PREDICTING HOSPITAL LENGTH OF STAY IN INTENSIVE CARE UNIT

by

Namita Singh

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Computer Science

at

The University of Wisconsin-Milwaukee

May 2019

ABSTRACT

PREDICTING HOSPITAL LENGTH OF STAY IN INTENSIVE CARE UNIT

by

Namita Singh

The University of Wisconsin–Milwaukee, 2019
Under the Supervision of Professor Rohit J. Kate

In this thesis, we investigate the performance of a series of classification methods for the Prediction of the hospital Length of Stay (LoS) in Intensive Care Unit (ICU). Predicting LOS for an inpatient in an hospital is a challenging task but is essential for the operational success of a hospital. Since hospitals are faced with severely limited resources including beds to hold admitted patients, prediction of LoS will assist the hospital staff for better planning and management of hospital resources. The goal of this project is to create a machine learning model that predicts the length-of stay for each patient at the time of admission.

MIMIC-III database has been used for this project due to detailed information it contains about ICU stays. MIMIC is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising de-identified health data associated with ~40,000 critical care patients at Beth Israel Deaconess Medical Centre. It includes demographics, vital signs, laboratory tests, medications, and more.

Different machine learning techniques/classifiers have been investigated in this thesis. We experimented with regression models as well as classification models with different classes of varying granularity as target for LoS prediction. It turned out that granular classes (in small unit of days) work better than regression models trying to predict exact duration in

days and hours. The overall performance of our classifiers was ranging from fair to very good and has been discussed in the results. Secondly, we also experimented with building separate LoS prediction models built for patients with different disease conditions and compared it to the joint model built for all patients.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Rohit J. Kate, who has mentored me tirelessly. I am very thankful to him for his continuous support in my research activities, for his patience, guidance and encouragement.

Besides my advisor, I would like to thank my thesis committee members Prof. Jun Zhang and Prof. Zeyun Yu for their invaluable suggestion and guidance.

I would also like to thank Computer Science department at the University of Wisconsin- Milwaukee, for providing me the best state of art facilities and excellent mentoring at all stages of my learning as a graduate student.

Finally, I must express my profound gratitude to my parents, my husband and to my children for providing me with unfailing support and continuous encouragement throughout my years of study and research. This accomplishment would not have been possible without them. Thank you.

Chapter 1

Introduction

1.1 Background and Problem Statement

Predictive modeling is an increasingly important tool in the healthcare field since the modern machine learning (ML) methods can use large amounts of data to predict individual outcomes for patients [9]. Machine learning can provide many useful results like likelihood of readmissions, mortality predictions, recommend treatments, etc. The goal of this thesis is to develop a predictive model for length of stay for in hospital admissions. *Length of Stay* (LoS) in number of days is from the initial admit date to the date that the patient is discharged from any given hospital facility [38,40]. A good prediction for LoS of a patient in the ICU can help efficient resource planning and utilization of the ICU facilities.

Intensive Care Units (ICUs) are the most expensive part of a hospital [37]. For ICUs, the LoS is an important metrics since it helps the hospitals plan future bed allocations and usage, determining and scheduling specialists for patients with multiple diagnoses, determining health insurance plans and reimbursement schedules, and planning for discharge for elderly patients and overall provide increased satisfaction to the admitted patients and lesser waiting times to future patients.US hospital stays cost the system at least $377.5 billion per year [2]. Recently Medicare legislation has proposed fixed amount of insurance payment for certain procedures. Hence hospitals would like to reduce the LoS for these procedures for an increased optimization of the ICU bed management. The development of a predictive model for LoS thus becomes very useful in such scenarios.

There can be significant variation of LOS across various facilities and across disease conditions and specialties even within the same healthcare system [38]. For this thesis we choose the MIT MIMIC-III database because it is publicly available for research and secondly because of the robust amount of information it holds. Another advantage of using a publicly available data is that the results of the study can be replicated by other researchers.

MIMIC is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising de-identified health data associated with ~40,000 critical care patients. It includes demographics, vital signs, laboratory tests, medications, and more [2,28].

Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task[41,42,43]. When the examples in the training data contains both the inputs and desired outputs, it is called Supervised machine learning. On the other hand, when the examples in training data contain only the input and output is derived based on the patterns and structures, like grouping and clustering of data points, it is called unsupervised learning [41].

In this thesis we have focused on supervised machine learning and built regression and classification models for predicting LoS. Regression is used when the output is continuous and Classification is used when the output is restricted to a limited set of values. While the prediction of LoS initially appears to be a regression task, we have studied ways to categorize the continuous output [LoS] into classes and convert the problem into a classification task and then compare the results of both regression and classification. For Regression model we have used Linear Regression algorithm whereas for the classification models we we have compared the results of three different algorithms Naïve Bayes, Logistic Regression and Multilayer Perceptron also commonly known as Neural Network.

1.2 Literature Review

In this section we will discuss the current knowledge and progress so far in the computational methods developed for length of stay prediction. In the last few years, several machine learning approaches have been utilized for length of stay prediction. While some of them research studies [5,6,7,8,9] employed clinical data from certain hospitals which is not widely available for the research community, the rest used publicly available datasets like MIT's MIMIC [10,7] which included tens of thousands of records.

There have been a wide range of features that have been explored for the LoS prediction task, based on available clinical data or expert advice. A few studies have tried to predict the LoS on specific category of patients based on diagnosis conditions like cardiac [3, 20, 36] and diabetic [22] based on datasets obtained from hospitals that are not available publicly for other researchers. Support vector machines (SVM) [13], artificial neural networks (ANN) [14], naive Bayes [15], logistic regression [16], decision trees [17] are the most popular machine learning algorithms which have been utilized for length of stay prediction. Several studies have used statistical methods, such as the regression method [2,19] and various machine learning regression methods, including MLP neural networks and regression tree, to predict the length of stay of patients in hospitals. The results of some studies have demonstrated the accuracy of the methods used to predict the length of stay. In the studies of LoS prediction, the results have not been very convincing so far. A separate study that studies 30 prediction models built for LoS prediction using regression techniques conclude as mentioned below:

"We were disappointed in the predictive performance of the regression models and conclude that it is difficult to predict LoS of unplanned ICU admissions using patient characteristics at admission time only.[19]"

Although models to predict LoS exist, they often are based on disparate variables from

small cohorts in single institutions, and there is scant evidence to suggest benefit or even application in clinical practice [20]. Modeling ICU LoS as an outcome variable is complex. LoS is prone to outliers and there is no standard definition to categorize prolonged LoS or criteria for selecting predictive variables [21]. Because of the erroneous nature of the regression models, there have been efforts towards creation of predictive model for LoS using classification techniques [21,22,23,44]. These studies have tried to classify the training data based on certain feature selection or dimensionality reduction. The classification technique in some research work [8] is binary classification (stay is long i.e. less than 5 days or short for greater than 5 days), while in some other work [36] it has been 3 classes of short (<3 days), intermediate (3-5 days) and long (>5 days) duration. The classification model works better in terms of metrics (average AUC of 0.657) and prediction accuracy, as compared to the regression models. However, we wanted to expand the range of these classes by converting the regression problem into classification and attain more accurate predictions than wide range classification. Weiss & Indurkhya [26, 27] had explored the idea of mapping regression into classification with their rule-based regression system. They used the P-class algorithm for class discretization as a part of their learning system.

This work clearly showed that it is possible to obtain excellent predictive results by transforming regression problems into classification and then use a classification learning system.

1.3 Motivations and Objectives

While using a broad range of machine learning algorithms and training strategies have been well studied in the past, for regression techniques, the results have not been very promising. The classification techniques that were used in previous research

were very wide range i.e. binary (short or long stay) or maximum 3 classes (less than 3day, 3-5 days and >5 days). Our aim in this project was to enable the use of existent classification inductive learning systems on problems of regression [25] to predict Lo S at a more granular level in terms of the number of days. We achieve this goal by transforming regression problems into classification problems. This is done by transforming the range of continuous goal variable values into a set of intervals that will be used as discrete classes [25]. We have experimented with various set of intervals and compared the results of the different classification approaches.

No studies have been done to compare the various class intervals for the classification model of LoS prediction. Since the range of LoS in the MIMIC-III dataset is extremely large ranging from 0 to 299 days, using a binary or ternary classification may not be sufficient for a practical prediction model. Hence it seems relevant to experiment with various class intervals to deduce which classification would work best under such circumstances. Instead of getting the exact continuous variable as the LoS target, it would be practically sufficient to categorize it into days by clubbing all the predictions that fall in the 24 hours bracket to one day category. Since from the ICU optimization point of view and also from the insurance company perspective a per day prediction model would be helpful rather than exact hours prediction as in regression.

Hence the objective of the proposed research is to:

- Build and compare models for different class intervals to determine the optimal class intervals for LoS predictions. We experimented with the following class intervals: (i) 31 classes of one-day class intervals until 30 days and rest >30 days (ii) 16 classes of 2-day classes until 30 days and rest >30 days (iii) 11 classes of 3-day classes until 30 days and rest >30 days and (iv) 7 classes of 5 day classes until 30 days and rest >30 days. We compared the results of these class interval schemes using three different supervised learning algorithms available in the Weka machine learning software [51]: Naïve Bayes, Logistic Regression and Multilayer Perceptron.

- To determine whether LoS prediction models built separately for different diagnosis categories improve the performance over the joint model built for all diagnosis categories. We used these diagnosis categories: Blood, Circulatory, Congenital, Digestive, Endocrine, Genitourinary, Infectious, Injury, Mental, Muscular, Misc, Neoplasm, Nervous, Pregnancy, Prenatal, Skin, and Respiratory. We compared the results of these diagnosis-specific classification models against the joint model using three different supervised learning algorithms in Weka: Naïve Bayes, Logistic Regression and Multilayer Perceptron.

Chapter 2

Materials and Methods

In this section, we explain the different computational approaches we used in creation of predictive models for length of stay (LoS). We shall begin with the dataset and machine learning algorithms which we used, then, we will delve deeper into the implemented methodology.

2.1 Dataset

For the current research study, we used MIMIC-III dataset [10, 28]. MIMIC-III is a large, publicly-available database comprising de-identified health-related data associated with approximately sixty thousand admissions of patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012[28,29,30]. The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, nurse and physician notes, imaging reports, and out-of-hospital mortality. MIMIC supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development [28,29,30]. It is notable for three factors:

- it is publicly and freely available.
- it encompasses a diverse and very large population of ICU patients.
- it contains high temporal resolution data including lab results, electronic

documentation, and bedside monitor trends and waveforms.

Access to the MIMIC-III dataset requires taking a research ethics and compliance training course and filling out a research application form. Once the user completes the required trainings and tests, they provide the user, access to the dataset. The latest version of MIMIC-III dataset v1.4 released on 4th September 2016 has been used in this research.

Because of the exhaustive nature of the dataset used, it required considerable amount of data cleaning and feature extraction. The target variable was LoS and various other dependent variable were identified and selected as the features base on past work [2]. The extracted target variable from the database was a continuous variable, so we first built up a regression model. The target variable was then categorized in days to build different classification models that were studied on and compared against each other and to the regression model.

2.2 Machine Learning Methods

In the following sections, we discuss the various supervised algorithms used in our study including Linear Regression, Logistic Regression, Naive Bayes and Multilayer Perceptron. Finally, we will explain our proposed approach to build LoS predictive models.

2.2.1 Linear Regression:

Classification involves a nominal class value, whereas regression involves a numeric class. Linear regression is a classical statistical method that computes the coefficients or "weights" of a linear expression, and the predicted ("class") value is the sum of each attribute value multiplied by its weight[49].

2.2.2 Logistic Regression

Linear regression is one the machine learning methods that is used to model continuous value functions. A popular type of generalized linear regression is called logistic regression which models the probability of the variable being predicted as a linear function of a group of predictor variables. The logistic regression is used for binary classification when the output variable of a model is specified as a categorical binary [31].

2.2.3 Naïve Bayes

Naive Bayes is a classification method based on probability theory. In order to estimate joint probability distribution of the features and output, it makes a naive assumption that all the features are conditionally independent of each other given the output. Along with this assumption it uses Bayes theorem to compute probability of the output given the features in terms of the probability of the features given the output which is easier to estimate using the training data. Naive

Bayes is computationally a very fast machine learning method [32].

2.2.4 Neural Network

Neural network, also known as multilayer perceptron, is a networks of perceptrons, usually connected in a forward feed way [34]. They use backpropagation algorithm to learn from training examples and then classify instances. Infact, they can implement arbitrary decision boundaries using “hidden layer”. Multilayer Perceptron is slower than other methods, which is a disadvantage [33].

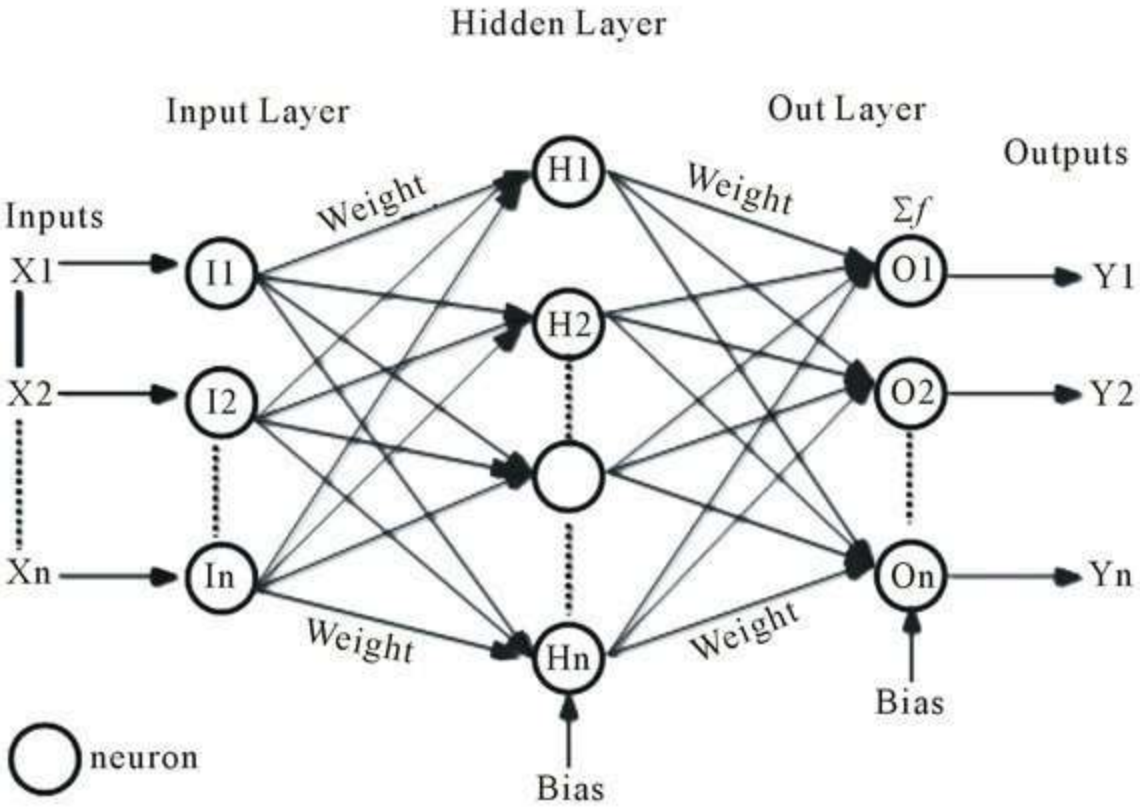

Figure 2.1: Multilayer Perceptron [34]

2.2.5 Model Evaluation

There are several evaluation measures, including accuracy, f-measure, precision, recall, sensitivity, specificity, AUC, etc. which can be used to evaluate prediction models. There are advantages and disadvantages of using them, for example, the accuracy only checks the correct classification on test data which could be misleading. Let's consider a scenario where we have 95% of data belonging to one majority class, if a classifier just classifies all the data in this class, then the final accuracy would be 95% without doing anything regarding the minority class and this misclassification will not be fairly represented in the accuracy of the model. However, the ROC Curve summarizes performance for all threshold levels whereas other measures are specific to the chosen classification threshold.

Area under the ROC curve (AUC) is then used to indicate performance with a single number. AUC is a very popular metric to evaluate the performance of classifiers. We used AUC as a measure to evaluate and compare the performance of the models. It can be helpful to indicate that the value of AUC ranges from 0.5 to 1. A random classifier has a 0.5 AUC and the perfect classifier has 1 AUC. A higher AUC shows better performance for a classifier.

2.2.5.1 Definitions

Following are the definitions of two important terms used in this work.

- **RMSE:** The RMSE is the commonly used metric to evaluate regression models. The RMSE is a commonly used measure of the differences between the values predicted by the model and the values observed, where a lower score implies better accuracy. For example, a perfect prediction model would have a RMSE of zero. The RMSE for this work is given as below, where (n) is the number of hospital admissions, (y-hat) the predicted LoS and (y) the actual LoS.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

- **Sensitivity:** Sensitivity also called the true positive rate (TPR) measures the measures the proportion of actual positives that are correctly identified as such [47]. It is given by the formula:

  TPR = TP / (TP + FN), where

  A TP (True Positive) test result is one that detects the condition when the condition is present.

  A FN (False Negative) test result is one that does not detect the condition when the condition is present.

  Specificity: Specificity also called the true negative rate (TNR) measures the proportion of actual negatives that are correctly identified as such. It is given by the formula:

  TPR = TN/ (FP + TN), where

  A TN (True Negative) test result is one that does not detect the condition when the condition is absent.

  A FP (False Positive) test result is one that detects the condition when the condition is absent.

- **False Positive Rate:** The false positive rate is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events [48].

- **Probability or Confidence:** All machine learning algorithms we used, give us a probability or confidence for each given instance which indicate how much they are confident about putting that instance in one specific class. These confidences are used to plot ROC Curve.

- **ROC Curve:** Using different thresholds on this confidence, different sensitivity and specificity measures can be obtained which are then plotted on a graph called ROC curve. By varying the decision threshold from 0 to 1, one can obtain an entire range of

true positive rates and false positive rates which when plotted on a graph is called a ROC curve. One of the noticeable properties of ROC curve is that it is independent of the class distribution. It means that, if the distribution of positive and negative instances changes in the dataset, its value does not change [44].

- **AUC**: AUC is an abbreviation for area under ROC curve. It is one number that summarizes ROC curve and is used to numerically evaluate classification models to determine which of the models predicts the classes best. The baseline value for AUC is 0.5. The closer the AUC of the model comes to 1, the better it is. So, models with higher AUC are preferred over models with lower AUCs.
- **10-Fold cross validation:** In this evaluation methodology the available data is randomly divided into k equal size folds, and each time the model is trained with k-1 folds and tested remaining fold, and this process is repeated for k times each time using a different fold for testing. The final performance is reported by taking average of the k metrics obtained from the k folds. k=10 is the standard and the most common value used for k.

2.4 Methodology

In this section, we will explain our methodology to build LoS predictive models.

2.4.1 Data Preparation

We have based data exploration and feature engineering on previous study [9] and code has been leveraged from the GitHub repository made available [50]. All pre-processing, data analysis, and machine learning were performed in accordance with MIMIC-III guidelines and regulations. The data preparation had two different stages as follows:

Data Exploration:

The first step in the data preparation was to pick up a subset of the MIMIC-III dataset for the proposed study. MIMIC-III dataset has 27 tables in csv format which entails details about, age, demographics, clinical studies and more. After a lot of study and analysis, we picked up the following tables for preparing our dataset by loading them into DataFrames using Pandas:

1. ADMISSIONS.csv: The ADMISSIONS table gives details about SUBJECT_ID (unique patient identifier), HADM_ID (hospital admission ID), ADMITTIME (admission date/time), DISCHTIME (discharge time), DEATHTIME, and more. The table had 58,976 admission events and 46,520 unique patients which seemed like a reasonable amount of data to do a prediction model study. We dropped rows pertaining to negative LoS since it means that patient died before prior to ICU admission. Also, the cases in which the patients died during the ICU stayed were dropped as such cases were not included in typical LoS prediction model by previous studies as well for creation of typical LoS model.
2. PATIENTS.csv: The PATIENTS table provided a de-identified date of birth and gender information.
3. DIAGNOSES_ICD.csv: The DIAGNOSES_ICD table consists of the patient and admissions IDs and an ICD9-Code. The ICD-9 Code is described as below:

"The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) is the U.S. health system's adaptation of international ICD-9 standard list of six-character alphanumeric codes to describe diagnoses. Standardizing codes improves consistency among physicians in recording patient symptoms and diagnoses for the purposes of payer claims reimbursement and clinical research.[35]"

4. ICUSTAY.csv: The ICUSTAYS table gives details about the HADM_ID (hospital admission ID), FIRST_CAREUNIT (details of the care unit patients like ICU, NICU, MICU, etc.) , INTIME(in time to the care unit), OUTTIME(out time from the care unit) and LoS(length of stay in the care unit) and more.

Feature Engineering

The **second step** here was to drop rows with negative LoS, usually related to a time of death before admission. The distribution of the length of stay looked like the figure below:

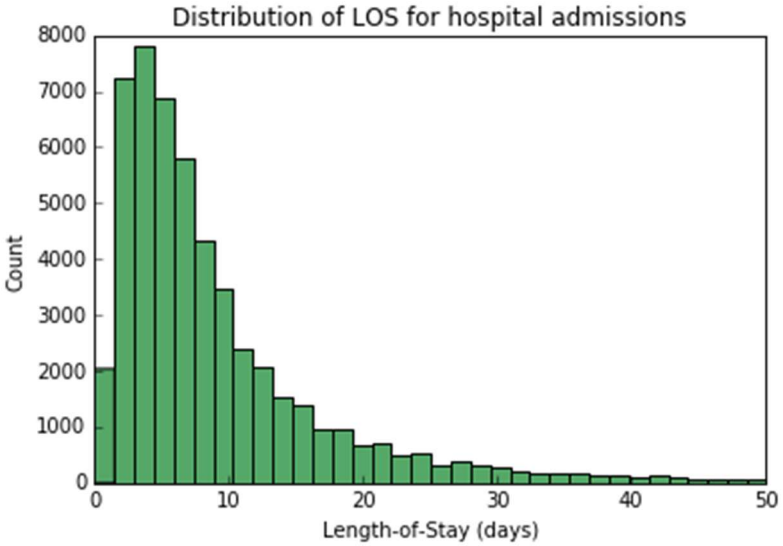

Figure 2.2: Distribution of LoS for hospital admissions.

The main challenge was to create the diagnosis categories from the DIAGNOSES_ICD table. There were 6,985 unique codes used in the MIMIC dataset and 631,048 ICD-9 diagnoses given to patients since most were diagnosed with more than one condition.

ICD-9 codes are standard list of six-character alphanumeric codes to describe diagnoses. For instance, the ICD_9 code of 403.01 falls in the range of "diseases of the circulatory system" and the .01 value further specifies "*hypertensive chronic kidney and related diseases"*. On investigation we found out that ICD-9 has 17 primary categories so it was decided to sort all the unique codes per admission into these categories. Reducing the ICD_9 codes from 6, 985 to 17 would make a better machine learning model for this study.

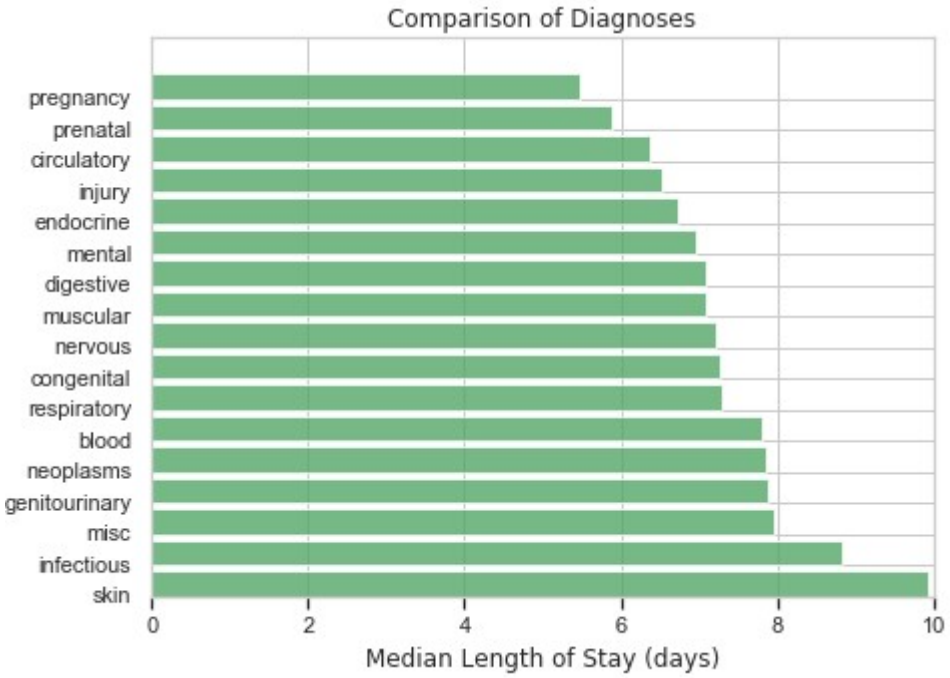

Figure 2.3: Comparison of Diagnoses

Lastly, we categorized the age into classes of newborn, young adult, middle adult and senior in order to obtain a better prediction model. The ethnicity counts in the ADMISSIONS table was more than 35+, it was compressed to 5 groups by combining into the higher-level main group. For example, Hispanic/Latino-Cuban, Hispanic/Latino-Salvadoran and Hispanic/Latino – Columbian were put in the one group as Hispanic/Latino.

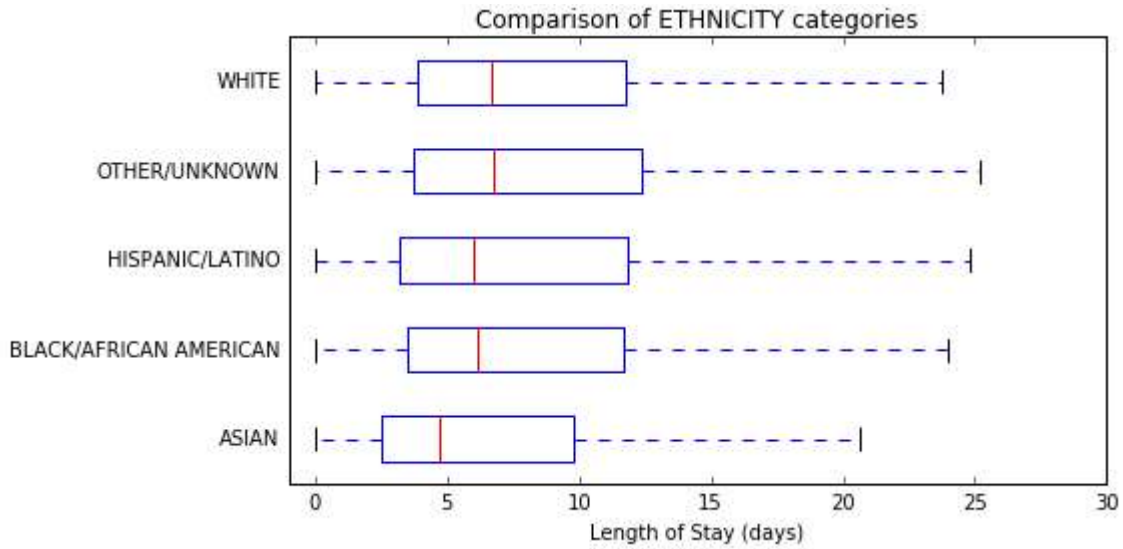

Figure 2.4: Distribution of LoS by Ethnicity

One interesting observation is the fact that Asians have the lowest median stay.

2.4.2 Features

For building the LoS predictive models, it is crucial to use informative features. The MIMIC-III dataset had the robust information for each admission in the form of 27 tables. From those we narrowed down to 4 tables, but still the information contained in those tables had to be further narrowed down. Based on previous studies [9],[50] and what is suggested by the MIMIC data team we finalized on 48 features and one target column. The total number of hospitals stays in the dataset was 53,104. Table 2.1 lists the feature names, their data type, the number of feature values and their description. A patient could have multiple diagnoses under the same ICD-9 categories, hence there could be multiple numeric values for a diagnosis category. For example, a person with cardiac condition could have 3 different clinical reports under ICD-9 category of Blood, hence the feature value for Blood will be 3 in this case. It is also important to note that a patient could have multiple diagnosis under separate categories. For example, a patient with cardiac arrest could possibly have diagnosis under both Blood and Circulatory.

Table 2.1: The MIMIC-III dataset features used to make LoS predictive model.

Feature name	Data type	Number of feature values	Description
Blood	Numeric	6	ICD_9 category
Circulatory	Numeric	16	ICD_9 category
Congenital	Numeric	10	ICD_9 category
Digestive	Numeric	12	ICD_9 category
Endocrine	Numeric	11	ICD_9 category
Genitourinary	Numeric	8	ICD_9 category
Infectious	Numeric	8	ICD_9 category
Injury	Numeric	22	ICD_9 category
Mental	Numeric	12	ICD_9 category
Misc	Numeric	9	ICD_9 category
Muscular	Numeric	8	ICD_9 category
Neoplasm	Numeric	11	ICD_9 category
Nervous	Numeric	9	ICD_9 category
Pregnancy	Numeric	13	ICD_9 category
Prenatal	Numeric	16	ICD_9 category
Respiratory	Numeric	9	ICD_9 category
Skin	Numeric	8	ICD_9 category
GENDER	Nominal	Binary	Male or Female
ICU	Nominal	Binary	ICU admission
NICU	Nominal	Binary	NICU admission
ADM_ELECTIVE	Nominal	Binary	Elective admission
ADM_EMERGENCY	Nominal	Binary	Emergency admission
ADM_NEWBORN	Nominal	Binary	Newborn admission
ADM_URGENT	Nominal	Binary	Urgent admission
INS_Government	Nominal	Binary	Government insurance
INS_Medicaid	Nominal	Binary	Medicaid insurance
INS_Medicare	Nominal	Binary	Medicare admission
INS_Private	Nominal	Binary	Private insurance
INS_Self Pay	Nominal	Binary	Self-payment type
REL_NOT SPECIFIED	Nominal	Binary	Religion not specified
REL_RELIGION	Nominal	Binary	Religious or not
REL_UNOBTAINABLE	Nominal	Binary	Religion unobtainable
ETH_ASIAN	Nominal	Binary	Asian ethnicity
ETH_BLACK/AFRICAN AMERICAN	Nominal	Binary	Black/African American ethnicity
ETH_HISPANIC/LATINO	Nominal	Binary	Hispanic/Latino ethnicity
ETH_OTHERS	Nominal	Binary	Ethnicity as others.
ETH_WHITE	Nominal	Binary	White ethnicity
AGE_middle_adult	Nominal	Binary	Age category as middle_adult
AGE_newborn	Nominal	Binary	Age category as newborn
AGE_senior	Nominal	Binary	Age category as senior
AGE_young_adult	Nominal	Binary	Age category as young_adult
MAR_DIVORCED	Nominal	Binary	Marital status as divorced
MAR_LIFE PARTNER	Nominal	Binary	Marital status as life partner
MAR_MARRIED	Nominal	Binary	Marital status as married
MAR_SEPARATED	Nominal	Binary	Marital status as separated
MAR_SINGLE	Nominal	Binary	Marital status as single
MAR_UNKNOWN	Nominal	Binary	Marital status as unknown
MAR_WIDOWED	Nominal	Binary	Marital status as widowed
LOS	Numeric	0-299	Regression model
	Nominal	31	1-day classification

Nominal	16	2-day classification
Nominal	11	3-day classification
Nominal	7	5-day classification

2.4.3 Predictive Models

In every experiment in this study, we utilized Weka's 10-fold cross validation. For each model, Weka randomly shuffles the order of available instances and divides the data in 10 equal folds. We used this strategy to be able to perform a fair and meaningful comparison between models. While the joint model consisted of 53104 records, the number of records for each of the diagnosis varied from 169(Pregnancy) to 41851(Injury).

2.4.4 Test Bed and Experimental Setup

For all experimental results presented in this section, we used 64-bit Windows 10 operating system on a PC with 2.40 GHz Intel Dual core CPU, 4MB cache and 8GB of RAM. Data extraction and feature engineering was done using Pandas and scikit-learn libraries for Python based on previous study [9] and code provided at GitHub repository [50]. Data modeling was done using Weka data mining library (version 3.6.13) [23] which has been freely available to the research community.

Regression Setup: The dataset extracted after our feature engineering consisted of 53104 records. It is said joint because it contains the records of ICU stays corresponding to all the diagnosis categories. Many experiments were performed on this joint dataset. The final dataset consisted of 48 features and one target column for LoS. The range of length of stay varied from 0 to 299.

Classification Setup (Joint Model): In order to perform regression by classification on the extracted dataset, we club the numeric values within 24 days period to a one-day class. Table2.2 shows the conditions used to classify the numeric values into nominal classes.

Table 2.2: Classification strategy for one-day classes

Classification condition for LoS	**Class**
LoS>0 and LoS <=1	D1
LoS>1 and LoS <=2	D2
…	…
LoS>29 and LoS <=30	D30
LoS>30	D99

Thus, the dataset used for creating the classification model consisted of 47 features, same as that of joint model and the target variable as LoS which in this case is nominal and consists of 31 classes corresponding to the conditions mentioned in Table2.3.

The second classification strategy was to create 2-day uniform classes until 30 days and one for all case where LoS was >30 days. Table 2.3 shows the classification rules for the two-day classifications.

Table 2.3: Classification strategy for two-day classes

Classification condition for LoS	**Class**
LoS>0 and LoS <=2	D2
LoS>2 and LoS <=4	D4
LoS>4 and LoS <=6	D6
LoS>6 and LoS <=8	D8
LoS>8 and LoS <=10	D10
LoS>10 and LoS <=12	D12
LoS>12 and LoS <=14	D14
LoS>14 and LoS <=16	D16
LoS>16 and LoS <=18	D18
LoS>18 and LoS <=20	D20

LoS>20 and LoS <=22	D22
LoS>22 and LoS <=24	D24
LoS>24 and LoS <=26	D26
LoS>26 and LoS <=28	D28
LoS>28 and LoS <=30	D30
LoS>30	D99

The third classification strategy was to create 3-day uniform classes until 30 days and one for all case where LoS was >30 days. Table 2.4 shows the classification rules for the two-day classifications.

Table 2.4: Classification strategy for three-day classes

Classification condition for LoS	**Class**
LoS>0 and LoS <=3	D3
LoS>3 and LoS <=6	D6
LoS>6 and LoS <=9	D9
LoS>9 and LoS <=12	D12
LoS>12 and LoS <=15	D15
LoS>15 and LoS <=18	D18
LoS>18 and LoS <=21	D21
LoS>21 and LoS <=24	D24
LoS>24 and LoS <=27	D27
LoS>27 and LoS <=30	D30
LoS>30	D99

The fourth classification strategy was to create 5-day uniform classes until 30 days and one for all case where LoS was >30 days. Table 2.5 shows the classification rules for the two-day classifications.

Table 2.5: Classification strategy for five-day classes

Classification condition for LoS	**Class**
LoS>0 and LoS <=5	D5
LoS>5 and LoS <=10	D10
LoS>10 and LoS <=15	D15
LoS>15 and LoS <=20	D20
LoS>20 and LoS <=25	D25
LoS>25 and LoS <=30	D30
LoS >30	D99

Lastly, we also experimented with the 3-class strategy as done in previous work [36] by Daghistani by grouping patients in three groups based on their LOS: short (<3 days), intermediate (3-5 days) and long (>5 days). The previous work was however done for cardiac adult patients using data from King Abdulaziz Cardiac Center (KACC). Table 2.6 shows the rules used in our classification approach.

Table 2.6: Classification strategy for three classes

Short	Intermediate	Long
<3 days	3-5 days	>5 days

Classification Setup (Diagnosis-specific Model): The diagnosis specific dataset corresponding to the 17 diagnosis categories were derived from the Joint model dataset mentioned above. For each of the diagnosis categories the null value records (means that diagnosis is not present) for the specific diagnosis were deleted so that the final dataset would correspond of only relevant records of that diagnosis. Based on this strategy, 17 different datasets corresponding to the 17 diagnosis categories were created for creating separate classification models for each diagnosis. The same classification approach as defined in tables 2.2, 2,3, 2,4 and 2.5 were applied for building the predictive models for diagnosis-specific categories.

Chapter 3

Results and Discussion

To analyze our proposed predictive length of stay [LoS] models, several experiments were performed. The results of the joint regression model are shown in Section 3.1. In Section 3.2, we analyzed our experiments with classification using one-day, two-day, three-day, five-day classes and also when only 3 classes are used. Finally, in Section 3.3 we further compared the performance of three machine learning methods for the one-day classification for the diagnosis specific models against the joint model.

3.1 Joint Regression predictive models

RMSE was used as metrics for the joint regression model. Lots of regression experiments were carried out with feature reduction and dimensionality reduction of target variable. The best results we got was an RMSE of 2.58 days using the Linear Regression model which was only marginally better than the RMSE of the ZeroR algorithm. The ZeroR is the Zero Rule algorithm used as a baseline for comparison. For a regression predictive modeling problem that predicts a numeric target value, the ZeroR simply predicts the mean of the training dataset.

Table 3.1: Comparison of Linear Regression model against ZeroR

ALGORITHM	RMSE
ZEROR	9.6446
Linear Regression	8.5154

Since the Linear Regression gave results that were only marginally better than the mean of the training dataset with an undesirably large error for predicting LoS, we decided to use the classification methodology for a better predictive modeling.

3.2 Classification Approach

To convert the regression problem to classification we experimented with various classification strategies. We experimented with creating one-day classes, two-day classes, three-day classes and five-day classes. We then compared the performance of these different classifications using three different machine learning models, Naïve Bayes, Logistic Regression and Multilayer Perceptron. Table 3.2 shows the results of these experiments.

Table 3.2.1: Comparison of the one-day, two-day, three-day and five-day classifications

Diagnosis	Instances	One-Day Classification			Two-Day Classification			Three-Day Classification			Five- Day Classification		
		Naïve Bayes	Logistic Reg.	MLP	Naïve Bayes	Logistic Reg.	MLP	Naïve Bayes	Logistic Reg	MLP	Naïve Bayes	Logistic Reg	MLP
Joint	53104	0.650	0.690	0.661	0.648	0.686	0.672	0.648	0.691	0.676	0.646	0.680	0.656

We also used the 3 classes strategy as used in previous work [36] to classify LoS as short (<3 days), intermediate (3-5 days) and long (>5 days) as defined in table 2.6. The results are shown in Table 3.3

Table 3.2.2: Comparison of classifiers for 3 class LoS classification as short, intermediate and long stay

Diagnosis	Instances	Naïve Bayes	Logistic Regression	Multilayer Perceptron
		AUC	AUC	AUC
Joint	53104	0.693	0.745	0.789

- From table 3.2, we don't see much difference in the performance of the various models, but because one-day model is more granular, it is more accurate and hence preferable to others.
- The logistic regression gives the best AUC metrics compared to the Naïve Bayes and Multilayer Perceptron models for the day-based classifications. However, for the 3-class classification in table 3.3, the Multilayer perceptron is better than the other two.
- For computationally slow algorithms like the logistic regression and multilayer perceptron it took even longer for the one-day classifications due to the high numbers of nominal classes involved. For two-day classifications the classes

were reduced by twice, for three-day classification the classes reduced by thrice and for five days the classes reduced 5 times, hence the computational speed of these algorithms improved as the target classes decreased.

- We decided to further refine our one-day classification model by creating diagnosis-specific models and comparing it against the joint model for one-day classification.

3.3 Comparative study of machine learning methods for joint vs diagnosis specific models

We created separate models for each of the 17 diagnosis categories and compared it against the joint model. We used three different machine learning algorithms Naïve Bayes, Logistic Regression and Multilayer Perceptron for this purpose. Naïve Bayes and Logistic Regression do not have any major parameters to tune in Weka, so we used the default parameters for these algorithms. Weka uses one hidden layer as the default for the Multilayer Perceptron algorithm and it was used as such.

Table 3.3: Comparison of the diagnosis specific model against the joint model

Diagnosis	Records	**One Day Classification**		
		Naïve Bayes	Logistic Regression	Multilayer Perceptron
Joint	53104	0.650	0.690	0.661
Blood	15692	**0.656**	**0.696**	**0.663**
Circulation	37537	0.620	0.656	0.592
Congenital	3109	**0.703**	**0.745**	**0.662**
Digestive	18407	0.624	0.645	0.593
Endocrine	31862	0.616	0.637	0.630
Genitourinary	18381	0.603	0.622	0.574
Infectious	11918	**0.676**	**0.702**	**0.689**
Injury	41851	**0.719**	**0.727**	**0.710**
Mental	14686	0.622	0.647	0.639
Misc	14329	0.611	0.630	0.635
Muscular	8805	0.611	0.615	0621
Neoplasm	7481	0.591	0.620	0.611
Nervous	13788	0.615	0.680	0.673
Pregnancy	169	0.600	0.620	0.593
Prenatal	10241	**0.660**	**0.722**	**0.709**
Respiratory	21126	0.605	0.634	0.620
Skin	5694	0.594	0.612	0.602

- For all the diagnosis categories, Logistic regression gives better AUC than Naïve Bayes and Multilayer perceptron models.
- Although the joint model in itself provides good results for LoS prediction. It can be observed from the analysis that separate model for certain disease categories provides more improved prediction for the LoS compared to Joint model. For

instances: Blood, Congenital, Infectious, Injury and Prenatal. Figure3.3.1 shows the LoS distribution for Joint model.

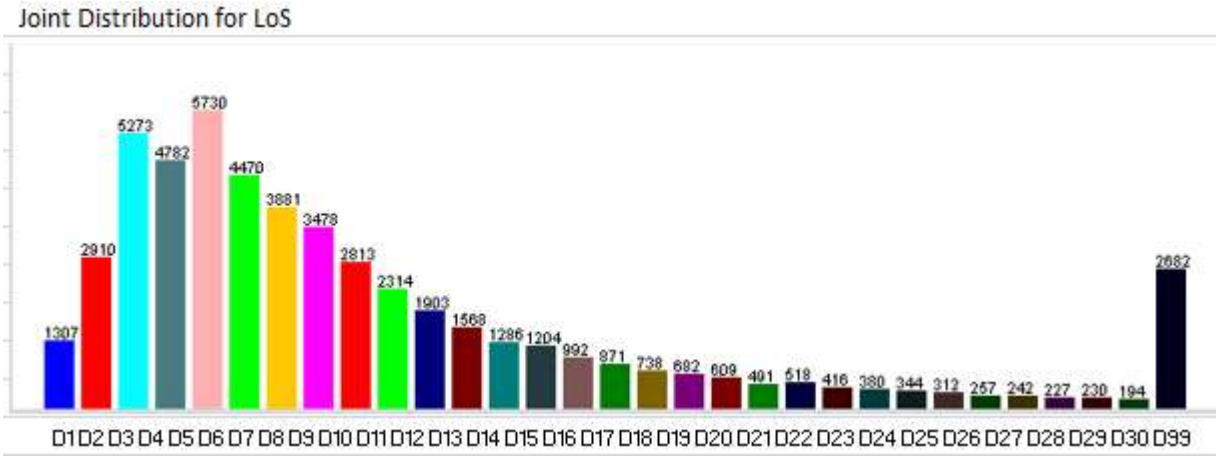

Figure 3.3.1: LoS Class Distribution for All Examples

- Not all the diagnosis specific model performs better than the joint mode. This could be attributes to the high level of information overlap amongst the classes, which we observed during data exploration phase. For example, a person with digestive diagnosis could fall under any or more of other classes due to multiple diagnoses. In other words, these diagnosis categories are not distinct enough with each other and hence their examples have similar distributions as seen in Figure 3.3.2 and Figure 3.3.3

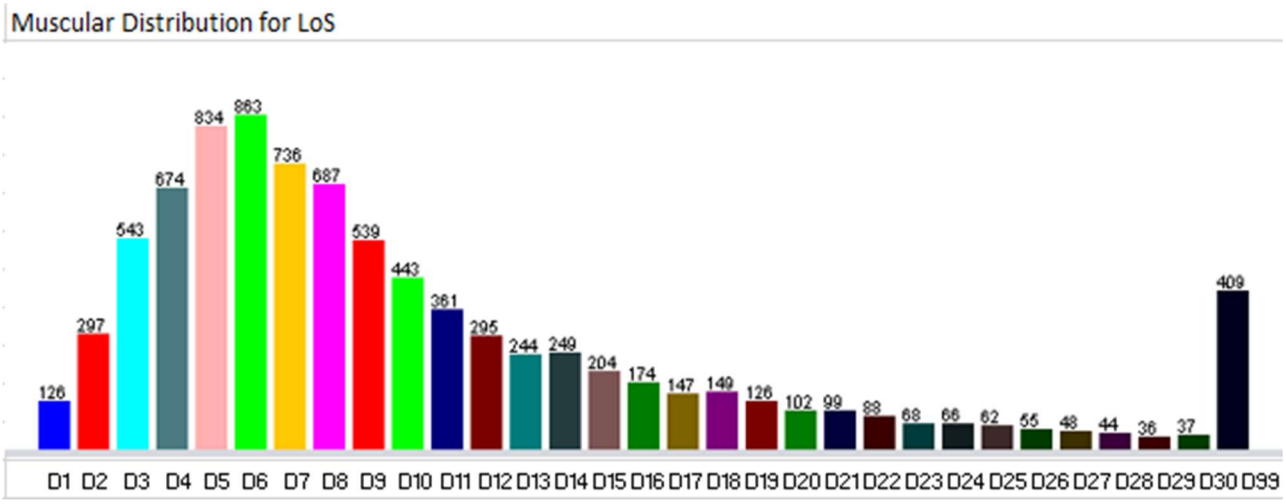

Figure 3.3.2: LoS Class Distribution for Muscular Diagnosis examples

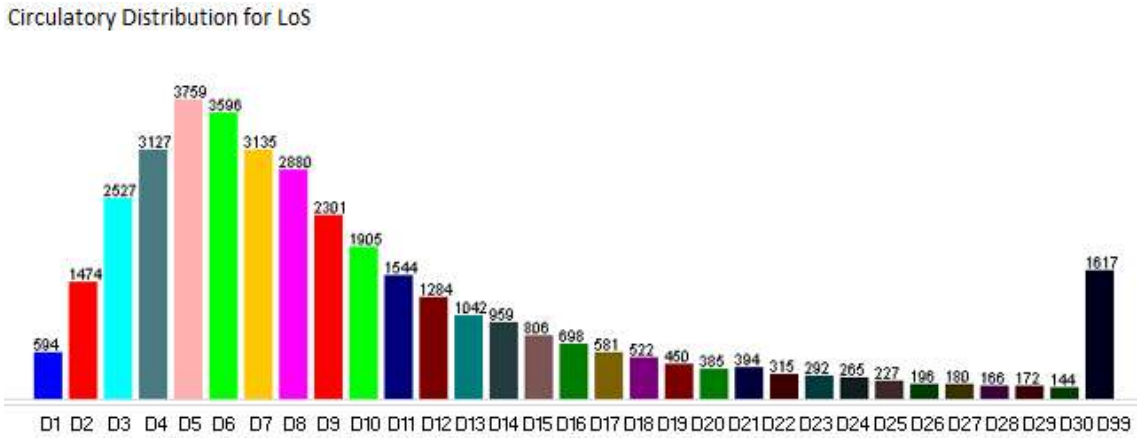

Figure 3.3.3: LoS Class Distribution for Circulatory Diagnosis examples

- The same reason applies to why some independent diagnosis categories such as Blood, Prenatal, Congenital, Infectious and Injury are showing results better than the joint model. This is due to the different class distributions with less overlap in other categories that make these categories stand out. It should be noted that these models are doing better than joint model even though the joint model gets several more training examples, this clearly shows that the distribution of examples is different in different diagnosis categories. This can be seen in Figure3.3.4, Figure 3.3.5, Figure 3.3.6 and Figure 3.3.7

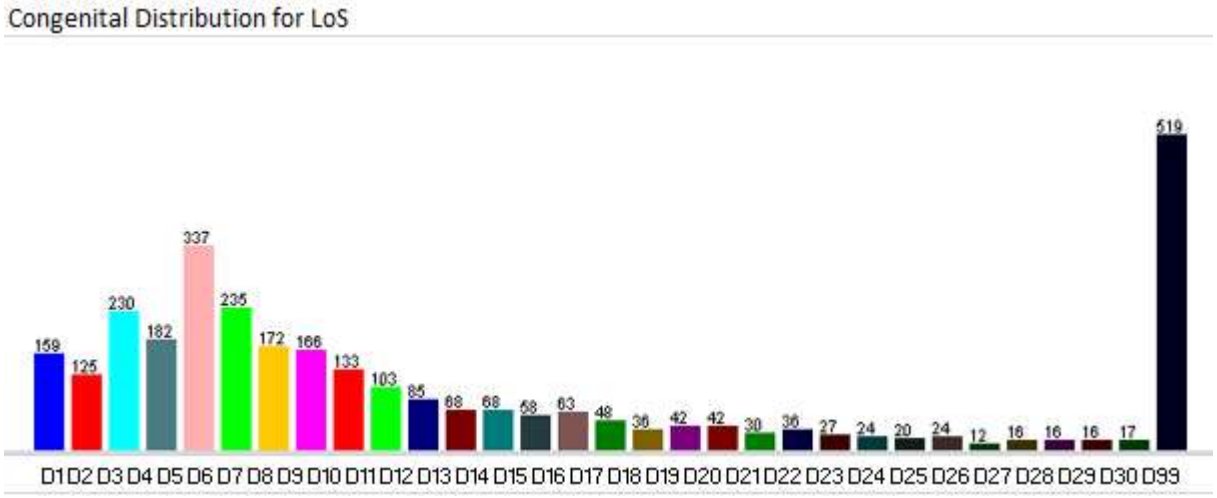

Figure 3.3.4: LoS Class Distribution for Congenital Diagnosis examples

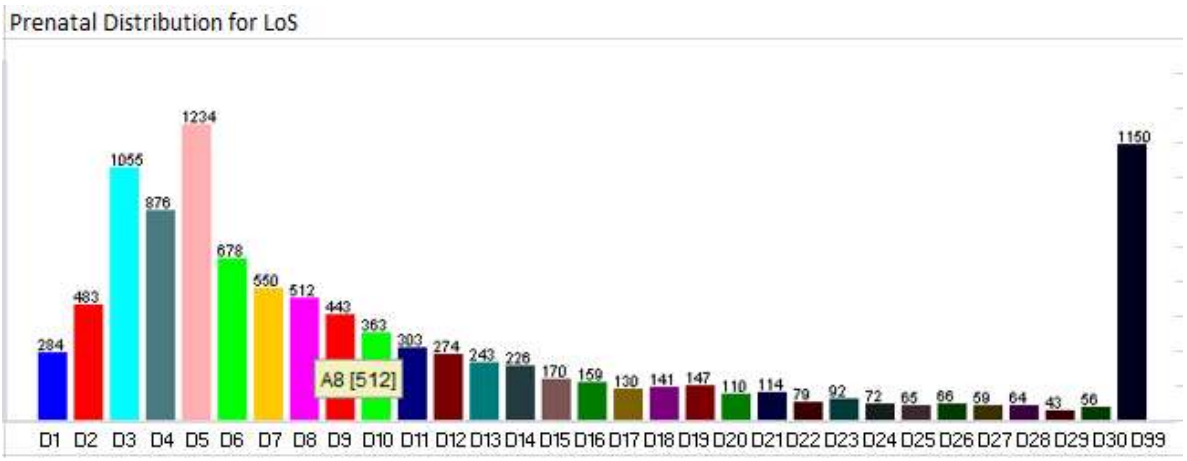

Figure 3.3.5: LoS Class Distribution for Prenatal Diagnosis examples

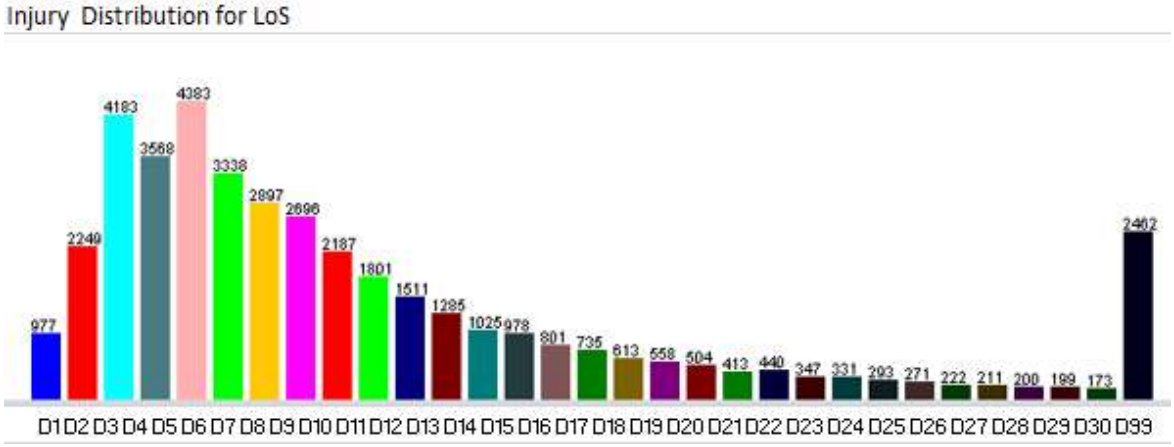

Figure 3.3.6: LoS Class Distribution for Injury Diagnosis examples

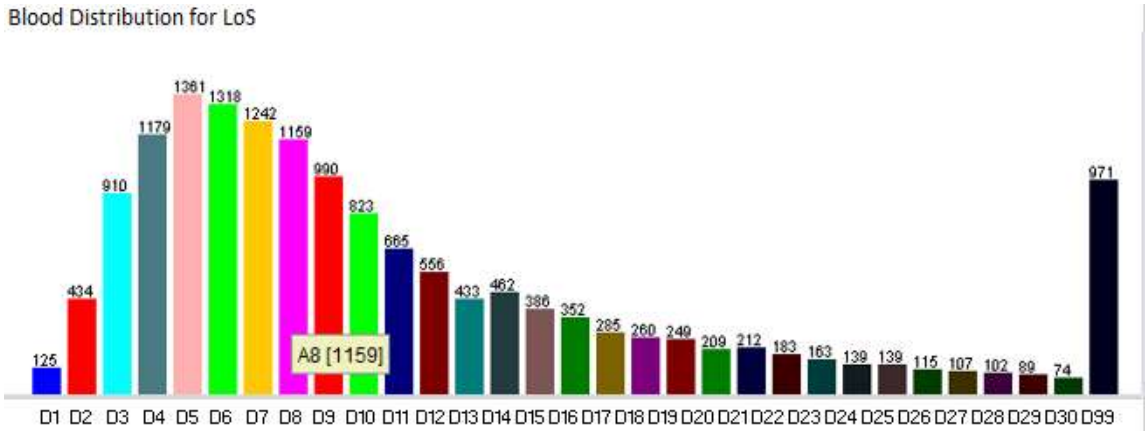

Figure 3.3.7: LoS Class Distribution for Blood Diagnosis examples

- For some of the disease categories the training examples are significantly lower compared to the Joint model. That may affect the results of the prediction model in certain cases, for example Pregnancy, Neoplasm, Skin etc. This is because the joint model learns better with more training examples but the learning curve is not

plateaued in the diagnosis specific models due to insufficient number of examples. This can be seen in Figure3.3.8

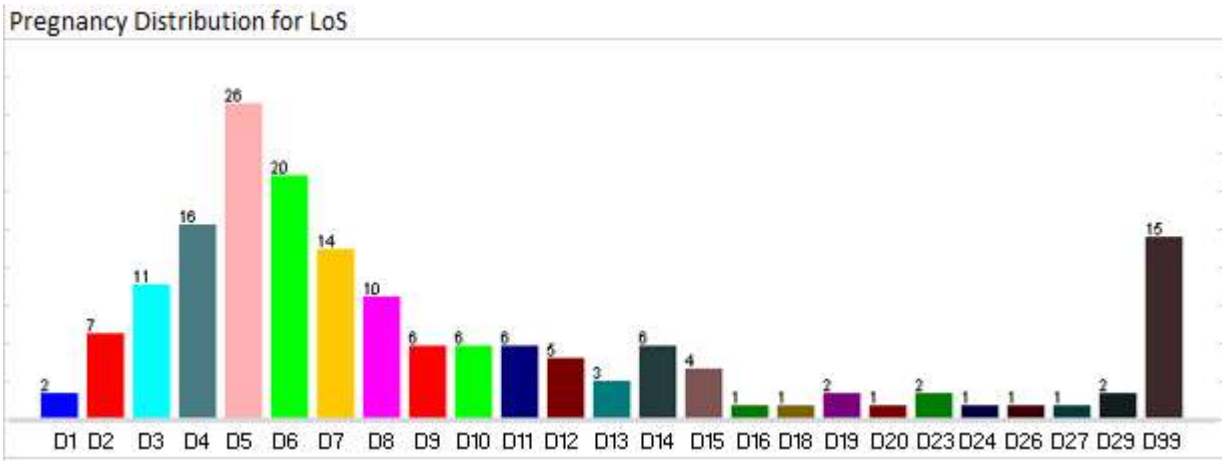

Figure 3.3.8: LoS Distribution for Pregnancy Diagnosis Model

Overall, these result show that for distinct diagnosis categories with sufficient number of training examples, it is better to build diagnosis specific models. For the rest, joint model is the better option.

Chapter 4

Conclusion and Outlook

We concluded that it was practically sufficient to predict as one-day class distribution instead of numeric prediction in hours from the ICU bed management and insurance company's perspective. And hence a classification approach to LoS prediction is more suitable than regression.

In this thesis we have done an empirical comparison of supervised learning algorithms for regression and classification used for building length of stay (LoS) predictive models. All the classification algorithms are able to predict LoS with various degrees of accuracy. Logistic Regression gave the best performance compared to other classification techniques used. However Multilayer Perceptron gives better results than Logistic Regression for the three-class classification. LoS prediction models built specifically for certain diagnosis categories (congenital, prenatal an injury) show a higher accuracy compared to the joint model.

An important aspect of this thesis is the use of variety of performance criteria to evaluate the learning methods. In this study we tried to first understand the challenges faced in the regression problem as also reported in previous studies [9] and then find ways to overcome it through classification methodology which had not been experimented with varying granularity in previous studies. We experimented with various class distributions of one-day, two-day, three-day and five-day classes and concluded that the one-day class is the best option given its fine granularity and yet comparable results with other strategies.

We further tried to refine the model by building 17 separate diagnosis specific models and comparing it against the joint model. Few of the diagnosis categories like Congenital, Infectious, Injury and Prenatal did perform better than the joint model,

whereas the results of other models were comparable to the joint model. This is because of the independent or non-overlapping nature of these diagnosis categories that some diagnosis-specific model performs better than joint. These results also give us clues that if these categories are further drilled down or diluted by considering there full 6-digit ICD codes, it could give us further improvement in results.

Bibliography

[1] Systematic Review of Data Mining Applications in Patient-Centered Mobile-Based Information Systems. Fallah M, Niakan Kalhori SR
Healthc Inform Res. 2017 Oct; 23(4):262-270.
https://www.ncbi.nlm.nih.gov/pubmed/29181235

[2] Predicting hospital length-of-stay at time of admission, Daniel Cummings, Dec 15 2018. http://www.medicalnewstoday.com/articles/282929.php.

[3]Use of data mining techniques to determine and predict length of stay of cardiac patients. Hachesu PR1, Ahmadi M, Alizadeh S, Sadoughi F.June1, 2013.
https://www.ncbi.nlm.nih.gov/pubmed/23882417

[4] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng. 2007;160:3–24.
http://www.cs.bham.ac.uk/~pxt/IDA/class_rev.pdf

[5] Maharlou H, Niakan Kalhori SR, Shahbazi S, Ravangard R. Predicting Length of Stay in Intensive Care Units after Cardiac Surgery: Comparison of Artificial Neural Networks and Adaptive Neuro-fuzzy System. Healthc Inform Res. 2018;24(2):109–117. doi:10.4258/hir.2018.24.2.109
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5944185/

[6] Stecker MM, Stecker M, Falotico J. Predictive model of length of stay and discharge destination in neuroscience admissions. Surg Neurol Int. 2017;8:17. Published 2017 Feb 6. doi:10.4103/2152-7806.199558
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5309446/

[7] Understanding & Predicting Length of Stay (LOS) using Machine Learning
https://dexur.com/a/ml-research-los/6/

[8] T. Gentimis, A. J. Alnaser, A. Durante, K. Cook and R. Steele, "Predicting Hospital Length of Stay Using Neural Networks on MIMIC III Data," *2017 IEEE 15th Intl*

Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Orlando, FL, 2017,pp.1194-1201.doi:10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191 https://ieeexplore.ieee.org/document/8328535

[9] Predicting hospital length-of-stay at time of admission, Daniel Cummings, Dec 15, 2018. https://towardsdatascience.com/predicting-hospital-length-of-stay-at-time-of-admission-55dfdfe69598

[10] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635. https://mimic.physionet.org/

[11] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

[12] I. H. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[13] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[14] S. Haykin. Neural networks, a comprehensive foundation. 1994.

[15] D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[16] D. W. Hosmer Jr and S. Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

[17] J. R. Quinlan.*C4. 5: programs for machine learning*. Elsevier, 2014.

[18] Verburg, Ilona & Atashi, Alireza & Eslami, Saeid & Holman, Rebecca & Abu-Hanna, Ameen & de Jonge, Everest & Peek, Niels & de Keizer, Nicolette. (2016). Which Models Can I Use to Predict Adult ICU Length of Stay? A Systematic Review. Critical Care Medicine. 45. 1. 10.1097/CCM.0000000000002054.

[19] Verburg, Ilona W M et al. "Comparison of regression methods for modeling intensive care length of stay." PloS one vol. 9,10 e109684. 31 Oct. 2014, doi:10.1371/journal.pone.0109684. https://www.ncbi.nlm.nih.gov/pubmed/25360612/

[20] Almashrafi, A., Elmontsri, M., and Aylin, P. Systematic review of factors influencing length of stay in ICU after adult cardiac surgery. BMC Health Serv Res. 2016; 16: 318

[21] Predicting Intensive Care Unit Length of Stay After Cardiac Surgery Litton, Edward et al. Journal of Cardiothoracic and Vascular Anesthesia , Volume 32 , Issue 6 , 2683 - 2684 . https://www.jcvaonline.com/article/S1053-0770(18)30237-4/fulltext

[22] A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients. Demokritos Athens, Eman Marzban, Georgios Giannoulis, Ayush Patel, Rajender Aparasu and Ioannis A. Kakadiaris. https://www.iit.demokritos.gr/sites/default/files/paper_6.pdf

[23] Spratt, Heidi et al. "A structured approach to predictive modeling of a two-class problem using multidimensional data sets." Methods (San Diego, Calif.) vol. 61,1 (2013):73-85.doi:10.1016/j.ymeth.2013.01.002. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3661737/

[24] Length of Stay Prediction and Analysis through a Growing Neural Gas Model. Luigi Lella, Antonio di Giorgio, Aldo Franco Dragoni. http://ceur-ws.org/Vol-1389/paper2.pdf

[25] Regression by Classification by Luis Torgo and Joao Gama.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.5374&rep=rep

1&type=pdf

[26] Weiss, S. and Indurkhya, N. (1993): Rule-base Regression. In Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1072-1078.

[27]Weiss, S. and Indurkhya, N. (1995): Rule-based Machine Learning Methods for Functional Prediction. In Journal Of Artificial Intelligence Research (JAIR), volume 3, pp.383-403.

[28] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016).DOI:10.1038/sdata.2016.35.Available at: http://www.nature.com/articles/sdata201635

[29]MIMIC Physionet: Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov P, Mark RG, Mietus JE, Moody GB, Peng C, and Stanley HE. Circulation. 101(23), pe215–e220. 2000.

[30] MIMIC Code Repository. Johnson, Alistair EW, David J. Stone, Leo A. Celi, and Tom J. Pollard. "The MIMIC Code Repository: enabling reproducibility in critical care research." Journal of the American Medical Informatics Association (2017): ocx084

[31]M. Kantardzic. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[32]J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[33]MORE DATA MINING WITH WEKA, University of Waikato. https://www.futurelearn.com/courses/more-data-mining-with-weka/0/steps/29142

[34] A Gentle Introduction To Neural Networks Series—Part 1. David Fumo, Aug 4,

2017.https://towardsdatascience.com/a-gentle-introduction-to-neural-networks-series-part-1-2b90b87795bc

[35] ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) by Margaret Rouse, December 2014.

[36] https://searchhealthit.techtarget.com/definition/ICD-9-CM

[37] Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH.doi:10.1016/j.ijcard.2019.01.046https://www.ncbi.nlm.nih.gov/pubmed/30685103

[38] Burchardi H, Moerer O. Twenty-four hour presence of physicians in the ICU. Crit Care.2001;5(3):131–137.doi:10.1186/cc1012. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC137273/

[39] Predicting Length of Stay in Hospitals By Sheri Gilley for Microsoft • January 20, 2017 https://gallery.azure.ai/Solution/Predicting-Length-of-Stay-in-Hospitals-1

[40] Predicting Hospital Length of Stay, Implemented with Microsoft Machine Learning Services. https://microsoft.github.io/r-server-hospital-length-of-stay/

[41] Machine Learning from Wkipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Machine_learning

[42] The definition "without being explicitly programmed" is often attributed to Arthur Samuel, who coined the term "machine learning" in 1959, but the phrase is not found verbatim in this publication, and may be a paraphrase that appeared later. Confer "Paraphrasing Arthur Samuel (1959), the question is: How can computers learn to solve problems without being explicitly programmed?" in Koza, John R.; Bennett, Forrest H.; Andre, David; Keane, Martin A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. Artificial Intelligence in Design '96. Springer, Dordrecht. pp. 151–

170. doi:10.1007/978-94-009-0279-4_9.

[43]Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2

[44]Maharlou H, Niakan Kalhori SR, Shahbazi S, Ravangard R. Predicting Length of Stay in Intensive Care Units after Cardiac Surgery: Comparison of Artificial Neural Networks and Adaptive Neuro-fuzzy System. Healthc Inform Res. 2018;24(2):109–117. doi:10.4258/hir.2018.24.2.109 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5944185/

[45] Research article: An attention based deep learning model of clinical events in the intensive care unit. Author: Deepak A. KajiID1, John R. ZechID1, Jun S. Kim1, Samuel K. Cho1,2, Neha S. Dangayach2,3, Anthony B. Costa2, Eric K. Oermann2 https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0211057&type=printable

[46] A. Patel, M. Johnson, and R. Aparasu, "Predicting in-hospital mortality and hospital length of stay in diabetic patients," Value in Health, vol. 16, no. 3, pp. A17–A17, 2013.

[47]Sensitivity and Specificity from Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Sensitivity_and_specificity

[48] False Positive Rate from Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/False_positive_rate

[49]University of Waikato, New Zealand. CC Creative Commons Attribution 4.0 International License. https://www.futurelearn.com/courses/data-mining-with-weka/0/steps/25396

[50] Github repository for Hospital-los-predictor. Author: Daniel Codes. Dec 17, 2018. https://github.com/daniel-codes/hospital-los-predictor

[51] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench.

Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.