auerbach AD, Wetterneck TB, Arora VM, Zhang J, Schnipper JL. "Hospital readmission in general medicine patients: a prediction model." J

[4] Gen Intern Med. 2010 Mar;25(3):211-9. doi: 10.1007/s11606-009-1196-1. Epub 2009 Dec 15.

[5] Donze J, Aujesky D, Williams D, Schnipper JL. "Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model." JAMA Intern Med. 2013 Apr 22;173(8):632-8. doi: 10.1001/jamainternmed.2013.3023.

[6] J. Frank Wharam and Jonathan P. Weiner, "The Promise and Peril of Healthcare Forecasting." Am J Manag Care. 2012;18(3):e82-e85)

[7] HIMSS Analytics, "Clinical Analytics: Can Organizations Maximize Clinical Data?" June 7, 2010.

[8] HIMSS Analytics, "Clinical Analytics in the World of Meaningful Use," Feb. 2011. Accessed at http://www.himss.org/files/himssorg/content/files/20110221_Anvita.pdf.

[9] Institute for Healthcare Technology Transformation, Analytics: The Nervous System of ITEnabled Healthcare, accessed at http://ihealthtran.com/iHT2analyticsreport.pdf.

[10] Terry, "EHR Data Not Ready for Prime Time, Studies Show," iHealthBeat, Feb. 9, 2012, accessed at http://www.ihealthbeat.org/insight/2012/ehr-data-not-ready-for-prime-time-studiesshow.