Big Data Research 13 (2018) 65–75

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Novel Approach to Predict Hospital Readmissions Using Feature Selection from Unstructured Data with Class Imbalance ☆

Arun Sundararaman *, Srinivasan Valady Ramanathan, Ramprasad Thati

Health Analytics Solution Factory, Accenture, India

ARTICLE INFO

Article history:
Received 16 October 2017
Received in revised form 12 May 2018
Accepted 12 May 2018
Available online 1 June 2018

Keywords:
Predictive analytics
Unstructured data
Discharge summary
Class imbalance
Domain related stop words
Feature selection

ABSTRACT

Feature selection for predictive analytics continues to be a major challenge in the healthcare industry, particularly as it relates to readmission prediction. Several research works in mining healthcare data have focused on structured data for readmission prediction. Even within those works that are based on unstructured data, significant gaps exist in addressing class imbalance, context specific noise removal which thus necessitates new approaches readmission prediction using unstructured data. In this work, a novel approach is proposed for feature selection and domain related stop words removal from unstructured with class imbalance in discharge summary notes. The proposed predictive model uses these features along with other relevant structured data. Five iterations of predictions were performed to tune and improve the models, results of which are presented and analyzed in this paper. The authors suggest future directions in implementing the proposed approach in hospitals or clinics aimed at leveraging structured and unstructured discharge summary notes.

Definitions

AUC: area under the curve; it is used in classification analysis in order to determine which of the used models predicts the classes best.

Precision: also called positive predictive value is the fraction of relevant instances among the retrieved instances.

Recall: also called sensitivity is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

Specificity: measures the proportion of negatives that are correctly identified as such.

F-Score: is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score.

1. Introduction

Clinical data is getting increasingly complex on different dimensions viz., volume, variety and velocity impacting the quality of care [1]. Unique characteristics of medical data that makes it complex include varied data formats, high level of missing values, privacy of data etc. This increasing complexity is leading to the need for sophisticated methods and techniques such as clinical data mining or predictive analytics for clinical decision support. Predictive analytics is a branch of data mining that supports prediction of outcome and probability of such outcome, based on insights from trends and patterns from historical data. A predictive model seeks to predict the results of a response or dependent variable based on a set of variables known as predictor or independent or criterion variables. Recent advances in predictive mining have given rise to a trend where more and more clinical systems are adopting predictive analytics as part of their clinical decision support modules. In clinical or medical informatics predictive analytics predominantly deals with models to predict patients' health, to support clinicians in diagnostic, therapeutic, or other medical decision tasks [2]. Emphasizing the significance of predictive analytics in evidence based medicine, Sanjeev Sood [3] lists multiple application areas where such techniques are becoming an essential element of clinical systems, early detection of emerging diseases or spotting outbreak of epidemics etc.

The role of discharge summary and insights contained therein in the study of readmission have assumed recent significance. A recent study [4] establishes that high-quality discharge summaries were associated with reduced risk of readmission for patients with heart failure.

Over the years, different data mining techniques have been introduced for application to medical-related fields increasing the complexity of techniques behind the predictive system. Examples

☆ This article belongs to Special Issue: Medical Data Analytics.

* Corresponding author.
E-mail address: arunts@gmail.com (A. Sundararaman).

https://doi.org/10.1016/j.bdr.2018.05.004
2214-5796/

of such techniques include, but not restricted to, genetic algorithms or artificial neural networks or fuzzy sets or inductive logic programming [5].

Researchers and practitioners in predictive analytics are advised and encouraged to focus on selecting the most appropriate approach, given the widespread availability of several new computational methods and tools [6]. The need for such novel approaches assumes significance due to the inherent complexities and special characteristics associated with medical data listed in the previous paragraph. Of the several novel possibilities that exist, use of mixed-data for predictions is considered extremely significant. Most traditional predictive models are limited to handling datasets that contain either numeric or categorical attributes. However, data sets with mixed types of attributes are common in real life data mining applications and hence the need to introduce a new framework that focuses on generating insights from mixed data [7].

In this work, the authors study prediction of readmission post hospital discharge, which falls in the domain of predictive analytics within medical informatics. These predictions were carried out incrementally over 5 iterations. A super data set comprising nearly 11318 Congestive Heart Failure (CHF) admissions was adopted from MIMIC III database. MIMIC-III is a freely accessible critical care database widely available for research purposes [8]. Adequate volume of data, diversity of patients, availability of associated clinical notes, authenticity and originality of data were some of the factors considered in deciding on the choice of MMIC III database. Few data transformation steps that were followed in building this dataset are explained in detail in subsequent chapters. Predictive algorithms based on logistic regression were built on a set of variables from "structured" data and compared with results from other machine learning techniques viz., Structured Vector Machine (SVM) and Random Forest. In parallel, text mining was performed on medical notes related to readmission. The objective of this text mining exercise was to generate insights from the said "unstructured" data and use them for predictions.

Uniqueness of this work were incorporated in iterations 3 and 5. In iteration 3, class imbalance was addressed using a novel 2 steps feature selection process from unstructured, described in detail in Section 2. Another unique step i.e. domain related stop words removal was introduced in this iteration, to achieve higher level of noise reduction. In iteration 4, this model was extended to build a predictive model using both the structured and unstructured data for classification of a hospital readmission. It was observed that the predictive model based on mixed-data provides significant improvement over distinct data sets (structured and unstructured data sets), details of which are discussed in detail in subsequent chapters.

2. Background of the work

The complexity and volume of medical data keeps increasing the challenge of quality of care requiring the data to be translated into a standardized format for the clinician's uptake [1]. Predictive analytics helps handle such a situation of complexity and volume of data, by deriving formal intelligence from such huge volumes of structured and/or unstructured data; the intelligence so derived is used by clinicians in decision-making process to enhance the effectiveness of disease treatment and preventions [9]. Clinical predictive analytics deals with learning models from structured/semi-structured/unstructured data and aims to predict variety of things around patient care, such as investigative, curative, or observing/nursing tasks. Such models based on historic information coupled with data mining techniques support clinicians to transition from population-based to personalized medicine [2]. With abundance of data, another challenge that needs to be handled relates to the unique characteristics of medical data. K.J. Cios and G.W. Moore [10] deal in detail the uniqueness of medical data in the context of key difference of data mining in medicine as compared to other fields. The summary of the same is presented in this paragraph, which is very essential to set the context of special considerations that need attention while building predictive analytics models for clinical purposes. Factors contributing to uniqueness of health data include raw medical data being voluminous and heterogeneous, needs physician's interpretation, carries information in unstructured free-text, sensitivity, poor mathematical characterization of medical data, privacy and confidentiality requirements.

2.1. Predictive algorithms

Applications using predictive analytics algorithms for clinical data need to consider the special characteristics of medical data described in the previous paragraph. Owing to the uniqueness of medical data listed above, clinical predictive applications have distinguishing features. In clinical prediction, data sets can be small and be derived from non-reproducible situations. The predictive models need to be designed to handle complexities such as data being affected by several sources of uncertainty (e.g. measurement errors, missing data or coding errors). Predictive analytics for clinical data needs to cope up with these problems by carefully applying variable and model selection, by correctly evaluating the resulting models and by explicitly encoding this knowledge and using it in data analysis [11]. Set of guidelines that may apply to construction of clinical predictive models presented in [11] are listed below:

- Model the probabilities and not crisp class membership. Prefer methods that report confidence intervals.
- Avoid over-fitting. Never test models on data that was used in their construction.
- Test the resulting model on an independent separate data set.
- The project is not finished when a good model is found. Think how to include the model within some clinical information or decision support system.

Different predictive algorithms such as fuzzy logic or genetic algorithms or neural networks or decision trees have been widely used for medical data mining. Different statistical methods k-Nearest Neighbor or logistics regression or Bayesian classifiers have been applied to clinical data to support the above listed algorithms. Published literature recognizes the need to adopt more complicated regression models to achieve substantial predictive accuracy, in the study of hospital readmission prediction [12].

2.2. Mixed-data mining using unstructured data

With the evolving volume and complexity of data, particularly in the Healthcare domain, there is growing need to consider unstructured data in modeling predictions. A recent study [13] recommends devising new predictive models appropriate to unstructured data. A similar study [14] observed that using data imputation techniques on unstructured data resulted in significantly improved results over prior research. Use of text mining to generate finer insights from clinical notes and use of those insights for timely decisions is proving to be extremely important in clinical decisions support. Other related works suggest that benefits of using text mining is to get decision points more quickly and in enabling organizations to explore interesting patterns, models, directions, trends and rules contained in text form of "unstructured" data. [15,16]. Emphasis on the need to expand the horizons to encompass both text and structured numerical data is increas-

ing, more specifically in the context of predictive mining (using mixed-data mining) of electronic medical records [18].

2.3. Class imbalance & feature selection

Clinical data and text content are among the major real world examples of data sets that pose imbalance between the classes. A data set with class imbalance problem may yield strong overall accuracy but very poor performance on the minority class. Michael Wasikowski proposes feature selection as an evolving technique to resolve class imbalance problem [23]. 3 different approaches exist to combat the class imbalance problem viz., resampling methods, new algorithms and feature selection methods. As per published literature, majority of the research so far has focused on resampling methods and a small extent of research on new algorithms and least attention to feature selection [23].

Feature selection contributes to boosting prediction accuracy by reducing dimensionality of the dataset. The key criteria for successful feature selection lies in the ability to minimize the number of selected features while retaining, as much as possible, the overall prediction information [21]. Most of the published literature focus on methods that are applicable to structured data such as filter, wrappers, hybrid and embedded. Feature selection has been used to yield improved results in text mining domain [24].

2.4. Discharge summary

Study of intelligence and insights from hospital discharge summaries has been a subject of interest both for medical practitioners and information science specialists. Medical experts propose that improving the quality of discharge summary (e.g. suggestions on follow-up tests or clarifications on medication etc.) help in avoiding readmissions. In a recent study [20] involving randomized controlled trial found that high-quality discharge summaries were associated with reduced risk of readmission for patients with heart failure. A more recent study [17] involved use of logistic regression and n-fold cross validation techniques on information derived from unstructured medical data.

There is a comprehensive definition of stop words and their handling governed by the context of their domain application area [25]. Filtering texts using domain vocabulary is recognized as an approach to solve high dimensionality problem [26]. However, this aspect of domain words removal in the text mining continues to be less explored.

2.5. Gaps in existing work

A stated gap with data mining application is that it considers only part of the information i.e. only numerical or structured data; however, a good deal of information is present in texts or unstructured data such as clinical reports or medical notes or bed side notes.

Based on the discussions above, the key gaps in existing works are as below:

- Need to achieve higher and targeted noise reduction in unstructured data.
- Information Gain from unstructured data.
- Traditional approaches to class imbalance problem are around oversampling & under sampling. There is a stated need to focus on feature selection and explore more optimal feature selection methods [22].
- Domain related stop words removal has not been explored in detail to assess their impact on prediction accuracy.

In summary, this work seeks to address the above gaps using insights from unstructured data along with structured data for hospital readmission prediction. This also introduces a novel approach to noise reduction in unstructured data.

The research question that this work attempts to answer is "does use of insights generated from unstructured text data in clinical notes help improve accuracy of hospital readmission prediction?"

3. Approach and methodology

In order to address the above gaps and seek answers to the question mentioned in the previous section, this work followed a novel approach in building a predictive model using both the structured and unstructured data for classification of hospital readmission. Though multiple research studies were published on hospital readmission prediction, a vast majority of them focus more on using structured data while leaving the associated unstructured data to limited or no use. Besides such structured information, large volumes of useful information are being captured and made available in unstructured form (e.g. clinical notes or nurses observations etc.). However, making effective use of these packets of information for data mining, gaining insights on patterns and deriving intelligence to predict outcomes has remained a large gap in medical informatics. This problem has been true both among medical/Information Technology practitioners and researchers.

This study approaches the ability to predict hospital readmissions from patient descriptive, clinical notes, discharge summary, diagnosis and procedures and admission events.

The work involved multiple steps of predictions i.e. using only structured data, using only unstructured data and using mixed-data. The objective was to compare the results in these iterations to examine if use of unstructured or mixed-data improved prediction accuracy. Fig. 1 provides a macro view of these iterations.

Based on this approach the research question is restated as below:

Research Question: **Is C > (A or B)?** Given, A = Prediction based on structured data, B = Prediction based on unstructured data and C = prediction based on mixed data.

The potential alternative outcomes and inferences may be

1) If C < A or if C < B, use of mixed-data does not improve prediction accuracy
2) If C > (A or B), use of mixed-data improves prediction accuracy in select cases
3) If C < (A and B), use of mixed-data improves prediction accuracy

We approached the problem by following a layered methodology to perform machine learning using Logistic regression, structured vector machine and random forests on structured data and extending to generate more features from unstructured text, by text mining. Text mining generated term frequency of N-gram features. We also used a unique method of selecting N-grams features in positive class and not in negative class and vice-versa.

Accenture's Health Analytics Solution Factory in Advanced Technology Centers-India comprises industry focused health data scientists and techno functional professionals engaged in exploring innovative solutions in health Analytics. This research work was carried out by these researchers and practitioners as an ongoing industry innovation initiative.

3.1. Data source and data transformation

It is extremely important to use reliable and large data set collected over a period of time and from a large cross section of

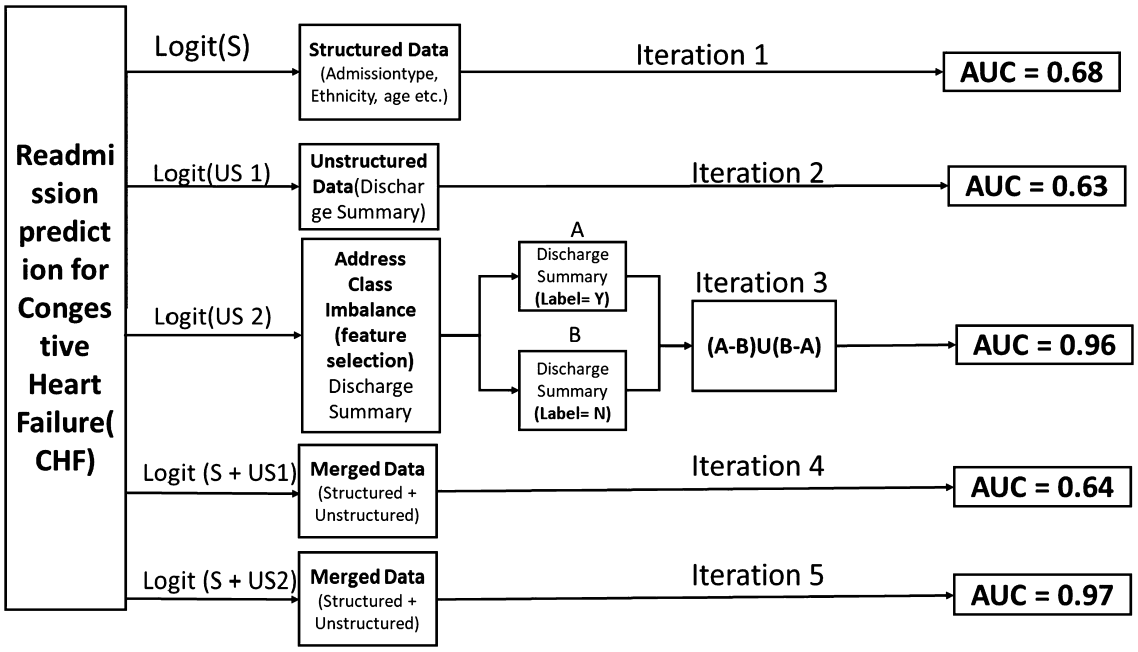

Fig. 1. Overview of research work.

the population to build and test the proposed predictive model. The robustness of the predictive model is largely dependent on the quantity of data and quality of data (limited missing values or noise data). MIMIC-III, a freely accessible critical care database was collected at Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA, USA during the period from 2001 to 2012. The latest version of MIMIC is MIMIC-III, which comprises over 58,000 hospital admissions for 38,645 adults and 7,875 neonates. MIMIC III is widely adopted database in the research community [8].

MIMIC III database is used for the predictive models detailed in this study. This data includes unstructured notes for every hospital admission viz. discharge summary, nursing notes and progress notes. This study used Hospital admissions data of CHF consisting of 11318 admissions across 8416 patients. The data selected for this study meets the volume and variety characteristics and thus is considered appropriate for big data research. In practice too, real-world data involving discharge summaries, nursing notes and progress notes in a hospital are typically generated from a complex Electronic Medical Records System and are ingested into a Data Lake or other Big Data architecture for further research and analysis.

The process of building the predictive model may be largely split into 3 phases, viz. data preparation or feature engineering, training the model and prediction. Before using raw data, a set of transformations and exclusions were performed to make the data ready for use in the model training and prediction. The unstructured features were generated from the discharge summary. This constitutes the data preparation phase. Details of such preparation and transformations are listed in Table 2. The rest of the variables are used without any transformations. The independent univariate descriptive of the numeric and categorical variables are listed in Table 1.

Structured data: Hospital readmission measures have been touted not only as a quality measure but also as a means to bend the healthcare cost curve. The Affordable Care Act (ACA) established the Hospital Readmission Reduction Program (HRRP) in 2012. Under this program, hospitals are financially penalized if they have higher-than-expected risk-standardized 30-day readmission rates for acute myocardial infarction, heart failure, and pneumonia [19]. Re-admission is a critical business challenge from a cost and patient satisfaction perspective in the presence or absence of ACA. A patient is considered to be readmitted if the subsequent admission occurs in less than 30 days and a flag for readmission is generated keeping this in consideration and whether the admission caused a readmission. Fatal Patient admissions are kept out of the model since patient expiry cannot lead to readmission. Length of stay at the hospital, number of intensive care units, number of previous admissions to the hospital, number of previous readmissions, and length of stay at ICU are calculated from the data for every admission.

Unstructured data: Text preprocessing of the unstructured text is an important step in noise reduction. The steps followed include removal of punctuation, whitespace, URL, natural language stop words, numeric contents. The discharge summaries were lemmatized to bring words to a common root word. Discharge summary containing lot of linguistics or common domain specific words and abbreviations presents us with unique challenge of introducing noise in the data. For e.g. the words “Patient” or “discharge” may be present in the discharge summary, although not a natural language stop word. The domain specific stop word removal helps in noise reduction and focuses on features that increase information gain [25,26]. Some of the features like number of past illness and number medications were generated using regular expressions. Discharge summaries are converted into term frequency of N-gram features after pre-processing for e.g. arrhythmia.cardiomyopathies.sudden. This provides a high dimension sparse feature matrix. Some examples of domain specific stop words are Hospital, Doctor, Discharge, Heart rate, Temperature, Chloride, Ambulatory etc.

3.2. Feature selection

As discussed in the previous section, Features needs to be selected from a high dimensional sparse feature matrix. From Table 1 the event rate of hospital readmission for CHF is 9% which is highly imbalanced. Zheng et al. [24] suggest that existing measures used for feature selection are not very appropriate for imbalanced data sets. The authors propose a feature selection framework, which selects features for positive and negative classes separately and then explicitly combines them. The authors show simple ways of con-

Table 1
Univariate descriptive of the numeric and categoric variables.

Variable	Label: Y	Label: N	Entire data
Admission type	ELECTIVE:4.6% URGENT:1.93% NEWBORN:0% EMERGENCY:93.4%	ELECTIVE:11.6% URGENT:3.0% NEWBORN:0.1% EMERGENCY:85.1%	ELECTIVE:11.14% URGENT:3.0% NEWBORN:0.1% EMERGENCY:85.77%
Insurance	Medicare:79.1% Private:13.42% REST:7.4%	Medicare:73.7% Private:18.72% REST:7.58%	Medicare:74.12% Private:18.31% REST:7.57%
Language	ENGL:78.38% SPAN:6.02% REST:15.6%	ENGL:73.37% SPAN:6.82% REST:19.81%	ENGL:75.60% SPAN:6.75% REST:17.65%
Admission location	EMERGENCYROOM ADMIT:44% TRANSFERFROM HOSP/EXTRAM TRANSFER:21.3% CLINIC REFERRAL/PREMATURE:18.6% REFERRAL/NORMAL DELI:14.9% REST:0.2%	CLINIC REFERRAL/PREMATURE:23.7% TRANSFER FROM HOSP/EXTRAM TRANSFER:14.6% REFERRAL/NORMAL DELI:8.4% EMERGENCYROOM ADMIT:5.16% REST:48.4%	CLINIC REFERRAL/PREMATURE:19.05% TRANSFER FROM HOSP/EXTRAM TRANSFER:20.86% REFERRAL/NORMAL DELI:14.4% EMERGENCYROOM ADMIT:44.6% REST:1.2%
Discharge location	SNF:27.5% HOME HEALTH CARE:27.18% REHAB/DISTINCT PART HOSP:16.6% LONG TERM CARE HOSPITAL:14.2% REST:14.52%	HOME HEALTH CARE:29.36% SNF:26.54% REHAB/DISTINCT PART HOSP:16.67% LONG TERM CARE HOSPITAL:6.9% REST:20.53%	HOME HEALTH CARE:29.19% SNF:26.62% REHAB/DISTINCT PART HOSP:16.76% LONG TERM CARE HOSPITAL:7.47% REST:19.96%
Marital status	MARRIED:45.05% SINGLE:23.66% WIDOWED:19.9% REST:11.39%	MARRIED:48.9% SINGLE:20.6% WIDOWED:22.14% REST:8.36%	MARRIED:48.67% SINGLE:20.84% WIDOWED:21.97% REST:8.52%
Ethnicity	WHITE:67.69% BLACK/AFRICAN AMERICAN:19.22% REST:13.09%	WHITE:72.21% BLACK/AFRICAN AMERICAN:10.39% REST:17.4%	WHITE:71.86% BLACK/AFRICAN AMERICAN:11.07% REST:17.07%
ICU hours	Mean:3.9 hours SD:3.93 hours	Mean:3.29 hours SD:4.11 hours	Mean:3.34 hours SD:4.1 hours
Previous admission	Mean:1.19 SD:2.05	Mean:0.47 SD:1.24	Mean:0.52 SD:1.33
Previous readmission	Mean:0.36 SD:0.80	Mean:0.11 SD:0.43	Mean:0.13 SD:0.47
Age group	66-80:39.13% 51-65:23.43% 80-100:22.07% REST:15.37%	66-80:38.56% 51-65:23.0% 80-100:21.4% REST:17.04%	66-80:38.61% 51-65:23.03% 80-100:21.51% REST:16.85%
LOS	Mean:11.54 days SD:11.20 days	Mean:10.72 days SD:9.85 days	Mean:10.78 days SD:9.97 days
ICU visits	Mean:1.91 SD:2.1	Mean:1.01 SD:1.34	Mean:1.08 SD:1.43
Count of Past ICD's	Mean:10 SD:7	Mean:9 SD:6	Mean:9 SD:7
Medications count	Mean:3 Sd:4	Mean:3 SD:4	Mean:3 SD:4
N gram feature	acut.chronic.systol aerosol.inhal.puff glargin.unitml.solut	saphen.vein.graft spontan.echo.contrast stenosis.aortic.regurgit	acut.chronic.systol spontan.echo.contrast stenosis.aortic.regurgit
Readmission flag			Class Label 'Y': 91% Class Label 'N': 9%

verting existing measures so that they separately consider features for negative and positive classes [24]. Odds Ratio compares the odds of a feature occurring in one category with the odds for it occurring in another category. In this study, in iteration 3 of the model we select features that are unique to re-admissions/non-re-admissions and combine them as a feature set to give maximum information gain and thereby reduce noise. This novel approach to select features for predicting hospital re-admissions is unique in its application.

3.3. Predictive modeling

The models were built in multiple iterations as presented in the section 1 of the paper. These models were built using machine learning packages in R software. The models were trained using 10-fold cross validation to ensure the generalization of the models. Logistic regression is used to model the problem across iterations.

Logistic regression: Logistic Regression is a discriminative classifier that models the posterior p(y1|X) directly given the input

Table 2
Variables with business rules.

Variable	Approach
Age group	DOB-admit date and then converted numeric to bins eg: 0–20, 21–40, 41–65 etc.
Previous admissions	count of previous distinct admit date to current admit date
Previous readmissions	calculated 30 days' readmission for each patient, if patient readmitted within 30 days then flagged as Y, then count of Y till current admission
Length of stay	Discharge date – Admit date
ICU visits	count of distinct ICU Admit time for each admission
ICU hours	ICU Discharge time-ICU Admit time for each icu visit, then sum of all the values for each admission
Past medical history (count of past ICD's)	Count of past diseases from the discharge summary E.g.: count of numbers (starting of the line) 1. Congestive heart failure (with an ejection fraction of 15% to 20%). 2. Type 2 diabetes with neuropathy. 3. Hypertension. 4. Diverticulosis 5. Alzheimer's dementia
Medications count	Count of medications from the discharge summary E.g.: count of '– 'symbol (starting of the line) – Metoprolol 50 mg p.o. b.i.d. – Captopril 6.25 mg p.o. t.i.d. – Aspirin 325 mg p.o. q.d. – Pantoprazole 40 mg p.o. q.d. – Heparin 5000 units subcutaneously b.i.d.

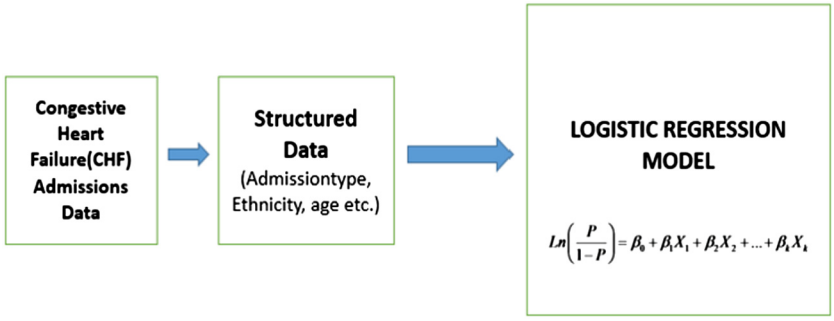

Fig. 2. Iteration 1 – model approach.

features. That is, it learns to map the input (X) vector directly to the output class label y1 (risk in our case). When the response is a binary (dichotomous) variable, logistic regression fits a logistic curve to the relationship between X and y1. The class decision for the given probability is then made based on a threshold value. The threshold is often set to 0.5, i.e. if p(y1|X)! 0.5, then we predict that the next readmission of the patient is likely within 30 days; otherwise not.

Multicollinearity: The correlation between predictor variables is an important factor affecting the model accuracy. The correlation was verified using Pearson correlation matrix and the collinear columns were removed. In the case of unstructured text the correlations were calculated among term frequency of the N-grams. The trigrams like act.bed.admissions and admission.act.bed turned out to be collinear because the counts were similar for common tokens.

1) Iteration 1: In this initial iteration of model, only structured data such as admission type, age etc. were used as input features to the model. In the raw data set, the imbalance was very high i.e. positive readmission cases at 9% and negative readmission cases at 91%. The class imbalance was addressed using SMOTE over sampling on the minor class. This resulted in a 46%, 54% ratio between instances of the two classes after applying the over sampling technique. Random under sampling approach was also considered; however, under sampling often tends to yield a biased sample, which is not representative of the population, impacting accuracy of results with actual test data set. The model approach of iteration 1 is given in Fig. 2.
2) Iteration 2: In this iteration, we convert the unstructured text through a series of pre-processing steps such as case conversion, stemming, whitespace and punctuation removal, domain and natural language stop word removal to generate term frequency of trigrams. The choice of N as 3 in N-grams was made based on the significance of variables by p-value. The value of n as 2 and 4 increased the information loss. The effect of term frequency of trigrams on re-admission was modeled. The model approach of iteration 2 is given in Fig. 3.
3) Iteration 3: In this iteration of model, class imbalance was addressed using feature selection. The text preprocessing steps similar to those in iteration 2 were performed. An additional step of domain related stop words removal was introduced in this iteration, to achieve higher level of noise reduction. The trigram features were reduced by considering features or trigrams in Label 'Y' (re-admissions) and not in Label 'N' (Non re-admissions) and vice versa. These features were combined

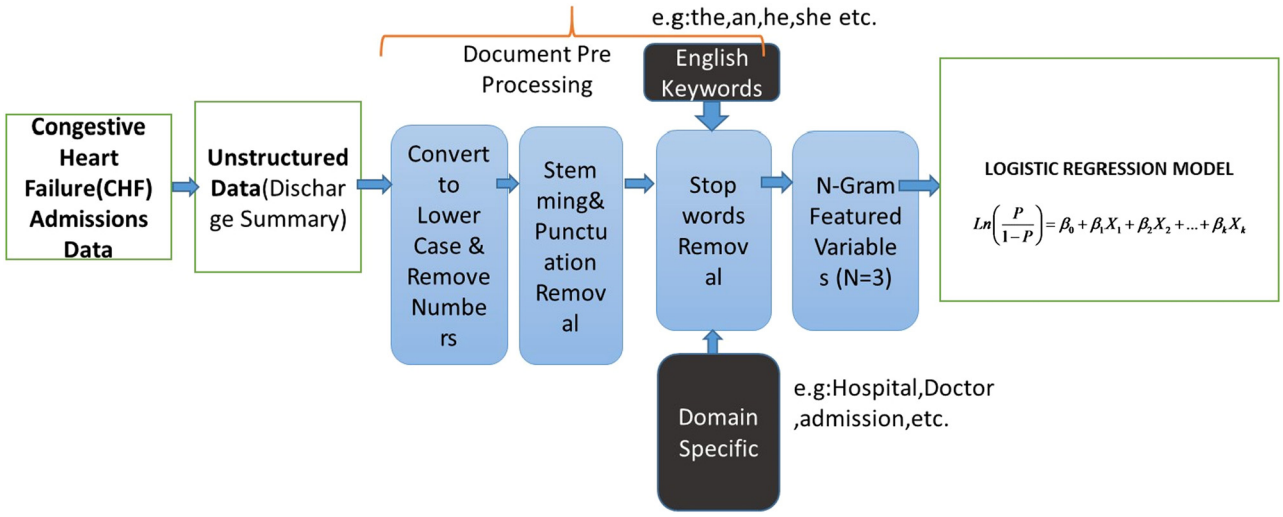

Fig. 3. Iteration 2 – model approach.

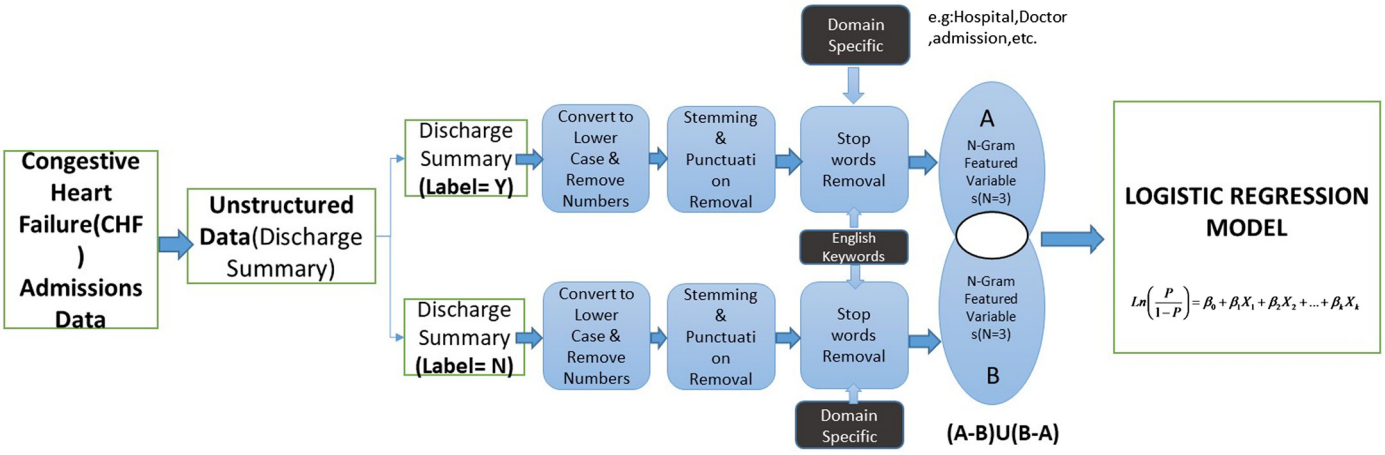

Fig. 4. Iteration 3 – model approach.

to make a prediction on re-admission. A diagrammatic representation of the model is presented in Fig. 4.

4) Iteration 4: This iteration combined structured (Iteration 1) and unstructured data (iteration 2) into the model to explore potential accuracy improvement by using mixed data. The mixed data contains structured variables and trigram features as in iteration 2 for modeling. The model approach is given in Fig. 5.
5) Iteration 5: This iteration combines the feature sets of iteration 1 and iteration 3 to leverage mixed data of structured and unstructured to make predictions. The model approach is represented in Fig. 6.

4. Results, analysis and findings

4.1. Iterations and results

The focus of the iterations was to observe the model's incremental performance across measures such as Area Under the Curve (AUC), precision, recall, specificity and F Score. The output of models across five iterations are tabulated in Table 3.

The precision is the measure of correctness achieved in positive prediction i.e. of observations labeled as positive, how many are actually labeled positive. The recall is a measure of actual observations which are labeled (predicted) correctly i.e. how many observations of positive class are labeled correctly. It is also known as 'Sensitivity'. The precision recall curve of the iterations is shown in the Fig. 7.

The frequency distribution of individual probability of re-admissions for models in iteration 1, 2, 4 are given in Fig. 8.

The frequency distribution of individual probability of re-admissions for models in iteration 3 & 5 are given in Fig. 9.

4.2. Analysis & findings

The machine learning models from iterations and datasets described in Section 3 were evaluated and results were presented in the previous sub-section. In this sub-section, we present the Analysis and key findings from the results.

The prediction performance of a largely imbalanced dataset such as hospital re-admission is enhanced based on actions taken to overcome class imbalance. Fig. 7 & Table 3 from the results section show an accuracy of >90% for iteration 1, 2 & 4 when not handled for class imbalance; but, AUC, Precision and recall are very low. It may be inferred from these results that accuracy is good because of class 'N' (non-readmissions). Hence there is need for a method to address class imbalance and improve AUC, precision and recall.

On using feature selection to overcome class imbalance as described in Iteration 3 & 5, significant improvement was observed in AUC, precision, recall and F-score. These measures give more importance to class 'Y' (re-admissions) where there is more inter-

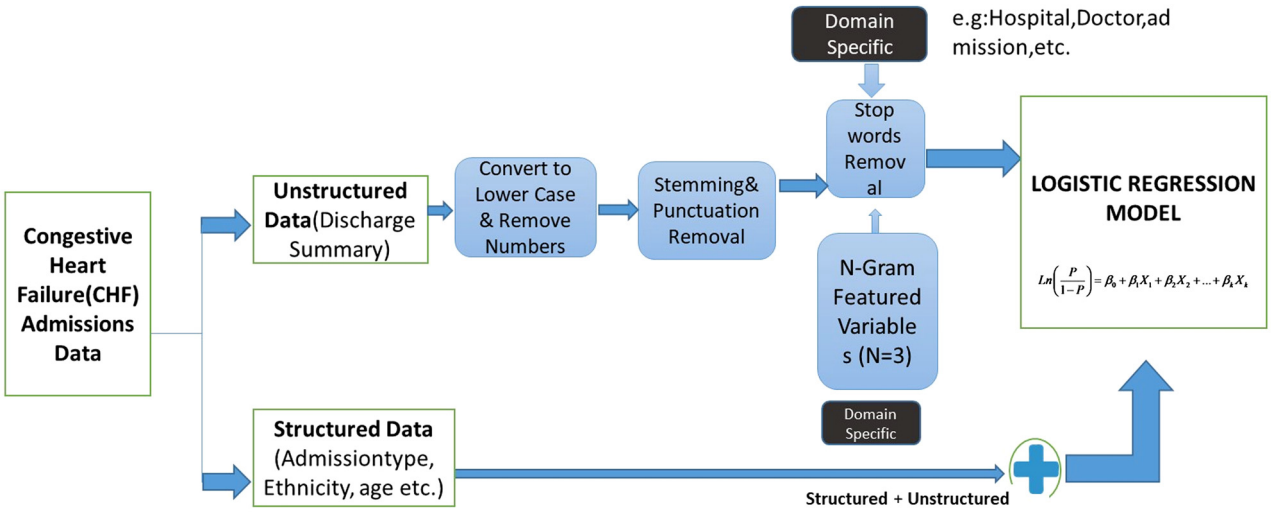

Fig. 5. Iteration 4 – model approach.

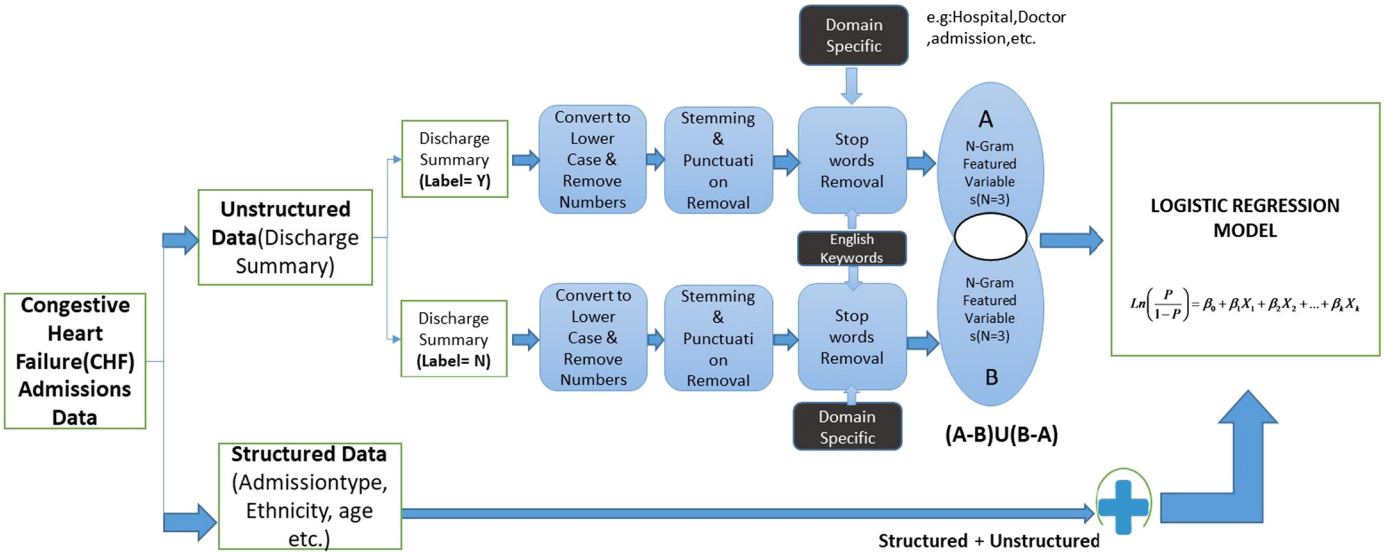

Fig. 6. Iteration 5 – model approach.

Table 3
Results from iterations.

Models	AUC	Accuracy	Precision	Recall/sensitivity	Specificity	F-Score
Iteration 1 (structured data)	0.68	0.91	0.48	0.13	0.99	0.20
Iteration 2 (unstructured data)	0.63	0.92	0.3	0.01	0.99	0.019
Iteration 3 (address class imbalance (feature selection))	0.96	0.98	1	0.86	1	0.92
Iteration 4 (mixed data – S + US1)	0.64	0.92	0.18	0.01	0.99	0.018
Iteration 5 (mixed data – S + US2)	0.97	0.98	0.99	0.87	1	0.93

est in improving the performance. The high values of precision and recall in the top quadrant can be seen in the Fig. 7.

Frequency distribution of individual admissions from iteration 1, 2 & 4 as depicted in Fig. 8 suggest that the probabilities are right skewed at class 'Y'. It also suggests the probabilities are scattered around the probability range, making the choice of decision threshold dependent on cost sensitivity. The choice to be made is between false negatives and false positives. Frequency distribution of individual admissions from iteration 3 & 5 as depicted in Fig. 9 suggests that the probabilities are distributed either to the left tail or the right tail and no values around the center. This shows discriminative power of the models in these iterations and makes the choice of decision threshold much easier.

From the above discussion, it is evident that models in iteration 3 & 5 outperforms other models. However, the incremental performance improvement from iteration 3 to iteration 5 is marginally around 1%. The difference between the approach being the inclusion of unstructured data along with structured data in iteration 5. This marginal improvement may apply only to the CHF re-admissions sample from MIMIC database. While for re-admission for other diagnosis or different data sources may have changes to the incremental performance. Although marginal from a statistical perspective, medical experts opine that even marginal improvements are treated as steps in right direction, since it is expected to have a positive impact on optimizing re-admission related costs and better patient experience by reducing risk of re-

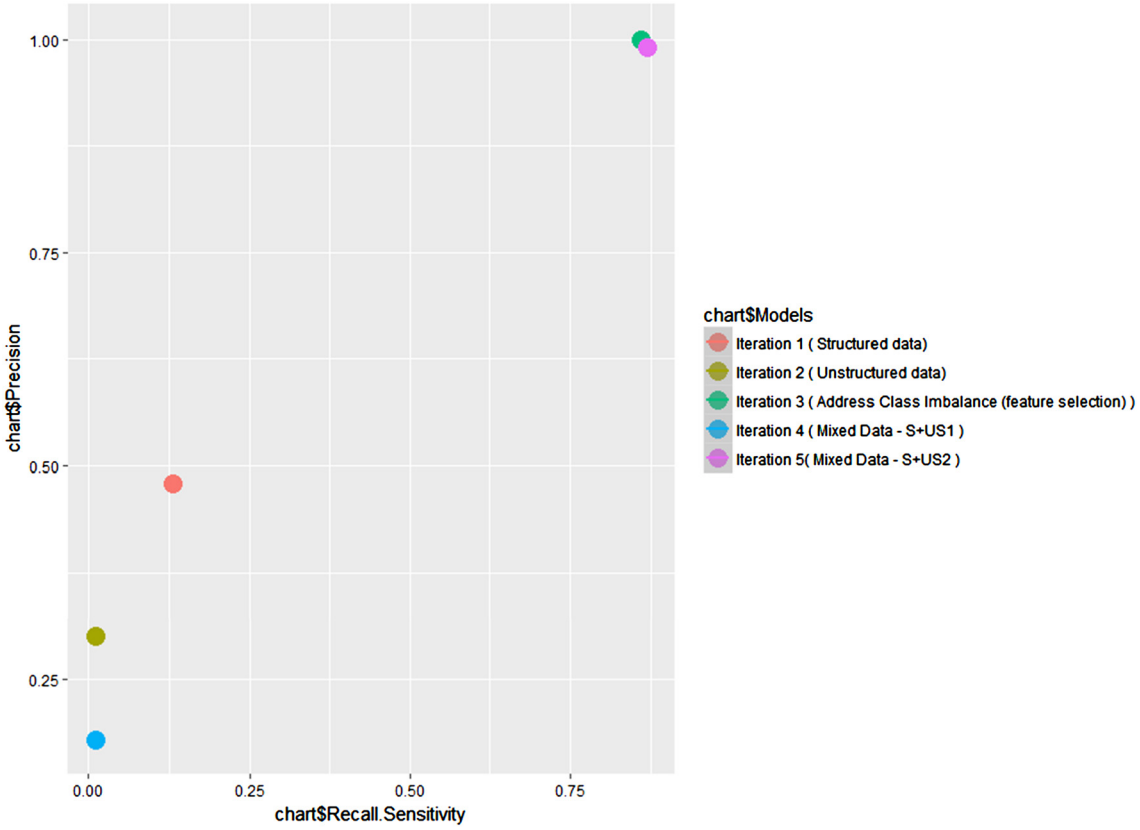

Fig. 7. Precision and recall chart of iterations. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

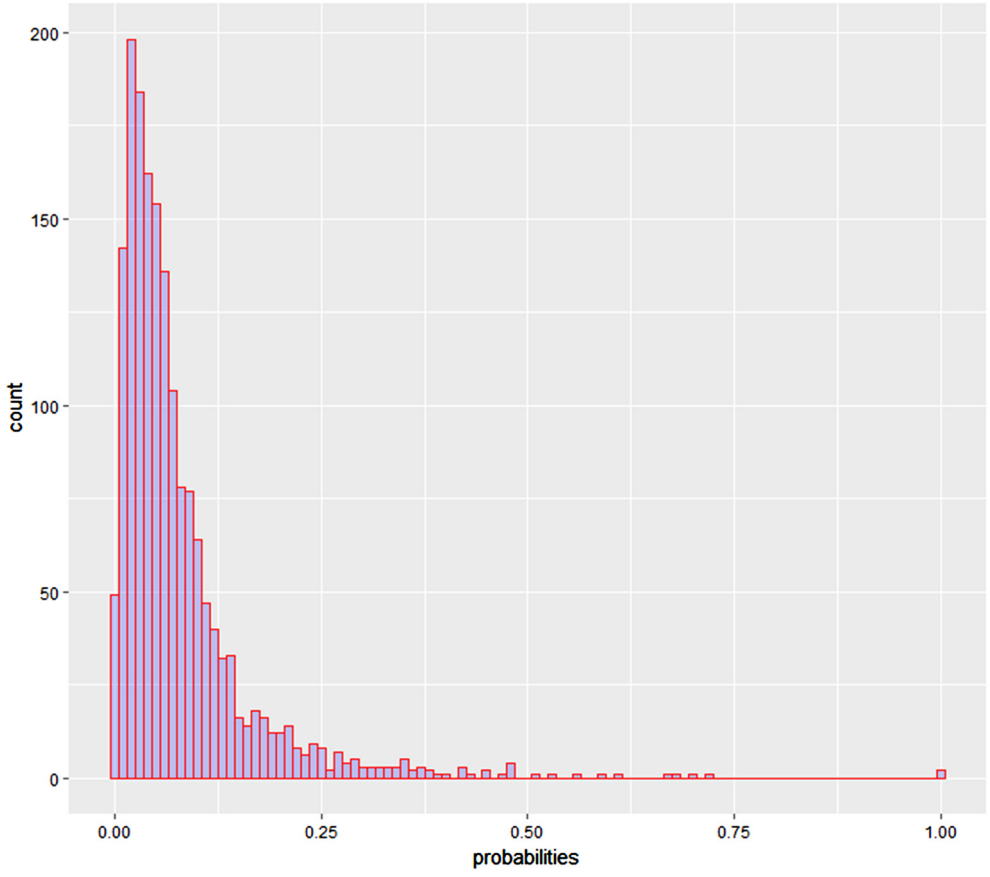

Fig. 8. Frequency distribution of probabilities in iteration 1, 2 & 4.

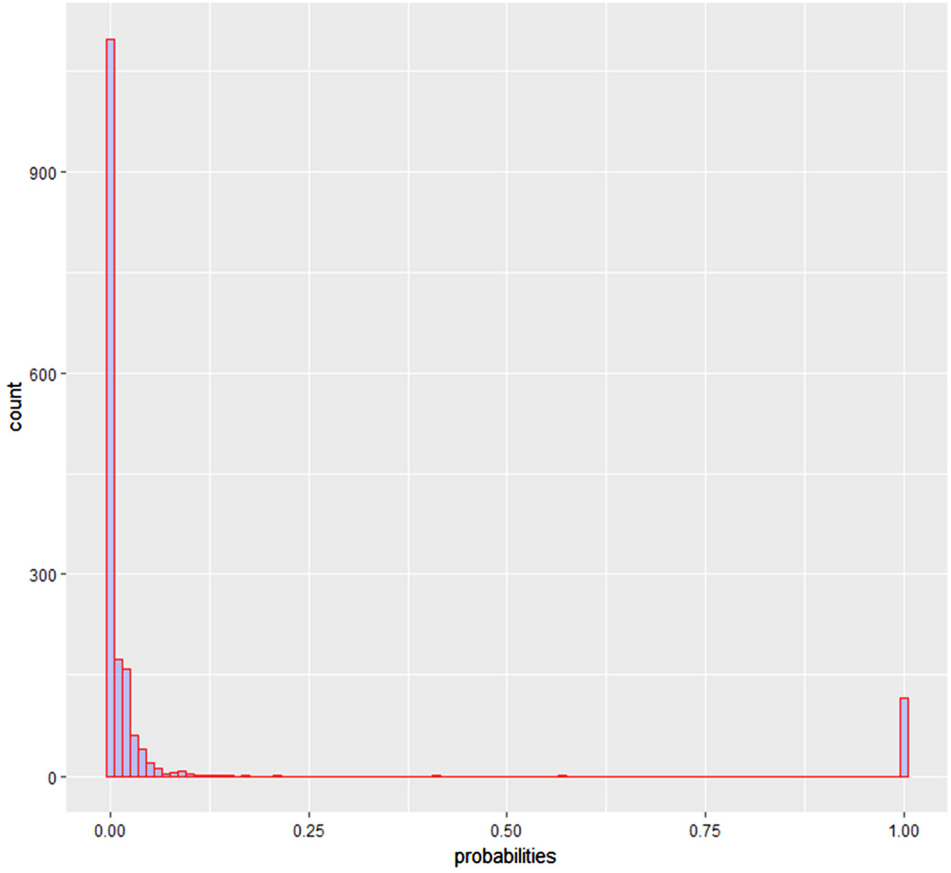

Fig. 9. Frequency distribution of probabilities in iteration 3 & 5.

Table 4
Gaps addressed through this work.

Gap	How addressed
Need to achieve higher and targeted noise reduction in unstructured data	In this study we selected features that are unique to re-admissions/non-re-admissions and combined them as a feature set to give maximum information gain and thereby reduce noise. This novel approach to select features for predicting hospital re-admissions is unique in its application
Information Gain from unstructured data	As above
There is a stated need to focus on feature selection and explore more optimal feature selection methods	The class imbalance was addressed using SMOTE over sampling in this model on the minor class. This resulted in a 46%, 54% ratio between instances of the two classes after applying the over sampling technique
Domain related stop words removal has not been explored in detail to assess their impact on prediction accuracy	This was addressed in iteration 3 (described in Section 3)

admission. Thus, it is recommended that iteration 5 be chosen for such research and applications.

5. Conclusion and future work

This study introduced 3 novel approaches in application of machine learning to hospital readmission, viz.,

1) Feature selection from mutually exclusive N-Grams to overcome class imbalance challenge.
2) Enhanced noise reduction using domain specific stop word removal.
3) Approach to extend features using both structured and unstructured data to improve model performance.

This work has addressed the gaps stated in existing work (summarized in Section 2) as listed in Table 4.

Future work and research directions are recommended as below.

- The trigram features of the unstructured text tend to change dynamically which requires re-training the model very frequently. Hence an automatic method to include new features in the model dynamically would be an ideal future step.
- The study used discharge summaries for prediction of readmissions. The possibility of predicting readmission probability for every progress of a patient recorded in the progress notes provides an immense potential for on-demand risk identification.

- Quality of data in discharge summary is expected to improve readmission prediction and experts have recommended the need to create high quality discharge summaries [17]. Automated mechanisms to route intelligence from prediction models to Health Information Systems Applications as a means to "closing-the-feedback-loop" needs to be developed.
- While the class imbalance problem was addressed in this work, it is observed that class imbalance is not the only problem responsible for the decrease in performance of learning algorithms [22]. Other data related problems such as distribution of the data within each class, degree of data overlapping among the classes, small disjunction and rare cases problems are also relevant in machine learning. Future work may be explored to handle feature extraction addressing these issues.
- Experiments in a recent work took patient data from a single institution and produced a statistical risk prediction model optimized for that institution [27], because different hospitals may have different characteristics in their patient populations. This approach may be applied to the readmission problem by localizing to individual facilities within a hospital group if this approach improves prediction accuracy for select diseases and/or conditions.

References

[1] Backere, et al., Automated generation and deployment of clinical guidelines in the ICU, in: Proceedings of the 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems, 2010, pp. 197–202.

[2] R. Bellazzi, F. Ferrazzi, L. Sacchi, Predictive data mining in clinical medicine: a focus on selected methods and applications, in: WIREs Data Mining Knowledge Discovery, 2011, pp. 416–430.

[3] S. Sood, Leveraging data analytics in healthcare – some interesting case reports, Indian J. Med. Inform. 6 (2012).

[4] Mohammed Salim Al-Damluji, et al., Association of discharge summary quality with readmission risk for patients hospitalized with heart failure exacerbation, Circ. Cardiovasc. Qual. Outcomes 8 (1) (2015) 109–111.

[5] R.H. Lin, An intelligent model for liver disease diagnosis, J. Artif. Intell. Med. 47 (2009) 53–62.

[6] D. Mishra, A.K. Das Mausumi, S. Mishra, Predictive data mining: promising future and applications, Int. J. Comput. Commun. Technol. 2 (1) (2010).

[7] M.L. Kassi, A. Berrado, L. Benabbou, K. Benabdelkader, Towards a new framework for clustering in a mixed data space: case of gasoline service stations segmentation in Morocco, in: 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, 2015, pp. 1–6.

[8] A.E.W. Johnson, T.J. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data (2016), https://doi.org/10.1038/sdata.2016.35.

[9] E. AbuKhousa, Predictive data mining to support clinical decisions: an overview of heart disease prediction systems, in: International Conference on Innovation in Information Technology, 2012, pp. 267–272.

[10] K.J. Cios, G.W. Moore, Uniqueness of medical data mining, Artif. Intell. Med. 26 (2002) 1–24.

[11] R. Bellazzi, B. Zupan, Predictive data mining in clinical medicine: current issues and guidelines, Int. J. Med. Inform. (2006) 81–97.

[12] J. Futoma, J. Mooris, J. Lucas, A comparison of models for predicting early hospital readmissions, J. Biomed. Inform. 56 (2015) 229–238.

[13] A. Gandomi, M. Haider, Beyond the hype: big data concepts, methods, and analytics, Int. J. Inf. Manag. 35 (2) (2015) 137–144.

[14] K.J. Nishant, et al., Soft computing based imputation and hybrid data and text mining: the case of predicting the severity of phishing alerts, J. Expert Syst. Appl. 39 (12) (2012) 10583–10589.

[15] R. Hale, Text mining: getting more value from literature resources, Drug Discov. Today 10 (6) (2005) 377–379.

[16] S. Chakrabarti, Mining the Web: Analysis of Hypertext and Semi Structured Data, Morgan Kaufman, 2000.

[17] Salim Al-Damluji, et al., Association of discharge summary quality with readmission risk for patients hospitalized with heart failure exacerbation, Circ. Cardiovasc. Qual. Outcomes 8 (1) (2015) 109–111.

[18] Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, Data sources for prediction: databases, hybrid data and the web, in: Fundamentals of Predictive Text Mining, Springer, 2015, pp. 147–164.

[19] Colleen K. McIlvennan, Zubin J. Eapen, Larry A. Allen, Hospital Readmissions Reduction Program, 2016.

[20] Sérgio Curto, Joao P. Carvalho, Cátia Salgado, Susana M. Vieira, João M.C. Sousa, Predicting ICU Readmissions Based on Bedside Medical Text Notes, 2002.

[21] Rita Domingues Viegas, Feature Extraction for Modeling Patients' Outcomes: an Application to Readmissions in ICUs, 2015.

[22] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas. Handling imbalanced datasets: a review, GESTS Int. Trans. Comput. Sci. Eng. 30 (2006) 25–36.

[23] Michael Wasikowski, Combating the Class Imbalance Problem in Small Sample Data Sets, Ph.D. Thesis, Department of Electrical Engineering & Computer Science and the Graduate Faculty of the University of Kansas School of Engineering, 2009.

[24] Z. Zheng, X. Wu, R. Srihari, Feature selection for text categorization on imbalanced data, ACM SIGKDD Explor. Newsl. 6 (2004) 80–89.

[25] C.S. Suresh, B. Vikranth, K.U. Siba, K.P. Prasant, in: Proceedings of the 5th International Conference on Frontiers in Intelligent Computing, vol. 1, Springer, 2018, p. 515.

[26] T. Vivek, T. Basant, S.T. Ramjeevan, G. Shailendra, Pattern and data analysis in healthcare setting, in: Advances in Medical Technologies and Clinical Practice Book, IGI, 2017.

[27] A.S. Yu, A. v. Esbroeck, F. Farooq, G. Fung, V. Anand, B. Krishnapuram, Predicting readmission risk with institution specific prediction models, in: Proceedings of the 2013 IEEE International Conference on Healthcare Informatics, ICHI '13, 2013, pp. 415–420.