Fall 12-16-2019

A New Scalable, Portable, and Memory-Efficient Predictive Analytics Framework for Predicting Time-to-Event Outcomes in Healthcare

Ramesh Manyam

A NEW SCALABLE, PORTABLE, AND MEMORY-EFFICIENT PREDICTIVE ANALYTICS FRAMEWORK FOR PREDICTING TIME-TO-EVENT OUTCOMES IN HEALTHCARE

by

RAMESHBABU MANYAM

Under the Direction of Yanqing Zhang, PhD

ABSTRACT

Time-to-event outcomes are prevalent in medical research. To handle these outcomes, as well as censored observations, statistical and survival regression methods are widely used based on the assumptions of linear association; however, clinicopathological features often exhibit nonlinear correlations. Machine learning (ML) algorithms have been recently adapted to effectively handle nonlinear correlations. One drawback of ML models is that they can model idiosyncratic features of a training dataset. Due to this overlearning, ML models perform well on the training data but are not so striking on test data. The features that we choose indirectly influence the performance of ML prediction models. With the expansion of big data in biomedical informatics, appropriate feature engineering and feature selection are

vital to ML success. Also, an ensemble learning algorithm helps decrease bias and variance by combining the predictions of multiple models.

In this study, we newly constructed a scalable, portable, and memory-efficient predictive analytics framework, fitting four components (feature engineering, survival analysis, feature selection, and ensemble learning) together. Our framework first employs feature engineering techniques, such as binarization, discretization, transformation, and normalization on raw dataset. The normalized feature set was applied to the Cox survival regression that produces highly correlated features relevant to the outcome.The resultant feature set was deployed to "eXtreme gradient boosting ensemble learning" (XGBoost) and Recursive Feature Elimination algorithms. XGBoost uses a gradient boosting decision tree algorithm in which new models are created sequentially that predict the residuals of prior models, which are then added together to make the final prediction.

In our experiments, we analyzed a cohort of cardiac surgery patients drawn from a multi-hospital academic health system. The model evaluated 72 perioperative variables that impact an event of readmission within 30 days of discharge, derived 48 significant features, and demonstrated optimum predictive ability with feature sets ranging from 16 to 24. The area under the receiver operating characteristics observed for the feature set of 16 were 0.8816, and 0.9307 at the 35th, and 151st iteration respectively. Our model showed improved performance compared to state-of-the-art models and could be more useful for decision support in clinical settings.

INDEX WORDS: Ensemble Learning, Gradient Decision Tree Algorithm, Extreme Gradient Boosting, XGBoost, Advanced Machine Learning Algorithms, Deep Learning, Artificial Intelligence, Neural Networks, Feature Engineering, Feature Selection, Recursive Feature Elimination, Coronary Artery Bypass Grafting Surgery, CABG, Cox Proportional Hazards Model, Survival Regression, Time-to-event Outcomes, Predictive Analytics Modeling.

A NEW SCALABLE, PORTABLE, AND MEMORY-EFFICIENT PREDICTIVE ANALYTICS FRAMEWORK FOR PREDICTING TIME-TO-EVENT OUTCOMES IN HEALTHCARE

by

RAMESHBABU MANYAM

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2019

A NEW SCALABLE, PORTABLE, AND MEMORY-EFFICIENT PREDICTIVE ANALYTICS FRAMEWORK FOR PREDICTING TIME-TO-EVENT OUTCOMES IN HEALTHCARE

by

RAMESHBABU MANYAM

Committee Chair: Yanqing Zhang

Committee: Zhipeng Cai

Pavel Skums

Jose N Binongo

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2019

DEDICATION

I dedicate my dissertation work to my loving, caring and supportive family.

To my parents, Veerakota Subbarao and Syamalamba, who never went to college, yet made sure I pursue formal studies. My father had been the smartest mind and silent learner with disability I ever witnessed in my life. He's my first and forever mentor who implanted the seeds of learning while my mother fought for bread and butter day in and day out.

To my grand ma, Manikyamba, who taught me how to survive in this world, how to aim and achieve goals with passion single-handedly, and make it count at the end. To my elder siblings (Nagalakshmi, Nagabhushanam, Venkateswara rao, and Aruna), who encouraged me to pursue studies to become a first college-grad in the family, while they themselves took care of family's day-to-day needs. To my brother-in-laws (Koteswara rao and Nageswara rao), my initial role-models, who guided me thoughtfully with both moral and financial support.

To my beloved beautiful wife, Dr. Suneetha B Manyam, who instills hard-working passion in me, shows me how to balance family, work, social and spritual lives, and still have fun every moment. To my lovely kids, Manvitha, Tejesh and Kautilya, my treasures and pleasures, who are the true inspiration for my strength in pursuing this thesis until it sees the light out of the tunnel.You four allowed me the space to scream, smile, vent, cry, and laugh. You would never know just how each of you shaped me into the human I have become. I feel blessed to have you all beside me thorough this challenging journey of learning and earning a doctoral degree with this dissertation. This effort and accomplishment belong to all of you.

ACKNOWLEDGMENTS

My Ph.D dissertation work has seen the light with consistent encouragement and support of many wonderful personalities all along. I would like to acknowledge and thank each and everyone of those who helped me make it happen.

I would like to express sincere gratitude to Dr. Yanqing Zhang, my dissertation committee chair and advisor, for his patient guidance, consistent encouragement, empathetic dedication, and valuable inputs and insights. I would also like to thank dissertation committee members, Dr. Zhipeng Cai, Dr. Pavel Skums and Dr. Jose Binongo for taking the time to offer their advice and assistance in keeping my progress on schedule.

I am greatly indebted to my team of mentors and colleagues at Emory University. Thank you so much, Dr. Lance Waller, Dr. Gari Clifford, Dr. James Blum, Dr. William B Keeling, Dr. Joshua Rosenblum, Jim Kinney, Supraja Koppisetty, Tweedy Robert, and Seth Carter. Your guidance, support and encouragement mean a lot to me.

I also want to thank our computer science department faculty and staff members: Dr. Martin Fraser, Dr. Yi Pan, Dr. Rajasekhar Sunderraman, Dr. Robert Harrison, Dr. Xiaojun Cao, Dr. Yingshu Li, Dr. Xiaolin Hu, Dr. K.N. King, Dr. Wei Li, Tammie Dudley, Jamie Hayes, Paul Bryan, Adreinne Martin, Celena Pittman, Venette Rice, Elizabeth Hazzard and all GSU faculty, staff and fellow-students - with whom I travelled a few years and learnt a lot about computer courses and life as well. To all my GSU faculty and staff members, I would be indebted forever. To Dori Neptune, Dr. Lisa Armitstead, and Jeff Steely, I would like

to express my sincere thanks for your support and encouragement during the Three Minute Thesis Competition at GSU.

I would like to thank all my former colleagues and friends at Georgia Department of Transportation, including Rama, Krishnaveni, Hema, Lavanya, Mangala, Prabhakar, Prashant, Pavan, Santosh, Jeboy, Jacob, Sai, Raj and Deepti. I also want to thank all my colleagues at my previous work places, Kennesaw State University, University of Tennessee, University of North Georgia.

I would like to acknowledge the consistent moral support and encouragement of my friends: SeshuKumar, Maharaj, Gautam, Riyaz, Sreenivas, Raju, Lanero, Vandemataram, Sureshbabu, Tanuja, Ashutosh, Nirmalya, Som, Praveen, ArunU, ArunJ, Satish, Vivek, Narendra, Mohan rao, Narasimha rao, Sucheta, Dhara, Gauri, Arun, Sarada, Shyam, Rajitha, Vijay, Kalyani, Ravi, Prasoona, Pavan, Aparna and their families. Special thanks to all the well-deserved family members of my relatives, Chandra Mohan, Prasad and Suresh for their hospitality, good wishes, and warm support during ups and downs as well. Thank you so much for your encouragement all along. I will continue to cherish, and improve further.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

In this chapter, we present an overview of the rationale of our proposed research and indicate why it's worth doing for healthcare quality improvement as well as knowledge advancement. Specifically, we discuss i) the prevailing problem of time-to-event outcomes in medical research that became a critical measure of quality and cost of healthcare in recent years, and ii) the best feasible approaches for more accurate assessment, and better risk-prediction methodologies to handle such outcomes. We tackle these problems in the context of developing a new scalable, portable and memory-efficient predictive analytics framework using highly tuned and optimized machine learning algorithms. In the following sections, we describe the motivation, significance and implications of this research study.

1.1 Role of Artificial Intelligence, Machine Learning, and Deep Learning in Healthcare

The US healthcare system generates approximately one trillion gigabytes of data annually, and this amount doubles-up every two years [1]. These enormous quantities of data have been accompanied by a rise in cost-effective, and large-scale computing power. Together, they raise the possibility that artificial intelligence - and machine learning, in specific - can generate insights both to improve the discovery of new predictive data analytics techniques and to make the delivery of current ones more effective.

Figure 1.1 provides a quick history and evolution of Artificial Intelligence and its subsets. In 1976, Maxmen [2] predicted that artificial intelligence (AI) in the 21st century would

usher in "the post-physician era," with health care provided by paramedics and computers. More than four decades later, today, the extinction of physician seems unlikely, though. However, as outlined by Hinton [3], in a related viewpoint in 2018, the emergence of a radically different approach to AI, called deep learning, has the potential to effect major changes in clinical medicine and health care delivery. As Verghese et al. [4] suggested, "clinicians should seek a partnership where the machine predicts, and the human explains and decides on action." Clinicians are trained to reason and act under uncertainty. AI does not reduce the importance of this professional competency. On the contrary, AI is best suited to augmenting human intelligence, bringing more data into focus to support human judgement.

Despite concerns of convincing clinicians and patients to 'trust a black-box', AI and Deep Learning (DL) have gradually entered the mainstream clinical medicine - especially in handling heterogeneous data sets in image-intensive fields such as radiology, radiotherapy, pathology, ophthalmology, dermatology, and image-guided surgery. In many cases, interpretation of images by deep learning neural network (NN) systems has outperformed individual clinicians - when measured against a consensus of expert readers or gold standards such as pathologic findings [5]. Clinically relevant applications have widened beyond image processing to include risk stratification for a broad range of patient populations, and health care organizations are capitalizing on deep learning and other machine-learning tools to improve logistics, quality management, and financial oversight [6]. However, NN models are computationally intensive and may take a longer time to train and converge to a solution. To

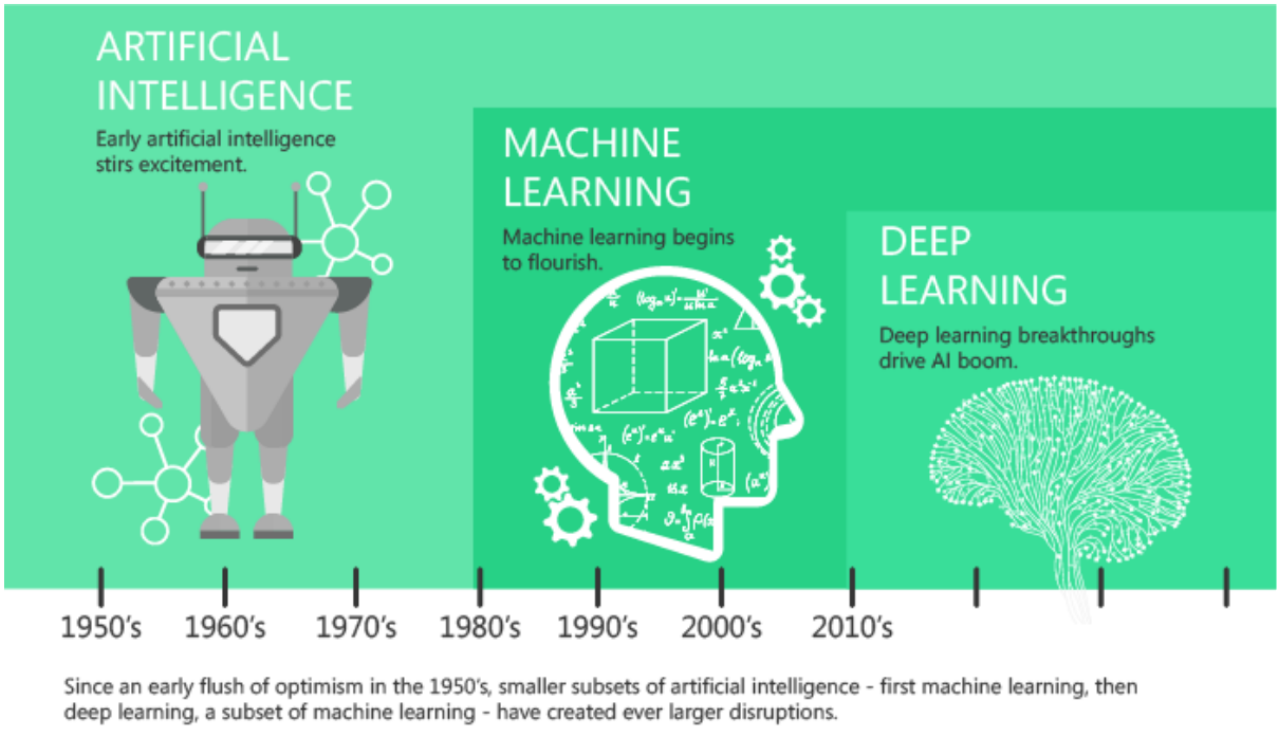


Figure 1.1 A Quick History of Machine Learning

handle small-to-medium-scale tabular data of time-to-event outcomes, the highly-effective, faster, scalable and portable advanced machine learning (ML) techniques, such as ensemble learning algorithms are better choice, and this forms the basis for our research study.

1.2 Hospital Readmissions

Healthcare costs are constantly rising around the globe with each passing year. Hospital readmissions, that occur shortly after discharge, are identified as one of the significant contributors to healthcare expenses. For instance, the estimated annual cost of readmissions for Medicare is $26 billion and $17 billion is considered avoidable. Centers for Medicare and

Medicaid Services (CMS) penalized over 2,500 hospitals by more than $564 million in 2017 for excessive 30-day hospital readmission rates [7].

In 2011, 3.3 million hospital readmissions, with an associated cost of $41.3 billion, made reducing hospital readmissions a national priority of the Affordable Care Act reform. Subsequently, CMS created 'Hospital Readmission Reduction Program' [HRRP] in 2012. The HRRP is a value-based reimbursement program that cuts-down payments to hospitals with excessive rates of readmissions [8]. This program tracks readmissions for Medicare patients admitted initially for six targeted conditions: heart attack, heart failure, pneumonia, chronic obstructive pulmonary disease (COPD), elective hip and knee replacement, and coronary artery bypass graft (CABG). Figure 1.2 shows the chronological timeline for readmission penalties for the above six medical procedures.

Among these 6 medical procedures, CABG is of particular interest due to its higher short-term readmission rates relatively and mean hospital charges. Among CABG patients, average medicare readmission rate is 1 in 5 [9, 10]. However, in many situations, these readmissions are potentially predictable, and thus avoidable.

1.3 Coronary Artery Bypass Grafting Surgery

Heart disease is the leading cause of death around the world. For both male and female populations, heart disease turns out to be the first leading cause of death in USA as well [11], as shown in Figure 1.3. According to the Center for Disease Control and Prevention [7], around 610,000 Americans die from heart disease every year, that's 1 of every 4 deaths.

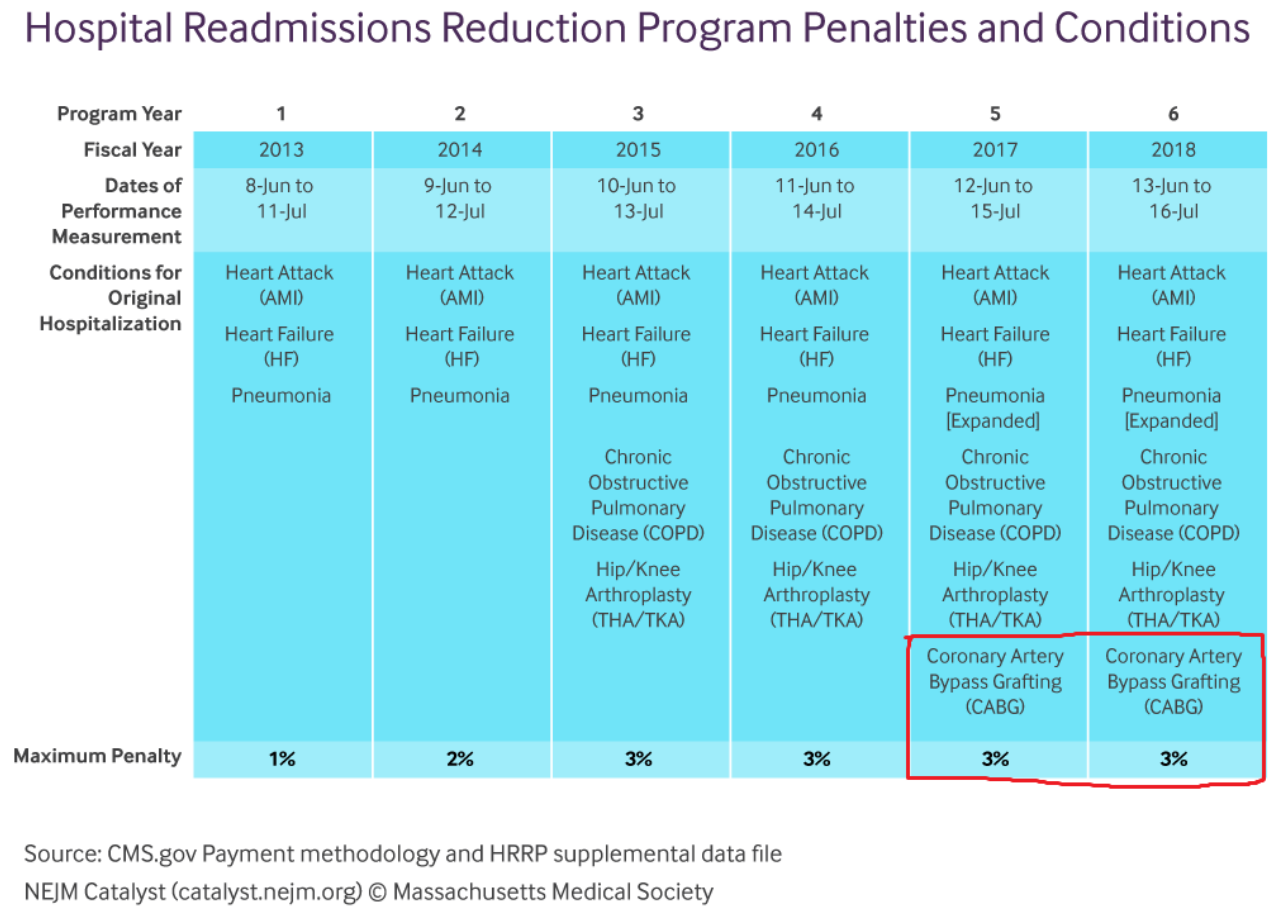


Figure 1.2 Hospital readmissions Reduction Penalties

Someone dies from heart disease, stroke, or cardiovascular disease every 40 seconds in the USA [10]. Heart disease costs the United States about $200 billion each year. This total includes the cost of healthcare services, medications, and lost productivity. Coronary Heart Disease (CHD) is the most common type of heart disease, killing over 370,000 people annually. CHD is a disease in which a waxy substance called plaque (plak) builds up inside the coronary arteries [13]. Coronary arteries are the blood vessels that supply oxygen-rich blood to heart, as shown in Figure 1.4. Over time, plaque can harden or rupture (break open).

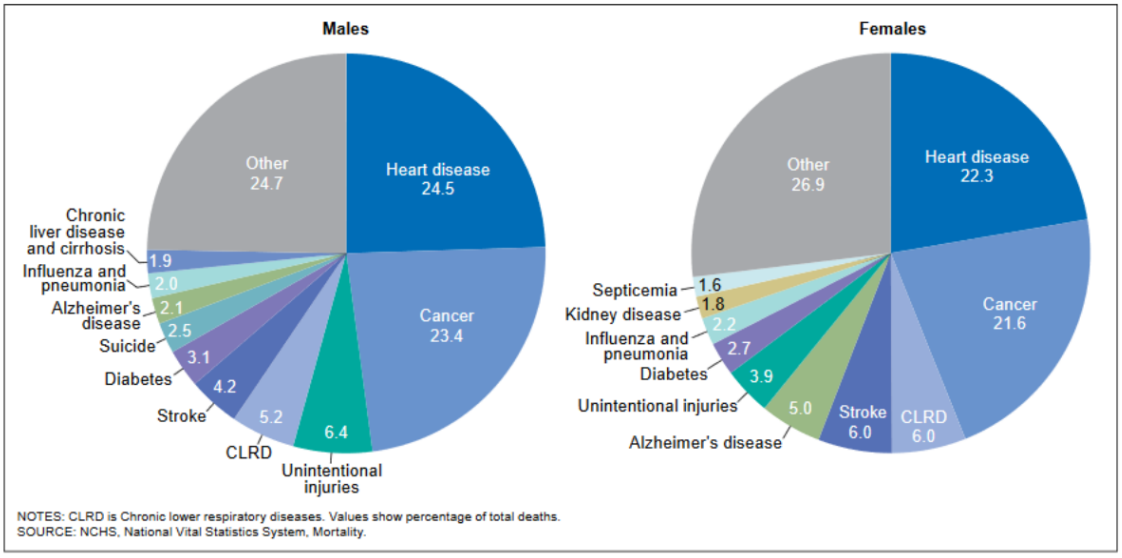


Figure 1.3 Percentage Distribution of Diseases

Hardened plaque narrows the coronary arteries and reduces the flow of oxygen-rich blood to the heart. This can cause chest pain or discomfort called angina (an-JI-nuh or AN-juh-nuh). If the plaque ruptures, a blood clot can form on its surface. A large blood clot can mostly or completely block blood flow through a coronary artery. This is the most common cause of a heart attack.

Coronary Artery Bypass Graft (CABG, pronounced like the word 'cabbage') surgeries are among the most commonly performed operations for CHD. During CABG, a healthy artery or vein from the body is connected, or grafted, to the blocked coronary artery. The grafted artery or vein bypasses the blocked portion of the coronary artery. This creates a new path for oxygen-rich blood to flow to the heart muscle. In Figure 1.4, figure A shows the location of the heart. Figure B shows how vein and artery bypass grafts are attached to the heart.The mean price for CABG was $151,271 and ranged from $44,824 to $448,038

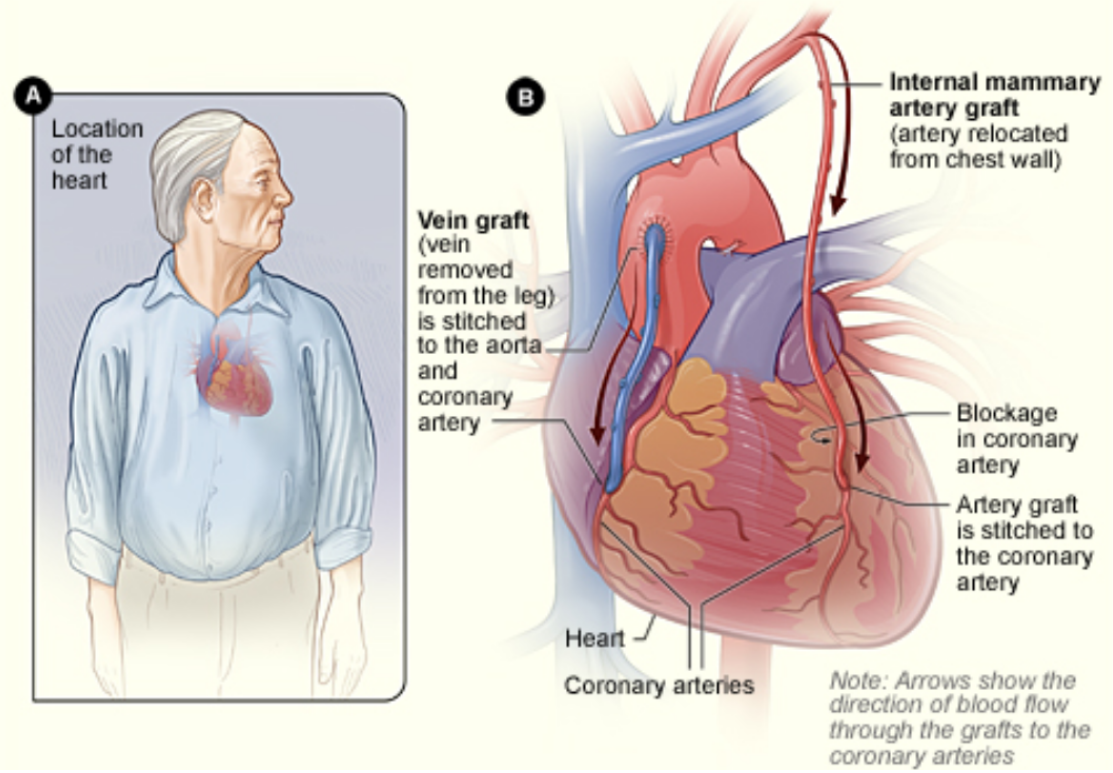


Figure 1.4 Coronary Artery Bypass Grafting

[14]. Readmission events of CABG surgery population forms the pilot study cohort of our proposed predictive analytics framework.

1.4 Objective

The primary focus of this research study is to develop a new scalable, portable, memory-efficient, and user-friendly predictive analytics framework fitting four components (feature engineering, survival analysis, feature elimination, and ensemble learning methods). The objectives of this new ensemble learning based model are to evaluate the predictive ability more accurately with effective feature engineering and feature selection techniques and to improve performance by using less features compared to state-of-the-art models. For testing

and validation of our proposed framework, we consider running experiments with a study cohort comprising of i) demographics data and pre-, peri-, and postperative electronic health records of cardiac surgery patient population, ii) time-varying perioperative clinical data, drawn from the Society of Thoracic Surgeons (STS) database pertaining to a multi-hospital academic health system.

1.5 Motivation for the study

Hospital readmissions contribute substantially to the overall healthcare cost. Coronary artery bypass graft (CABG) is of particular interest due to its relatively high short-term readmission rates and mean hospital charges. Readmission is an important source of patient dissatisfaction as well. Because hospital readmissions are costly and potentially avoidable, reducing hospital readmission rates has been proposed as a means to improve quality of care and reduce costs [4]. Therefore it is important to be able to accurately identify those patients at high risk for early readmission after open-heart surgery. With more accurate predictive analytics model, clinicians i) can more appropriately decide on patient's readiness for discharge and ii) can recommend better follow-up care measures for patients who are at higher risk of readmission. This is where predictive healthcare transforms into preventive healthcare.

1.6 Contribution

The current dissertation work presents a multidisciplinary research efforts to investigate the advanced machine learning, deep learning, and ensemble learning algorithms, survival analysis methods, automatic feature selection and feature engineering techniques, and to

analyze the complicated and unstructured healthcare data in a better manner so as to develop effective decision support tools for clinical settings.

We constructed a new scalable and portable predictive analytics framework for predicting time-to-event outcomes in healthcare (for instance, hospital readmission events following CABG surgery) using advanced feature engineering, feature selection methods, and ensemble learning techniques - prominently though "eXtreme Gradient Boosting" (XGBoost) that employs a gradient boosting decision tree algorithm. We have established simple feature engineering techniques to process the real data sets of unstructured electronic health records and employed automatic feature selection (Recursive Feature Elimination) techniques to minimize redundancy which led to improved prediction accuracy compared to state-of-the-art models. We carried out experiments with a study cohort of adult cardiac surgery patient population, drawn from two data sources: the Society of Thoracic Surgeons (STS) database and clinical data warehouse of a multi-hospital academic health system. The results of our model demonstrated improved predictive ability in terms of 'Area Under the Receiver Operating Characteristics (AUROC)', and concordance index metrics.

1.7 List of Additional Contributions and Accomplishments

The following list provides additional contributions of our research work such as, papers presented at scientific conferences, submitted to journal publication, received travel awards.

- Manyam, R. B., Zhang, Y., Keeling, W. B., Binongo, J., Kayatta, M., & Carter, S. (2018). Deep Learning Approach for Predicting 30 Day Readmissions after Coronary

Artery Bypass Graft Surgery. *arXiv preprint arXiv:1812.00596*.

- Manyam, R. B., Zhang, Y., Binongo, J., Rosenblum, J.M., & Keeling, W. B. (2019, June). *Unraveling the impact of time-dependent perioperative variables on 30-day readmission following CABG*. Presented at the 45th Annual Meeting of Western Thoracic and Surgical Association, Olympic Valley, CA, USA. Abstract retrieved from https://meetings.westernthoracic.org/abstracts/2019/CF6.cgi

- Manyam, R. B., Zhang, Y., Binongo, J., Rosenblum, J.M., & Keeling, W. B. (2019, June). *A New Scalable, Portable Predictive Analytics Framework for Predicting Time-to-event Outcomes in Healthcare*. Presented at the 2nd Machine Learning in Science and Engineering Symposium, Atlanta, GA, USA.

- Manyam, R. B., Zhang, Y., Binongo, J., Rosenblum, J.M., & Keeling, W. B. (In Review). Unraveling the impact of time-dependent perioperative variables on 30-day readmission following CABG. Manuscript submitted to the *Journal of Cardiovascular and Thoracic Surgery*.

- Manyam, R. B., Zhang, Y., Binongo, J., Rosenblum, J.M., & Keeling, W. B. (2019, October). *A Simple, Scalable And Portable Machine Learning Model With Effective Feature Selection For Accurately Predicting 30-day Readmission After Discharge Following CABG*. Submitted to the 100th Annual Meeting of the Association of American Thoracic Surgeons, New York, NY, USA.

- Manyam, R. B., Zhang, Y., Binongo, J., Rosenblum, J.M., & Keeling, W. B.. (2019, November). *Explainable Decision Tree Framework for Effectively Assessing the Risk Factors for 30-day Readmission Following CABG*. Submitted to the 34th AAAI Conference of Artificial Intelligence, NEw York, NY, USA

- In addition to the above contributions, our research work secured second prize in Georgia State University's (GSU) Spring 2019 Three Minute Thesis (3MT) competition, featured on GSU merit pages, Computer Science department, Graduate School, and Provost office websites, and received travel awards at two scientific conferences: 2nd Machine Learning in Science and Engineering Symposium (June, 2019), and Machine Learning for Health (ML4H) Workshop at the Thirty-second Conference on Neural Information Processing Systems (NeurIPS, December 2018). We enclosed a few pictures taken at the conference presentations in Appendix.

1.8 Thesis Outline

The remainder of this dissertation is structured as follows. Chapter 2 presents an overview of the background information and concepts relevant to the scope of our research focus. Chapter 3 provides a comprehensive review of state-of-the-art models that deal with the assessment of risk factors associated with 'time-to-event' outcomes in acute care medical procedures in general, and prediction of hospital readmission events following cardiac surgery, in specific. Chapter 4 describes our proposed model, a new scalable, portable, and memory-efficient predictive analytics framework, demonstrates the experiments that we conducted and the

results obtained. Chapter 5 highlights the summary of our research findings, limitations, conclusion, and recommendations for further research and enhancements to our predictive anlytics framework.

CHAPTER 2
BACKGROUND STUDY

In this chapter, we present an overview of background information and concepts relevant to the design of our proposed framework. We review topics such as linear regression, logistic regression, univariable, multivariable, survival regression analyses, Cox Proportional Hazard (CPH) model, and machine learning methods. Additionally, we discuss ensemble learning algorithms and feature selection methods that form the basis of our new predictive analytics framework.

2.1 Univariable, Bivariable and Multivariable Analysis

Data in statistics is usually classified based on how many variables are in a particular study. For example, 'height' might be one variable and 'weight' might be another variable. Depending on the number of variables being analyzed at, the data might be univariable, bivariable or multivariable. Univariable involves the analysis of a single variable, while multivariable analysis examines two or more variables.

2.1.1 Univariable Analysis

Univariable analysis is the simplest form of analyzing data. As the name "Uni" implies, data has only 'one' variable in this kind of analysis. It doesn't deal with causes or relationships (unlike regression) and it's major purpose is to describe; it takes data, summarizes that data and finds patterns in the data.

2.1.2 Bivariable Analysis

Bivariable analysis is the analysis of exactly two variables. It usually involves an independent variable (for example, X) , and a dependent variable (Y).

2.1.3 Multivariable Analysis

Multivariable (MVA) deals with analysis of more than one statistical outcome variable at a time. The technique is used across multiple dimensions while taking into account the effects of all variables on the responses of interest. This MVA is more valuable technique, especially when working with correlated variables. MVA involves a dependent variable and multiple independent variables.

2.1.4 Independent vs. Dependent Variables

We can interpret the correlation between the dependent variable and the independent variable as resulting from a cause-effect relationship from independent (cause) to dependent (effect) variable. Sometimes, it may be more reasonable to refer to "independent variables" as "predictors", and "dependent variables" as "response-," "outcome-," or "criterion-variables." Suppose, we want to find out the relationship between caloric intake and weight. As shown in Table 2.1, 'Caloric Intake' would be 'independent' variable and 'Weight' would be 'dependent' variable.

Table 2.1 Independent vs Dependent Variable

Calorie Intake (X)	Weight in Lbs
4000	200
2500	150
2000	135
3500	175
1500	110

2.2 Regression Analysis

Regression is a statistical technique used to model and analyze the relationships between variables and how they contribute and relate to each other to producing a particular outcome together. Regression analysis helps to understand how a typical value of a (outcome|dependent) variable changes when any one of the (predictor|independent) variables is varied, while the other (predictor|independent) variables are unchanged. (i.e., which among the independent variables are related to the dependent variable). Linear and Logistic regressions are usually the first modelling algorithms that people learn for Machine Learning and Data Science.

2.2.1 Linear Regression

A linear regression refers to a regression model that is completely made up of linear variables. Single Variable Linear Regression (SVLR) is a technique used to model the relationship between a single input independent variable (feature variable) and an output dependent variable using a linear model i.e a line, for instance. The more general case is Multi-variable Linear Regression (MVLR) where a model is created for the relationship between multiple independent input variables (feature variables) and an output dependent variable (classifier

variable). The model remains linear in that the output is a linear combination of the input variables. We can model a multi-variable linear regression as the following:

Y = a_1*X_1 + a_2*X_2 + a_3*X_3 ..... a_n*X_n + b

Where a_n are the coefficients, X_n are the variables and b is the bias. As we can see, this function does not include any non-linearities and so is only suited for modelling linearly separable data.

2.2.2 Logistic Regression

Logistic regression is a class of regression where the independent variable is used to predict the dependent variable. When the dependent variable has two categories, then it is a binary logistic regression. When the dependent variable has more than two categories, then it is a multinomial logistic regression. When the dependent variable category is to be ranked, then it is an ordinal logistic regression. The idea of logistic regression is to find a relationship between features and probability of particular outcome. For example, when we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail. This type of a problem is referred to as 'binomial logistic regression', where the response variable has two values 0 and 1 or pass and fail or true and false. Multinomial logistic regression deals with situations where the response variable can have three or more possible values.

Figure 2.1 provides the expression of Logistic regression.

The left hand side of the equation is called the logit or log-odds function, and p(x)/(1-p(x)) is called odds. The odds signifies the ratio of probability of success to probability of

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Figure 2.1 Log-odds function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

Figure 2.2 Sigmoid Equation

failure. Therefore, in logistic regression, linear combination of inputs are mapped to the log(odds) - the output being equal to 1. If we take an inverse of the above function, we get Sigmoid equation, as expressed in Figure 2.2:

The Sigmoid function produces an S-shaped curve (Figure 2.3), and it always gives a value of probability ranging from 0<p<1. Logistic regression has been popular with medical research in which the dependent variable is whether or not a patient has a disease, in our research application case [as described in Chapter 4], for example, it could be a 'readmission event within 30 days of discharge following CABG'.

For logistic regression, the predicted dependent variable is the estimated probability that a particular subject will be in one of the categories (for example, the probability that a specific patient has the disease, given his/her set of scores on the predictor variables).

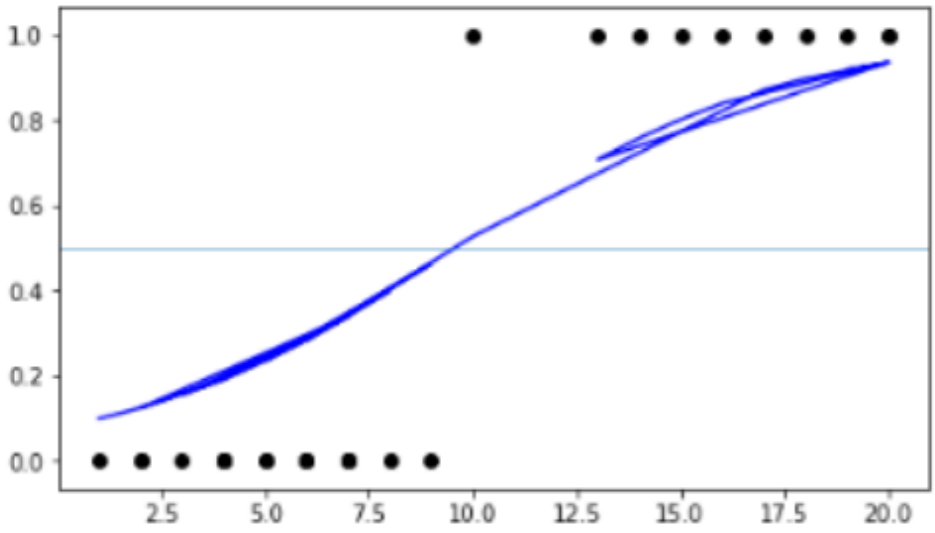


Figure 2.3 Sigmoid function

2.3 Survival Analysis

Survival analysis is a kind of statistical modeling where the main goal is to analyze and model time until the occurrence of an event of interest (for example, hospital readmission). The challenging characteristic of survival data is the fact that 'time-to-event' of interest for many instances is unknown because the event might not have happened during the period of study or missed tracking of occurrence, caused by other events. This concept is called 'censoring' which makes the survival analysis different [15]. The special case of censoring is when the observed survival time is less than or equal to the true event time called 'right-censoring' - which is one of the concepts of our focus in this research study.

In Figure 2.4, an illustrative example is given for a better understanding of the definition of censoring and the structure of survival data. Six instances are observed in this study for 12 months and the event occurrence information during this time period is recorded. So, we

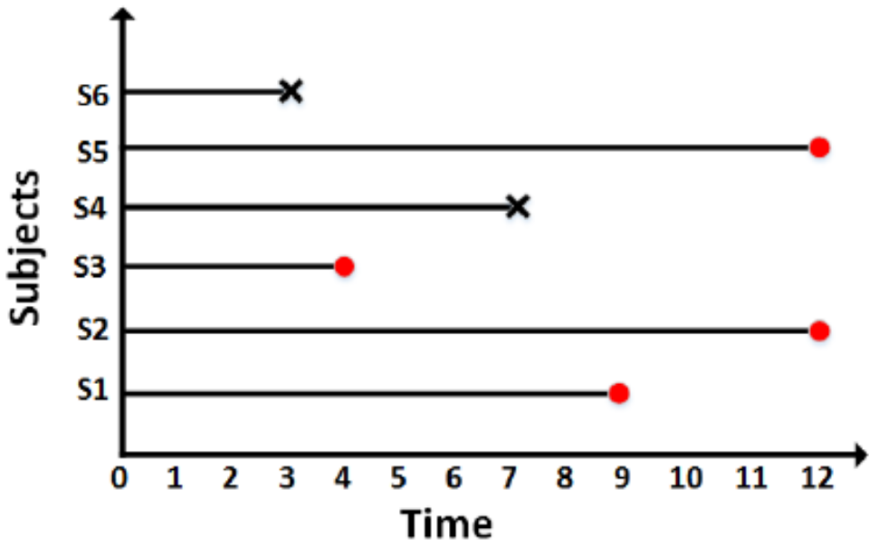


Figure 2.4 Illustration of Censoring Problem

can see that only subjects S4 and S6 have experienced the event (marked by 'X') during the follow-up time and the observed time for them is the event time. For subjects S1, S2, S3 and S5, the event did not occur within the 12 months period, and hence, these subjects are considered to be censored and marked by red dots in the figure. More specifically, subjects S2 and S5 are censored since there was no event occurred during the study period, while subjects S1 and S3 are censored due to the withdrawal or being lost to follow-up within the study time period.

Since the censored data is present in survival analysis, the standard statistical and machine learning approaches are not appropriate to analyze and predict time-to-event outcome because those approaches will miss the censored/right-censored instances.

Survival modeling provides different statistical approaches to analyze such censored data in many real-world applications. In survival analysis, a given instance i, represented by a

$$h(t) = \lim_{\delta(t)\to 0} \frac{Pr(t \le T \le t + \delta(t) | T \ge t)}{\delta(t)}$$

Figure 2.5 Survival Hazard Function

triplet (X i, i, Ti) where Xi refers to the instance characteristics and Ti indicates 'time-to-event' of the instance. If the event of interest is observed, Ti corresponds to the time between baseline time and the time of event happening, in this case i = 1. If the instance event is not observed and it's time-to-event is greater than the observation time, Ti corresponds to the time between baseline time and end of the observation, and the event indicator is i = 0. The goal of survival analysis is to estimate the time to the event of interest (T) for a new instance X j [15].

2.3.1 Survival and Hazard Function

Survival and hazard functions are the two fundamental functions in survival analysis. The survival function is denoted by

S(t) = Pr(T >t)

which signifies the probability that an individual has 'survived' up to time t. The initial value of survival function is 1 when t = 0 and it monotonically decreases with t . The hazard function corresponds to the probability that an individual dies at time t given that he or she has survived up to that point. Figure 2.5 shows the definition of the hazard function.

Survival and hazard function are non-negative functions. While the survival function decreases over time, the shape of a hazard function can be in different forms: increasing, decreasing, constant, or U-shaped.

2.3.2 Survival Regression using Cox's Proportional Hazard Model

The Cox proportional hazards model [16] is the most commonly used multivariable approach for analysing survival time data in medical research. It is a survival analysis regression model, which describes the relation between the event incidence, as expressed by the hazard function and a set of covariates. In brief, the hazard is the instantaneous event probability at a given time, or the probability that an individual under observation experiences the event in a period centred around that point in time, for instance, hospital readmission within 30 days after discharge. Mathematically, the Cox model is written as

h(t)= h0(t) $\times$ expb1x1+b2x2++bpxp

where the hazard function h(t) is determined by a set of p covariates (x1, x2, ..., xp), whose impact is measured by the size of the respective coefficients (b1, b2, ..., bp). The term h0 is called the baseline hazard, and is the value of the hazard if all the xi are equal to zero (the quantity exp(0) equals 1). The 't' in h(t) reminds us that the hazard may vary over time.

An appealing feature of the Cox model is that the baseline hazard function is estimated nonparametrically, and so unlike most other statistical models, the survival times are not assumed to follow a particular statistical distribution. The Cox model is essentially a multiple linear regression of the logarithm of the hazard on the variables xi, with the baseline hazard being an 'intercept' term that varies with time. The covariates then act multiplicatively on the hazard at any point in time, and this provides us with the key assumption of the PH model: the hazard of the event in any group is a constant multiple of the hazard in any other.

Proportionality implies that the quantities exp(bi) are called 'hazard ratios'. A value of 'bi' greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the ith covariate increases, the event hazard increases and thus the length of survival decreases. In other words, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival. This proportionality assumption is often appropriate for survival time data, but it is important to verify that it holds. The purpose of the model is to simultaneously evaluate the effect of several factors on survival. Thus, it allows us to examine how specific factors influence the rate of a particular event happening (e.g., hospital readmission or disease recurrence) at a particular point in time, for instance 'T' days. This rate is commonly referred as the hazard rate. Predictor variables are usually termed covariates in the survival-analysis literature [17]. Lifelines has an implementation of the Cox proportional hazards regression model (implemented in R under coxph) [18]

2.3.3 Evaluation Metrics for Survival Analysis

Since the censored instances exist in survival data, the standard evaluation metrics such as mean squared error and R-squared are not appropriate for evaluating the performance of survival analysis (19). In survival analysis, the most popular evaluation metric is based on the relative risk of an event for different instances called concordance index or c-index. C-index is a generalization of the area under the ROC curve (AUC) that can take into account censored data. It represents the global assessment of the model discrimination power. Similar to AUC, C-index of 1.0 corresponds to the best model prediction, while

$$\frac{1}{N} \sum_{i,\delta_i=1} \sum_{j,y_i<y_j} I[S(\hat{y}_i|X_i) < S(\hat{y}_j|X_j)]$$

Figure 2.6 Concordance Equation

C-index of 0.5 represents a random prediction.

This measure is defined in Figure 2.6. where N refers to the all comparable instance pairs and S is the survival function. The main motivation for using c-index in survival analysis is originated from the fact that the medical doctors and researchers are often more interested in measuring the relative risk of a disease among patients with different risk factors, than the survival times of patients.

2.4 Statistical Methods Vs Machine Learning Methods

In general, the survival analysis models can be divided into two main categories: (1) statistical methods including non-parametric, semi-parametric and parametric and (2) machine learning based methods such as survival trees, Bayesian methods, neural networks and random survival forests [20].

Figure 2.7 illustrates the multidisciplinary nature of machine learning (ML), data mining, data science, and related areas [21]. However, the two fields of interest here, statistics and ML are converging more and more even though the below diagram shows them almost exclusive. While both machine learning and statistics share the same goal, the differences lie in terms of 'learning from data' (drawing knowledge or insights from the data), the volume

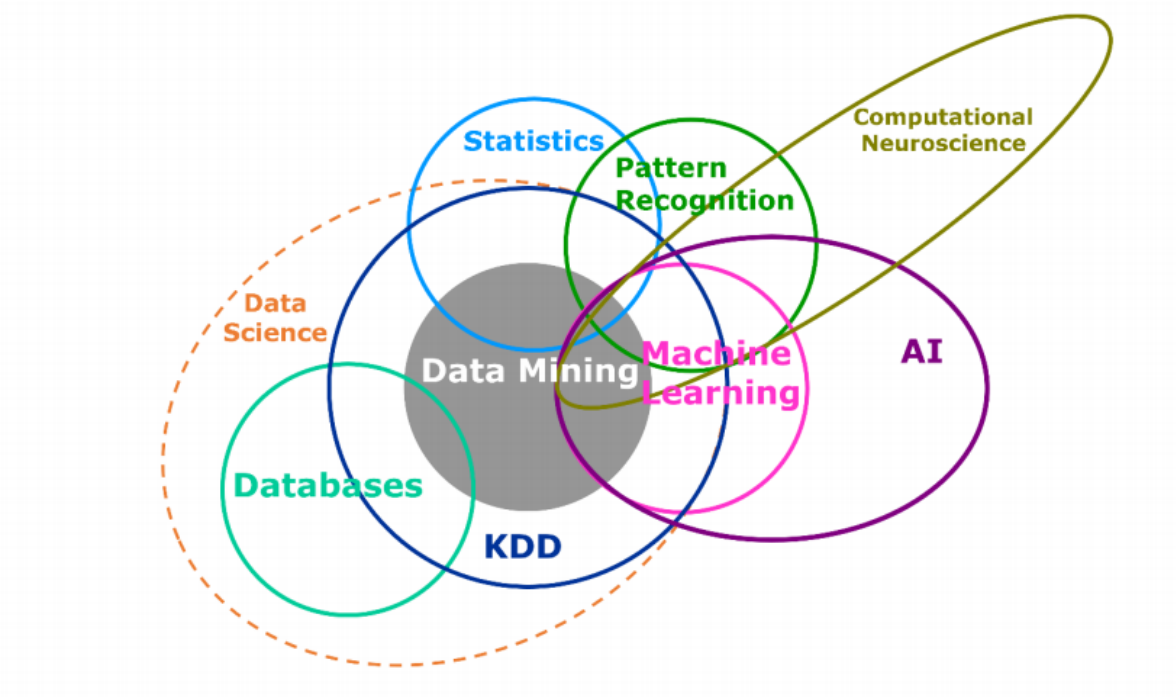


Figure 2.7 Multidisciplinary Nature of Machine Learning

Table 2.2 Mapping between Common Vocabulary Terms in Data Analysis Fields

Machine Learning Term	**Multidisciplinary Synonyms**
Feature, input	Independent Variable, Variable, Column
Case, Instance, Example	Observation, Record, Row
Label	Dependent Variable, Target
Train	Fit
Class	Categorical Target Variable Level

of data involved and human involvement for building a model. Statistical modeling are generally applied for smaller data with less attributes or they end up over fitting. Machine learning does really well with wide (more number of attributes) and deep (more number of observations). The general terminology and what it means in terms of both the methods are listed in Table 2.2.

2.5 Machine learning, Neural networks, and Deep learning

In machine learning, one develops and studies methods that give computers the ability to solve problems by learning from experiences. The goal is to create mathematical models that can be trained to produce useful outputs when fed input data. Machine learning models are provided experiences in the form of training data, and are tuned to produce accurate predictions for the training data by an optimization algorithm. The main goal of the models is to be able to generalize their learned expertise, and deliver correct predictions for new, unseen data.

There are several kinds of machine learning, loosely categorized according to how the models utilize its input data during training. Basically, these subfields are categorized into supervised-, semi-supervised-, and unsupervised-learning algorithms, as such. These are illustrated in Figure 2.8. In unsupervised learning the computer is tasked with uncovering patterns in the data without manual/human guidance. Clustering is a prime example. Most of today's machine learning systems belong to the class of supervised learning. Here, the computer is given a set of already labeled or annotated data, and asked to produce correct labels on new, previously unseen data sets based on the rules discovered in the labeled data set. From a set of input-output examples, the whole model is trained to perform specific data-processing tasks. Suppose, we have input data (X), and output variable (Y). The algorithm learns the mapping function from input to output.

$f : X \rightarrow Y$

Y typically represents an instance from a fixed set of class. The objective of the learning

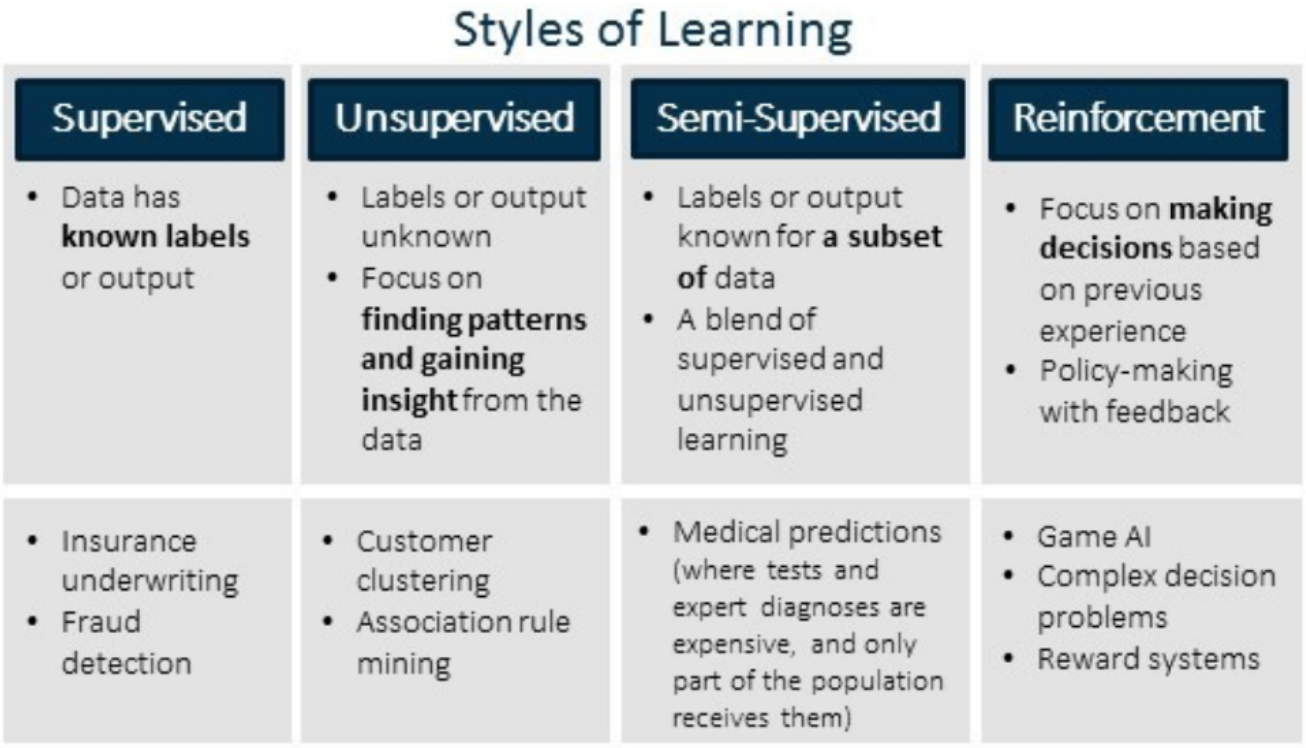


Figure 2.8 Categories of Machine Learning

algorithm is to approximate the mapping function well so that when a new input data comes, the model can predict the output class for it. For example, an algorithm trained on the labelled dataset of a specific diagnosis (for instance, an abnormal overweight, fat content etc, that indicate signs of obesity), learns to identify patients having diabetes or not.

A model's generalization ability is typically estimated during training using a separate dataset, namely, the validation set, and it is used as feedback for further tuning of the model. After several iterations of training and tuning, the final model is evaluated on a test set, used to simulate how the model will perform when faced with new, unseen data. The taxonomy of ML algorithms and methods are depicted in the Figure 2.9 [21].

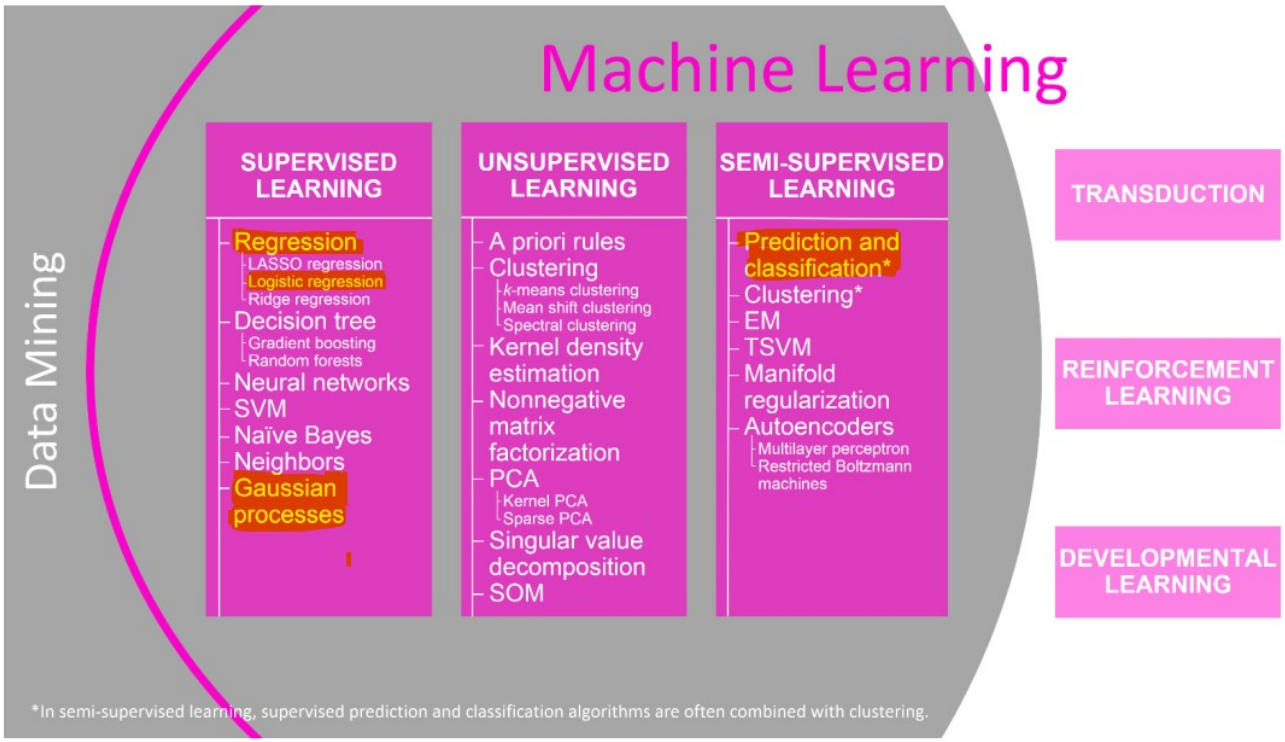


Figure 2.9 Taxonomy of Machine Learning

2.6 Deep Learning

Traditionally, machine learning models are trained to perform useful tasks based on manually designed features extracted from the raw data, or based on features learned by other simple machine learning models. Fig 2.10 summarizes the layers of learning algorithms and approaches that have been evolving since 1950s. In deep learning, the computers learn useful representations and features automatically and directly from the raw data, bypassing the manual (difficult) step. By far the most common models in deep learning are various variants of artificial neural networks, but there are others. The main common characteristic of deep learning methods is their focus on feature learning: automatically learning represen-

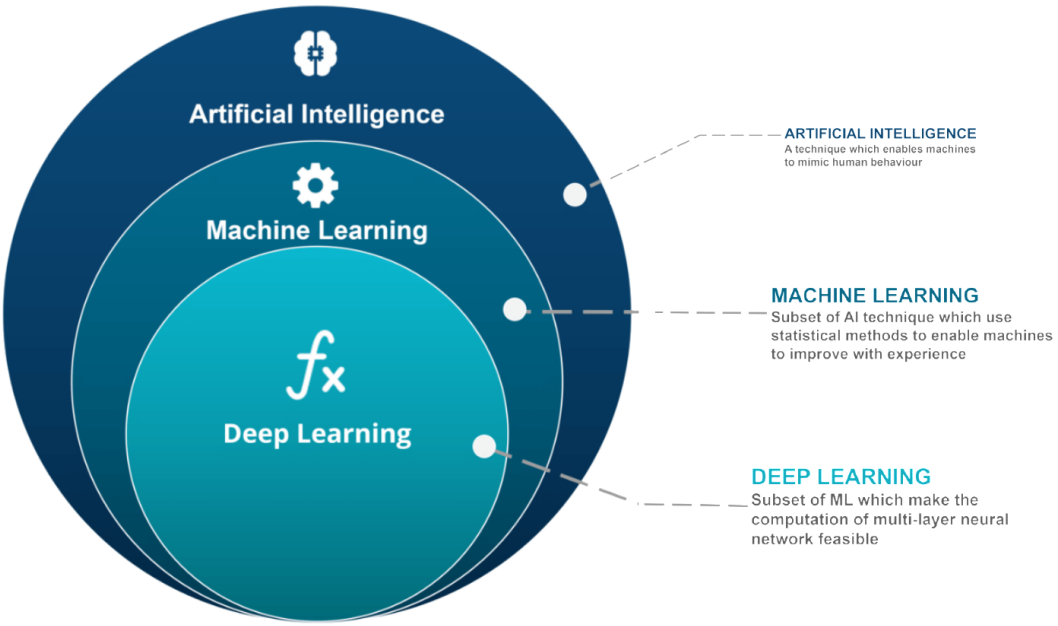


Figure 2.10 AI, ML, DL

tations of data. This is the primary difference between deep learning approaches and more "classical" machine learning. Discovering features and performing a task is merged into one problem, and therefore both improved during the same training process. See [21-23] for general overviews of the field. In a simple definition, deep learning or deep machine learning refers to use of a neural network with multiple layers of hidden nodes between input and output where the deep architectures are constructed by several levels of non-linear operations, as shown in Figure 2.11.

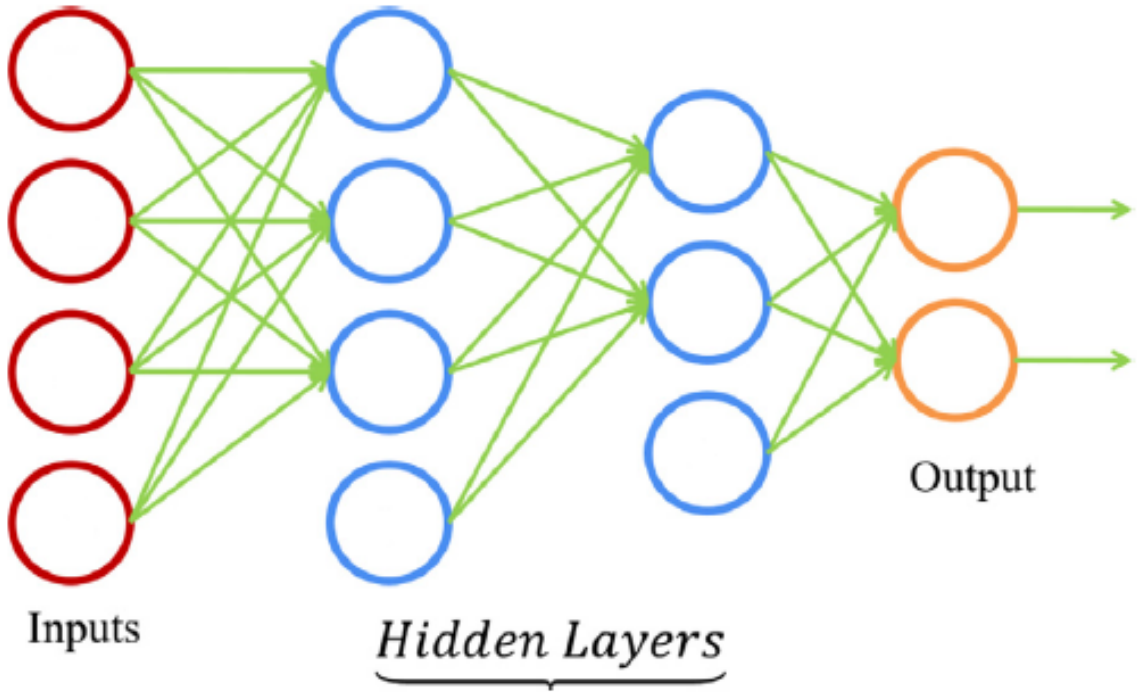


Figure 2.11 Layers of Neural Network

2.7 Ensemble Learning

This section introduces key concepts that are being used in our proposed ensemble learning predictive analytics framework.

In general, 'ensembling', as the name implies, is a technique of combining two or more algorithms of similar or dissimilar types called base learners. Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking) [24]. This approach helps obtaining better predictive performance compared to a single model.

Ensemble methods can be divided into two groups. The first category is sequential ensem-

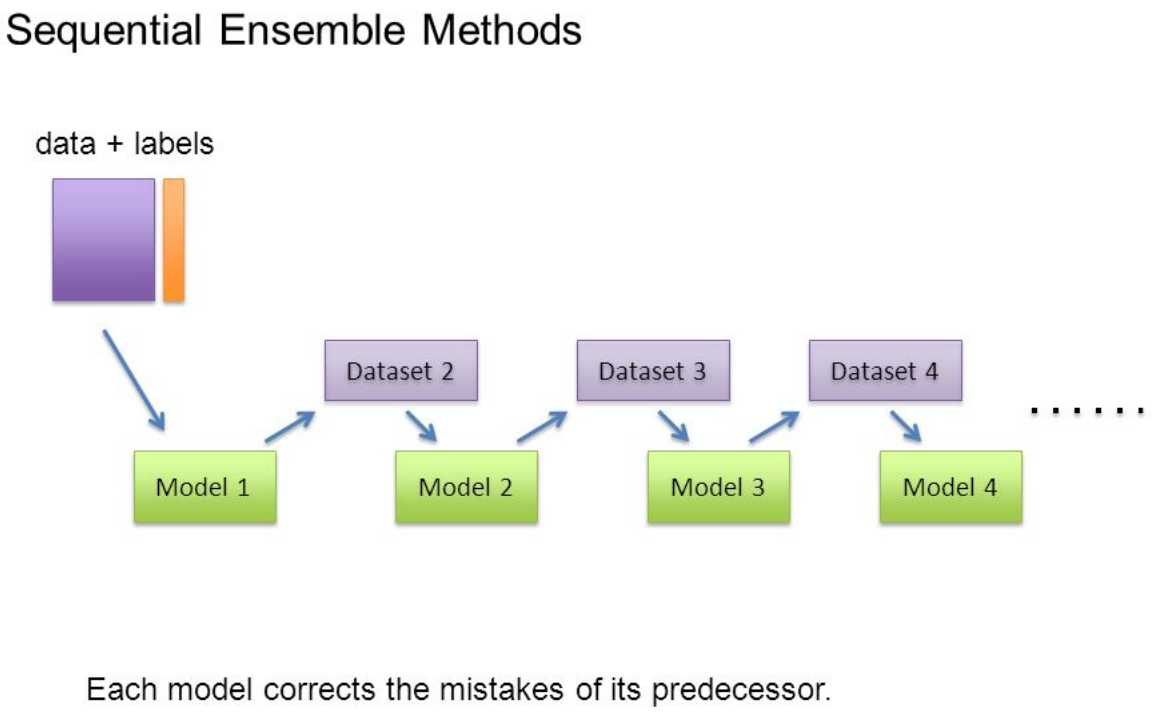


Figure 2.12 Sequential Ensemble Learning

ble methods where the base learners are generated sequentially (e.g. AdaBoost). The basic motivation of sequential methods is to exploit the dependence between the base learners. The overall performance can be boosted by weighing previously mislabeled examples with higher weight. Figure 2.12 depicts an example of sequential ensemble methods. The second category is parallel ensemble methods where the base learners are generated in parallel (e.g. Random Forest). Figure 2.13 depicts a simple parallel ensemble method. Now, we discuss below, two of the most common ensembling methods that we propose to use in our model.

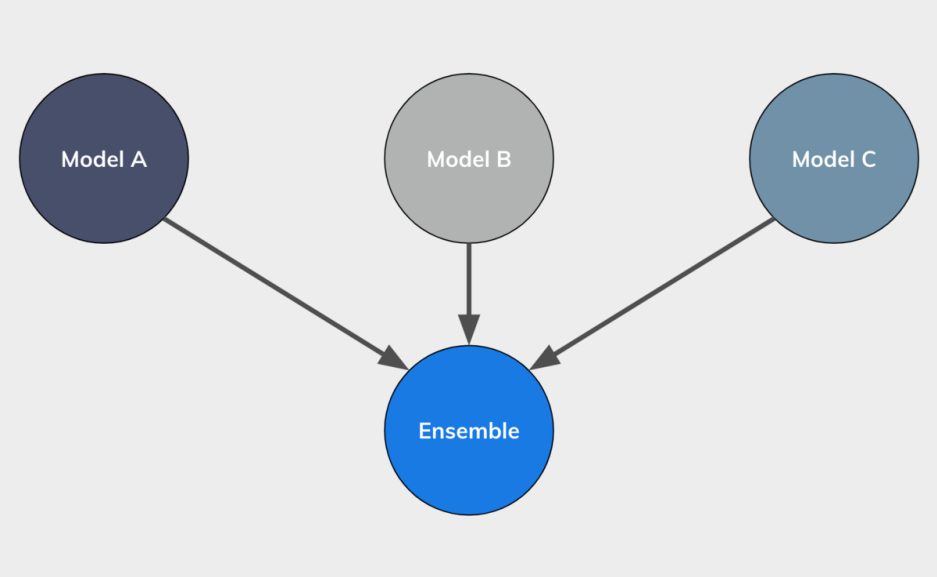


Figure 2.13 Parallel Ensemble Learning

$$f(x) = 1/M \sum_{m=1}^{M} f_m(x)$$

Figure 2.14 Bagging Function

2.7.1 Bagging

Bagging stands for bootstrap aggregation. One way to reduce the variance of an estimate is to average together multiple estimates. For example, we can train M different trees on different subsets of the data (chosen randomly with replacement) and compute the ensemble as described in Figure 2.14.

The idea behind bagging is to combining the results of multiple models (for instance, all decision trees) in order to get a generalized result. Bagging (or bootstrap aggregating) technique uses these subsets (bags) to get a fair idea of the distribution of complete set. The

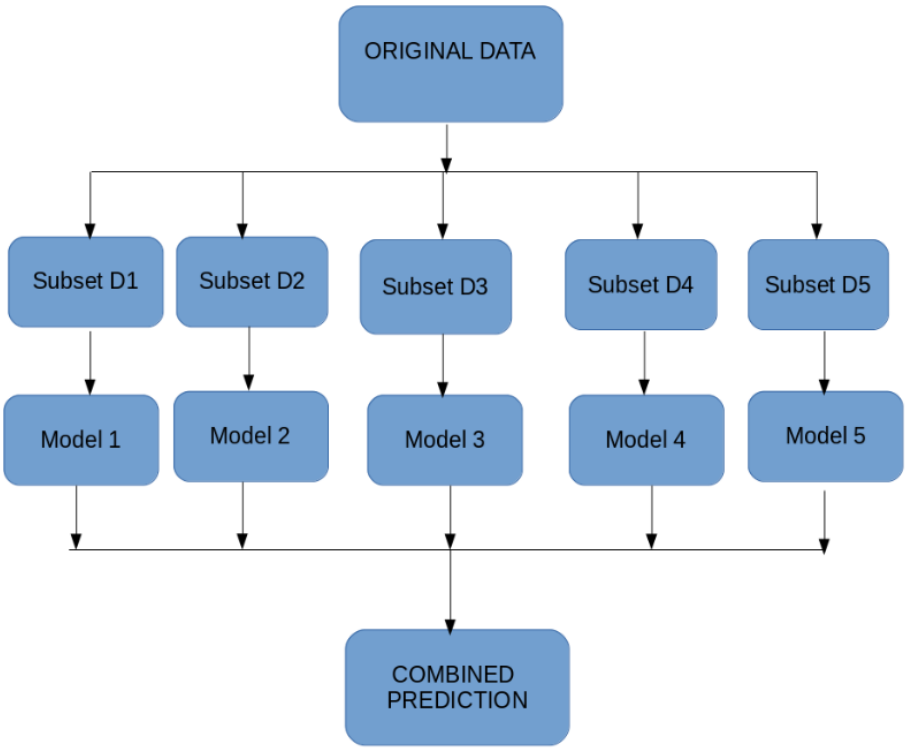


Figure 2.15 Illustration of Bagging

size of subsets created for bagging may be less than the original set. Multiple subsets are created from the original dataset, selecting observations with replacement. A base model (weak model) is created on each of these subsets. The models run in parallel and are independent of each other. The final predictions are determined by combining the predictions from all the models, as illustrated in Figure 2.15.

2.7.2 Boosting

Boosting is a sequential technique in which, the first algorithm is trained on the entire dataset and the subsequent algorithms are built by fitting the residuals of the first algorithm, thus giving higher weight to those observations that were poorly predicted by the previous model.

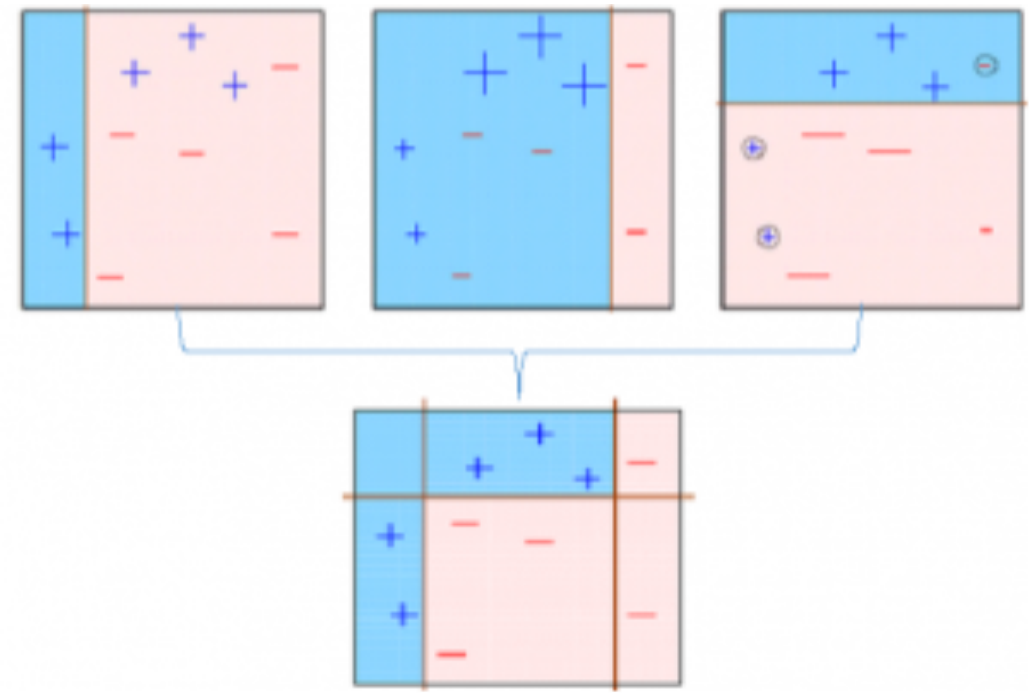


Figure 2.16 Illustration of Boosting

As illustrated in Figure 2.16, boosting technique can be described in the below steps.

A subset is created from the original dataset. Initially, all data points are given equal weights. A base model is created on this subset. This model is used to make predictions on the whole dataset. Errors are calculated using the actual values and predicted values. The observations which are incorrectly predicted, are given higher weights (Here, the three misclassified blue-plus points will be given higher weights). Another model is created and predictions are made on the dataset (This model tries to correct the errors from the previous model). Similarly, multiple models are created, each correcting the errors of the previous model. The final model (strong learner) is the weighted mean of all the models (weak learners).

Thus, the boosting algorithm combines a number of weak learners to form a strong learner. The individual models would not perform well on the entire dataset, but they work well for some part of the dataset. Thus, each model actually boosts the performance of the ensemble. It's really important to note that boosting is focused on reducing the bias. This makes the boosting algorithms prone to overfitting. Thus, parameter tuning becomes a crucial part of boosting algorithms to make them avoid overfitting.

2.7.3 XGBoost

In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered one of the best choices in recent times.

XGBoost (eXtreme Gradient Boosting) is an advanced implementation of the gradient boosting algorithm [69].The XGBoost uses the gradient boosting decision tree algorithm. Gradient boosting is an approach in which new models are created sequentially that predict the residuals or errors of prior models, which are then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.This approach supports both regression and classification predictive modeling problems.

XGBoost has proved to be a highly effective ML algorithm, extensively used in machine learning competitions and hackathons. XGBoost has high predictive power and is almost 10 times faster than the other gradient boosting techniques. It also includes a variety of regular-

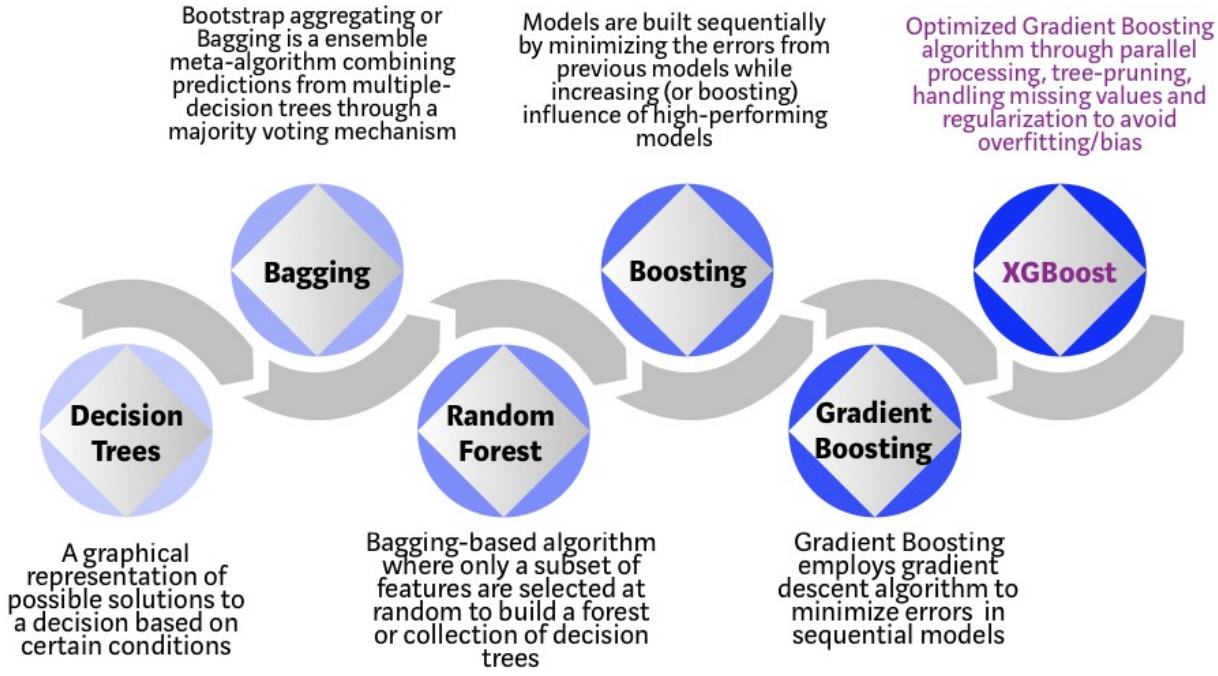


Figure 2.17 Evolution of XCBoost

ization which reduces overfitting and improves overall performance. Hence it is also known as 'regularized boosting' technique. Figures 2.17 and 2.18 illustrate the evolution of tree-based algorithms over the years, and the optimization techniques of XGBoost, respectively [69].

2.8 Feature Selection Methods

Feature selection (FS) is the process of selecting a subset of the most relevant features to help build a more accurate predictive analytics model [72]. In general terms, feature selection is also referred as variable selection or attribute selection. The main objectives of FS are to improve the prediction performance of the predictors, provide faster and more cost-effective predictors by eliminating the least important ones, and identifying the best combination of features that contribute the most to the target outcome. Also, the lesser the number of

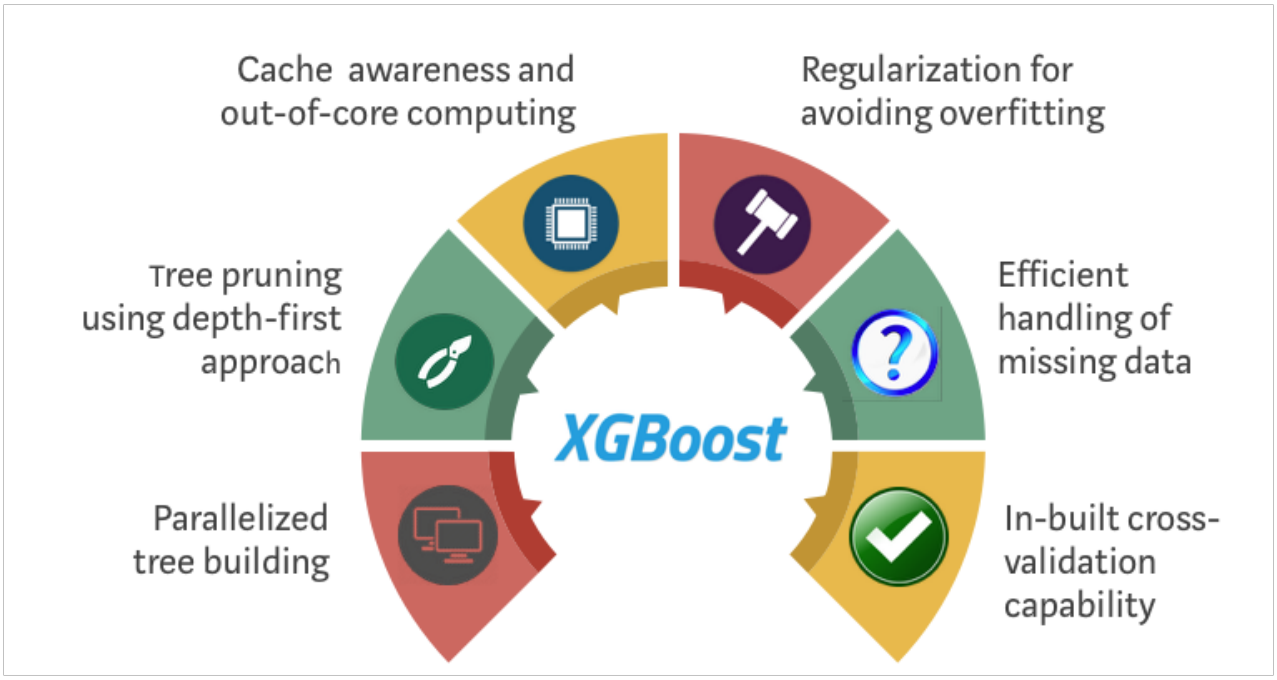


Figure 2.18 Features of XCBoost

features, the lesser the complexity of the model. Many research studies have shown that feature selection can be used to increase the classification performance [73-76].

There are three general categories of FS algorithms: filter methods, wrapper methods and embedded methods [77]. Filter methods employ a statistical measure to assign a score to each feature (for example: Chi squared test or correlation coefficient scores). The features are ranked by the score and either selected to be kept or removed from the model. Embedded methods learn which features best contribute to the model's accuracy while it is being created. Regularization algorithms (for instance, LASSO, Elastic Net and Ridge Regression) are the most common type of embedded FS methods.

Wrapper methods consider the selection of a set of features as a search problem, where

various combinations of subsets of features are prepared, evaluated and compared to each other. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy. An example of a wrapper method is the recursive feature elimination (RFE) algorithm, that our framework utilizes. The RFE technique evaluates the input feature set by recursively removing the redundant and the least important features, and builds a model on those features that remain. This method uses the model accuracy to identify which features (and combination of features) contribute the most to predicting the target outcome.

CHAPTER 3
REVIEW OF THE STATE-OF-THE-ART

In this chapter, we present a comprehensive review of the existing predictive analytics models that deal with time-to-event outcomes in healthcare. Furthermore, we review and discuss prior research studies that focused on the assessment of risk factors for short-term hospital readmission events following acute care medical procedures in general, and prediction of short-term readmission outcomes following cardiac surgery, in specific.

3.1 Statistical Models

In 2018, Deo et al. [25] reported development of a decision support tool, i.e., a Risk Calculator to predict 30-day readmission tool after CABG. This study uses statistical analysis methods and R language environment [26]. In this research paper, categorical data were presented as counts (percentages) while continuous data were presented as median (interquartile range). The study cohort was stratified into those who did and did not experience 30-day readmission. Categorical data was compared with the Chi square test while continuous data was analysed with the Wilcoxon rank-sum test or two-tailed 't' test. Their initial model incorporated 37 pre- and post-procedural variables. Subsequently, the risk calculator was developed using 16 preoperative and three postoperative factors. The final model was also developed with hierarchical regression using the hospital identifier as the random effects term. The results demonstrated a c-statistic of 0.65. The Center for Medicare Services presented a model with a c-statistic of 0.62 when applied to the development model and 0.63 when applied to the validation samples [27]. While the CMS model was limited to preoperative data, the study

of Deo et al. take into account important postoperative events that influence readmission.

Benuzillo et al. [28] proposed a model for 'Predicting readmission risk shortly after admission for CABG surgery'. This research considers the Society of Thoracic Surgeons (STS) database and the study cohort includes isolated CABG procedures performed in four Intermountain Healthcare hospitals. In this model, a total of 50 clinically relevant risk factors were considered, and grouped into demographic characteristics, prior histories, biometric data, lifestyle data, laboratory results, preoperative cardiac status, preoperative medications, hemodynamic data, and previous cardiac intervention data. The data were analyzed using Stata 13 statistical software (StataCorp LP, College Station, TX). The methods used were Chi-square test (or the Fisher exact test when appropriate) for categorical variables and the two-sample t-tests (or the Wilcoxon rank sum test when appropriate) for continuous variables in bivariate analysis. Predictors with a $P < 0.05$ were included in a multivariable logistic regression model and removed stepwise, using backward elimination procedures.The Hosmer-Lemeshow test was used to evaluate model calibration.The area under the curve c-statistic obtained from the derivation sample was 0.63, 0.65 in the validation sample. In addition, the model was shown to calibrate well according to the Hosmer-Lemeshow goodness of fit test 9.31 ($P = 0.32$). However, the predictive power of the model presented here was limited by the exclusion of postoperative risk factors such as renal failure or postoperative infections, as it considers a small of number of commonly available electronic data. Zywot et al. [29] proposed a 'Preoperative Scale to Determine All-Cause Readmission after CABG', and this model uses, STATA, version 12, statistical software (StataCorp LP, College Station, TX) for

data analysis. Statistical analysis of categorical variables was performed using Chi square tests and cohort comparison was performed using t-tests and ANOVA. Predictive factors were added in a stepwise manner to a mulitivariable logistic regression and odds ratios (OR) and 95% confidence intervals (CI) were calculated. They devised 100-point readmission risk score based only on preoperative factors using the State Inpatient Database (SID) for New York, California, Washington and Florida. While their model has a higher c-statistic of 0.702, Deo's [25] model appears more intuitive to use in regular clinical practice. Zywor et al. presented a 100-point risk scale without specifically providing data of how to convert this into a clear probability for the individual patient , while the model of Deo et al. gives a simpler comparison of total score and probability. However, the findings of Zywot's model are similar to those of previous studies on readmission rates among post-CABG patients.

Feng et al. [30] carried out a retrospective study on CABG readmission rates and risk factors, with data drawn from California, Florida, and New York State Inpatient Databases. In this study, statistical tests and analyses were performed using SAS version 9.3 (SAS Institute, Cary, North Carolina). Fisher's exact test or Chi-square analysis was used to test statistical significance for categorical variables, including insurance status, readmission rates, readmission diagnoses, and demographic data. Continuous variables were analyzed with analysis of variance (ANOVA) or Kruskal-Wallis tests, for non-normally distributed variables, as appropriate. All p-values were two-sided with statistical significance evaluated at an alpha level of 0.05. mulitivariable logistic regression analyses were used to estimate odds ratios (OR) with 95% confidence intervals for 30- and 90-day readmissions for each

insurance cohort (compared to private insurance), while controlling for other demographic variables, comorbidities, postoperative complications, hospital LOS, discharge disposition status, and other potential confounders. According to this model, independent risk factors for readmission within both 30 days and 90 days after discharge from CABG are age, female sex, black or Hispanic race, low median household income (based on zip code), discharge with home health care, non-routine transfers, and length of stay (LOS). Additionally, the presence of certain complications during the initial postoperative hospitalization, including cardiovascular, intraoperative, and pulmonary complications, were associated with higher risk of readmission at both 30- and 90-day intervals. The results of this study are in consistent with previous studies that have demonstrated readmission rates ranging from 8% to 21%. [31, 32, 33, 43]. The limitation of the study is its use of an administrative dataset, which may lack relevant clinical information or intraoperative data that could change the comorbidity burden of given patients or explain adverse perioperative outcomes. Follow-up data was not available beyond the readmission findings that they presented; the database was also limited as it can only trace patients readmitted to the hospital within the same state.

In 2017, Kilic et al. [34] developed and validated a risk score for readmission after ault cardiac operations. Data analyses in their model were performed with STATA, version 11, statistical software (StataCorp LP, College Station, TX). This model also uses student's t test for continuous variables, Chi-square test for categorical variables. The inclusion of each variable into the model was evaluated in a stepwise fashion using the likelihood ratio test, with a p value less than 0.05 considered significant. The predictive capability of the

composite risk score in evaluating the risk of 30-day readmission was then evaluated in the validation cohort using logistic regression analysis, the Akaike information criterion, the Chi-square test, the c-index, and the Hosmer-Lemeshow goodness-of-fit test. The model showed predictive capability in the validation cohort, with an Akaike information criterion of 795, c2 result of 40 (p ¡ 0 .001), C-index of 0.64, and nonsignificant Hosmer-Lemeshow test result of 0.57. multivariable logistic regression analysis of this study identifies six predictors as the most significant, and they are chronic lung disease (CLD), race, insurance type (Private, Medicare, Medicaid, Other), type of operation (CABG, Valve, CABG- Valve, Aortic, other), length of stay in hospital, postoperative acute renal failure. The major limitation of this study is that readmissions to other hospitals were not captured in the analysis. Another limitation was that other variables that may be associated with 30-day readmission were not available in their database. This includes variables such as proximity to the hospital, a social support system, and medication compliance. Furthermore, readmissions were captured only if they occurred within 30 days, although the incidence of specific causes of readmission may peak at later times.

In 2016, Espinoza et al. [35] studied about '30-day readmission score after cardiac surgery'. In their model, statistical analyses were performed using SPSS version 21 (IBM® SPSS ® Statistics). univariable analysis used Chi-square or Fisher's exact, while multivariable model analysis was done through stepwise logistic regression. This model demonstrated the area under the curve (AUC) 0.663, and 0.639 for the test and validation data sets, respectively. Although, this model includes a large cohort of patients that is well represented

by differences in gender, age, and all types of cardiac surgical procedures, the data are derived from a single center. The model needs to be tested prospectively at multiple centers to substantiate its broad applicability. Furthermore, only 2 out of 5 risk factors included in this scoring model are considered to be 'modifiable risk factors' (preoperative anemia and peak glycemic serum level).

Swaminathan et al. [36] studied on ' Gender Differences in Hospital Outcomes after CABG' In tius model, all analyses were conducted using SAS, version 9.2 (SAS Institute, Cary, North Carolina) and SPSS, version 20 (IBM corporation, Armonk, New York). In univariable analysis, Chi-squared and independent t-test were used for categorical and continuous variables, respectively. multiivariable logistic regression analysis was used to adjust for baseline differences between male and female groups by adjusting for univariable predictors of examined outcomes ($p < 0.01$). Their findings reported that women have worse in-hospital outcomes than men; however, the gender gap is slowly closing. Fanari et al. [37] tried 'predicting readmission risk following CABG at the time of admission'. In their model, Fractional Polynomial (FP) Regression (FPR) was used to assess nonlinearity of continuous variables, while Hierarchical Logistic Regression (HLR) was used to model readmissions (a patient may have had more than one). Derivation models were developed by a combination of forward selection and backward elimination of variables.They developed 3 models, one at admission and two at discharge. The c-indices were 0.673, 0.7, and 0.714 for the admission, discharge and STS models respectively. Internal validation of the models demonstrated c-index of 0.641, 0.659 and 0.670 respectively. Although these models showed that clinical

risk factors could be predictive of CABG readmission, they also had some limitations. First, none of them added previous utilization as an important potential predictor of readmission. Second, they all used registry data that would need abstracting and that may be subjective to inaccurate coding and/ or not inclusive of all comorbidities and complications that may impact and predict readmission.

Shahian et al. [38] linked the Society of Thoracic Surgeons data with Medicare readmissions claims to devise a 30-day readmission model. They found dialysis, increased creatinine levels, severe chronic lung disease, preoperative atrial fibrillation, insulin dependent diabetes mellitus, female gender, immunosuppressive therapy,recent myocardial infarction,low body surface area in men and obesity in women, as factors associated with higher likelihood of 30-day readmission after CABG. The c-statistic of their model was 0.648. The final list of covariates selected by the surgeon panel included all covariates that were either (1) selected at the 0.05 level in the original full sample for at least 1 calendar year; or (2) were selected in at least 50% of bootstrap replicates at the 0.05 level in at least 1 calendar. Logistic regression was used to identify readmission risk factors. To estimate hospital-specific performance, the selected covariates were entered into a hierarchical logistic regression model with hospital-specific random intercept parameters. These hierarchical models were used to estimate hospital-specific risk-standardized readmission rates (RSRRs) using methodologies identical to the existing CMS readmission measures [39-41]. The data for this model was drawn STS National Database linked to Medicare claims records. However, not all Medicare CABG admissions could be linked to an STS record, and some Medicare admissions were,

by design, excluded from analysis (eg, unknown discharge date). Similarly, some Medicare hospitals were excluded because of extremely low CABG volumes, possibly indicating they were not actually CABG providers.

Maniar et al. [42] performed 'prospective evaluation of patients readmitted after cardiac surgery and analyzed outcomes to identify risk factors'. The analyses of continuous and categorical variables were performed using t tests, and chi-square tests respectively. Through SAS software, version 9.3 (SAS Institute Inc, Cary, NC), multiivariable Logistic regression analysis identified independent risk factors for readmission. This study considers patients patients underwent cardiac surgical procedures at a single tertiary care institution. This single-center experience was limited by sample size, precluding the use of more sophisticated matching between study patients and control patients, and required the use of 2 multiivariable logistic regression models to allow for testing of statistically and clinically significant covariates identified by univariable algorithms. Price et al. [43] attempted 'Risk Analysis for Readmission after CABG' and proposed a strategy to reduce readmissions. This model considers data that were extracted from New York State Cardiac Surgery Reporting System (CSRS), and the STS Adult Cardiac Surgery Database. Univariable screening was performed using chi-square or independent t-tests. Multivariable logistic regression was then also used to further evaluate the subgroup of variables meeting univariable statistical significance ($p < 0.05$). The results demonstrate a c-index of 0.651. However, given the small sample considered in this model, this study may have likely been underpowered to uncover all risk factors and limited in its ability to perform extensive multiivariable analysis.

Currie et al. [44] proposed a 'predictive model for readmission within 30 days after CABG', and used patient-specific perioperative parameters for individual risk assessment for early readmission. Chi-squared and independent t-test were used for categorical and continuous variables, respectively. A logistic regression was performed with variables having tests with p-value 0.02. Hannan and colleagues [45] performed analysis of more than 33,000 patients undergoing CABG in New York state, demonstrated 30-day readmission rates ranging from 8.3% to 21.1% across institutions, and found that the most common causes of readmission were infection and heart failure. The model uses stepwise logistic regression, and all analyses were performed with SAS software version 8.2. Zhongmin Li et al. [56] also retrospectively analyzed 30-day readmissions among 11,823 patients discharged alive after isolated CABG in 2009 in the State of California and found risk factors similar to Hannan's previous findings [45]. But, they also identified other risk factors such as postoperative atrial fibrillation and recent myocardial infarction, both of which support Espinoza's findings [35].

This study reported c-statistic of 0.65. This is not uncommon for models that predict complications or readmissions, but it means that unavailable patient-level predictors of readmission could have improved the discrimination and/or that unmeasured hospital quality of care/process measures would have substantially improved the ability to predict readmissions.

In 2000, Stewart et al. [46] attempted studying about 'Predictors of 30-Day Hospital Readmission After Coronary Artery Bypass', with data drawn from a single institution. All patients of the final cohort (n=485) were personally contacted by telephone on the 30th postoperative day to determined if patient had been readmitted. Statistical analysis was

performed using STATA software.. Continuous variables were compared using student's t tests. Categorical or dichotomous variables were compared using Chi-square or Fisher's exact tests. multiivariable logistic regression was performed with all variables included in the model to control for potential confounding or effect modification. The most significant finding of this study was that women were readmitted more than twice as often as men within 30 days of CABG. According to this study, the most common reasons for readmission were atrial fibrillation, leg wound infection, followed by congestive heart failure or angina.

Kansagara et al [47] carried out a systematic review of 'Risk Prediction models for Hospital Readmission' in 2011. Their study reviewed 7843 citations and found ound 26 readmission risk prediction models of medical patients tested in a variety of settings and populations. The most common outcome used was 30-day readmission. Half the models were largely designed to facilitate calculation of risk-standardized readmission rates for hospital comparison purposes. The other half were clinical models that could be used to identify high-risk patients for whom a transitional care intervention might be appropriate. Of these, nine were tested in large US populations and had poor discriminative ability (c-statistics 0.55 – 0.65). Seven models could potentially be used to identify high-risk patients for intervention early during a hospitalization (c-statistics 0.56 – 0.72), and five could be used at hospital discharge (c-statistics 0.68 – 0.83). Six studies compared different models in the same population and two of these found that functional and social variables improved model discrimination. Though most models incorporated medical comorbidity and prior utilization variables, few examined variables associated with overall health and function, illness severity, or social

determinants of health. Kangara's review concluded that most readmission risk prediction models, whether designed for comparative or clinical purposes, perform poorly. Though in certain settings such models may prove useful, efforts to improve their performance are needed as use becomes more widespread.

3.2 Survival Regression Models

One of the survival models that is most commonly used in the statistical and medical research literature is the Kaplan-Meier estimator [78], which has the advantage of being able to learn very flexible survival curves, but the disadvantage of not incorporating patients' covariates. Hence it is useful at the population level but not useful at the individual level. As we have noted already the Cox proportional hazard model [16] is capable of incorporating patients' covariates, but assumes that the hazard rate is constant and that the log of the hazard rate is a linear function of covariates. Other models make different assumptions about the underlying stochastic processes and about the relationship between the covariates and the parameters of the assumed process. For instance, Lee et al. and Doskum et al. [79, 80] assume a Wiener process, while [81] assumes a Markov Chain.. A merit of these models is that, because they formulate survival analysis as the problem of determining the distribution of the first time at which the prescribed stochastic process hits a prescribed boundary, they are able to incorporate competing risks. The drawback of these models is that they are tied to the specific form of stochastic process that they assume. The models might be of limited use unless we have already learned the underlying stochastic process. In the healthcare

setting this means learning the underlying disease process, which would seem to be an even more complicated problem than survival analysis itself - especially since the states of the disease or diseases are typically hidden and not directly observable. Alternatively, Fine and Gray [82] modified the traditional proportional hazard model by direct transformation of the cumulative incidence function, but this model is also severely limited by strong assumptions on the form of the hazard rates and on the way in which the parameters depend on covariates.

Recently, the survival analysis problem has also received significant interest in the machine learning literature. These include random survival forests by IIshwaran et al [83], deep exponential families by R. Ranganath and Blei [84], dependent logistic regressors by Yu et al. [85], and semi-parametric Bayesian models based on Gaussian processes by Fernndez et al.[86]. These methods are capable of incorporating the individual patient's covariates, but none of them appears readily adaptable to the problem of competing risks. Recently, Alaa and van der Schaar [87] used deep multi-task Gaussian process to develop a nonparametric Bayesian model for survival analysis with competing risks, but their model still relies on assumption that the latent stochastic process follows Gaussian process.

3.3 Neural Network Models

Faraggi and Simon [88] developed the first application of neural networks to survival analysis. In contrast to the standard CPH model, this work uses a feed-forward network to learn the relationship of the covariates to the hazard function. Recently, Katzman et al. [89, 94] and Luck et al. [90] have followed a similar general approach, using more sophisticated

network architectures and loss functions. These works have improved on the CPH model by relaxing the specific functional relationship between covariates and the hazard function in the standard CPH model while maintaining the other central assumption - that the hazard rate is constant over time. Consequently, these studies do not fully utilize the potential capacity of deep neural networks to learn complex representations of risk and in particular to capture the time-dependent influence of covariates on survival. In the other research, Lee et al. [91] introduced a different approach called DeepHit which employs deep architecture to estimate the survival times distribution. They used neural network including two types of sub-networks: (1) a single shared sub- network and (2) family of cause-specific sub-networks. They evaluated their method based on real and synthetic datasets which illustrate that DeepHit leads to better performance in comparison with state of the art methods. Although these survival models developed by using deep learning are well suited for high-dimensional survival data, they are not the best choice when labeled instances are scarce, it seems more effort s should be accomplished to improve them in such situations. The other drawback of these deep learning based survival model is related to interpretability. Majority of the studies discussed above did not provide interpretable framework for treatment recommendation or survival risk analysis while their proposed methods were based on deep representation of original features (risk factors) through multiple non-linear transformations [93-98].

In the context of our research area, sophisticated statistical prediction models (classifiers) have been applied to CABG outcomes readmission risk prediction problem, ranging from simple univariable analysis to multiivariable logistic regression and Bayesian statistics.

However, most regression models require statistical assumptions (eg, linearity, additivity, distributional), which may not be justified [48], and the management of missing data is problematic. Bayesian models assume that prediction variables are independent and also require categorical data that typically can assume only two values. However, they do not require iterative training and easily accommodate missing features. Each of these methods has inherent limitations when applied to a complex biological process, and a high degree of predictive accuracy has yet to be achieved. Neural networks are a form of artificial intelligence that may obviate some of the problems associated with traditional statistical techniques, and it has been asserted by Steen et al. [49] that they will represent the next major advance in predictive modeling. Previously, Lippman et al. described the results of pilot studies of CABG risk prediction from databases of approximately 1,000 and 40,000 patients and a limited set of variables [50,51].

To the best of our knowledge, the applicability of multilayer perceptron neural networks (MLP) to coronary artery bypass grafting risk prediction was first assessed using the STS database of 80,606 patients who underwent coronary artery bypass grafting in 1993 by Lippman et al [52]. In this study, the results of traditional logistic regression and Bayesian analysis were compared with single-layer (no hidden layer), two-layer (one hidden layer), and three-layer (two hidden layer) MLP neural networks. These networks were trained using stochastic gradient descent with early stopping. A committee classifier combining the best neural network and logistic regression provided the best model calibration in the above study. The Receiver Operating Characteristic (ROC) curve areas for predicting mortality

were approximately 76% for all classifiers, including neural networks. Calibration (accuracy of posterior probability prediction) was slightly better with a two-member committee classifier that averaged the outputs of a MLP network and a logistic regression model. Unlike the individual methods, the committee classifier did not overestimate or underestimate risk for high-risk patients.

Despite optimism that artificial intelligence techniques might be the next major advance in risk prediction for coronary bypass, results in this analysis of more than 80,000 patients from the above model confirms the suspicion of many investigators that all prediction systems have inherent limitations [53-55]. Neural networks alone failed to improve upon the ROC curve area of logistic regression or Bayesian analysis, suggesting an absence of complex nonlinear relationships, at least among the variables presented to the network.

NNs are being used in the areas of prediction and classification of outcomes in medicine - areas where regression models have traditionally been used. In most applications, outcomes (termed outputs by users of NNs) are characterized by the presence or absence of an event. In this situation, a NN is regarded as an alternative to traditional logistic regression methods. However, in the situation where outcomes are characterized by the time to an event, the application of NNs to predict clinical events requires development of a strategy to address the time course of a disease process. In cases where the event of interest does not occur, outcomes are regarded as (right)-censored. Simple exclusion of censored observations from the available training set would limit the amount of data available for network development and could lead to significant biases in event predictions. Several strategies have been developed to

extend NN prediction methods to accommodate right-censored data. These are methods due to Faraggi and Simon (57), Liestol et al. (58), and a modification of the Buckley-James method (59). Because Cox regression analysis is an accepted solution to the problem of analyzing censored data, the performance of Cox regression models, when compared to those of NNs, can provide a useful perspective on the utility of a NN approach.

Xiang et al. [92] compared the performance of neural network methods and Cox regression for censored survival data. The results of Xiang's study suggested that NNs could serve as effective methods for modeling right-censored data. However, the performance of the NN is somewhat variable, depending on the method that is used. The limitation of their study was the use of simulated data.

With the advent of big data in biomedical and health field in recent times, Wang et al [61] presented the scope of applications of deep learning in biomedical informatics. In the recent past, EHRs were analyzed using traditional machine learning techniques, whereas recently the progress in the field of deep learning let to the application of deep learning techniques to EHRs [62]. Electronic Health Records (EHR) are the rich source of patient information, including both structured and unstructured data, with various sources such as, diagnosis, laboratory test results, medications, procedures, and other clinical data. DL obtains better results than conventional Machine Learning (ML) models applied in processing EHRs. DL have the capability to learn complex patterns, extract data-driven features, and high-dimensional data, and is becoming popular in biomedical data analysis. As more data becomes available, DL systems can evolve and deliver where human interpretation is difficult.

NN software has generated a great deal of interest partly because it effectively places advanced modeling tools in the hands of users of personal computers. One of the advantages of NNs is that they can detect complex patterns among the inputs. In contrast, one disadvantage of NN models is that they can model idiosyncratic features of the training dataset. When this happens the NN has overlearned, and will appear to perform extremely well on the training dataset but further testing on new data will generally be far less successful. In fact, overlearning was apparent in our testing datasets. For example, comparing training to testing results for Design 5, the Faraggi-Simon, Liestol-Andersen-Anderson and Buckley-James methods [57-59] declined by 0.041, 0.079 and 0.019, respectively. Overfitting can be reduced by using the m-item out validation approach, which is time intensive (Lachenbruch and Mickey, 1968).

Another disadvantage is that NNs are 'computationally intensive' and may take a long time to train and converge to a solution. Using the quasi-Newton minimization algorithm, the Liestol-Andersen-Andersen and Buckley-James methods converged to a solution very fast. In contrast, the Faraggi-Simon method, which used the simple gradient minimization algorithm, took a long time to train. Additionally, the NN may converge not to the optimal solution, but rather to a local minimum, so that the resulting NN will perform sub-optimally. Some methods such as genetic algorithms can be used to avoid local minima. However, the genetic algorithm requires a lot of computer memory and is slow in convergence.

One advantage of Cox regression, is that the regression coefficients can be interpreted as the likelihood of an outcome given some value(s) of the risk factor (e.g.,odds ratios or relative

risks). NN weights usually do not lend themselves to such interpretation. The NN methods presented here represent only a few of the many options in survival modeling. It may be more valuable to explore issues of methods, applications, and potential for improvement than to draw conclusions about whether one method is inherently 'superior' to another.

In this research, we propose a new scalable, portable and memory-efficient predictive analytics framework, using four components: feature engineering, time-to-event (survival) analysis, feature selection, and ensemble learning algorithms. The motivation for this study comes from literature gap and application needs in several domains (e.g., healthcare, manufacturing and finance) in dealing with time-to-event outcomes.

CHAPTER 4

A NEW SCALABLE, PORTABLE AND MEMORY-EFFICIENT PREDICTIVE ANALYTICS FRAMEWORK FOR PREDICTING TIME-TO-EVENT OUTCOMES IN HEALTHCARE

In this chapter, we describe our proposed model, a newly constructed scalable, portable, and memory-efficient predictive analytics framework that requires minimal resources, computationally as well as memory-wise. The primary objectives of this new ensemble learning based model are to evaluate the predictive ability more accurately with effective feature engineering and feature selection techniques and to improve performance by using less features compared to state-of-the-art models. Our framework comprises of four components (feature engineering, survival analysis methods, feature selection techniques, and ensemble learning methods) working together to produce a more robust and accurate predictive analytics model, as illustrated in Figure 4.1.

In this research study, our new contributions include i) design and development of more appropriate feature engineering rules to process the unstructured input feature set of medical data, ii) incorporation of time-dependent perioperative granular data points, i.e, multiple rows of records for each subject, comprising of both the time-fixed and time-dependent health records of patient population iii) application of better automatic feature selection techniques, and iv) parameter tuning and optimization of the ensemble learning algorithms - in order to minimize overlearning issues on the training data and improve predictive performance when tested with new data. In other words, in the proposed framework, we attempted to improve performance by a four-step process, i.e., more granular data, automatic feature selection,

algorithm tuning, and ensemble methods. While this model's approach can be extended to any general application area of time-to-event outcomes (for example, a device failure, a project success or a student dropout), we focus on a healthcare application set-up here to explain all the modules of the framework in the following sections.

4.1 The Conceptual Framework

Our proposed framework employed the technique of ensemble learning methods, i.e., combining the decisions from multiple models to improve performance. As shown in Figure 4.2, the first component, feature engineering, formed the foundational basis as it handled data extraction and processing tasks. The second component encompassed the selection of the best possible feature set that influence the outcome, and this role was handled by Cox Proportional Hazard (CPH) survival analysis model, and an advanced feature selection technique, 'Recursive Feature Elimniation (RFE). The third and fourth components, model training and validation were handled by an appropriately tuned and optimized ensemble learning method, 'eXtreme Gradient Boosting'(XGBoost). This framework analyzes the entire input feature set, evaluates the best possible feature set and processes the most significant feature vectors relevant to the outcome and generates better predictive performance metrics. The components of the framework are described in the following sections in more detail.

The model first processes a consolidated raw dataset or feature set (for instance, electronic health records of cardiac surgery patients), and employs imputation and one-hot encoding processes to handle missing and categorical variables, using XGBoost libraries.

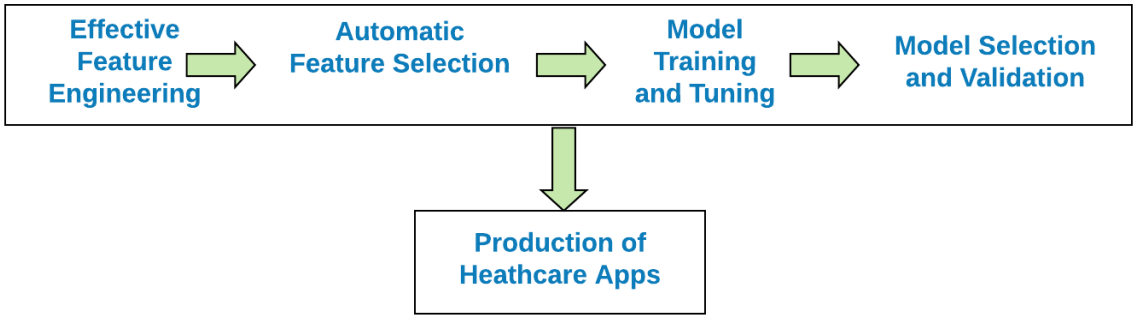


Figure 4.1 Workflow of the Proposed Framework

Feature engineering techniques, such as feature binarization, discretization, transformation, normalization, and weighting are then applied to form the informative features from the raw dataset. This whole feature set is then fed to the Cox survival regression model that produces highly correlated features relevant to the outcome, using univariable and multivariable analyses. To further bring down the number of redundant or irrelevant features, we employed RFE, an automatic feature selection technique that considers smaller and smaller sets of features recursively, eliminates the least important ones, and identifies the best combination of features that contribute the most to the target outcome.

The resultant feature set was deployed to XGBoost. XGBoost uses a gradient boosting decision tree algorithm that employs a boosting technique in which new models are created sequentially that predict the residuals or errors of prior models, which are then added together to make the final prediction. Within the XGBoost, we first load the selected feature set, prepare data, train the model with training data set; run k-fold cross-validation with subsets of data (for k=3,5,and 10), perform parameter tuning to reduce overlearning; validate the model with validation data set; and then collect the evaluation metrics, such as,

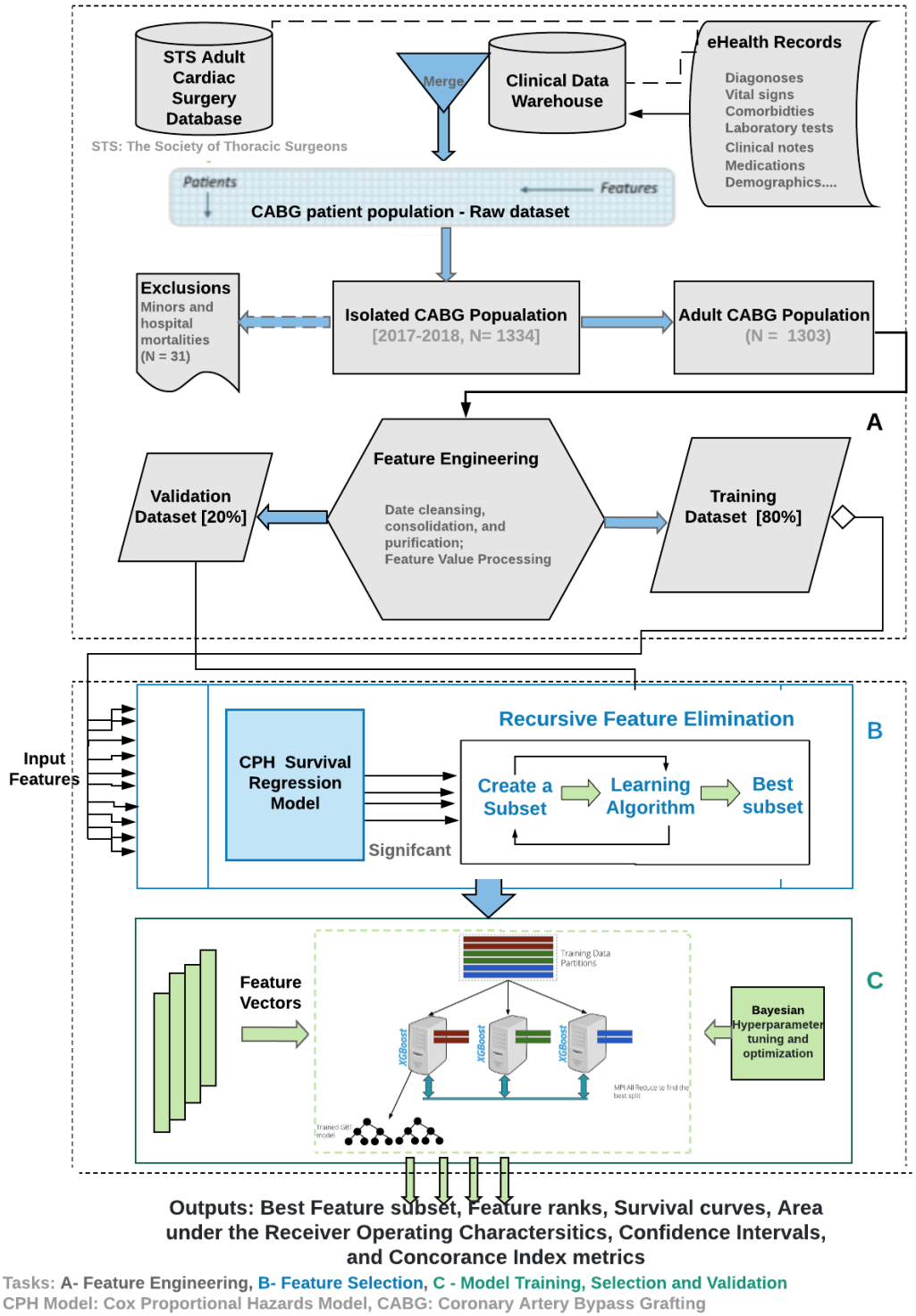


Figure 4.2 Proposed Ensemble Learning Framework

area under the receiver operating characteristics (AUROC), ro concordance index metrics, and logarithmic loss plots for both the training and validation data sets, and survival curves for the most significant and highly correlated features that impact the likelihood of the target event.

4.2 Feature Engineering and Feature Selection

Feature engineering is the act of injecting knowledge into a machine learning model. Through this component, we carried out the tasks, such as, removal of unnecessary and redundant features from the raw data set; inclusion of new features by combining the existing ones; transformation of features, modification of features and feature values. These processes are explained in the subsequent subsections.

4.2.1 Data Sources and Study Cohort

As shown in Figure 4.2, the primary database we used in this study was Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database, pertaining to a multi-hospital academic health system. Also, we utilized institution's Clinical Data Warehouse (CDW) as a secondary data source in order to cross-refer and fill-in any missing data values as well as to extract additional data elements (that were not usually recorded in STS database) such as, time-dependent perioperative variables, insurance provider data as well. The study cohort included all patients who were discharged alive following isolated CABG, and the number of initial subjects was 1334. Minors (of $<$ 18 years old) and in-hospital mortalities that occurred during the index admission were excluded, and 1303 patients were included for final

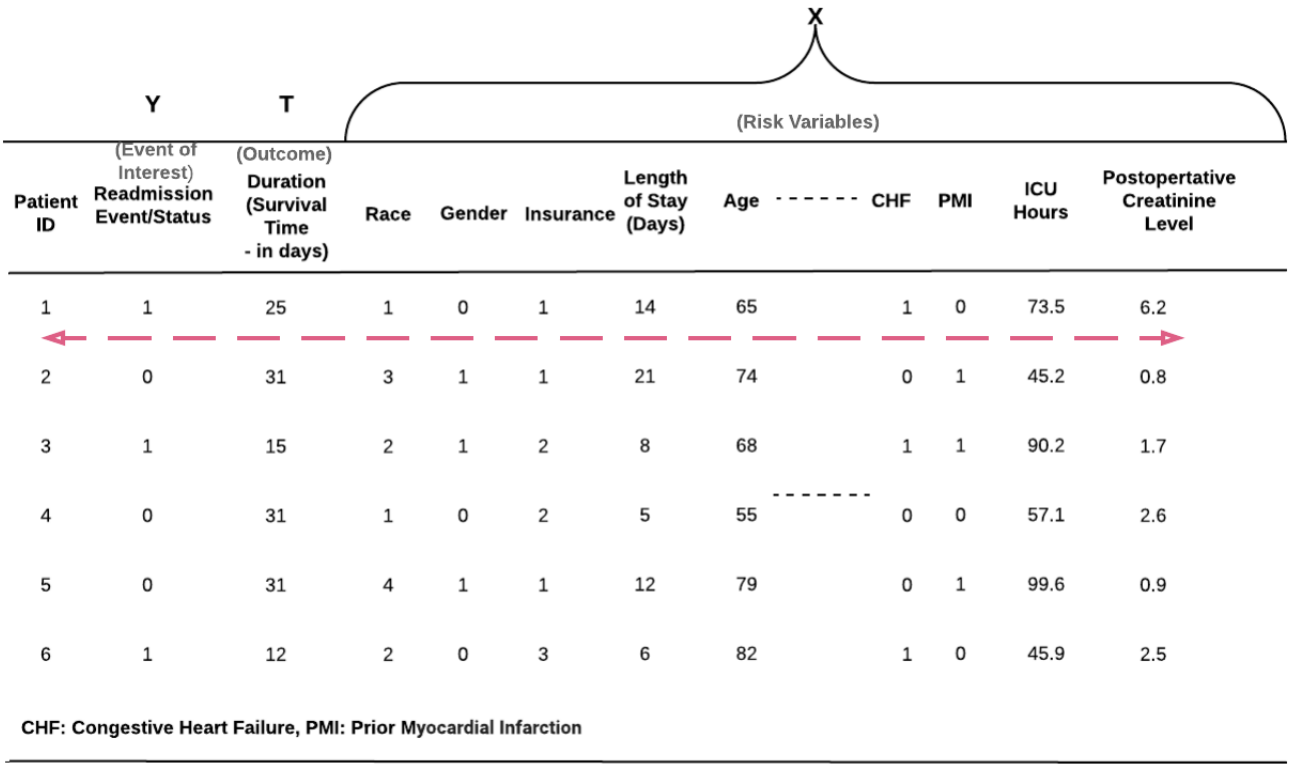


Figure 4.3 Illustration of Data Feed Formatting in Prior Models

analysis. The entire study cohort was then divided randomly into training and validation sets with a 80:20 ratio.

4.2.2 Study Features and Outcomes

For our experiments, one of the primary events of interest that we considered was the same hospital all-cause readmission within 30 days of discharge following CABG. The initial list of variables were either (1) selected at a p value less than 0.05 level in the original full sample,

or (2) added in accordance with a well-established list of risk factors associated with prior studies [28 - 32], or 3) recommended by surgeons. As listed in Tables 1-4 in Appendix A, this study included 82 potential readmission risk factors, including patient demographics (such as age, race, gender); preoperative comorbidities (diabetes, congestive heart failure, prior myocardial infarction); pre- and post-operative vital signs (blood pressure, heart rate); laboratory values (white blood cell count, glucose, hemoglobin, creatinine), intraoperative covariates (number of diseased coronary vessel systems, operative approach, Cardiopulmonary Bypass Utilization, total time in operation room), and postoperative events (such as atrial fibrillation, pneumonia, renal failure, reoperation for bleeding, prolonged ventilation, new dialysis). The notable addition in this study was the inclusion of granular data points of time-varying covariates, as illustrated in Figures 4.3 and 4.4. For each patient in the cohort, multiple readings of preoperative laboratory values (such as creatinine, glucose, hemoglobin A1c, hematocrit, white blood cell count for preoperative day 1, day 2) within 48 hours prior to CABG operation were considered for analysis. Similarly, multiple recordings of postoperative laboratory values (such as creatinine, glucose, hemoglobin, hematocrit for postoperative day 1, day 2, day 3), and vital signs (such as blood pressure, heart rate for postoperative day 1, day 2, day 3) within 72 hours post-CABG were considered in the analysis.

Not every individual has a full set of all of the measurements recorded in the STS primary database due to version-upgrades or data-entry misses. So, we cross-referred institution's CDW for missing and/or additional data values and filled-in data where appropriate, including CDW-specific information, such as, prior heart failure, preoperative chronic lung disease,

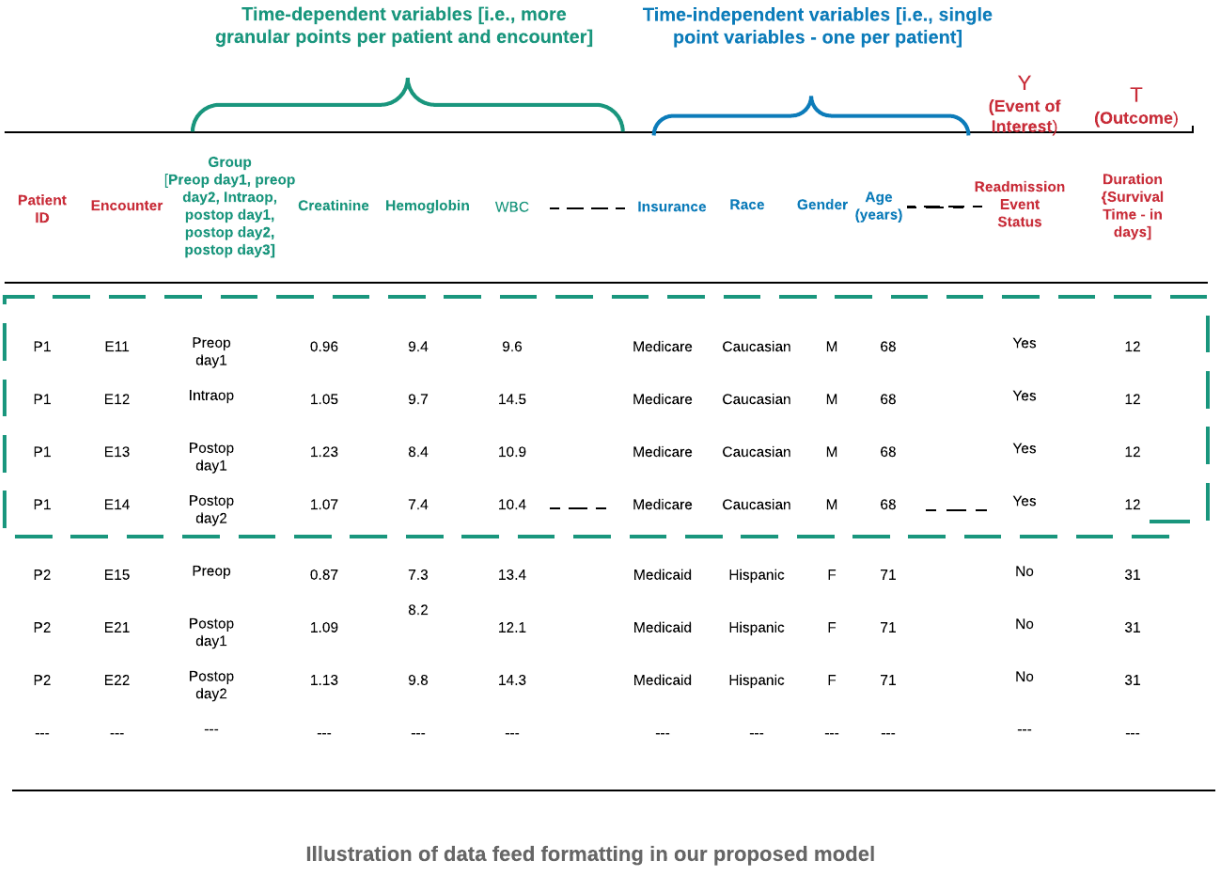


Figure 4.4 Illustration of Data Feed Formatting in the Proposed Framework

preoperative insurance type (Medicare, Medicaid, Private, and Other). This process led to minimize the amount of missing data to less than 5% in the overall study cohort.

4.2.3 Feature Data Extraction and Purification

One of the most valuable parts of machine learning is predictive modeling, i.e., development of models that get trained with historical data to make predictions on new data. So, we first tried to create new and different perspectives on our data in order to best expose the

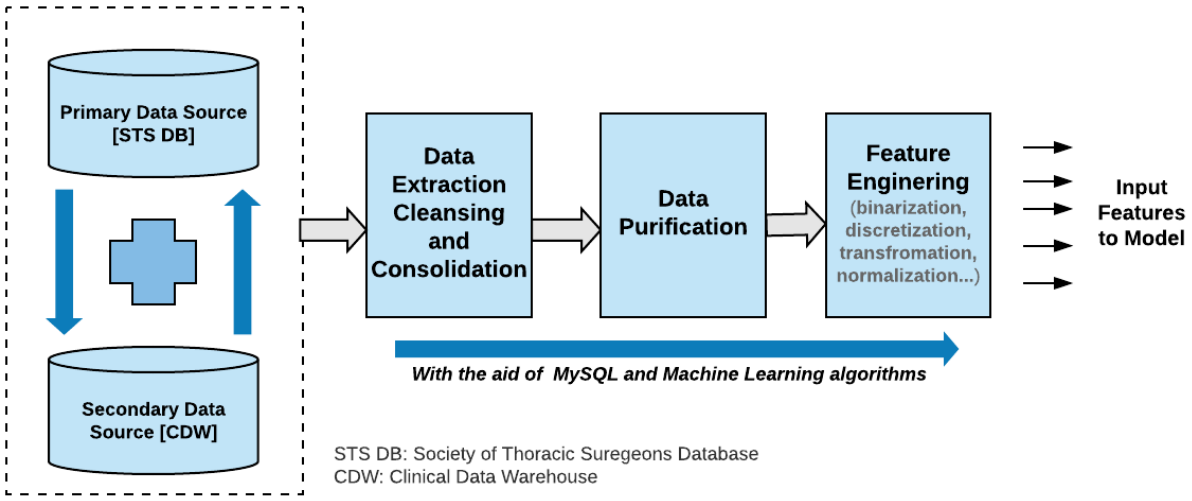


Figure 4.5 Feature Engineering Process in Ensemble Learning Framework

structure of the underlying problem to the learning algorithms - and thus, help improve performance. A five step process of data integration was performed which included merging, cleansing, consolidation, normalization, and rescaling in order to fit the data well into our proposed framework, as depicted in Figure 4.5. Below is the list of algorithmic steps that we devised for the above processes. These tasks were performed with the aid of Structured Query Language (SQL) scripts, using an open source MySQL workbench software [65] and Python language [66].

Algorithm 1: Data pre-processing

1. Extract the data set of patients who underwent isolated CABG procedures during the study period, from the primary data source,, i.e., STS database.

2. Link and merge the dataset with the secondary data source, CDW for additional data

elements, time-dependent variables and granualr data poitns for each subject of the study (Table 4 in Appendix A).

3. De-identify the data set, excluding the patient-specific protected health information (PHI)

4. Delete non-relevant subjects, variables, based on the criteria described in study cohort section.

5. Rescale and/or normalize categorical values in order to fit well into model (Figure 4.4 and Table 1-4 in Appendix A]

6. Handle missing data for columns with < 5% data missing [by applying missing data algorithm described in section 4.2.4]

7. Apply XGBoost's one hot encode to categorical data

    - Encode string output values as integers for classification [through 'Label Encoder' string class of XGBoost]
    - Map integer values into binary variables

4.2.4 Handling of Missing Data for Selected Feature Set

Handling missing data is important as many machine learning algorithms do not support data with missing values. Accurate estimation of the missing values is a critical component in data analysis [70-71]. In our study population, missing data mostly occurred due to unrecorded or unobserved instances in STS database. This kind of missing data falls into

'Missing Completely At Random' (MCAR) category [67]. However, after we applied variable selection criteria to the whole data set, the missing data was very minimal. We applied below ML algorithm to handle missing data, because ML imputation methods outperform statistical methods in the prediction outcome [68]. Finally, all the variables included in the CPH survival regression model were complete. Hence, only complete cases or subjects were included in the models we developed. Also, we have not removed any rows with missing data, because it might change the distribution of actual data. We used below two step-processes and algorithms for handling missing data. We customized sparsity-aware spilt finding handled in XGBoost [69].

Purification by skipping data

- There are two kinds of skipping that we might want to do when we have missing data.
- We can either skip data points that have missing data or skip features that have missing data.
- Somehow we have to make a decision of whether to skip a data point, skip features, or skip some data points and some features and that's a kind of complicated decision to make.
- **Idea 1:** Skip data feature where any feature contains a missing value [make sure only a few data points are skipped]
- **Idea 2:** Skip an entire feature if it's missing for many data points [make sure only a few unnecessary or redundant features are skipped]

Purification by imputing data

- Simple imputation may introduce bias. So, we designed an algorithm, as described below. This customized approach adapts to missing data at hand and makes it robust for better outcomes from the study.

Algorithm 2 for handling missing data

- Determine the amount of missing data for each column/data field.
- If missing data is > 5%, apply case deletion of data columns
- Else, apply multiple imputation through XGBoost
- For categorical variables, use the most frequent and 'mode' values
- For continuous variables - apply imputer class to impute with mean/median, based on the nature of data fields [with XGBoost modules
- Split data into training and validation sets randomly in 80:20 ratio [using either XGBoost methods or SQL scripts]

4.2.5 Feature Processing

Feature processing of input data can result in a lift in performance on algorithms that use weighted inputs or distance measures. We performed feature processing techniques, such as, feature binarization, discretization, transformation, normalization on input data. A traditional rule of thumb when working with ML methods is to rescale feature values to

the bounds of Objective/activation functions. For instance, if we use sigmoid activation functions, it helps rescaling data to values between 0 and 1; for the Hyperbolic Tangent (tanh), values between -1 and 1 would help. So, based on LR model and CPH survival regression, as we used sigmoid function, we rescaled and normalized feature values for most of the categorical variables [that have arbitrarily assigned values at the source database]. Figure 4.6 provides the list of feature engineering rules that we implemented in this study.

4.3 Model Derivation and Evaluation

To the best of our knowledge, ensemble learning approach coupled with feature engineering and automatic feature selection techniques is a new contribution in the field of predictive analytics modeling for cardiac surgery outcomes. In this section, we describe the research problem that our framework aims to address, followed by evaluation measures to be derived from this framework. Then we enumerate the results obtained from the model.

4.3.1 Research Problem Statement

Our model can be applied to the research application that is illustrated in Figure 4.7. In this context, we define T as a random point in time that a patient is readmitted to the hospital, or is censored from the study by any reason, where 0 < T <= 31. In this study, we focus on readmissions that occur within 30 days after discharge following cardiac surgery procedure. The study begins whenever a patient is discharged from a facility and ends at the 30th day, setting status of the event =1 if the patient was readmitted to the hospital within the interval and event status = 0 otherwise. In case of no rehospitalization during the 30-day

Variable	Categorical String values - assigned numerical values @ source	Normalized and/or Rescaled values	Comment
Gender	Male - 1 Female -2	0 (Male) 1 (Female)	Rescaled to 0 and 1
Race	White/Caucasian - 1 American_Indian_Alaskan_Native - 2 Native_Hawaiian_Pacific_Islander - 3 Black_African_American -4 Hispanic_Latino_Spanish -5 Asian -6 Other -7	1 (White/Caucasian/Alaskan, Hawaiian, Native Indian) 2 (Asian) 3 (Hispanic) 4 (African American) 5 (Others)	Based on the data distribution, combined all USA-originated into one category. And it resulted into a significant factor with better 'p' values
Cerebrovascular Disease	Yes - 1 No - 2 Unknown - 3	0 (No) 1 (Yes)	Blanks and Unknowns (< 1%) are assigned to 'mode' values '0'
Chronic Lung Disease	No -1 Mild -2 Moderate -3 Severe - 4 Severity unknown -5	0 (No) 1 (Mild) 2 (Moderate) 3 (Severe)	Blanks are considered as 'No' (< 1%); 'Severity Unknown' values are merged into 'Mild - most frequent case.
Cardiopulmonary Bypass Utilization	None -1 Combination -2 Full -3	0 (None) 1 (Combination) 2 (Full)	
Admission Status	Elective -1 Urgent -2 Emergent -3	0 (Elective) 1 (Urgent)	No data available for Emergent or Emergent salvage categories.
Discharge Location	Home -1 Extended Care/Transitional Care /Unit/Rehab -2 Other acute care hospital -3 Nursing Home - 4 Hospice - 5 Left AMA - 6 Other - 777	0 (Home) 1 (Non-home)	Home or Non-home feature played pivotal role on readmission event in previous studies
Alcohol Use	<= 1 drink/week -1 2-7 drinks/week -2 >= 8 drinks/week -3 None - 4 Unknown -5	0 (None) 1 (<=1 drink) 2 (2-7 drinks) 3 (>=8 drinks)	Entries such as Unknowns, blanks (< 1%) are assigned to mode value '0'.

Figure 4.6 Illustration of Feature Engineering Rules

interval, the event indicator becomes zero and the readmission time random variable T gets 31. Such observations are called right-censored, which form the only type of censoring in our study. We assume that censoring is non-informative due to the fact that the censoring random variable U is the arranged end of the study and does not have any information about the distribution of T.

4.3.2 Time-to-event Outcomes

From the feature set that we derived in the preceding section, we identified inpatient, non-elective, planned or unplanned, and all-cause hospital readmissions to the hospital within 30 days of discharge from the index procedure. We identified readmissions from primary data source as well as confirmed with CDW data source as well. Readmission to any hospital other than the index hospital-chain either has not been traced or included. The primary outcome evaluated form the model is all-cause readmission event within 30 days of hospital discharge following the cardiac surgery procedure. In the subsequent section we demonstrate the results obtained from the model.

4.4 Experiments and Results

This section describes the experiments we conducted and the results obtained with the study cohort described in section 4.2.1. A total of 1303 patients met inclusion criteria and were included for analysis. All of these patients had multiple records of perioperative time-dependent data measured based on their encounters. With each patient in the study cohort having multiple rows grouped by encounter, the study size was 3852 records of data in

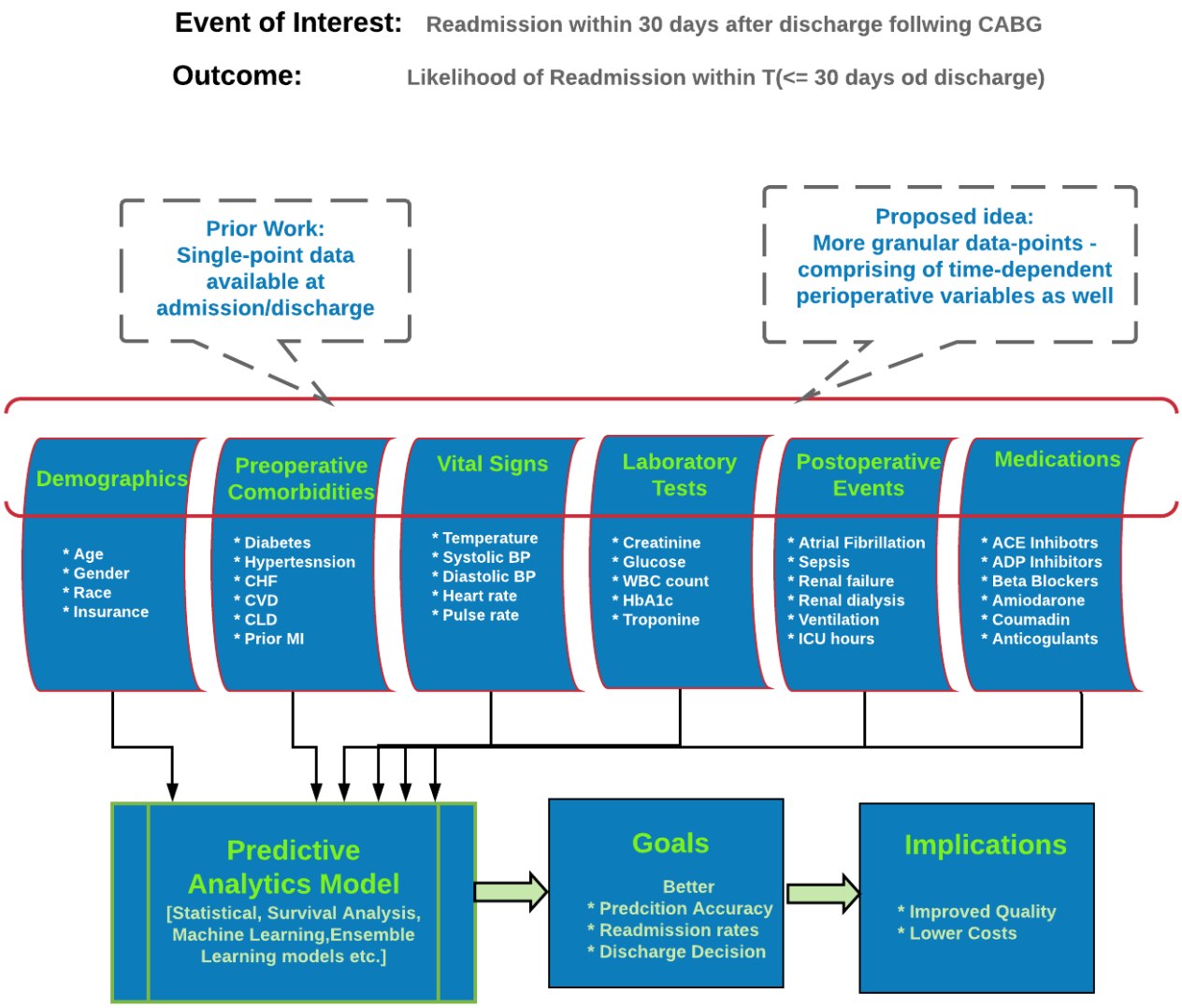


Figure 4.7 Illustration of Research Application

total. Out of 1303 patients, 92 suffered 30-day readmission after isolated CABG (7.1%). Of 82 potential variables analyzed for this study, 72 potential readmission risk variables were included in the model evaluation.

Risk Variable	Univariate Analysis			Multivariate Analysis	
	β coeff.	Hazard Ratio	p value	β coeff.	Hazard Ratio
Race	0.2667	1.3056	<0.001	0.1219	1.1297
Gender	0.3448	1.4117	<0.01	0.1162	1.1232
Hospital	0.2275	1.2555	<0.001	0.0580	1.0598
Preoperative Insurance	0.2717	1.3123	<0.05	0.1214	1.1290
Admission Status	0.3411	1.4064	<0.05	-0.0024	0.9976
Alcohol Use	-0.2620	0.7695	<0.001	-0.1371	0.8719
Diabetes	0.7115	2.0370	<0.001	0.2826	1.3266
Cardiogenic Shock	0.9220	2.5143	<0.01	-0.6368	0.5290
Preoperative Congestive Heart Failure	0.8461	2.3306	<0.001	0.2150	1.2399
Prior Myocardial Infarction	0.2428	1.2748	<0.01	-0.0507	0.9506
Hemo Data -EF	-0.0329	0.9676	<0.001	0.0138	0.9863
Home Oxygen	0.5328	1.7038	<0.05	0.1889	1.2079
Total Albumin	-0.2624	0.7692	<0.001	-0.1761	0.8386
Operative Approach	-0.0355	0.9651	<0.05	0.0169	1.0171
Intra Aortic Balloon Pump	1.2238	3.4002	<0.001	0.7835	2.1891
Intraoperative Blood Products	0.3873	1.4729	<0.01	-0.3118	0.7321
Total OR Time	0.0022	1.0022	<0.01	-0.0019	0.9981
Total ICU Time	0.0024	1.0024	<0.001	0.0018	1.0018
Postoperative Ventilation Time	0.0020	1.0020	<0.001	-0.0008	0.9992
Postoperative Atrial Fibrillation	0.1884	1.2073	<0.05	0.5857	1.7962
Postperative Pulmonery Ventilation Prolonged	0.8180	2.2660	<0.001	-0.1187	0.8880
Postoperative Reoperation for Bleeding	1.3732	3.9481	<0.001	0.8980	2.4547
Postoperative Reoperation for Other Cardiac Reasons	1.5566	4.7426	<0.01	-0.7952	0.4515
Postoperative Return to OR for Non-cardiac Reason	1.5658	4.7864	<0.001	1.5131	4.5406
Postoperative Sepsis	1.0466	2.8480	<0.05	-1.0211	0.3602
Postoperative Pneumonia	1.1815	3.2592	<0.001	0.2607	1.2978
Postoperative Renal Dialysis Required	1.7732	5.8899	<0.001	-0.3581	0.6990
Postoperative Renal Failure	1.4485	4.2565	<0.001	-2.5311	2.2824
Postoperative Dialysis Required after Discharge	2.1037	8.1967	<0.001	0.5588	0.0796
Postoperative HIT	-0.8387	0.4323	<0.001	-0.4751	0.6218
Discharge on Aspirin	0.6166	1.8526	<0.05	0.4162	1.5162
Discharge on Lipid Lowering Statin	0.3137	1.3685	<0.01	0.4884	1.6298
Length of Stay	0.0585	1.0602	<0.001	0.0313	1.0318
Discharge Location	0.5923	1.8081	<0.001	0.2175	1.2429

Figure 4.8 Significant Feature Set: Characteristics of time-independent covariates evaluated from univariable and multivariable analyses of the Cox Proportional Hazards survival regression model

Risk Variable	Univariate Analysis			Multivariate Analysis	
	β coeff.	Hazard Ratio	p value	β coeff.	Hazard Ratio
Preoperative Hemoglobin A1c Level	0.1622	1.1761	<0.001	0.0168	1.0169
Preoperative Creatinine	0.1815	1.1990	<0.001	0.1374	1.1473
Preoperative Blood Glucose	0.0074	1.0075	<0.001	0.0044	1.0045
Preoperative Hematocrit	-0.0464	0.9547	<0.001	0.0525	1.0539
Preoperative WBC Count	0.0644	1.0665	<0.01	0.0643	1.0664
Highest Intraoperative Glucose	0.0030	1.0030	<0.05	0.0016	1.0016
Lowest Intraoperative Hemoglobin	0.1025	0.9026	<0.001	-0.1178	0.8889
Postoperative Creatinine	0.2149	1.2397	<0.001	0.1845	1.2026
Postoperative Glucose	0.0036	1.0036	<0.05	0.0011	1.0011
Postoperative Hemoglobin	-0.1029	0.9022	<0.001	0.0412	1.0421
Postoperative Hematocrit	-0.0605	0.9413	<0.001	0.1243	0.8831

Figure 4.9 Significant Feature Set: Characteristics of time-dependent pre and perioperative covariates evaluated from univariable and multivariable analyses of the Cox Proportional Hazards survival regression model

4.4.1 Significant Feature Set from COX PH Model

Out of the 72 features, 45 were identified as significant with a p-value of <0.05 when subjected to univariable analysis using CPH survival regression methods. 34 of these were time-independent pre-, peri-, and postoperative covariates (Figure 4.8), while 11 features were time-dependent covariates (Figure 4.9). Variables highlighted in red belong to the most significant feature set that impact the target outcome with a hazard ratio of greater than 1.0.

Multivariable analysis determined the 28 most significant factors, and 9 (32.1%) covariates were identified as the most significant in the time-dependent group while the remaining 19 covariates (67.9%) were time-independent. The time-dependent covariates included Preoperative Hemoglobin A1c level with Hazard Ratio = 1.0169, Preoperative Creatinine (HR=1.1473), Preoperative Blood Glucose (HR=1.0045), Preoperative Hematocrit

(HR=1.10539), Preoperative WBC count (HR=1.0664), Highest Intraoperative Glucose (HR=1.0016), Postoperative Creatinine (HR=1.2026), Postoperative Glucose (HR=1.0011), Postoperative Hemoglobin (HR=1.0421).

Postoperative Creatinine was the most significant risk factor for 30-day readmission in this model followed by Postoperative Hemoglobin, and Postoperative Glucose in time-dependent group. For instance, the exponential coefficient (exp(coef) = exp(0.1845) = 1.2036) for 'Postoperative Creatinine' covariate, also known as Hazard Ratio (HR), indicated the impact of covariate on readmission. It can be inferred that, an increase of 1.0 gm/dL in creatinine level has a hazard of readmission factor of 1.3439. The most significant risk contributors in time-independent group were Postoperative Return To OR For Non-cardiac Reason, Reoperation for Bleeding, Renal Failure, and Postoperative Atrial Fibrillation.

4.4.2 Survival Curves of Significant Features

With a fitted CPH model, survival plots for some of the most significant were created (Figure 4.10). These plots illustrated the impact of a covariate on 30 day readmission, as the covariate changes over time, while all the other covariates remain equal. For instance, the graph for Postoperative Creatinine implied that creatine greater than the baseline level of 1.2 increases the risk of readmission. Similarly, the survival curve for Length of Stay implied that a patient with longer stay at the hospital (beyond approximately 8 days) had a higher risk of readmission. The features highlighted (in red color) in Figures 4.8 and 4.9 were selected for the final predictive model. Features such as alcohol use, cardiogenic shock, total albumin, intraoperative blood products, postoperative pulmonary ventilation prolonged,

postoperative reoperation for other cardiac reasons, postoperative sepsis, postoperative renal dialysis required, postoperative dialysis required after discharge, postoperative HIT, lowest intraoperative glucose, postoperative hematocrit were excluded from the final model on the basis of the results of the CPH multivariate survival regression analysis. The area under the ROC curve for the training data set (n=1042) was 0.951, and it was 0.873 in the validation data set (n=261), as shown in Figure 4.11.B. Excluding the time-dependent variables, the area under the ROC curve drops to 0.868 and 0.658 for the training and validation data sets, respectively (Figure 4.11.A).

4.4.3 Concordance Index Measures

In this subsection, we present the results of concordance index metrics obtained from experiments with state-of-the-art models, such as Cox PH model [18], Deep Survival analysis model, Tensor Flow Deep Survival model [94, 95], XGBoost [69], and the proposed ensemble learning model (Figure 4.12). All the above models were subjected to the same study cohort. Our proposed framework demonstrates improved concordance index statistics (that represent a generalization of the area under the receiver operating characteristics curve taking into account survival data as well) of 0.873 for validation data set.

4.4.4 Recursive Feature Elimination, Cross-Validation and Parameter Tuning

This section presents experimental details and results of 'Recursive Feature Elimination' (RFE) feature selection algorithm, cross validation, and XGBoost parameter tuning and optimization techniques. These methods were applied to the feature set of 45 readmission

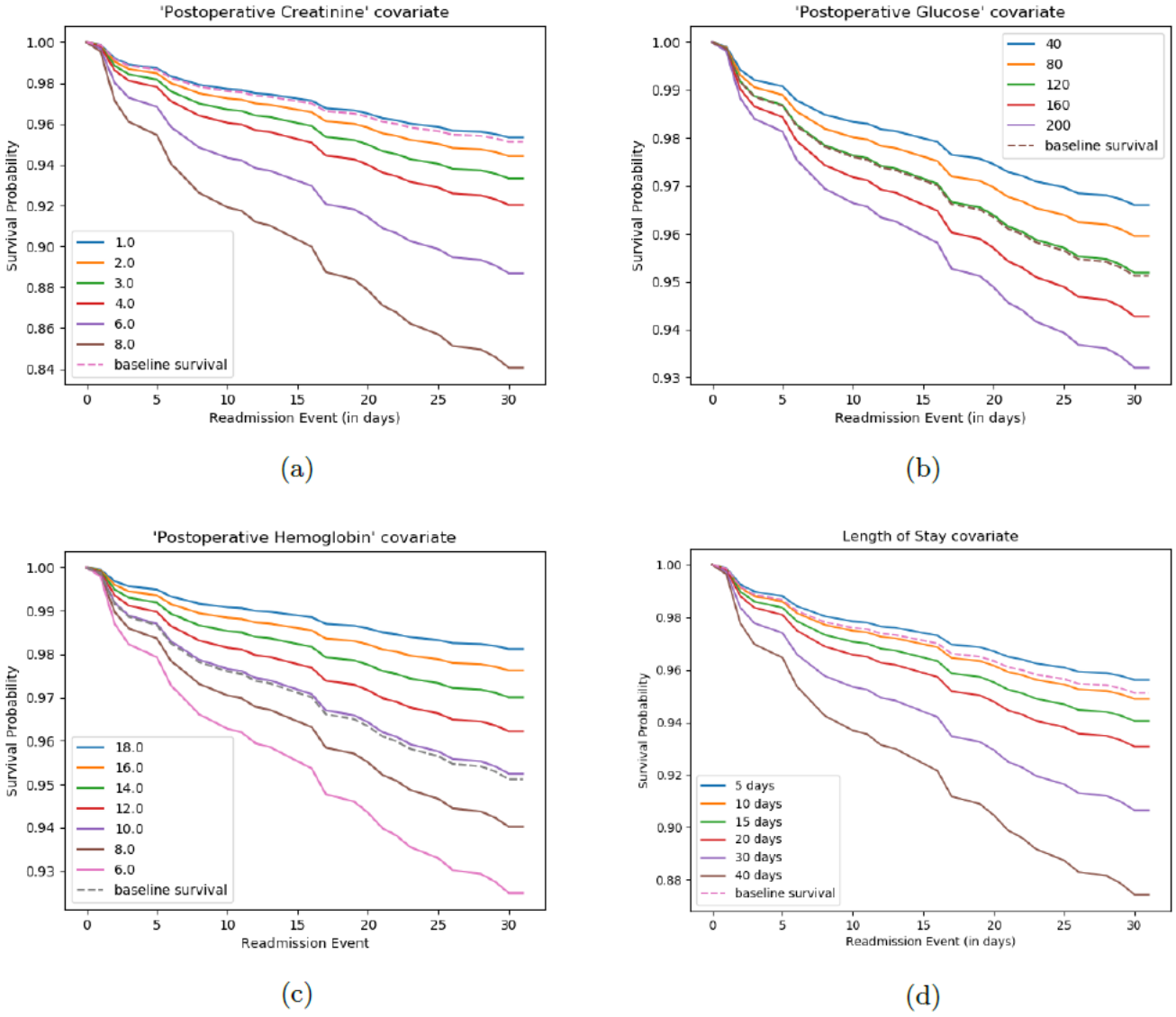


Figure 4.10 Survival curves for the most significant covariates evaluated from the CPH multivariable model: Postoperative Creatinine (a), Postoperative Glucose (b), Postoperative Hemoglobin (c), and Length of Stay (d). These plots depict the impact of a covariate on 30-day readmission event, as the covariate changes while everything else holds equal

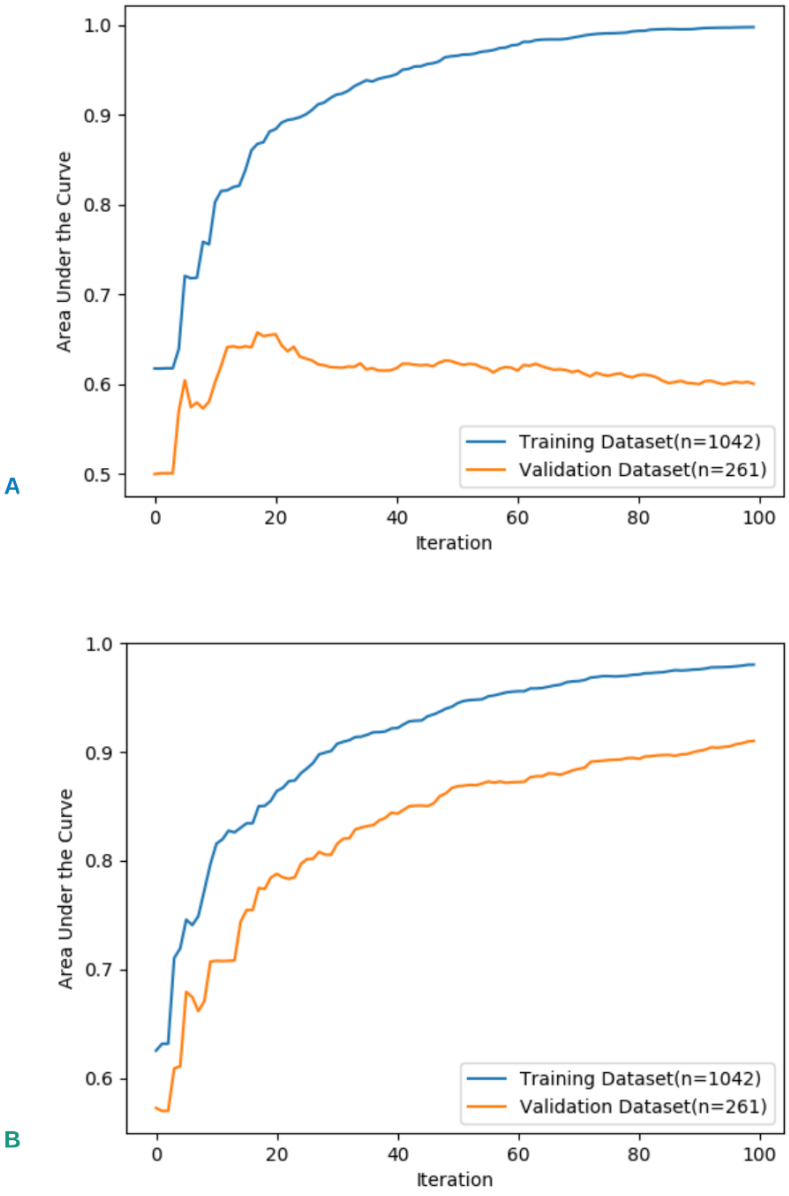


Figure 4.11 Performance metrics of the prediction model before (A), and after (B) the inclusion of time-dependent covariates.(A: AUROC statistics for Training dataset = 0.868, and Validation dataset = 0.658, best values at 17th iteration; B: AUROC statistics for Training dataset = 0.951, and Validation dataset = 0.873, best values at 55th iteration)

Dataset	Model 1 [CoxPH]	Model 2 [DeepSurv]	Model 3 [TFDeepSurv]	Model4 [XGBoost]	Proposed Model [Ensemble Learning]
Training (n=1042)	0.726	0.749	0.786	0.876	**0.951**
Validation (n=261)	0.713	0.654	0.652	0.749	**0.873**

Figure 4.12 Performance comparison matrix: C-Index measures with 95% confidence interval for training and validation datasets

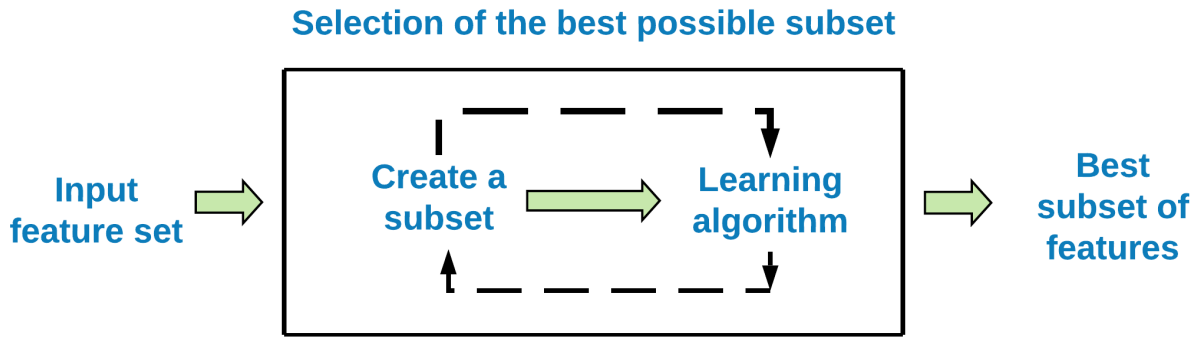
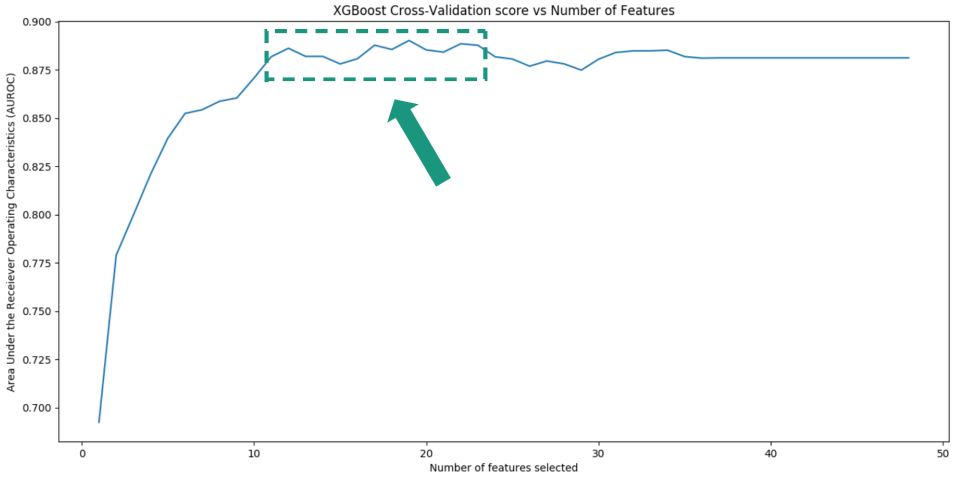


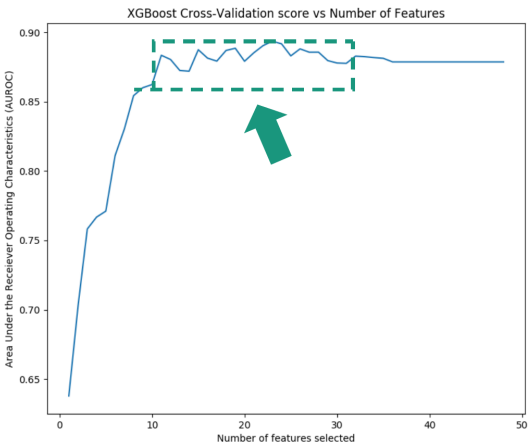
Figure 4.13 Illustration of Recursive Feature Elimination algorithm

risk factors (that resulted with survival regression described in the preceding section 4.4.1) to further improve the performance of the model.

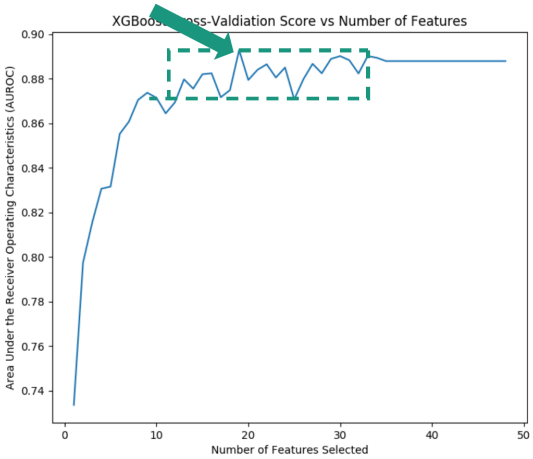
RFE method, an automatic feature selection technique, evaluated the input feature set by recursively removing the redundant and the least important features, and built a model on the features that contribute the most to the model accuracy (Figure 4.13). Figure 4.14 demonstrates the RFE cross validation scores versus the number of features, when the feature



A. 3-fold Cross Validation



B. 5-fold Cross Validation



C. 10-fold Cross Validation

Figure 4.14 Illustration of Cross Validation scores versus Number of Features obtained with RFE Feature Selection Method

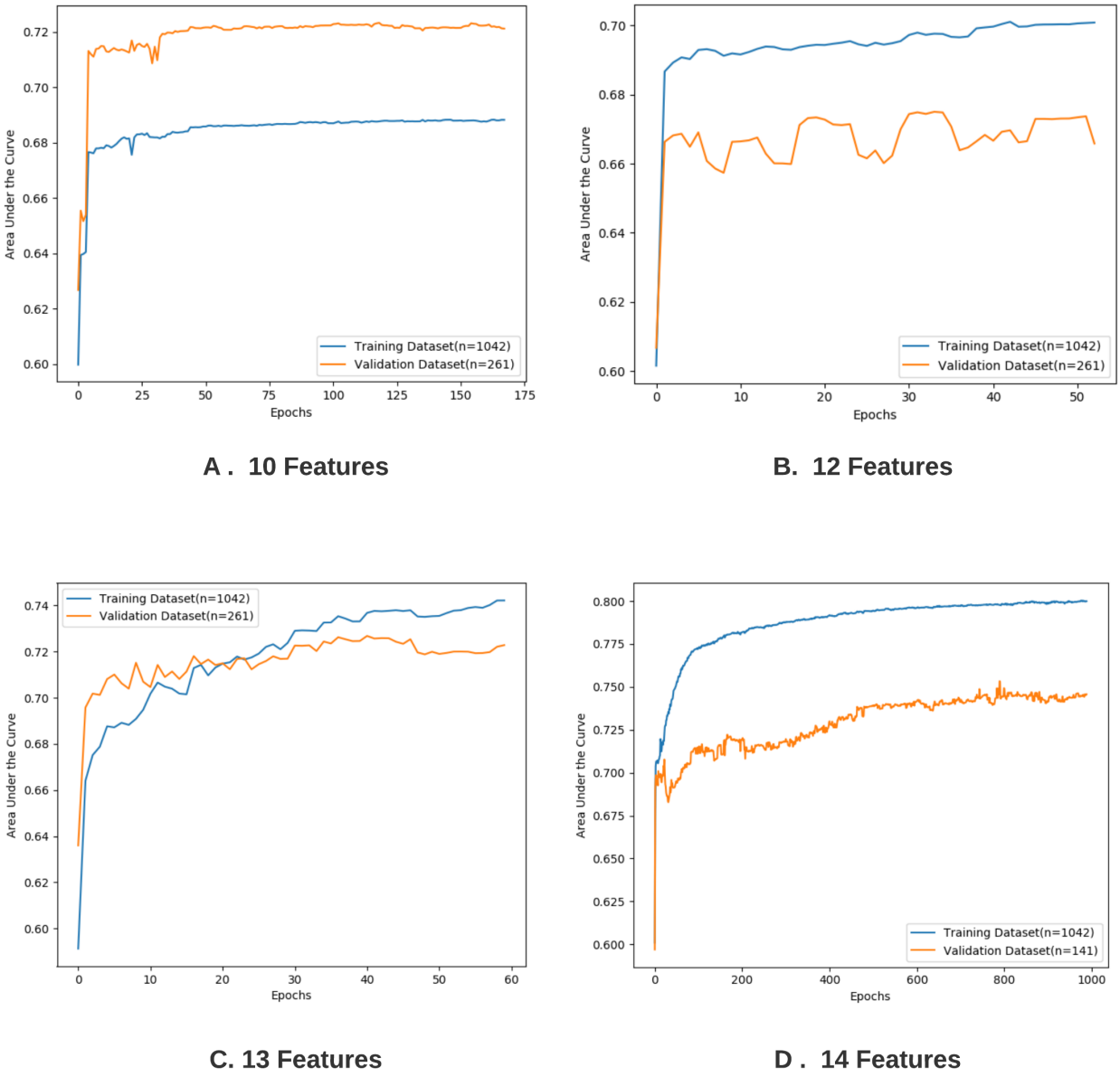


Figure 4.15 Performance Metrics of the Prediction Model: Area Under the Receiver Operating Characteristics (AUROC) measures for number of features n = 10, 12, 13, and 14

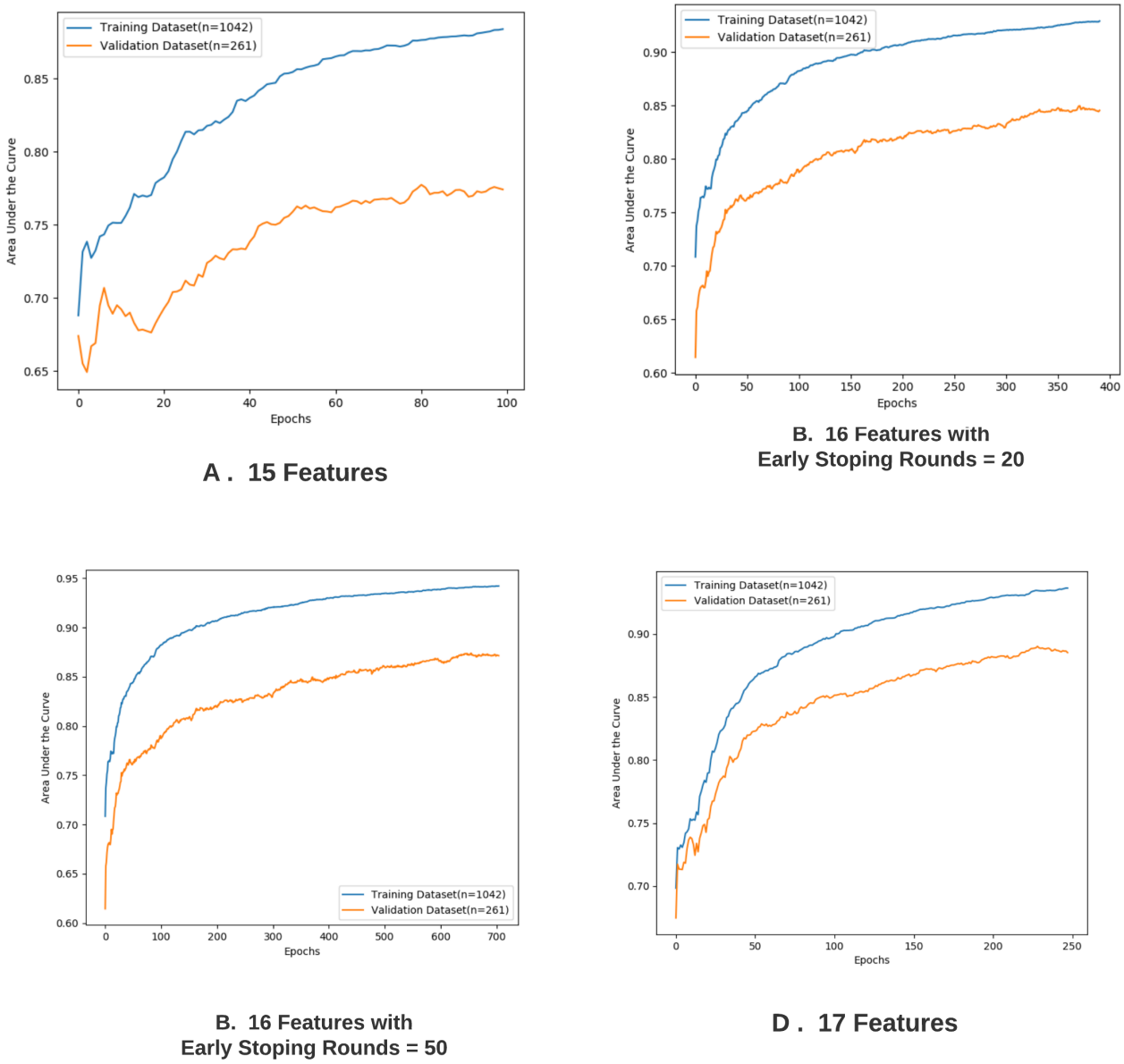


Figure 4.16 Performance Metrics of the Prediction Model: AUROC measures for number of features n = 15, 16, and 17

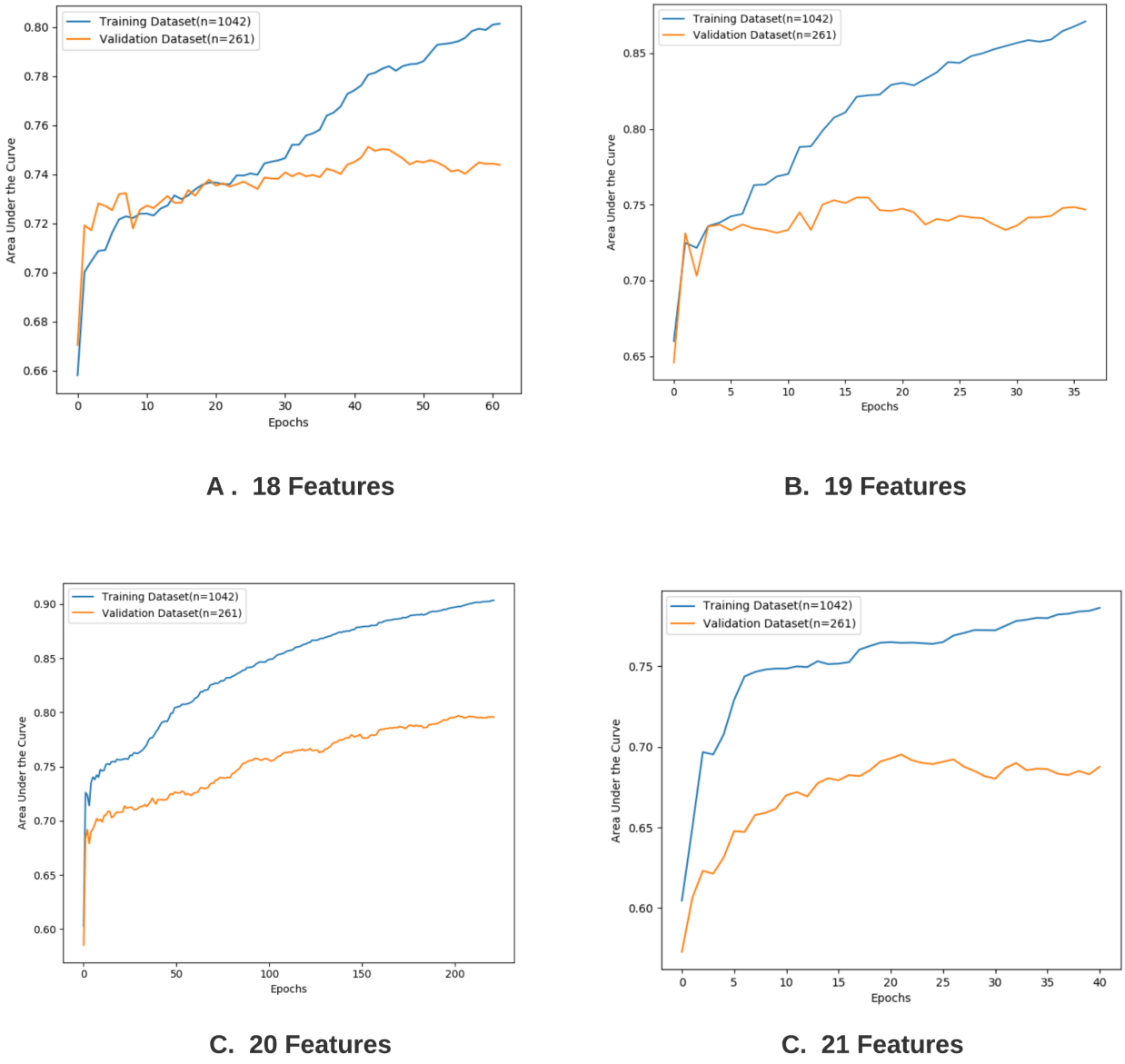


Figure 4.17 Performance Metrics of the Prediction Model: AUROC measures for number of features n = 18, 19, 20, and 21

set was subjected to k-fold cross-validation, with k=3, 5, and 10. The green-highlighted region indicate the range of optimum number of features that tend to yield best possible predictive accuracy of around 0.89 for the validation data set of the study cohort.

Figures 4.15 through 4.19 illustrate the performance metrics of the prediction model in terms of area under the receiver operating characteristics (AUROC) for feature sets of 10 through 30. Using RFE feature selection algorithm and 10-fold cross-validation, the model demonstrated optimum predictive ability of around 0.89 when the number of features were between 20 to 24 (Figure 4.20.A). Of these 24, 16 features resulted with importance scores of greater than or equal to 9 (Figure 4.20.C). These were: Preoperative Creatinine, Preoperative Hemoglobin A1c Level, Lowest Intraoperative Hemoglobin, Postoperative Hematocrit, Postoperative Creatinine, Postoperative Hemoglobin, Total Albumin, Alcohol Use, Postoperative HIT, Postoperative Return to OR for Non-Cardiac Reason, Intra-Aortic Balloon Pump, Race, Preoperative Insurance, Discharge Location, Preoperative Congestive Heart Failure, Postoperative Reoperation for Bleeding.

For the 16 selected features, the AUROC measures observed were 0.9492 and 0.8816 at the 35th iteration, and 0.9995 and 0.9307 at the 151st iteration (Figure 4.20.B) for the training and validation datasets respectively. In addition to the AUROC measures, the proposed prediction model builds individual decision trees using the selected set of 16 features. Three instances of such decision trees are depicted in Figures 4.21 though 4.23. These boosting decision tree plots show how the model arrived at its final decisions and what splits it made to arrive at those decisions. In other words, these trees provide better understanding and

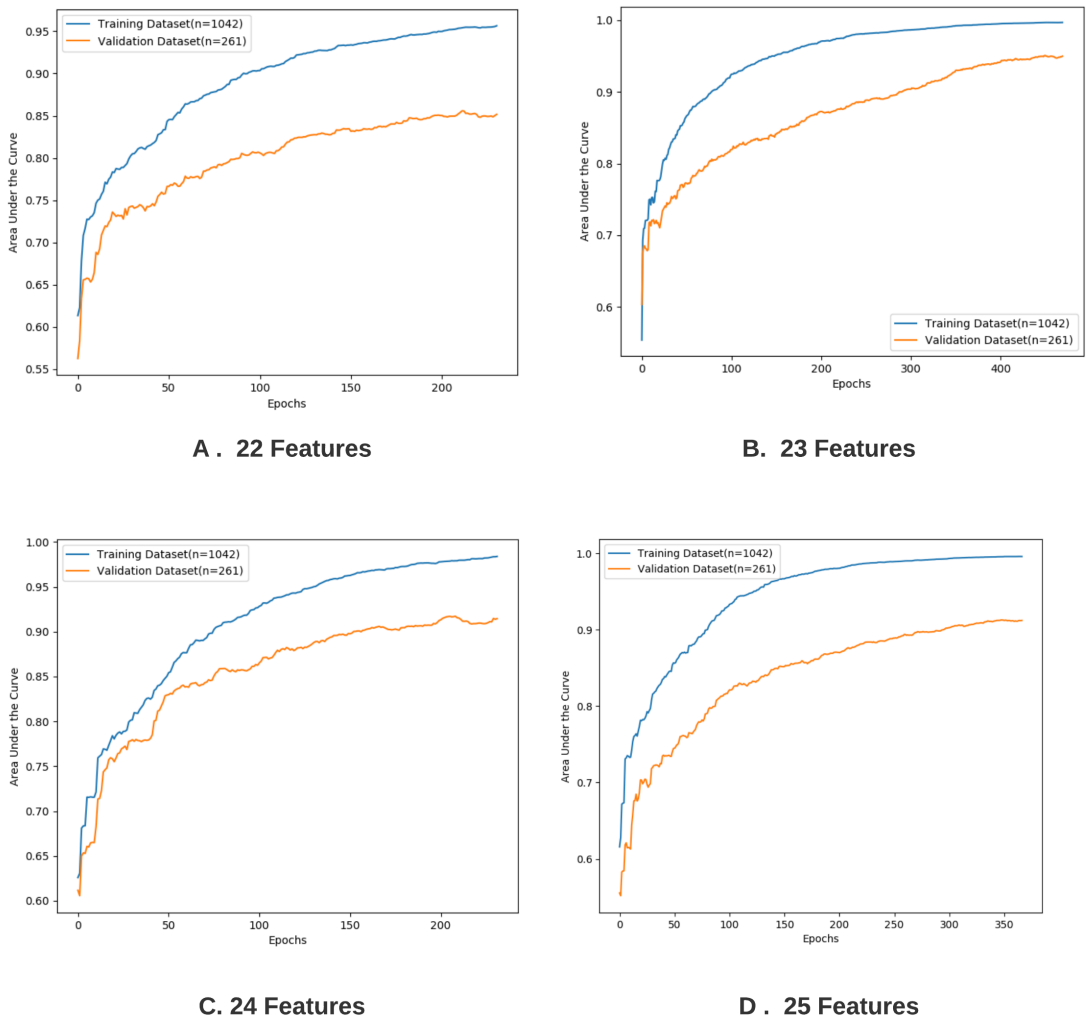


Figure 4.18 Performance Metrics of the Prediction Model: AUROC measures for number of features n = 22, 23, 24, and 25

interpretability about the impact of each feature (for example: Length of Stay, Postoperative Creatinine, Postoperative Hemoglobin etc.) on the target outcome, i.e., the impact of the predictor variables on 30-day readmission after discharge following CABG.

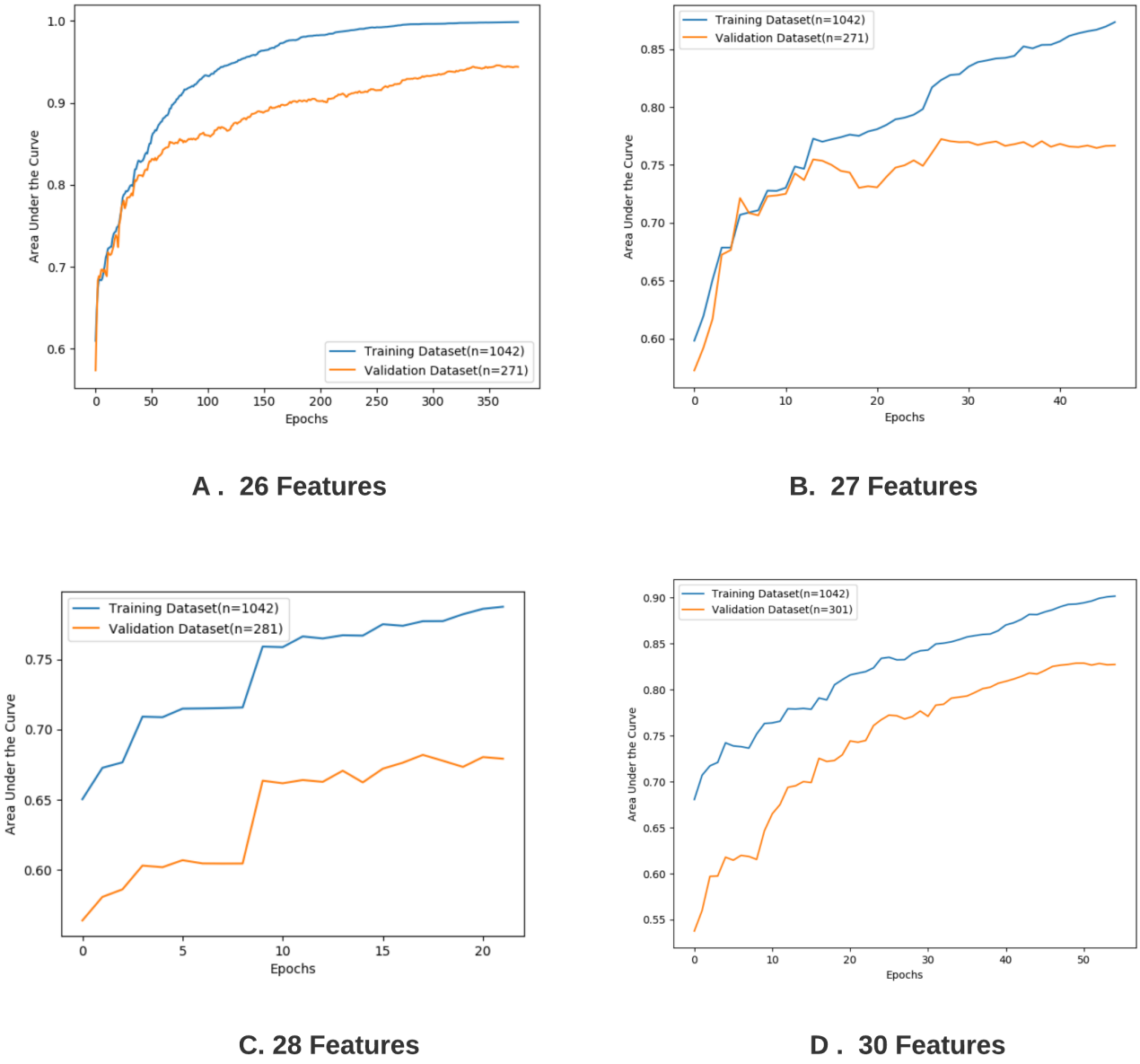
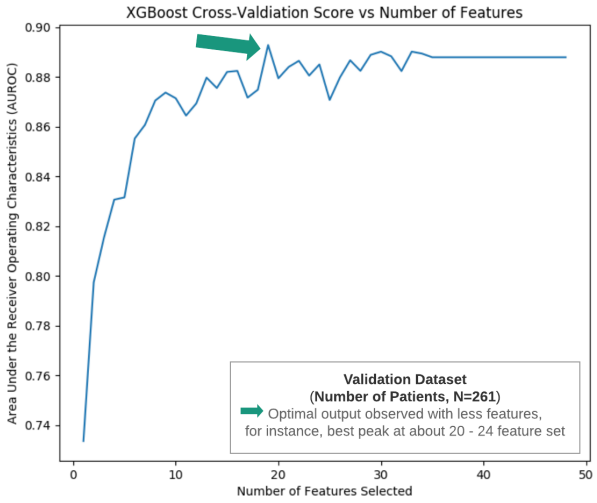
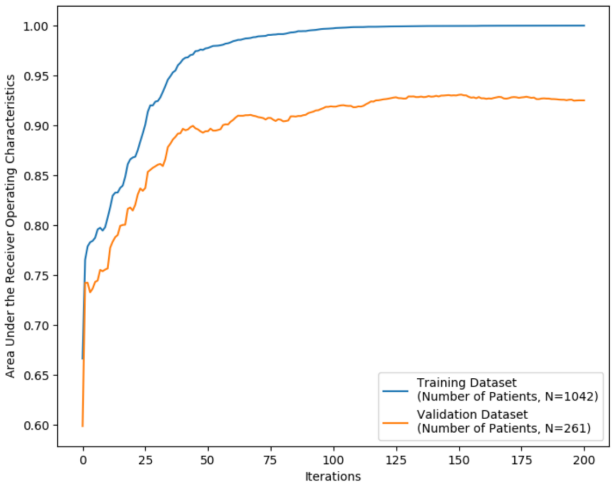


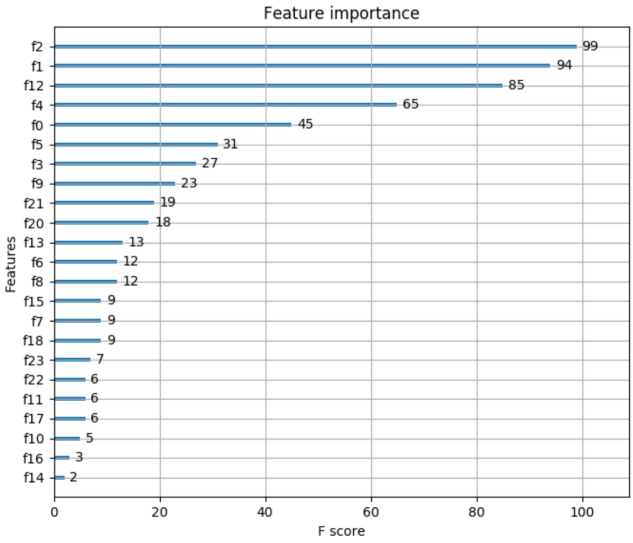
Figure 4.19 Performance Metrics of the Prediction Model: AUROC measures for number of features n = 26, 27, 28, and 30



A. Predicitive Ability of the Model: Area Under the ROC Curve Versus Number of Features Selected



B: Performance Metrics of the Prediction Model with top 16 Significant Features Training dataset AUROC = 0.9492, Validation dataset AUC = 0.8816 at 35th iteration



C. Most Significant 16 Features in the order of Feature Importance scores:

f2 = Preoperative Creatinine
f1 = Preoperative Hemoglobin A1c Level
f12 = Lowest Intraoperative Hemoglobin
f4 = Postoperative Hematocrit
f0 = Postoperative Creatinine
f5 = Postoperative Hemoglobin
f3 = Total Albumin
f9 = Alcohol Use
f21 = Postoperative HIT
f20 = Postoperative Return to OR for Non-Cardiac Reason
f13 = Intra Aortic Balloon Pump
f6 = Race
f8 = Preoperative Insurance
f15 = Discharge Location
f7 = Preoperative Congestive Heart Failure
f18 = Postoperative Reoperation for Bleeding

Figure 4.20 Performance Metrics of the Proposed Ensemble Learning Predictive Analytics Framework

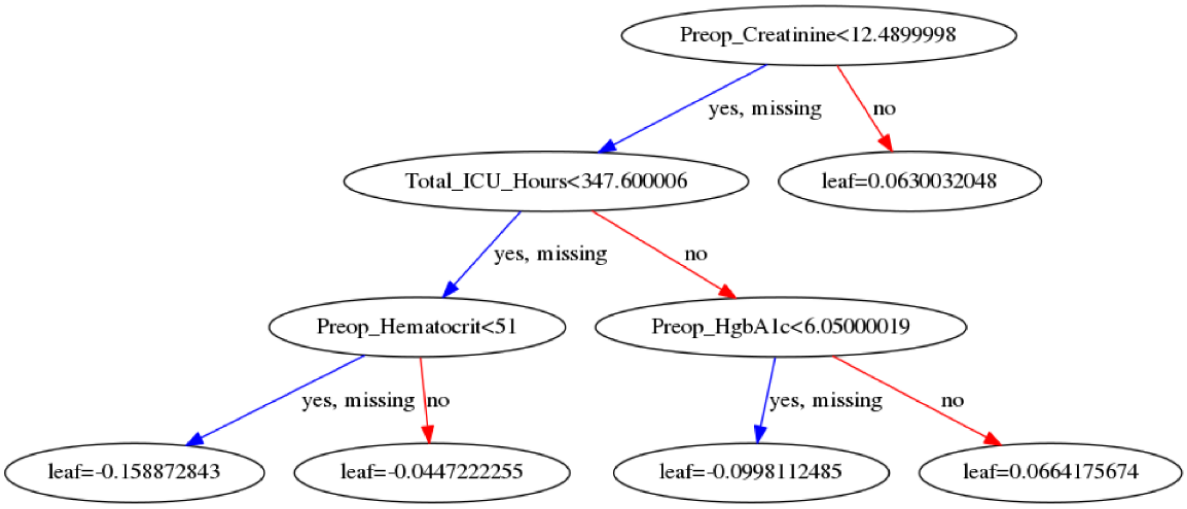


Figure 4.21 Illustration of 1st decision tree instance derived from the model's best feature set

Figure 4.22 Illustration of 2nd decision tree instance derived from the model's best feature set

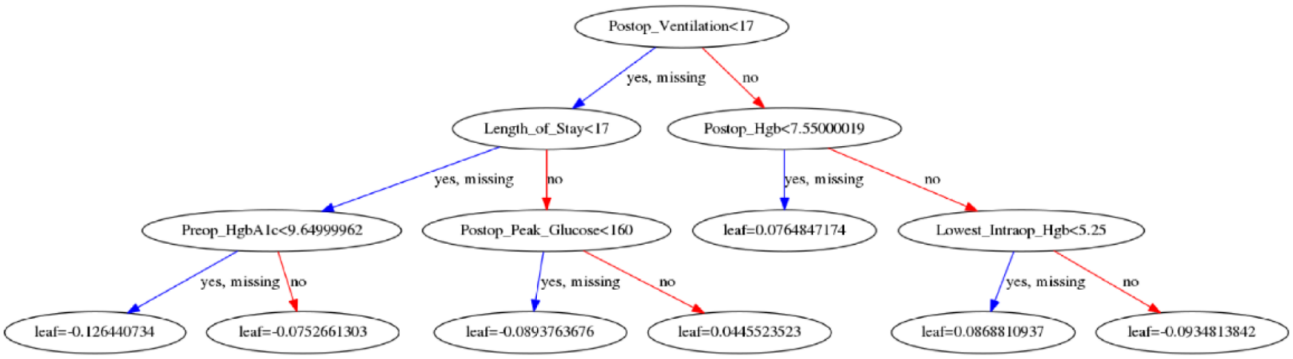


Figure 4.23 Illustration of 3rd decision tree instance derived from the model's best feature set

CHAPTER 5
SUMMARY, CONCLUSION, AND FUTURE SCOPE

This chapter provides a summary of the review of findings, conclusions garnered from the results, limitations of the proposed prediction model, and directions for further research and enhancements.

5.1 Summary

In undertaking this study, we sought to develop a new simple, scalable, portable and user-friendly open-source predictive analytics model by using advanced ensemble learning methods and automatic feature selection techniques more effectively. This approach facilitated analyzation of huge amounts of medical data, elimination of redundant features, reduction of overlearning on training data and hence, improvisation of prediction accuracy over new test data.

With our proposed ensemble learning model, we attempted to improve performance using a four-step process, i.e., by feeding more clean and structured data though simple feature engineering rules, applying effective feature selection methods, algorithmic tuning, and ensemble learning techniques, as illustrated in Figure 5.1. Also, majority of the previous time-to-event prediction models in medical studies used 'single-point' or time-fixed data (for example, one row per each patient) available at either patient's index-admission or discharge. Our new model utilized more granular data points (i.e. multiple rows of records) for each subject of the study, for example, data consisting both the time-dependent and time-independent perioperative covariates for each patient grouped by encounter, to predict

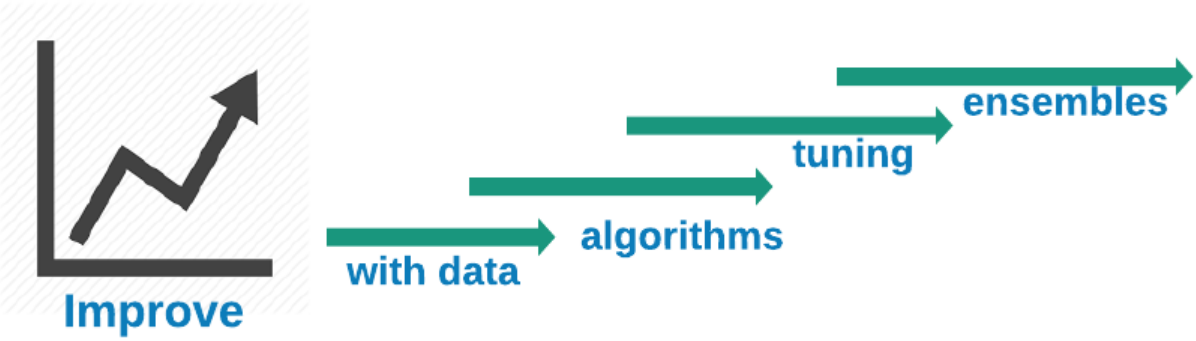


Figure 5.1 Snapshot of performance improvement techniques

hospital readmission within 30 days of discharge following CABG. Furthermore, our study identified that the best feature set (comprising of 16 out of the 82 input features) that contributes the most to the target outcome would in turn produce better prediction accuracy for the model, as summarized in Figure 5.2. The apparent merits of this framework are: Enhanced learning test data, improved performance and accuracy with less features, scalability and Portability, user friendliness with automatic feature engineering and feature selection methods.

5.2 Limitations of the Study

Our study is not without limitations. One of those is the smaller and limited study cohort size. The study cohorts were derived from a single-institution's multi-hospital facility and may not have captured data regarding target outcomes at other hospitals. This was reflected in the admittedly low rate of readmission events. Also, the study cohort size was relatively

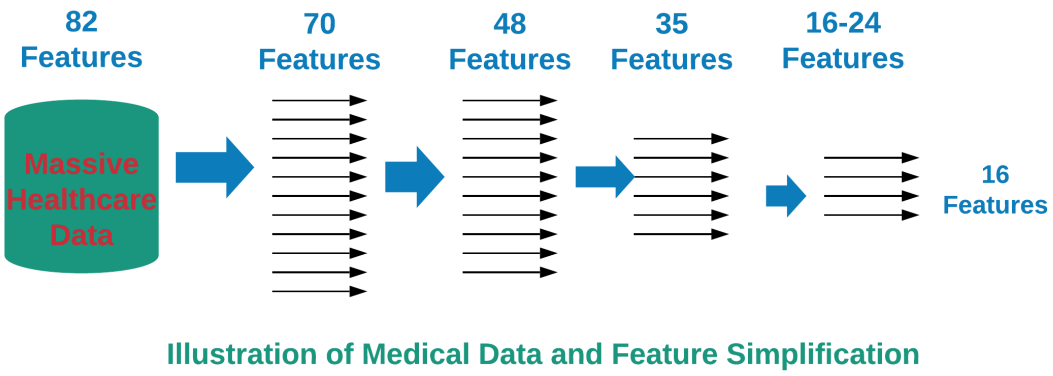


Figure 5.2 Workflow of feature engineering, feature selection and simplification

smaller with 1334 patients with a study period of 20 months. The current model should be tested prospectively at multiple hospitals to substantiate its broader applicability. This study considered time-dependent pre-, perioperative covariates, within 48 hours prior to CABG procedure and up to 72 hours of postoperative period. In a way, this facilitated minimal missing data (¡1%) for all the covariates, and provided more consistent supplemental data drawn from local clinical data warehouse. This model was more suited to handle tabular form of medical data, and the wave-from or time-series data were untested or unexplored in this study.

5.3 Recommendations for Further Research and Enhancements

Perhaps the most intriguing extension of the present study would be to test this model's performance in a variety of settings and populations in order to make it reproducible and

more generalizable. For instance, this framework can be extended to handle time-to-event outcomes in applications areas of manufacturing (eg: to predict a device failure), academics (eg: to predict a student drop-out), business or finance (eg: to predict a project success or failure, to track customer behavior, stocks behaviour etc.). Further potential performance-hikes can be explored with other feature selection methods and hyper-parameter tuning and optimization techniques as well. Another potential direction for future research in ensemble learning predictive analytics modeling could be handling of wave-form or time-series medical data as well.

5.4 Conclusion

Time-to-event outcomes in healthcare can be predicted more accurately with less features by using open-source advanced ensemble learning techniques. Prediction accuracy could be further improved by feeding huge amounts of highly-correlated features of patient medical data, including time-varying laboratory values, vital signs, and medications. Our new prediction model demonstrated excellent accuracy and could be more useful for decision support in clinical settings.

REFERENCES

1. May,T. (2018). "The fragmentation of health data," *Medium*. Retrieved July 31, 2018 from http://medium.com

2. Maxmen, J. S. (1976). The post-physician era: Medicine in the twenty-first century.

3. Hinton, G. (2018). Deep learning—a technology with the potential to transform health care. *Jama*, 320(11), 1101-1102.

4. Verghese, A., Shah, N. H., & Harrington, R. A. (2018). What this computer needs is a physician: humanism and artificial intelligence. *Jama*, 319(1), 19-20.

5. Shah, N. D., Steyerberg, E. W., & Kent, D. M. (2018). Big data and predictive analytics: recalibrating expectations. *Jama*, 320(1), 27-28.

6. Naylor, C. D. (2018). On the prospects for a (deep) learning health care system. *Jama*, 320(11), 1099-1100.

7. Centers for Disease Control and Prevention. (2018). Multiple cause of death 1999-2017 on CDC WONDER online Database, released December, 2018. Data are from the Multiple cause of death Files, 1999-2017 as compiled from data provided by the 57 vital statistics jurisdictions through the vital Statistics cooperative Program.

8. Centers for Medicare and Medicaid Services. *Readmissions Reduction Program*. Retrieved January 15, 2018 from https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html. .

9. Giacomino, B. D., Cram, P., Vaughan-Sarrazin, M., Zhou, Y., & Girotra, S. (2016). Association of hospital prices for coronary artery bypass grafting with hospital quality and reimbursement. *The American journal of cardiology*, 117(7), 1101-1106.

10. McIlvennan, C., Eapen, Z., & Allen, L. (2015). Hospital readmissions reduction program. Circulation : *Journal of the American Heart Association.*, 131(20), 1796-1803.

11. Heron, M. Deaths: Leading causes for 2014 pdf icon. *National vital statistics reports*, 65(5).

12. Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, 360(14), 1418-1428.

13. Osnabrugge, R. L., Speir, A. M., Head, S. J., Jones, P. G., Ailawadi, G., Fonner, C. E., ... & Rich, J. B. (2014). Cost, quality, and value in coronary artery bypass grafting. *The Journal of thoracic and cardiovascular surgery*, 148(6), 2729-2735.

14. National Heart, Lung, and Blood Institute. (n.d.). *Coronary Artery Bypass Grafting.* Retrieved Jan 15, 2018 from https://www.nhlbi.nih.gov/health-topics/coronary-artery-bypass-grafting.

15. Wang, P., Li, Y., & Reddy, C. K. (2017). Machine learning for survival analysis: A survey. *arXiv preprint arXiv:1708.04649.*

16. Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187-202..

17. Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part II: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3), 431.

18. Davidson-Pilon C. *Lifelines*. (n.d.). Retrieved January 15, 2018 from https://github.com/CamDavids /lifelines.

19. Heagerty, P. J. , & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61 (1), 92–105.

20. Langova, K. (2008). Survival analysis for clinical studies. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub*, 152(2), 303-307.

21. Hall, P., Dean, J., Kabul, I. K., & Silva, J. (2014). *An overview of machine learning with SAS® enterprise miner™*. SAS Institute Inc.

22. Nezhad, Sadati, Yang, & Zhu. (2019). A Deep Active Survival Analysis approach for precision treatment recommendations: Application of prostate cancer. *Expert Systems With Applications*, 115, 16-26.

23. Lundervold, A. S., & Lundervold, A. (2018). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*.

24. Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11, 169-198.

25. Deo, S. V., Raza, S., Altarabsheh, S. E., Deo, V. S., Elgudin, Y. E., Marsia, S., ... & Kolte, D. (2018). Risk Calculator to Predict 30-Day Readmission After Coronary Artery Bypass: A Strategic Decision Support Tool. *Heart, Lung and Circulation*.

26. Team, R. C. (2013). R: A language and environment for statistical computing.

27. Horwitz, L., Partovian, C., Lin, Z., Herrin, J., Grady, J., Conover, M., ... & Bernheim, S. (2011). Hospital-wide (all-condition) 30-day risk-standardized readmission measure. *Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation*. Retrieved September, 10, 2018.

28. Benuzillo, J., Caine, W., Evans, R., Roberts, C., Lappe, D., & Doty, J. (n.d.). Predicting readmission risk shortly after admission for CABG surgery. *Journal of Cardiac Surgery.*, 33(4), 163-170.

29. Zywot, A., Lau, C. S., Glass, N., Bonne, S., Hwang, F., Goodman, K., ... & Paul, S. (2018). Preoperative Scale to Determine All-Cause Readmission After Coronary Artery Bypass Operations. *The Annals of thoracic surgery*, 105(4), 1086-1093.

30. Feng, T. R., White, R. S., Gaber-Baylis, L. K., Turnbull, Z. A., & Rong, L. Q. (2018). Coronary artery bypass graft readmission rates and risk factors-A retrospective cohort study. *International Journal of Surgery*, 54, 7-17.

31. Li, Z., Amstrong, E. J., Parker, J. P., Danielsen, B., & Romano, P. S. (2012). Hospital variation in readmission after coronary artery bypass surgery in California. *Circulation: Cardiovascular Quality and Outcomes*, 5(5), 729-737.

32. Lancey, R., Kurlansky, P., Argenziano, M., Coady, M., Dunton, R., Greelish, J., ... & Williams, T. (2015). Uniform standards do not apply to readmission following coronary artery bypass surgery: a multi-institutional study. *The Journal of thoracic and cardiovascular surgery*, 149(3), 850-857.

33. Anderson, J. E., Li, Z., Romano, P. S., Parker, J., & Chang, D. C. (2016). Should risk adjustment for surgical outcomes reporting include sociodemographic status? A study of coronary artery bypass grafting in California. *Journal of the American College of Surgeons*, 223(2), 221-230.

34. Kilic, Magruder, Grimm, Dungan, Crawford, Whitman, & Conte. (2017). Development and Validation of a Score to Predict the Risk of Readmission After Adult Cardiac Operations. *The Annals of Thoracic Surgery*, 103(1), 66-73.

35. Espinoza, J., Camporrontondo, M., Vrancic, M., Piccinini, F., Camou, J., Benzadon, M., & Navia, D. (2016). 30-day readmission score after cardiac surgery. *Clinical Trials and Regulatory Science in Cardiology*, 20, 1-5.

36. Swaminathan, R. V., Feldman, D. N., Pashun, R. A., Patil, R. K., Shah, T., Geleris, J. D., ... & Singh, H. S. (2016). Gender differences in in-hospital outcomes after coronary artery bypass grafting. *The American journal of cardiology*, 118(3), 362-368.

37. Fanari, Z., Elliott, D., Russo, C. A., Kolm, P., & Weintraub, W. S. (2017). Predicting readmission risk following coronary artery bypass surgery at the time of admission. *Cardiovascular Revascularization Medicine*, 18(2), 95-99.

38. Shahian, D. M., He, X., O'Brien, S. M., Grover, F. L., Jacobs, J. P., Edwards, F. H., ... & Han, L. (2014). Development of a clinical registry-based 30-day readmission measure for coronary artery bypass grafting surgery. *Circulation, 130(5)*, 399-409.

39. Keenan, P. S., Normand, S. L. T., Lin, Z., Drye, E. E., Bhat, K. R., Ross, J. S., ... & Wang, Y. (2008). An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. *Circulation: Cardiovascular Quality and Outcomes*, 1(1), 29-37.

40. Krumholz, H. M., Lin, Z., Drye, E. E., Desai, M. M., Han, L. F., Rapp, M. T., ... & Normand, S. L. T. (2011). An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. *Circulation: Cardiovascular Quality and Outcomes*, 4(2), 243-252.

41. Lindenauer, P. K., Normand, S. L. T., Drye, E. E., Lin, Z., Goodrich, K., Desai, M. M., ... & Krumholz, H. M. (2011). Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia. *Journal of Hospital Medicine*, 6(3), 142-150.

42. Maniar, H. S., Bell, J. M., Moon, M. R., Meyers, B. F., Marsala, J., Lawton, J. S., & Damiano Jr, R. J. (2014). Prospective evaluation of patients readmitted after cardiac surgery: analysis of outcomes and identification of risk factors. *The Journal of thoracic and cardiovascular surgery*, 147(3), 1013-1020.

43. Price, J. D., Romeiser, J. L., Gnerre, J. M., Shroyer, A. L. W., & Rosengart, T. K. (2013). Risk analysis for readmission after coronary artery bypass surgery: developing a strategy to reduce readmissions. *Journal of the American College of Surgeons*, 216(3), 412-419.

44. Currie, K. B., & Lancey, R. (2011). A predictive model for readmission within 30 days after coronary artery bypass grafting. *Journal of the American College of Surgeons*, 213(3), S107.

45. Hannan, E. L., Zhong, Y., Lahey, S. J., Culliford, A. T., Gold, J. P., Smith, C. R., ... & Wechsler, A. (2011). 30-day readmissions after coronary artery bypass graft surgery in New York State. *JACC: Cardiovascular Interventions*, 4(5), 569-576.

46. Stewart, R. D., Campos, C. T., Jennings, B., Lollis, S. S., Levitsky, S., & Lahey, S. J. (2000). Predictors of 30-day hospital readmission after coronary artery bypass. *The Annals of thoracic surgery*, 70(1), 169-174.

47. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15), 1688-1698.

48. Harrell, J. F., Lee, K. L., Matchar, D. B., & Reichert, T. A. (1985). Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer treatment reports*, 69(10), 1071-1077.

49. Steen, P. M. (1994). Approaches to predictive modeling. *The Annals of thoracic surgery*, 58(6), 1836-1840.

50. Lippmann, R. P., Kukolich, L., & Shahian, D. (1995). Predicting the risk of complications in coronary artery bypass operations using neural networks. *In Advances in neural information processing systems* (pp. 1055-1062).

51. Lippmann, R. P., & Kukolich, L. (1995, April). Using neural networks to predict the risk of cardiac bypass operations. In *Applications and science of artificial neural networks* (Vol. 2492, pp. 651-661). International Society for Optics and Photonics.

52. Lippmann, R. P., & Shahian, D. M. (1997). Coronary artery bypass risk prediction using neural networks. *The Annals of thoracic surgery*, 63(6), 1635-1643.

53. Silber, J. H. (1995). Report cards on cardiac surgeons. *The New England journal of medicine*, 333(14), 938.

54. Grover, F. L., Hammermeister, K. E., & Shroyer, A. L. W. (1995). Quality initiatives and the power of the database: what they are and how they run. *The Annals of thoracic surgery*, 60(5), 1514-1521.

55. Parsonnet, V. (1995). Risk stratification in cardiac surgery: is it worthwhile?. *Journal of cardiac surgery*, 10(6), 690-698.

56. Li, Z., Amstrong, E. J., Parker, J. P., Danielsen, B., & Romano, P. S. (2012). Hospital variation in readmission after coronary artery bypass surgery in California. *Circulation: Cardiovascular Quality and Outcomes*, 5(5), 729-737.

57. Faraggi, D., & Simon, R. (1995). A neural network model for survival data. *Statistics in medicine*, 14(1), 73-82.

58. Liestol, K., Andersen, P.K., Andersen, U., 1994. Survival analysis and neural nets. *Statist. Med.* 13, 1189-1200.

59. Buckley, J., James, I., 1979. Linear regression with censored data. *Biometrika* 66, 429436.

60. Xiang, A., Lapuerta, P., Ryutov, A., Buckley, J., & Azen, S. (2000). Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational statistics & data analysis*, 34(2), 243-257.

61. Wang, S., Fu, L., Yao, J., & Li, Y. (2018, May). The application of deep learning in biomedical informatics. In 2018 *International Conference on Robots & Intelligent System (ICRIS)* (pp. 391-394). IEEE.

62. Caroprese, L., Veltri, P., Vocaturo, E., & Zumpano, E. (2018, July). Deep learning techniques for electronic health record analysis. In 2018 9th *International Conference*

on Information, Intelligence, Systems and Applications (IISA) (pp. 1-4). IEEE.

63. Guru, V., Fremes, S., Austin, P., Blackstone, E., & Tu, J. (n.d.). Gender differences in outcomes after hospital discharge from coronary artery bypass grafting. *Circulation : Journal of the American Heart Association.*, 113(4), 507-516.

64. Vaccarino, V., Lin, Z., Kasl, S., Mattera, J., Roumanis, S., Abramson, J., & Krumholz, H. (2003). Sex differences in health status after coronary artery bypass surgery. *Circulation : Journal of the American Heart Association.*, 108(21), 2642-2647.

65. McLaughlin, M. (2013). *MySQL Workbench: Data Modeling & Development*. McGraw Hill Professional.

66. Team, P. C. (2015). Python: A dynamic, open source programming language. *Python Software Foundation*, 78.

67. Rahman, M. M., & Davis, D. N. (2013). Machine learning-based missing value imputation method for clinical datasets. In *IAENG Transactions on Engineering Technologies* (pp. 245-257). Springer, Dordrecht.

68. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2), 105-115.

69. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

70. Cai, Z., Heydari, M., & Lin, G. (2006). Iterated local least squares microarray missing value imputation. *Journal of bioinformatics and computational biology*, 4(05), 935-957.

71. Wang, Y., Cai, Z., Stothard, P., Moore, S., Goebel, R., Wang, L., & Lin, G. (2012). Fast accurate missing SNP genotype local imputation. *BMC research notes*, 5(1), 404.

72. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182.

73. Cai, Z., Goebel, R., Salavatipour, M. R., & Lin, G. (2007). Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. *BMC bioinformatics*, 8(1), 206.

74. Cai, Z., Zhang, T., & Wan, X. F. (2010). A computational framework for influenza antigenic cartography. *PLoS computational biology*, 6(10), e1000949.

75. Yang, K., Cai, Z., Li, J., & Lin, G. (2006). A stable gene selection in microarray data analysis. *BMC bioinformatics*, 7(1), 228.

76. Cai, Z., Ducatez, M. F., Yang, J., Zhang, T., Long, L. P., Boon, A. C., ... & Wan, X. F. (2012). Identifying antigenicity-associated sites in highly pathogenic H5N1 influenza

virus hemagglutinin by using sparse learning. *Journal of molecular biology*, 422(1), 145-155.

77. Wang, Y., & Ni, X. S. (2019). A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *arXiv preprint arXiv:1901.08433*.

78. Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282), 457-481.

79. Lee, M.-L. T., and Whitmore, G. A. 2006. Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statistical Science* 21(4):501–513.

80. Doksum, K. A., and Hyland, A. 1992. Models for variable stress accelerated life testing experiments based on wiener processes and the inverse gaussian distribution. *American Statistical Association and American Society for Quality* 34(1):74–82.

81. Longini, I. M.; Clark, W. S.; Byers, R. H.; Ward, J. W.; Darrow, W. W.; Lemp, G. F.; and Hethcote, H. W. 1989. Statistical analysis of the stages of hiv infection using a markov model. *Statistics in Medicine* 8(7):831–843.

82. Fine, J. P., and Gray, R. J. 1999. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94(446):496–509.

83. Ishwaran, H., & Kogalur, U. B. (2014). RandomForestSRC: Random forests for survival, regression and classification (RF-SRC). *R package version*, 1(0).

84. R. Ranganath, A. Perotte, N. E., and Blei, D. 2016. Deep survival analysis. *arXiv preprint arXiv:1608.02158*.

85. Yu, C. N.; Greiner, R.; Lin, H. C.; and Baracos, V. 2011. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Proceedings of the 24th Conference on Neural Information Processing Systems (NIPS 2011)*.

86. Fernndez, T.; Rivera, N.; and Teh, Y. W. 2016. Gaussian processes for survival analysis. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016)*.

87. Alaa, A. M., and van der Schaar, M. 2017. Deep multi-task gaussian processes for survival analysis with competing risks. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017)*.

88. Faraggi, D., and Simon, R. 1995. A neural network model for survival data. *Statistics in Medicine* 14:73–82.

89. Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2016). Deep survival: A deep cox proportional hazards network. *stat*, 1050, 2.

90. Luck, M.; Sylvain, T.; Cardinal, H.; Lodi, A.; and Bengio, Y. 2017. Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245*.

91. Lee, C., Zame, W. R., Yoon, J., & van der Schaar, M. (2018, April). Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI*

Conference on Artificial Intelligence.

92. Xiang, A., Lapuerta, P., Ryutov, A., Buckley, J., & Azen, S. (2000). Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational statistics & data analysis*, 34(2), 243-257.

93. Gao, D., Grunwald, G., Rumsfeld, J., Schooley, L., MacKenzie, T., & Shroyer, A. (n.d.). Time-varying risk factors for long-term mortality after coronary artery bypass graft surgery. *The Annals of Thoracic Surgery*, 81(3), 793-799.

94. Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 24.

95. Katzman J., *DeepSurv*. (2016). Retrieved January 15, 2018 from https://github.com /jaredleekatzman/DeepSurv.

96. Giunchiglia, E., Nemchenko, A., & van der Schaar, M. (2018, October). RNN-SURV: A Deep Recurrent Model for Survival Analysis. In *International Conference on Artificial Neural Networks* (pp. 23-32). Springer, Cham.

97. Huang, C., Zhang, A., & Xiao, G. (2018). Deep Integrative Analysis for Survival Prediction. *Pac Symp Biocomput*, 23, 343-352.

98. Li, Y., Wang, L., Wang, J., Ye, J., & Reddy, C. K. (2016, December). Transfer learning for survival analysis via efficient L2, 1-norm regularized Cox regression. In *2016 IEEE*

16th International Conference on Data Mining (ICDM) (pp. 231-240). IEEE.

APPENDICES

A Appendix

A.1 Appendix

Table 1. List of demographics and preoperative variables (time-independent) considered for the study cohort

Variable(s)	Data type and range of values at source database	Rescaled and normalized values for the model [Comments]
Demographics and Preoperative:	Categorical and continuous variables as listed below	Rescaled and Normalized to fit well with the model
Race	White/Caucasian - 1, American Indian/Alaskan Native -2, Native Hawaian/Pacific Islander - 3, Black/Afrcian American - 4, Hispanic or Latino or Spanish - 5, Asian - 6, Other - 7	1 (White/Caucasian/Alaskan/ Hawaian/Native Indian, Pacific Islander), 2 (Asian), 3 (Hispanic), 4 (African American), 5 (Others)
Age	25 to 88 years	Unchanged
Gender	Male - 1; Female -2	0 (Male); 1 (Female)
Hospital	Hospital 1, Hospital 2, Hospital 3	1 (Hospital 1), 2 (Hospital 2), 3 (Hospital 3)
Insurance	Medicaid, Medicare, Private - HMO/BlueCross/Commercial, Others	0 (Private/Others), 1 (Medicare), 2 (Medicaid)
Admission Status	Elective - 1,Urgent - 2, Emergent - 3, Emergent Salvage - 4	0 (Elective), 1 (Urgent), [No data available for emergent cases in the study cohort]
Body Mass Index (in Kg/m^2)	16.3 - 201.5	Unchanged
Alcohol Use	<= 1 drink/week - 1, 2-7 drinks/week - 2 >= 8 drinks/week - 3, None - 4, Unknown - 5	0 (None/Unknown), 1 (<= 1 drink/week) 2 (2-7 drinks/week), 3(>= 8 drinks/week)
Chronic Lung Disease	No -1, Mild -2, Moderate -3, Severe - 4, Severity Unknown - 5	0 (No), 1 (Mild), 2 (Moderate), 3 (Severe) [Values with 'Severity Unknown' are merged with mode value 'Mild')
Home Oxygen	No-2, Yes, PRN - 2, Yes, oxygen dependent -3, Unknown - 5	0 (No/Unknown), 1 (Yes, PRN, Yes, Oxygen dependent)
Tobacco Use	Never smoker - 1, Current everyday smoker -2, Current some day smoker - 3, Former smoker - 5, Smoking status unknown - 6	0 (No/Unknown), 1 (Former), 2 (Current, someday), 3 (Current, everyday)
Total Albumin (in g/dL)	1.9 to 9.0	
Total Bilirubin (in mg/dL)	0.1 to 4.5	
Cerebrovascular Disease Diabetes Dyslipidemia Hypertension Illicit Drug Use Immunocompromise Peripheral Arterial Disease Sleep Apnea	Yes -1, No - 2, Unknown - 3	0 (No/Unknown), 1 (Yes), [In this study cohort, data with 'unknown' values are < 1%, and are assigned with mode value '0']

Table 2. List of time-independent preoperative and intraoperative variables considered for the study cohort

Variable	Type and range of values at source database	Rescaled and Normalized Values for the model
Demographics and Preoperative continued:	Categorical and continuous variables variables as listed below	Rescaled and Normalized to fit well with the model
Previous Cardiac Intervention	Yes -1, No - 2, Unknown - 3	0 (No/Unknown) 1 (Yes)
Cardiogenic Shock	Yes, at the time of the procedure - 3 Yes, not at the time of the procedure but within prior 24 hours - 4 No - 2	0 (NO) 1 (Yes, at the time of procedure) [No data for category '4' in the study cohort]
Congestive Heart Failure	Yes -1, No - 2, Unknown - 3	0 (No/Unknown), 1 (Yes)
Prior Myocardial Infarction	Yes -1, No - 2, Unknown - 3	0 (No/Unknown), 1 (Yes)
Hemodynamics- Ejection Fraction	Yes -1 , No - 2	0 (No), 1 (Yes)
Anginal Classification within 2 weeks	CCS Class 0 - 1, CCS Class I - 2, CCS Class II - 3 CCS Class III - 4, CCS Class IV -5	0 (CCS Class 0), 1 (CCS Class I), 2 (CCS Class II), 3 (CCS Class III), 4 (CCS Class IV)
Cardiac Arrhythmia	Yes -1 , No - 2	0 (No),1 (Yes)
Intraoperative Variables:	Categorical and continuous variables variables as listed below	Rescaled and Normalized to fit well with the model
Cardiopulmonary Bypass(CPB) Utilization	None - 1,Combination - 2,Full - 3	0 (None),1 (Combination), 2 (Full)
Number of Diseased Coronary Vessel Systems	None- 1, One - 2, Two - 3, Three - 4	0 (None), 1 (One), 2 (Two), 3 (Three)
Blood Products	Yes -1 , No - 2	0 (No), 1 (Yes)
Number of Distal Anastomoses with Venous Conduits	Numerical value in the range of 0 to 5	Unchanged
Number of Distal Anastomoses with Arterial Conduits	Numerical value in the range of 1 to 6	Unchanged
Intra-Aortic Balloon Pump (IABP)	Yes -1 , No - 2	0 (No/Unknown), 1 (Yes)
Cardiopulmonary Bypass Used	Yes -1 , No - 2	0 (No), 1 (Yes)
Operative Approach	Full conventional sternotomy -1, Partial sternotomy -2, Limited (mini) thoracotomy, left -12, Port access -16	The study cohort consists of 4 types operative approaches listed.
Total OR Time (in minutes)	195 to 702	Unchanged

Table 3. List of time-independent postoperative variables considered for the study cohort

Variable	Type and range of values at source database	Rescaled and Normalized Values for the model
Postoperative:	Categorical and continuous variables variables as listed below	Rescaled and Normalized to fit well with the model
Length of Stay (in days)	2 to 77	Unchanged
Total ICU Time (in hours)	3.2 to 1307.6	Unchanged
Ventilation Time (in hours)	0 to 1308	Unchanged
Sternal Superficial Wound Infection	Yes, within 30 days of procedure - 3 Yes, >30 days after the procedure, but during hospitalization for surgery - 4, No - 2	0 (No), 1 (Yes)
Neuro Stroke	Yes, hemorrhagic - 3, Yes, ischemic - 4 Yes, undetermined type - 5, No -2	0 (No), 1 (Yes, hemorrhagic), 2 (Yes, ischemic), 3 (Yes, undetermined type)
Discharge_Location	Home - 1, Extended Care/Transitional Care Unit/Rehab - 2, Other acute care hospital -3, Nursing Home - 4, Hospice - 5, Left AMA - 6, Other - 777	0 (Home), 1 (Non-home)
Blood Products In Hospital Postoperative Events Atrial Fibrillation Pulmonary Ventilation Prolonged Other Cardiac Arrest Reoperation for Mediastinal Bleeding Reoperation for Other Cardiac Reasons Return To OR For Other Non-cardiac Reason Rhythm Disturbance Requiring Perm Device Sepsis Sepsis Positive Blood Cultures Surgical Site Infection Pneumonia Renal Failure Renal Dialysis Required Thrombosis in a Deep Vein Dialysis Required After Discharge Heparin Induced Thrombocytopenia,	Yes -1, No - 2	0 (No), 1 (Yes)
Discharge on Lipid Lowering Statin Discharge on Aspirin Discharge on Beta Blockers Discharge on Lipid Lowering other than Statin	Yes -1, No - 2, Contraindicated -3	0 (No, 1 (Yes), 2 (Contraindicated)

Table 4. List of time-dependent pre- and peri-operative variables considered for the study cohort

Variable name	Type and range of values at source database
Preoperative Hemoglobin A1c Level (%)	4.1 to 16
Preoperative Creatinine (mg/dL)	0.37 to 16.28
Preoperative Hematocrit (%)	13.7 to 54.9
Preoperative WBC count	0.97 to 29.7
Preoperative Glucose (mg/dL)	61 to 355
Highest Intraoperative Glucose (mg/dL)	86 to 377
Lowest Intraoperative Hemoglobin	4.6 to 34.5
Postoperative Creatinine (mg/dL)	0.4 to 15.6
Postoperative Glucose (mg/dL)	81 to 331
Postoperative Hemoglobin (g/dL)	6.1 to 15.5
Postoperative Hematocrit (%)	19.3 to 45.6
Postoperative Systolic Blood Pressure (mm Hg)	77 to 196
Postoperative Diastolic Blood Pressure (mm Hg)	7 to 116
Postoperative Heart Rate	47 to 173

Table 5. Illustration of feature value processing in our proposed framework

Variable	Categorical String values - assigned numerical values @ source	Normalized and/or Rescaled values	Comment
Gender	Male - 1 Female -2	0 (Male) 1 (Female)	Rescaled to 0 and 1
Race	White/Caucasian - 1 American_Indian_Alaskan_Native - 2 Native_Hawaiian_Pacific_Islander - 3 Black_African_American -4 Hispanic_Latino_Spanish -5 Asian -6 Other -7	1 (White/Caucasian/Alaskan, Hawaiian, Native Indian) 2 (Asian) 3 (Hispanic) 4 (African American) 5 (Others)	Based on the data distribution, combined all USA-originated into one category. And it resulted into a significant factor with better 'p' values
Cerebrovascular Disease	Yes - 1 No - 2 Unknown - 3	0 (No) 1 (Yes)	Blanks and Unknowns (< 1%) are assigned to 'mode' values '0'
Chronic Lung Disease	No -1 Mild -2 Moderate -3 Severe - 4 Severity unknown -5	0 (No) 1 (Mild) 2 (Moderate) 3 (Severe)	Blanks are considered as 'No' (< 1%); 'Severity Unknown' values are merged into 'Mild - most frequent case.
Cardiopulmonary Bypass Utilization	None -1 Combination -2 Full -3	0 (None) 1 (Combination) 2 (Full)	
Admission Status	Elective -1 Urgent -2 Emergent -3	0 (Elective) 1 (Urgent)	No data available for Emergent or Emergent salvage categories.
Discharge Location	Home -1 Extended Care/Transitional Care /Unit/Rehab -2 Other acute care hospital -3 Nursing Home - 4 Hospice - 5 Left AMA - 6 Other - 777	0 (Home) 1 (Non-home)	Home or Non-home feature played pivotal role on readmission event in previous studies
Alcohol Use	<= 1 drink/week -1 2-7 drinks/week -2 >= 8 drinks/week -3 None - 4 Unknown -5	0 (None) 1 (<=1 drink) 2 (2-7 drinks) 3 (>=8 drinks)	Entries such as Unknowns, blanks (< 1%) are assigned to mode value '0'.

Figure 3 Picture taken at the Three Minute Thesis (3MT) Competition at Georgia State University, Spring, 2019: Receiving 2nd prize from Jeff Steely (Dean of Libraries, GSU), and Lisa Armistead, Associate Provost of Graduate Programs, GSU

Figure 4 Picture taken at the Three Minute Thesis (3MT) Competition at Georgia State University, Spring, 2019: with Dr. Yanqing Zhang, Advisor

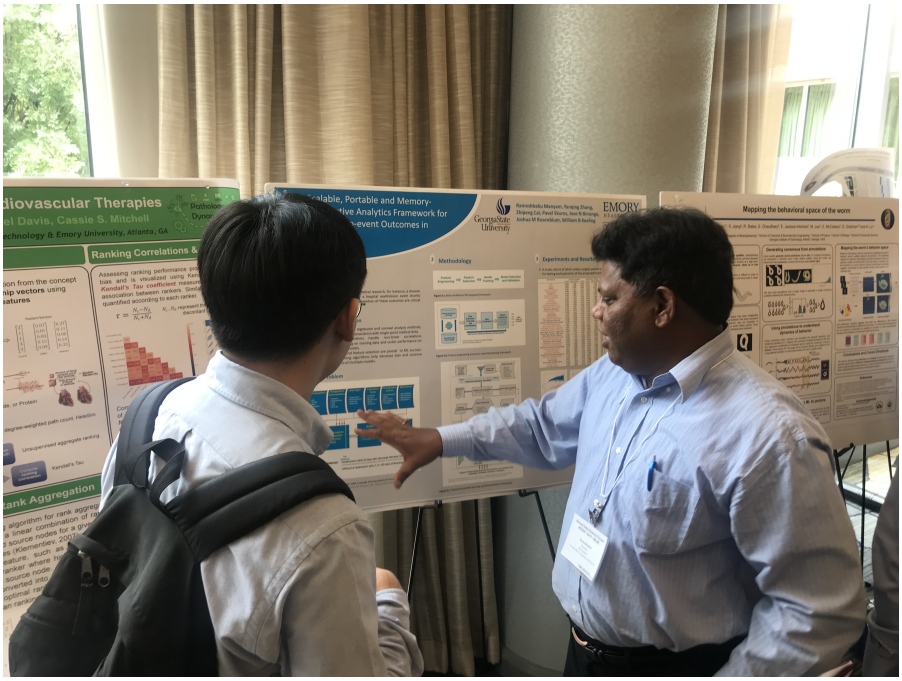


Figure 5 Picture taken at the 2nd Machine Learning in Science and Engineering Symposium, June 2019: Poster session

Figure 6 Photo taken at the 45th Annual Meeting of Western Thoracic and Surgical Association, Olympic Valley, CA, USA, June 2019

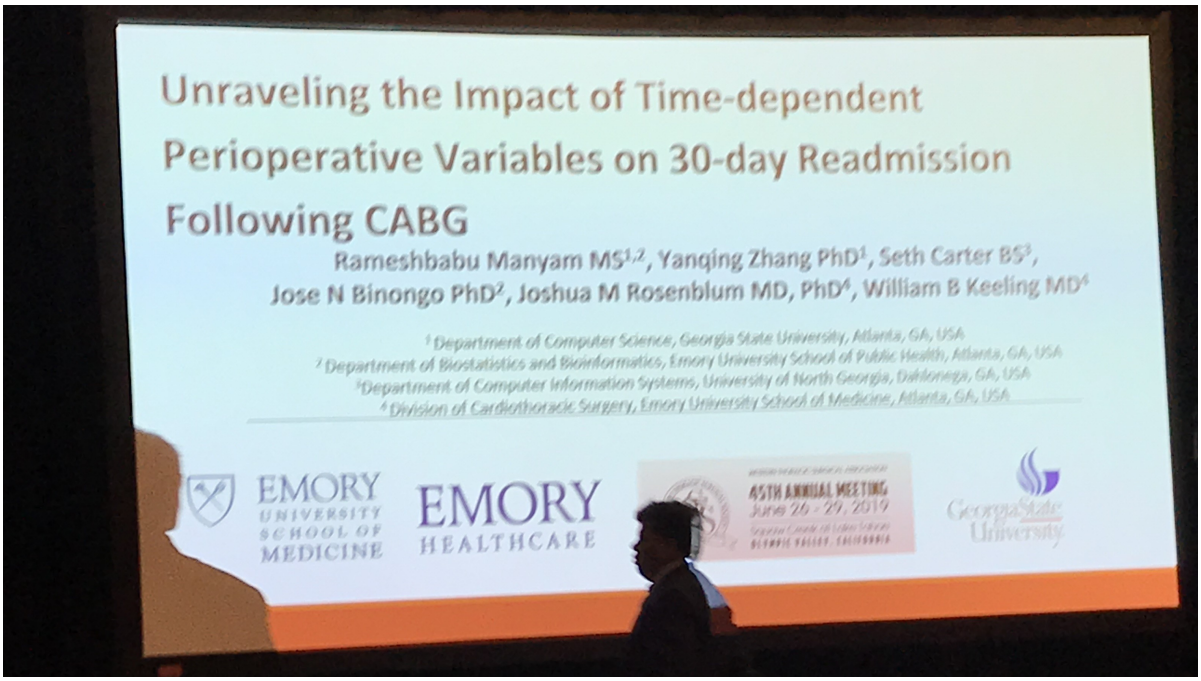


Figure 7 Photo taken at the 45th Annual Meeting of Western Thoracic and Surgical Association, Olympic Valley, CA, USA, June 2019

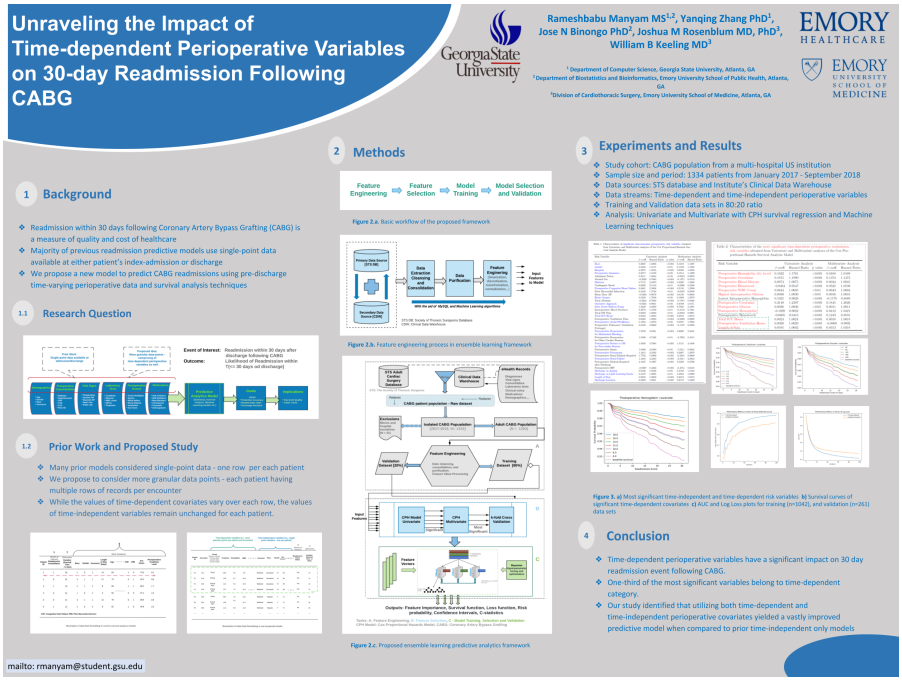


Figure 8 Poster Presented at the 45th Annual Meeting of Western Thoracic and Surgical Association, Olympic Valley, CA, USA, June 2019