Predications from Non-Medical Data using Data Mining

Prof. Srinatha D K, Pavan Kumar M, Vinay N, Mohan M
Professor, Alliance Collage of Engineering and Design, Bangalore
Scholar, Alliance Collage of Engineering and Design, Bangalore, pavankumarmc66@gmail.com
Scholar, Alliance Collage of Engineering and Design, Bangalore, vinayp1308@gmail.com
Scholar, Alliance Collage of Engineering and Design, Bangalore, mohanganes3@gmail.com

Abstract: **Organization data of Hospital Information System and Electronic Health Record are Classified by lower Privacy sensitivity, So it is not only easier for Portability and handling but also for higher information quality. In this paper we test the concept that the application of Data Ming techniques on data of this information can be used to address prediction problems in the Health IT domain. The novelty of this approach that consist in that medical data (Diagnoses, test results, doctor's notes etc) are not included in the predictor's dataset. There is also limited need for separation of patient cohorts based on specific health conditions.**

Methods: We perform the early readmission Probability of early readmission at the time of a patient's discharge. Real Hospital Information System data is extracted to perform data processing techniques. Then we apply a series of algorithms based on data mining (Support Vector Machine, Gaussian Naïve Bayes, K-nearest Neighbours, Logistic Regression and Deep Multilayer Neural Network) and to measure the emergent model's performance.

Keywords-electronic health records; Hospital information systems; early readmissions; machine learnings

I. INTRODUCTION

Organization of Hospital Information System and Electronic Health Record data concerning patient hospitalization, such as repeated incidents that lead to inpatient admission, transfer to Intensive Care Unit, outcome classification, length of stay etc. The data is high quality in terms of completeness and accuracy for several reasons such as service handling by personnel are more qualified and experienced in the use of computer systems and double checking due to secondary processes e.g. claim management etc.

As a experimental problem we select the prediction of the probability of previous readmission at the patient's time of discharge. Previous readmission if defined as short time for admission to a hospital typically before 30 days after discharge. This topic has been a source of controversy in the domain of healthcare organizations. Readmissions are reducing considered to have a direct and significant impact on the health system's cost efficiency. As the readmission of every occurrence means to new administrative admission overhead costs plus medical costs related to the high probability of the patient's health deterioration, Which makes to lead the need for more readmission most of the time.

Despite the fact that in Greece there is no official regulatory framework in place to penalize or otherwise deal with high rates of early readmissions, there is no doubt that, eventually, measures will be searched in the direction of alignment with National Health Systems that have already addressed this issue, e.g. the United States [1].

Individualized early readmission counter-measures include follow-up visits or phone calls, home care procedures, establishing active communication channels with primary –or family, when available– doctors, patient and/or next of kin-education programs etc. [2] These procedures, if applied to every discharged patient, can be costly and, by nature, hard to scale in a typical hospital context. This creates high demand for methods capable of assessing the risk of patient readmission, in order to apply selected countermeasures to targeted patient groups, thus increasing cost efficiency (effectively using the resources available for preventing readmissions).

Our experimentation with early readmission prediction, introduces a novel approach, as it does not utilize neither the inclusion of medical predictors (test results, diagnoses, doctors' notes etc.) in the predictors' dataset, nor separation of patient cohorts based on specific health conditions; both of these methods, to the extent of our knowledge, are used extensively in the existing relevant literature [3], [4], most of the times in combination with each other.

II. DATA SUMMARY AND PROCESSING

A. The Dataset

We acquired dataset from one of the Hospital. They provided access to limited hospitalization data, via database queries. The data doesn't have any identifying personal information, all the process took place in the premises under hospital staff supervision and appropriate data security measures were taken.

The initial working dataset consisted of 127,943 hospitalization records covering a time span from 2005 to 2018. The only information used was the existence of categorised discharge diagnosis as a rough indicator. The columns that were used our experiment are listed in table 1.

B. Feature Engineering

The dataset was manually cleaned in order to eliminate unwanted data:

- X Outliers due to bad data quality of patients that appeared in the dataset with age less than 0 or greater than 125 and incident cases not closed properly.
- x Extreme patient cases that would undermine the calculation process without adding to the predictive value of the training set. x Hospitalizations with declared outcome: "X-Death".

Classified values (GENDER, OUTCOME). New columns for each category were introduced with values [0, 1] and each sample took value of 1 in the column of the respective category it initially belonged. A new column READMISSION_30_days was introduced deriving from the transformation if initial READMISSION_DAYS c

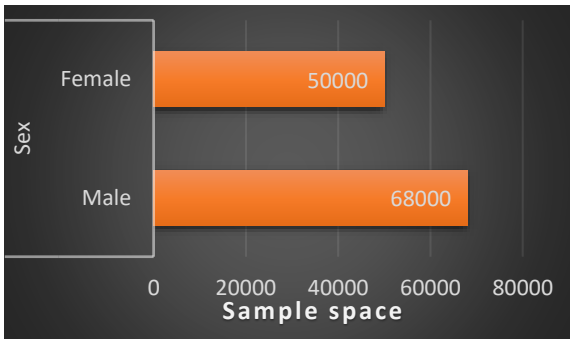
Value =1 initial data >30

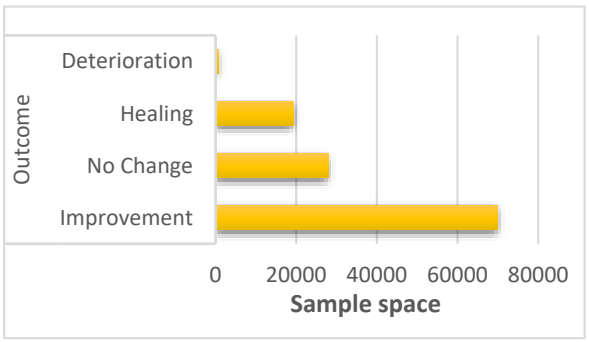
Value =0 initial data <=30

The Graphical representations of the main features of the dataset are presented , in the form of bar charts or density plots accordingly is presented in Fig.1 and Fig.2.

In order to address the great imbalance in the READMISSION_30_DAYS column, due to the fact that early readmissions occur in a fraction of the total hospitalizations, we applied random under-sampling techniques to the dominating class (READMISSION_30_DAYS=0) samples, randomly picking a subset of equal size to the minor class (READMISSION_30_DAYS=1). Thus, the final dataset consisted of 44,384 samples, equally representing each class (22,192 samples each).

Table I.	**Data columns (features)**
AGE	Patient's age in the date of discharge
GENDER	Patient sex
DURATION_DAYS	Length of stay (days) during current hospitalization
REACTION	Outcome (X -Death, H-Healing, I-Improvement, N-No change, D-Deterioration)
ICU_STAY	I-The patient had to be transferred in ICU during current hospitalization ,0- otherwise
EARLIER_ADNO	Number of previous admissions in the hospital
EARLIER_STAY	Cumulative length of stay (days) during previous admissions in the hospital
EARLIER_ICUSTAY	Cumulative days of ICU treatment during previous admissions in the hospital





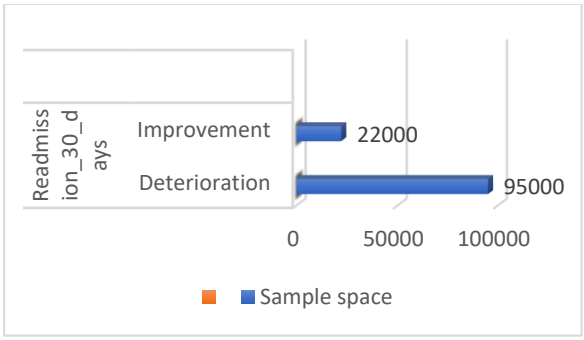


Figure 1. Features of the initial Dataset

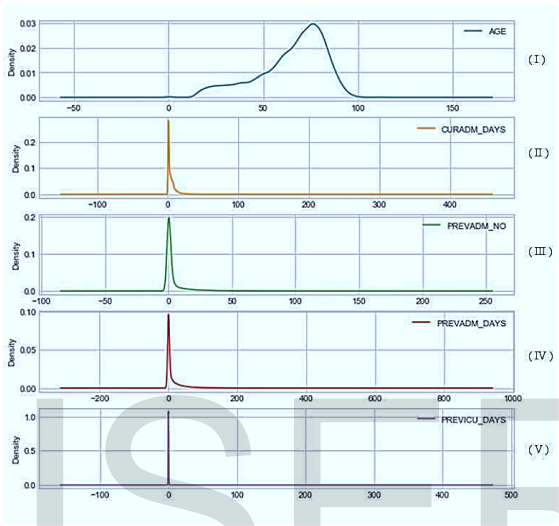


Figure 2. Density of numerical features of the initial dataset

III. METHODS

A. Classification algorithms

1) SVM (Support Vector Machine) is used to construct a set of hyper plane using mathematical prosecco in high dimensional spaces with larger distance to trailing points of any classes that is defined by the predictors.in general, larger margins could mean lower generalization errors.

2) K-NN (Nearest Neighbours) In neighbours-based classification approach labelling of each point of the parameters' hyper-space is computed based on a simple majority vote of the nearest known (already labelled) neighbouring points: a query point is assigned the class which has the most representatives in the geometrical hyper-space "vicinity". The most commonly used technique is k-NN, with "k" denoting the number of nearby already labelled points that will take part in the "voting" process [6]. Various implementations of this technique, depending on the context and the nature of the data at hand, use, in addition to different choice of k, different ways to measure the geometrical point-to-point distance (Euclidean, "Manhattan" etc.) as well as appointing different weights to the contributions of the neighbours e.g. proportional to the inverse of the distance.

3) NB (Naïve Bayes) Naive Bayes methods are based on the application of Bayes' theorem. The "naivety" of the approach is referring to the assumption of independence between every pair of features. Naive Bayes classifiers have been performing well in real-world situations, such as document classification and spam filtering. Naive Bayes classifiers uses less training data to estimate the necessary parameters in faster way [7]. Each distribution can be independently estimated as a one-dimensional distribution (decoupling). This helps addressing issues stemming from the "curse of dimensionality" (various phenomena that arise when analysing data in high dimensional spaces).

4) LR (Logistic Regression) Logistic regression is a linear model used for classification, despite its name. In the literature, Logistic regression is often referred as "logit", maximum-entropy classification (MaxEnt) or loglinear classifier. The logistic function is used to model the probabilities of the possible outcomes.

5) Deep Learning. Artificial neural networks (ANNs) are computing systems inspired by the biological neural networks that constitute animal brains, specifically the neuronal structure of the mammalian cerebral cortex. ANNs "learn" via a supervised process that occurs with each time the network is presented with a new input pattern (cycle or "epoch") through a forward activation flow of outputs and the backwards error propagation of weight adjustments. Reference provides a comprehensive and wide overview of the application of deep learning in fields as translational bioinformatics, medical

imaging, pervasive sensing, medical informatics, and public health. It is widely accepted, and theoretical background has been provided that deep networks are compactly representing functions that shallow architectures are not capable of [10].

IV. EXPERIMENTS

A. Experimental Setup

The dataset which is set for the final algorithmic procedure is transformed by the way of scaling each of the features numerically to the values of the range 0 to 1, The effect of intimidate attribute to nominate the processes that depends highly in computing geometrical distances.

At first we experiment each of four process from Section 3.1.Using built-in functions all the methods are being trained by the Scikit-Learn, a collection of data mining and analysing of tools for programming language Python. The purpose of this experiment is not to build models with highest performance possible and there is no more effort is allotted to optimize the models on hyper-parameters. The variables which are used for deriving models via the general-purpose recommendation of Scikit-Learn library (which is referenced for each algorithm in the notes of Table 2).

Table 2. RESULT REPORTS

Method	Performance	
	AU ROC [a]	*MCC* [b]
Support Vector Machine [c]	0.711	0.437
K-Nearest Neighbors [d]	0.785	0.57
Gaussian Naïve Bayes [e]	0.708	0.431
Logistic Regression [f]	0.723	0.456
Deep Learning Neural Network	0.748	0.498

a) *Area Under Receiver Operating Characteristic Curve*
b) *Matthews Correlation Coefficient*
c) *C=1.0, cache_size=200, class_weight=None, coef0=0.0, degree=3, gamma='auto', kernel='rbf', ,shrinking=True, tol=0.001*
d) *leaf_size=30, metric='minkowski',_neighbors=5,weights='uniform'*
e) *priors=None*
f) *C=1.0, class_weight=None, dual=False, fit_intercept=True,intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1, penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False*

In order to differentiate the models, To assess predictive power we have used 10-fold cross-validation i.e. we have divided our training data into 10 consecutive folds and having them into the shuffled ones in order to increase the randomness and As a validation of the 9 remaining folds form the training set we can use each fold. Each method is trained 10 times, every time on 90% of dataset and withholding is 10% for final evaluation.

Second, Using Keras we have constructed a deep Neural Networking, application Program Interface having a high level neural networks which is running on top of Tensor Flow, a software which is open source library for numerical computation is used for data flow graphs. In our experimental network it consists of four completely connected layers: one which is input layer fed among the training patterns, two complete hidden layers of 256 units of each using RELU (Rectified Linear Units) which is for functioning the activation and a single unit which outputs result value 0 or 1 based on a classic sigmoid function on output layer. The function loss is used in the network for the optimization process which is set to “binary cross-entropy” and ADAM respectively. We have provided with a Dropout systematize mechanism for each hidden layer of probability p=0.5, in order to facilitate the process of learning the networks. The Dropout technique on each training case which contains random excluding several layers and proven to perfect reduce of overfitting. The internal validation to 10% of the training data to which we have set, retaining (like the previous experiments) 10% for the evaluation done for final and the network for 100 epochs which we have trained. Fig. 3 (a, b) which shows the progress of accuracy and loss while training.

B. Results

As we don’t measure any performance of those applied methods which are in terms of accuracy classification, i.e. the number of correct predictions made which is divided by total number of predictions which are made. Which is that due, especially while handling datasets with unbalanced variables, this is of the metric which may be false or confusing, significantly reporting the better results than real predicted value of respective model (the appearance often called “Accuracy Paradox”). Rather, as evaluator’s performance, under the Receiver Operating Characteristic Curve (ROC) and the Matthews Correlation Coefficient (MCC) are the Area which is being used.

An area which is under the ROC curve (which is sometimes referred as Statistic) are widely used for literature as performance for metric which is more reliable rather than normal and simple accuracy. The curve of ROC which is representing clear positive rate (TPR, “sensitivity”) opposed to false positive rate (FPR, “fall-out” calculated as 1 – “specificity”). An exact classification can achieve a score 1, where as guessing randomly corresponded to 0.5. As a measure of the quality of binary classification the Matthews correlation coefficient is being used, which a correlation coefficient between those predicted and observed classifications and which returns a value in between −1 and +1, a perfect prediction represents as +1, 0 as not better than random and final total disagreement between observation and prediction is -1.

Table 2 which is reported the average performing of all methods being applied, which shows that each of them performed well above the random postulate. Finally, Fig. 3(c) provides a graphical representation for the ROC curve of the trained Deep

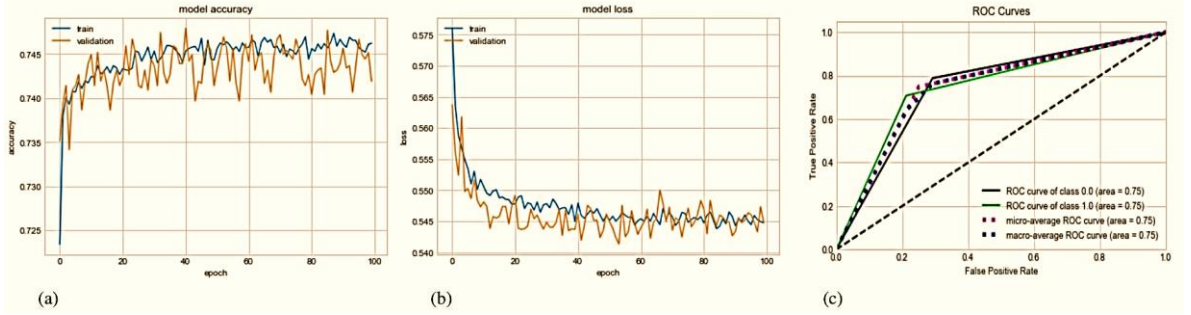


Figure 3. Progression of network accuracy (a), loss (b) during training and ROC curves (c) of the trained model

V. DISCUSSION

The data in this paper will provided a fact of concept that data mining methods can be successfully applied to non-medical patient hospitalization manifest and in order to achieve valid prediction of patient related outcomes such as probability of readmission due to less predictive possibilities by random guessing we using this tested algorithms outputs in models with more predictive possibilities better than random guessing. We certify various restriction related to both our approach and selected strategy. We have taken the data from only one hospital because the admissions in other hospitals are not calculated with this data and some bias related to the policies and the overall standard of the specific hospital. The primary objective of this paper was the data mining and model parameterization is use for minimal optimization in the process. To see the performance of the proposed models with different data handling selecting more features from the origin database and perform feature significance analysis and less aggressive balancing under-sampling optimization.

Ultimately, we prove that the same data valuable as the construction of a highly performing model and widely implemented ability to describe product environments, the abundance, datasets similar to the one used in this is over all credibility and easy for extraction, in production environments various Software Providers will lowers the implementation difficulty and application will be wide spread in software/production environments, which is essential in go over the gap between the real-world and research labs applications . By using the low privacy sensitivity of this kind of data will reduces the legal and ethical problems of their manipulation, portability and more effective in performance, we can use for research and production environments in order to increase efficiently.

REFERENCES

[1] "Readmissions-Reduction-Program," 2017. [Online]. Available: https://www.cms.gov/Medicare/Medicare-Fee-for-ServicePayment/AcuteInpatientPPS/Readmissions-ReductionProgram.html. [Accessed: 10-Feb-2018].

[2] K. J. Verhaegh, J. L. MacNeil-Vroomen, S. Eslami, S. E.

Geerlings, S. E. de Rooij, and B. M. Buurman, "Transitional Care Interventions Prevent Hospital Readmissions for Adults with Chronic Illnesses," Health Aff., vol. 33, no. 9, pp. 1531–1539, Sep. 2014.

[3] J. Futoma, J. Morris, and J. Lucas, "A comparison of models for predicting early hospital readmissions," J. Biomed. Inform., vol. 56, pp. 229–238, Aug. 2015.

[4] D. Kansagara et al., "Risk prediction models for hospital readmission: A systematic review," JAMA - J. Am. Med. Assoc., vol. 306, no. 15, pp. 1688–1698, Oct. 2011.

[5] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.

Cambridge: Cambridge University Press, 2000.

[6] L. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, p. 1883, 2009.

[7] H. Zhang, "The Optimality of Naive Bayes," Proc. Seventeenth Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2004, vol. 1, no. 2, pp. 1–6, 2004.

[8] D. Hosmer, S. Lemeshow, and R. X. Sturdivant, "ModelBuilding Strategies and Methods for Logistic Regression," in Applied Logistic Regression, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2005, pp. 91–142.

[9] D. Ravi et al., "Deep Learning for Health Informatics," IEEE J.

Biomed. Heal. Informatics, vol. 21, no. 1, pp. 4–21, Jan. 2017.

[10] Y. Bengio, "Learning Deep Architectures for AI," Found. Trends® Mach. Learn., vol. 2, no. 1, pp. 1–127, 2009.

[11] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, no. Oct, pp. 2825–2830, Jan. 2012.

[12] F. Chollet, "Keras," GitHub, 2015. [Online]. Available: https://github.com/fchollet/keras

[13] M. Abadi et al., "TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning," in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16), 2016, pp. 265–284.

[14] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," Proc. 27th Int. Conf. Mach.