Spring 5-18-2020

Emerging Technologies in Healthcare: Analysis of UNOS Data Through Machine Learning

Reyhan Merekar
CUNY Bernard M Baruch College

Emerging Technologies in Healthcare: Analysis of UNOS Data Through Machine Learning

by

Reyhan Merekar

Submitted to the Committee on Undergraduate Honors at Baruch College of the City University of New York in partial fulfillment of the requirements for the degree of Bachelor of Business Administration in Computer Information Systems with Honors

April 27th, 2020

Table of Contents

Acknowledgements

First and foremost, I would like to thank my Faculty Advisor, Professor Arturo Castellanos, for providing feedback and guidance throughout this project and most of my undergraduate career. We have worked on a few projects together, and it has been a privilege to learn from him throughout these four years. Additionally, I would like to thank my two Faculty Readers, Professors Kevin Craig and Zeda Li, for taking time to read my work and provide important feedback.

Special thanks to Claudio A. Bravo, Miguel Alvarez, and Mahek Shah for devoting time out of their busy schedules to provide valuable insight and context throughout this project.

Of course, I would like to thank my family for continuously supporting me through my undergraduate studies. All of this would not be possible without their belief in me.

Abstract

The healthcare industry is primed for a massive transformation in the coming decades due to emerging technologies such as Artificial Intelligence (AI) and Machine Learning. With a practical application to the UNOS (United Network of Organ Sharing) database, this Thesis seeks to investigate how Machine Learning and analytic methods may be used to predict one-year heart transplantation outcomes. This study also sought to improve on predictive performances from prior studies by analyzing both Donor and Recipient data. Models built with algorithms such as Stacking and Tree Boosting gave the highest performance, with AUC's of 0.6810 and 0.6804, respectively. In this work, a roadmap was created that justifies the need for these technologies in healthcare. In application, the data was prepared, models were built using advanced algorithms, and important variables were selected. These steps were continuously done with validation from experienced clinicians. To yield greater insights in this study, the dataset was split row-wise by factors such as LVAD Support, Donor/Recipient Gender Combinations, and Time Period; this rendered 8 new datasets for analysis. This work explores the trade-off between interpretability and performance in applying analytic methods in a real-world problem in this domain. Finally, forward looking industry implications are discussed.

Chapter 1: Introduction

Heart transplantation has been carried out since the 1970s, but still remains one of the riskiest procedures today. Formally, a heart transplant is defined as the surgical replacement of the heart of a diseased individual with that of a healthy donor (National Heart, Lung, and Blood Institute, "Heart Transplant"). Typically, patients who have end-stage heart failure, where the heart is severely damaged or weakened, undergo this procedure. Heart failure is caused by conditions such as coronary heart disease, hereditary conditions, and/or viral infections (National Heart, Lung, and Blood Institute, "Heart Transplant"). A patient in need of a heart transplant can locate donor organs through the United Network for Organ Sharing. This private, non-profit organization manages the United States' organ transplant system and provides a computerized national waiting list which assures equal access and fair distribution of organs as they become available (United Network for Organ Sharing, "About UNOS").

Several risk factors are associated with heart transplantation. The first is primary graft dysfunction (PGF), which occurs when the donor heart fails and is unable to function (National Heart, Lung, and Blood Institute, "Risk Factors"). This is an immediate issue and usually leads to a quick time of death for the patient. It is also a major contributor to mortality and additionally, may lead other complications (Iyer et al. 1-2). The patient's immune system may also reject the newly transplanted heart within the first six months of transplantation. To combat this, the patient must take additional medicine to suppress the immune system. Long term side effects associated with this medicine include diabetes, osteoporosis, and kidney damage (National Heart, Lung, and Blood Institute, "Risk Factors").

The presence of technology in healthcare, particularly data science, has begun to emerge within the past few decades. Professionals are beginning to explore the implications of using data to provide reliable solutions. Heart transplantation procedures are primed to increase in the coming years, which will be driven by the aging population in the United States. According to Primers in Medicine, the number of people older than 65 will "double by 2060" (Gedela et al. 19). Another key trend is heavy investment in AI and Machine Learning. As reported by Accenture, AI investment by healthcare firms will increase to $6B by 2021 (Collier et al. 2). As more people are at risk for end-stage heart failure and other diseases, AI and Machine Learning can be leveraged to increase predictive power for heart transplantation outcomes and disease detection.

This project aims to use Machine Learning[1] techniques to predict one-year heart transplantation outcomes using queried data from the UNOS registry from 1990-2016. It also strives to build on prior studies in this domain. Predictive models are constructed to help clinicians better understand underlying patterns in Donor and Recipient data. Since the underlying models will be able to predict the likelihood of a patient's survival after one year, they will allow the clinician(s) to take the appropriate course of action for treatment.

1. In this work, the terms Machine Learning and Data Mining are used interchangeably.

Chapter 2: Background & Additional Context

The imminence of technology within healthcare has grown rapidly and has been a driver of major industry shifts throughout the past decade. One example of such advancements is the emergence of AI, specifically, Machine Learning. Formally defined, it is the practice of using algorithms to build models that learn from any *n*-number of observations and try to emulate the underlying pattern of the phenomena (Beam and Kohane 1317). This concept allows the computer to autonomously make decisions without instructions from the researcher and is better suited for higher dimensional datasets where relationships may not necessarily be linear (Beam and Kohane 1317). The following trends emphasize why AI will be prevalent and contribute to the evolving landscape of healthcare/medicine.

Moving to Performance Based Medicine

The healthcare system in the United States is now moving toward a financial model based on value rather than volume. The onus is now on delivering excellent population health through treating patients like members. Rather than accounting for revenue due to patient volume, this value-based model shows each visit as an expense rather than a source of revenue (Burrill, "Health Care Outlook for 2019: Five Trends That Could Impact Health Plans, Hospitals, and Patients").

This shift will take time, as the current transition has not been entirely smooth. In the short term, healthcare firms may see financial hits before longer-term costs decrease. Despite this, the value-based model has been embraced as the best method in lowering healthcare costs while increasing the quality of care. Since patients are seeking the best care possible, Machine Learning can be a catalyst in providing that, helping people live healthier lives (NEJM Catalyst, "What is Value Based Healthcare?").

Virtual Visits and Wearables

The general population dreads visiting the doctor, and often, waits until a later date when the condition has become more severe to visit one. This mentality drives up costs for the patient. Virtual visits and telehealth serve as a basis to interact with a caregiver without attending the office. According to Steve Burrill of Deloitte, this technology helps them "see more patients, deal with rising clinical complexity, and support patients as they take a greater role in their own care." There is much room to leverage this practice, as currently, only 14% of caregivers are utilizing this (Burrill, "Health Care Outlook for 2019: Five Trends That Could Impact Health Plans, Hospitals, and Patients").

As the popularity of wearables (e.g., Apple Watch, FitBit) grow, so does the data they transmit. The Internet of Medical Things (IoMT) is the health spin-off of the Internet of Things. This phenomenon can be explained as the "collection of medical, drug delivery devices and applications that connect to healthcare IT systems through online computer networks" (D et al. 290). Medical devices equipped with Wi-Fi allow for machine-to-machine interactions. Such a phenomenon can help clinicians/healthcare professionals collect data points that may be used for disease prediction, patient status checks, and drug developments.

Leveraging Artificial Intelligence

Artificial Intelligence lies at the center of all of these trends. For example, in a virtual visit, software can be used to track a person's mood. Rather than meeting with someone, data from a patient's Electronic Health Record (EHR) can help manage illnesses. Data from wearables and tracking will be used to predict what a diagnosis may be, what drugs can be developed to help that will help the patient, and realistic timelines of treatment.

AI aims to mimic human cognition. Rather than fully replace doctors, it must be used in a way where professionals are working with AI to enhance clinical decisions. Now, with various analytic methods and collections of EHR's, this is becoming a reality. For example, it is already making waves in radiology, oncology, disease prediction/prevention, and outcome prediction with AI system IBM Watson. The system includes underlying Machine Learning models but is also a pioneer in the field. According to Jiang et al., "99% of the treatment recommendations from Watson are coherent with the physician decisions" (241). Several reviews have appeared in literature referencing similar analytic methods in healthcare; these have covered techniques, algorithms, and dataset evaluations. Additionally, research in this space is growing, as the number of published papers has increased by nearly 300% from 2008 to 2015 (Srivastava et al. 1665). This trend also contributes to the motivation for this study.

Emergence of Big Data in Healthcare

Big data has an incredible potential to yield significant value in healthcare. This will be driven by decreasing costs of data storage, access to powerful but remote cloud computing, proliferation of "smart" devices, and the increase in electronic communication. Take for example healthcare titan Kaiser Permanente, which consists of approximately nine million members. The firm has the capability to manage up to 44 petabytes of data through its EHR. This is "4,400 times the equivalent of the data stored in the Library of Congress" (Roski et al. 1115). This implies that there is a vast amount of data readily available for analysis in healthcare.

Big data is typically understood as a combination of three concepts – volume, velocity, and variety. Volume refers to the amount of data currently present in an enterprise; many experts assert that "90% of the data (stored)" has been created only over the last eight years (Sherman 4). Velocity measures the time sensitivity of data; reporting and analysis need to be immediate and

there is greater pressure to bridge the gap between when the data is acquired and when it is analyzed. Finally, variety is the idea that data is now collected from many different sources. It can be structured (e.g., tabular) or unstructured (e.g., emails, documents, PowerPoints). In this work, these opportunities are taken advantage of, and findings from prior studies are considered.

Chapter 3: Related Work

As mentioned previously, this project builds on prior studies in the field. Machine Learning and analytic methods have been explored not only for heart transplantation, but in other realms of clinical decision making as well. For example, a study was conducted that analyzed the risk of acute kidney injury (AKI), which is associated with chronic kidney disease and poses a high risk of mortality (Parreco and Chatoor 725). Another study used machine learning techniques to predict remission outcomes of Type 2 Diabetes following bariatric surgery (Johnston et al. 580).

Although these studies are not directly related to heart transplantation, they do have aspects in common. Those studies, along with this one, are Classification problems where the response variable is binary, the methodology is relatively similar, and the same evaluation method of Area Under the Curve (AUC) is utilized. Additionally, a common theme in Machine Learning studies in healthcare is the Logistic Regression model serves as a baseline for model comparison.

Moreover, directly related studies have been conducted with the same data in which Machine Learning was used to predict one-year mortality. One example of a similar study included a report from the Journal of Cardiac Failure, which compared results from traditional statistical techniques with more advanced techniques. This study employed six variables: age of recipient, creatinine, body mass index, liver function tests, aspartate transaminase, and hemodynamics. With the given variables, models were created with traditional statistical techniques and machine learning algorithms; these were all evaluated by the metric AUC. The implementation of Deep Learning models yielded the best AUC, which was roughly 0.66. At the end, it was deemed that the implementation of more advanced techniques failed to yield an

improved result when compared to traditional approaches, as there was a modest discrepancy between the two (Miller et al. 3-4). One possible drawback to this study was that the researchers only used the recipient data and did not emphasize the donor's data. The study was also only limited to univariate variables.

In another similar study, clinicians and data scientists utilized existing models to predict one-year mortality outcomes. The two models were IMPACT (Index for Mortality Prediction After Cardiac Transplantation) and IHTSA (International Heart Transplantation Survival Algorithm) which both implemented Machine Learning and Deep Learning techniques. The IMPACT model yielded an AUC of 0.608 with 18 variables being used, while the IHTSA model yielded an AUC of 0.64 with 32 variables (Medved et al. 3-6). One drawback to this study is that it was ambiguous as to which variables were used when each model was initially created.

Another similar study was carried out where analytic methods were leveraged to predict mortality outcomes, however, this study addressed a slightly different problem and differed in overall methodology. Various analytic techniques were still carried out, but this paper considered one, five, and nine-year mortality instead. To impute missing values for bias removal, the researchers used Synthetic Minority Oversampling TEchnique (SMOTE) which is used to "improve random oversampling" (Blagus and Lusa 2). This technique is well suited for lower dimensional data but is not ideal in higher dimensional settings. The best performance this study was able to uncover were AUC's of 0.624, 0.676, and 0.838 for one, five, and nine-year mortality, respectively (Dag et al. 47-49).

In this study however, Donor and Recipient data will be analyzed through advanced algorithms to improve overall predictive power in heart transplantation. Additionally, the data will be split in various ways to potentially yield further insights and variables are to be

eliminated from the full dataset for interpretability. This is further explained in following chapters.

Chapter 4: Data & Methodology

Data mining projects such as these have flexibility to be creative and dynamic. There is, however, an accepted framework known as CRISP-DM (**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) which is independent in industry and technology used. This framework aims to organize the steps in a data mining project. CRISP-DM has been seen as beneficial for organizations conducting large data mining projects in terms of cost, reliability, and manageability. This is the overarching methodology used for this project to keep a rigid structure and reduce errors. The relevant five elements of this framework are **Business Understanding, Data Understanding, Data Preparation, Model Building, and Evaluation.**

Figure 1. Phases of the CRISP-DM Model for Data Mining/Machine Learning. *CRISP-DM – a Standard Methodology to Ensure a Good Outcome - Data Science Central*. https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome.

Business Understanding

The initial phase of the framework involves defining the problem at hand solely from the perspective of the business. Following this, the problem is converted into one that can be solved by data mining (Wirth and Hipp 5). In this case, the project lies in the domain of healthcare/medicine, so that was kept in mind throughout the process. The clearly defined problem is to understand and predict one-year mortality after a heart transplant.

Data Understanding

The next phase of the framework is understanding the data. This begins with obtaining data and conducting elementary observations to pick up on any immediate trends. Here, data quality issues are also identified (Wirth and Hipp 5). The corresponding query from the UNOS database yields a 32018 by 558 data frame. Each row represents an individual patient, and each patient has descriptive variables. It was apparent that the quality of the data was not quite up to par; missing values were present which made conducting initial analyses difficult. The data types of some variables also were not correct.

Data Preparation

The data preparation phase spans all activities used to create the final data set; the one(s) that will be used in the model building process. This phase may be repeated many times in a given project. Some tasks include feature selection, data cleaning, and additional data transformation. In this project, the data had to be "cleaned" (e.g., adjusting data types, removing features, dealing with missing values) (Wirth and Hipp 5). One example of an adjusted data type was the response variable, "One_year_mortality_retransplant". This had to be changed to a "factor" data type. According to Python's pandas documentation, these data types take on a "limited, and usually fixed, number of possible values." Missing values were dealt with

differently depending on the algorithm used to build each model. Cleaning was validated with clinicians, as they provided more insight for understanding the variables. In this project, the target variable had categories “Died” or “Has not died” signified by the binary 1 and 0, respectively. In terms of feature selection, initially, many variables were omitted due to lack of variability cutting down the count to 420. Feature Selection was used again to reduce the number of variables to improve interpretability. This is further discussed at the end of the chapter.

Additionally, the data was iteratively prepared by splitting on factors such as time period of transplant, LVAD support, and Donor/Recipient gender combinations. With respect to time period, two new datasets were derived based on transplantation dates before and after 2006. For gender, each donor/recipient combination was analyzed, yielding four new datasets. The four combinations were male donor/male recipient, male donor/female recipient, female donor/male recipient, and female donor/female recipient. A Left Ventricular Assist Device (LVAD) is used when a patient is near heart failure but cannot acquire a transplant immediately. This device is installed to assume the role of providing blood to vital organs (Gedela et al. 19). Two datasets were created here to understand how the analyses would change depending on whether the recipient had an LVAD or not.

Model Building

In this phase, models are built and tested. Usually, there are several techniques that can be used to create models (Wirth and Hipp 6). The specific algorithms used for this study were Logistic Regression, Decision Trees/Random Forest, Tree Boosting, Neural Networks (Deep Learning), and Stacking. Logistic Regression is the most traditional algorithm used in medicine, so for the purposes of this project, it served as the baseline for comparison. All models were built using the H2O module in the Python programming language powered by Amazon Web Services.

Discussion of Relevant Algorithms

1. Logistic Regression

Logistic Regression is a widely used algorithm when the desired outcome is a question of classification. Rather than modeling a specific response directly, Logistic Regression models the probability that the desired outcome belongs in a particular category. As the algorithm deals with probability, the related sigmoidal function must only produce outputs between 0 and 1. Below is the general logistic function for a multivariate Logistic Regression (see Fig. B in Appendix for visual representation):

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}},$$

where $X = (X_1, X_2, \dots, X_p)$ are p predictors. In this work, the model derived from this algorithm served as a baseline for the performance of other "advanced" models (James et. al 130-133). With respect to models built with Logistic Regression, missing values were filled by mean imputation.

2. Decision Trees/Random Forest

Decision Tree is a hierarchical structure composed of branches and nodes. These are essentially a collection of 'if/else' statements that split decisions into binary classifications (in this case, "Died" or "Has not died"). Formally, this algorithm involves *stratifying* or *segmenting* the data set into a number of simple regions and making predictions (James et. al 306-312):

a. Divide the predictor space (the set of possible values for $X_1, X_2, \ldots, Xp$) into J distinct and non-overlapping regions, $R_1, R_2, \ldots, R_J$.
b. Create each binary partition based on *node purity*. This can be quantified by the Gini Index, where a *pure node* would produce a value closer to 0 ($\hat{p}_{mk}$ represents the proportion of the k_{th} classification in the m_{th} region):

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

c. For every observation that falls into the region Rj, the same prediction is made, which is simply the **mode** of the response values for the training observations in Rj.

Figure 2. Decision Tree Algorithm. James, Gareth, et al. An Introduction to Statistical Learning. *Springer New York, 2013. DOI.org (Crossref)*, pp. 306-312, doi:10.1007/978-1-4614-7138-7.

Although decision trees are relatively simple to interpret, their downfall lies in the fact that they are not robust. A small change in the data can yield quite a large change in the final estimated tree. To correct this problem, the Random Forest algorithm may be implemented, which is essentially a large collection of decision trees (Amornsamankul et al. 3). This algorithm corrects the instability of Decision Trees by bootstrap aggregating and decorrelating trees. This leads to stronger predictive power and lessens the risk of overfitting, which is touched on later in this chapter. With these models, missing values were filled by using the mode of respective variables.

3. Tree Boosting

Another way to combat Decision Tree instability is by using Boosting. Tree Boosting (e.g., Gradient Boosting Machine, Extreme Gradient Boosting) is widely used in machine learning to achieve optimal performance. It is an ensemble method that sequentially creates new members; the newest member is created to account for incorrectly labeled instances from previous learners to minimize the loss function (direct relationship to error). The results of new trees are then applied partially to the entire solution. The algorithm executes *M* boosting iterations to learn a function *F(x)* that outputs predictions $\hat{y} = F(x)$ while simultaneously minimizing a loss function $L(y, \hat{y})$. At each iteration, a new estimator $f(x)$ is added to correct the prediction of *y* for each instance in training. This is shown formally below:

a. Start with a function which approximates the true relationship of *x* and *y*:

$$F_{m+1}(x) = F_m(x) + f(x) = y$$
$$f(x) = y - F_m(x)$$

b. This fits the model $f(x)$ for the current boosting iteration to the errors above (difference of actual and predicted). This can be shown as a gradient descent algorithm when the loss function is the squared error:

$$L\big(y, F(x)\big) = \frac{1}{2}(y - F(x))^2$$

c. Let the summation of this loss function be denoted as *J*. The goal is to minimize *J* by adjusting $F(x_i)$, the function of a particular instance.

$$J = \sum_i L(y_i, F(x_i))$$
$$\frac{dJ}{dF(x_i)} = \frac{d(\sum_i L\big(y_i, F(x_i)\big))}{dF(x_i)} = F_m(x_i) - y_i$$

d. Thus, errors are equal to the negative gradient of the squared error loss function:

$$f(x) = y - F_m(x) = -\frac{d(\sum_i L\big(y_i, F(x_i)\big))}{dF(x_i)}$$

Figure 3. Tree Boosting Algorithm. Mitchell, Rory, and Eibe Frank. "Accelerating the XGBoost Algorithm Using GPU Computing." *PeerJ Computer Science*, vol. 3, July 2017, pp. 3-4. *DOI.org (Crossref)*, doi:10.7717/peerj-cs.127.

By adding a model that approximates this, the loss function is further minimized (Mitchell and Frank 3-4).

Often in the building phase, this algorithm will produce the strongest model in all facets of evaluation as it is inherently robust. Some of the drawbacks to this is that Tree Boosting tends to overfit but can also be corrected by adjusting tree sizes by pruning (Chen and Guestrin 3-5). With these models, missing values were filled by using the mode of respective variables.

4. Artificial Neural Networks (Deep Learning)

Neural Networks have recently been adapted as a viable data analytic method. These networks aim to emulate that of the human brain, as they contain "neurons" (linear or non-linear computing elements) interconnected in complex ways and organized in layers.

A simple perceptron[2] constructs a linear combination of the inputs called the net input. Thereafter, an activation function is linked to produce an output, this maps any real input to a bounded range. A functional link network introduces a hidden layer in the network. This uses nonlinear activation functions to produce a fully nonlinear model (in parameters). The resulting model is known as an MLP or multilayer perceptron. These models are flexible, general purpose, and non-linear and have the ability to yield multiple outputs from many inputs. Given enough data, this can approximate to a desired degree of accuracy. During building, numerical values were filled by using mean imputation, while categorical ones were omitted. An illustration of a sample neural network may be found in the Appendix (Fig. C) (Sarle 2-5).

5. Stacking

Stacking is an approach for building up classifier ensembles; this refers to a collection of classifiers in which their decisions are put together to classify new instances. The algorithm

2. A perceptron is a single-layer Neural Network. An interconnected system of perceptrons (MLP) yields a Neural Network.

combines multiple classifiers to induce a higher-level classifier with improved predictive performance (Sakkis et al. 1-2). In this study, two variants are used. “Best of Family” combines the best models from each algorithm, and “All Models” merges all models in a given iteration of building. Missing values were handled by the base algorithms of the ensemble.

Resampling Methods

Generally, in Machine Learning projects, the researcher splits a given dataset into a training and test set based on a chosen ratio. The training set is used to build models upon, while the test set acts as a proxy for how the model will perform on future, unseen data. This method, known as the Hold-Out Method, introduces additional error.

An important concept to understand in this domain is the *bias-variance tradeoff* (sometimes also referred to as the tradeoff between data-fit and complexity). The main point of model building is to showcase that it can generalize to unseen data. Maybe counter-intuitively, predictive performance is not maximized by learning the training data as precisely as possible. An extremely close fit to the training data is typically not ideal because the model will pick up random fluctuations in the data (i.e. noise) and miss the “broader regularities” in the dataset (Briscoe and Feldman 3-4). Using just the Hold-Out method can lead to high bias, low variance, or vice-versa. An illustration of the bias-variance tradeoff may be found in the Appendix (Fig. D).

Alternatively, resampling methods can be defined as iteratively fitting models on randomly drawn samples of given data. One of these methods, perhaps the most popular, is *k*-Fold Cross Validation. The dataset is divided into k folds (row-wise), where each $k - 1$ folds become training sets while the remaining fold acts as a test set. An error value is calculated for the “test” fold, and finally, the average of each test fold becomes the overall error value

(Rodriguez et al. 569). This was the resampling method of choice for this project, with $k = 5$. An illustration may be found below:

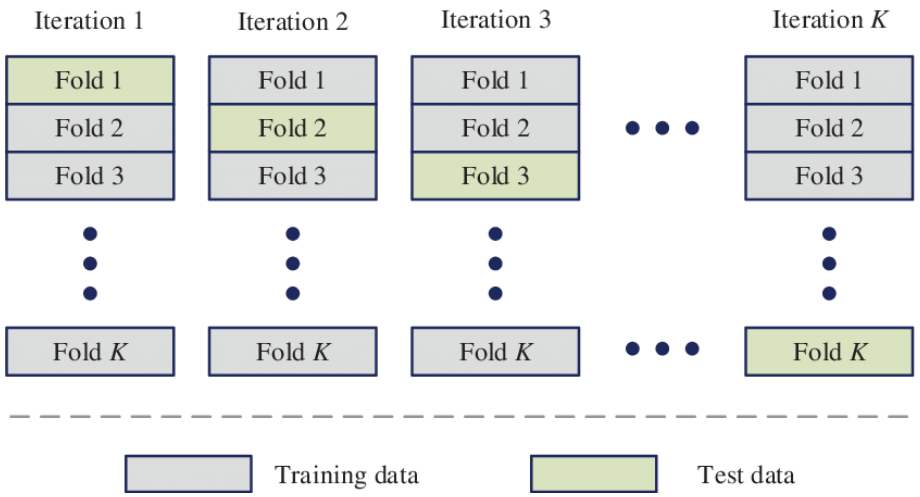

Figure 4. K-Fold Cross Validation. Ren, Qiubing, et al. "Tectonic Discrimination of Olivine in Basalt Using Data Mining Techniques Based on Major Elements: A Comparative Study from Multiple Perspectives." *Big Earth Data*, vol. 3, no. 1, Jan. 2019, p. 14. *DOI.org (Crossref)*, doi:10.1080/20964471.2019.1572452.

Compared to the simple Hold-Out method, *k*-Fold Cross Validation has advantages in both bias and variance. Bias is reduced in the sense that there are more observations "seen" by the model during training. Variance is also reduced due to the fact that there is smaller overlap between each training set. Given these considerations, it is empirically proven that $k = 5$ or $k = 10$ yield test errors that do not suffer from high bias nor very high variance (James et al. 181-184).

Model Evaluation

At this point in the framework, multiple models have been built and are ready to be compared. The models are constructed correctly and are assumed to be of the best quality possible. Before deployment of a model, it is vital to thoroughly evaluate each model, review steps taken to build the model, and justify it using the business context (Wirth and Hipp 6).

Confusion matrices are used in evaluation of problems of binary classification. It is a 2x2 table formed by counting the number of the four outcomes of a binary classifier. The two "classes" are actual and predicted with a positive/negative value for each. The four cells are filled with the number of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) (Amornsamankul et al. 5).

In this project, the outcomes of "Died" or "Has not died" are examined. True Positives in this case refer to the number of "Died" predicted when the actual value was also "Died." False Positives are the number of "Has not died" predicted when the outcome was actually "Died." False Negatives are the number of predicted "Died" when the actual was "Has not died." True Negatives are the number of predicted "Has not died" when the actual classification was "Has not died". In this case, False Negatives are more costly than False Positives, so the metric of Sensitivity (Recall) is also relevant in this study.

		Predicted Class	
		Died	Has not died
Actual Class	Died	**TP**	**FP**
	Has not died	**FN**	**TN**

Figure 5. One-year Mortality Confusion Matrix.

From this, prediction accuracy, sensitivity (recall), and specificity may be derived:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Key Metric: Area Under the Receiver Operating Characteristic Curve (AUC)

The Receiver Operating Characteristic Curve (ROC Curve) has been identified as a viable solution of visualizing a classifier's performance in order to select an optimal decision threshold. The term was coined during World War II when it was used in "Signal Detection Theory" for radars, but later made its way to diagnostic medicine. It was determined that an "ideal" threshold is almost always a trade-off between sensitivity (True Positives) and specificity (True Negatives). Since it may be difficult for a researcher to imagine an ideal "cut-off", this concept was visualized. The Y-Axis represents Sensitivity while the X-Axis represents Specificity. In theory, a researcher would want to achieve both high Sensitivity and Specificity, but this is far from practical in application. Hence, there is a trade-off, and either one of the two metrics can be optimized (Bradley 1145).

The area under the ROC Curve (referred to as AUC) is widely recognized as the measure of a diagnostic test's discriminatory power. The value of AUC ranges from 0.0 to 1.0, where a value of 1.0 implies 100% sensitivity and specificity. A value of 0.5 indicates no discriminative value. This entails 50% sensitivity and 50% specificity (Fan et al. 20). In terms of this project, ROC curves were visualized for each model, and ranked based on AUC. Models with AUC's that performed better than the baseline Logistic Regression were deemed the most useful.

Variable Analysis

As there are greater than 400 variables in the full dataset, only significant variables were considered in individual variable analyses. Each model produced a list of important variables and these were brought to clinicians for validation. Additionally, to further drill down the number of variables, the important variables from the top model from each dataset was taken. These were then counted up to understand the shared important variables throughout all models.

Chapter 5: Results

After the methodology was carried out, initial results were obtained for each model per dataset. These contain the top two models from each dataset, the baseline Logistic Regression, algorithms, AUC measures, Sensitivity, and dataset dimensions. A detailed table depicting all of these may be found below.

Table 1. Model Performance by Dataset and Algorithm.

Algorithm	**AUC**	**Sensitivity (Recall)**
Full Dataset (32,018 Observations, 420 Variables)		
Logistic Regression	0.63683332	0.4190646
Stacking, All Models	0.681168	*N/A*
Extreme Gradient Boosting	0.680403	0.41853034
Patients with no LVAD (10,912 Observations, 420 Variables)		
Logistic Regression	0.57672026	0.45374164
Stacking, All Models	0.65443	*N/A*
Extreme Gradient Boosting	0.653004	0.41827255
Patients with an LVAD (7,700 Observations, 420 Variables)		
Logistic Regression	0.58331451	0.5547305
Stacking, All Models	0.65692	*N/A*
Extreme Gradient Boosting	0.654315	0.43852264
Patients Recorded Before 2006 (13,406 Observations, 420 Variables)		
Logistic Regression	0.61300899	0.57854533
Stacking with All Models	0.674364	*N/A*
Extreme Gradient Boosting	0.67345	0.41904527
Patients Recorded After 2006 (18,612 Observations, 420 Variables)		
Logistic Regression	0.60626572	0.46481952
Stacking, All Models	0.668372	*N/A*
Extreme Gradient Boosting	0.66386	0.4589736

Female Donor Female Recipient (2,713 Observations, 420 Variables)		
Logistic Regression	0.54148685	0.5008384
Stacking, Best of Family	0.662545	*N/A*
Extreme Gradient Boosting	0.677088	0.53623605
Female Donor Male Recipient (2,730 Observations, 420 Variables)		
Logistic Regression	0.570355	0.47577724
Stacking, Best of Family	0.647494	*N/A*
Extreme Gradient Boosting	0.635834	0.42375335
Male Donor Female Recipient (2,713 Observations, 420 Variables)		
Logistic Regression	0.5310544	0.5507362
Stacking, All Models	0.651886	*N/A*
Extreme Gradient Boosting	0.664648	0.47671908
Male Donor Male Recipient (11,306 Observations, 420 Variables)		
Logistic Regression	0.57988709	0.45201996
Stacking, All Models	0.650807	*N/A*
Extreme Gradient Boosting	0.647327	0.41651163

After the initial analysis, a subset of the full dataset was taken. The data was filtered on the basis of important variables from the Extreme Gradient Boosting Algorithm on the full dataset and models were ran again. The results are displayed below:

Table 2. Model Performance: Full Dataset with Significant Variables and Algorithms.

Full Dataset with Significant Variables (32,018, 25)		
Algorithm	**AUC**	**Sensitivity (Recall)**
Logistic Regression	0.6603	0.4304224
Stacking, All Models	0.671926	*N/A*
Extreme Gradient Boosting	0.670706	0.4328194

Following that, more features were removed to make the model more interpretable. The top 16 variables were chosen based on important variables from the top model from each dataset. These were the significant variables that were most common throughout all models from all datasets. The models were run again, and the results are displayed below:

Table 3. Model Performance: Full Dataset with Clinically Significant Variables and Algorithms.

Full Dataset with Clinically Significant Variables (32,018, 16)		
Algorithm	**AUC**	**Sensitivity (Recall)**
Logistic Regression	0.6284	0.44540313
Stacking, Best of Family	0.6500	*N/A*
Extreme Gradient Boosting	0.6491	0.46050477

Clinically Significant Variables

Analysis of variables included validation with clinicians. They were able to provide additional context as to whether or not models included variables that made sense from a physiological standpoint. The top 16 variables included in the most interpretable model are displayed on the next page.

Table 4. Top 16 Clinically Significant Variables with Descriptions.

Top 16 Clinically Significant Variables with Descriptions	
Variable	**Description**
Creatinine of Recipient	Waste product filtered by kidney; a build-up may lead to cardiovascular disease.
Total Bilirubin of Recipient	Associated with liver complications; a strong predictor in heart failure.
Ischemic Time	The time an organ has spent cooling/warming before transplant.
Most Recent Creatinine Measurement of Recipient	Most recent evaluation of creatinine levels for a recipient at the time of listing.
Age of Donor	Median Age: 32 years old
Age of Recipient	Median Age: 55 years old
PVR of Recipient	Pulmonary vascular resistance for recipient measured at listing.
Right Ventricular Mass of Donor **Right Ventricular Mass of Recipient**	Both estimates of the myocardial mass of that ventricle based on echocardiographic measurements.
Systolic Pressure of Recipient	Pulmonary artery systolic pressure in mm/Hg at registration.
Predictive Heart Mass Ratio	Total heart myocardial mass estimated by echocardiogram of donor divided by that of the recipient.
Albumin Measurement of Recipient	Albumin measured in the recipient plasma at registration.
Distance	Distance between the donor and the recipient in nautical miles.
Cardiac Output of Recipient	Cardiac output of the recipient at registration.
Recipient Waiting Days	Total days on the UNOS waiting list.
GPT Measurement of Donor	GPT Level measured at time of transplant.

Chapter 6: Discussion & Insights

After analyzing all datasets, it was apparent that advanced methods such as Tree Boosting (Extreme Gradient Boosting) and Stacking led to higher AUC's than the baseline Logistic Regression. This is because Tree Based algorithms are usually more complex. Although they may be prone to overfitting, pruning and correctly adjusting the number of trees resolves this issue. Logistic Regression is simpler to understand and less prone to overfitting, however, may not always grant optimal predictive performance.

Running the Stacking algorithm on the full dataset yielded the highest AUC of about 0.6810. Although this is significant, Stacking is just a collection of models, and cannot be truly used for interpretation. This is also why there were no Sensitivity metrics for those models. Aside from Stacking, Boosting also gave high results. When applied to the full dataset, the model created from this algorithm had an AUC of roughly 0.6804. In practice however, it would not be feasible for a cardiologist to assess 420 parameters (variables) when trying to understand the future of a patient. This brings in the idea of real-world use where models must be *interpretable*. When the full dataset was cut down by column, 25 variables remained. These were the most important variables deemed by the Boosting model ran on the full dataset. All algorithms were run on that subset of data and the Stacking algorithm yielded an AUC of 0.6719. The Boosting algorithm was very similar in performance, with a 0.6707 AUC. After counting recurring variables from all model variable importance plots, analysis was reduced to 16 variables. The Stacking algorithm gave an AUC of .6500 and the Boosting algorithm gave an AUC of .6491. This would be far more interpretable for clinicians as the parameters were brought down to only 16 variables. Some performance is lost, but the difference is marginal (Fig. 6).

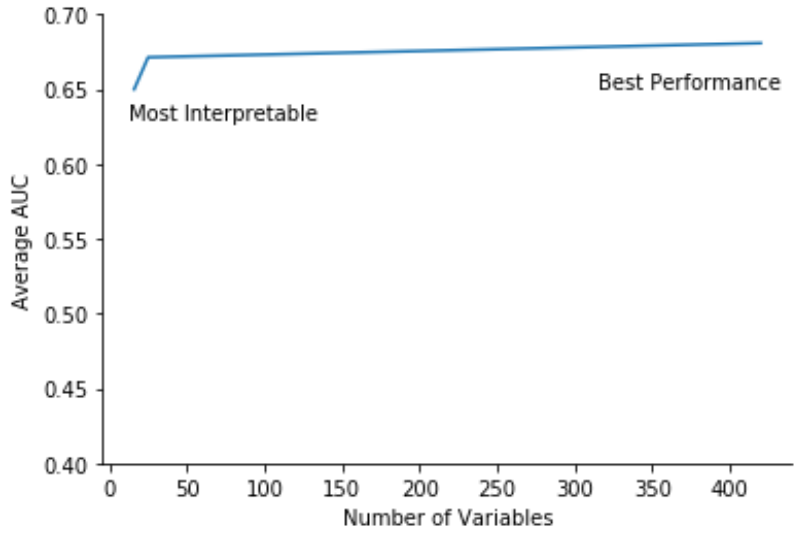

Figure 6. Visual Depiction of Trade-off Between Performance and Interpretability for Boosting Models. Average AUC on the Y-Axis corresponds to the average of the Top 2 models from the Full Dataset, Full Dataset with Significant Variables, and Full Dataset with Clinically Significant Variables.

The prior studies discussed in Chapter 3 did not achieve an AUC greater than 0.66 when using advanced learning techniques. Through optimal parameter tuning, the Tree Boosting algorithms were able to achieve higher performance. In those studies, however, the trend was leaning towards Deep Learning being a long-term solution in this field. In this study, running the Deep Learning algorithm on the full dataset only produced an AUC of 0.5706. This could have been again, due to data quality issues as neural networks do not perform well on rather sparse datasets. Even after cleaning, the data was not suitable to implement an effective Deep Learning model.

From a clinical standpoint, the aforementioned variables are reasonable to build models with going forward. Creatinine levels are often greater in patients that are hospitalized with heart failure, which validates this variable as a driver of mortality (Smith et al. 14). Liver function abnormalities such as high levels of Total Bilirubin are also associated with higher morality (van Deursen et al.). Age of Recipient is a good indicator of mortality; older people generally have poorer health. Ischemic time and Age of Donor are also reasonable because these represent

worse organ quality. Multiple studies have deemed that age is an independent risk factor for mortality and argued that longer ischemic times are associated with higher mortality outcomes (Kilic et al.).

Another point to analyze is the comparisons of the row-wise dataset splits. Comparing the LVAD support datasets, it was apparent that there was not much of a difference between performance of those top models. The same can be said for time period, but models run on observations before 2006 yielded a marginally higher AUC. For Donor/Recipient gender combinations, it was found that the dataset containing Female Donors and Female Recipients gave an AUC of 0.6770. This exceeded the performance of all other Donor/Recipient gender combinations.

Although the models ran on the row-wise splits did not perform better than models from the full dataset with respect to AUC, they are useful in terms of Sensitivity. Albeit not discussed as thoroughly as AUC, it is meaningful to analyze this metric in clinical decision making because in this case, False Negative Cases would be costly. The Tree Boosting model for the full dataset produced a Sensitivity of about .42. After running that algorithm on the Donor/Recipient gender combinations datasets, the average Sensitivity of the four was .463. Comparing those two metrics, Sensitivity of Donor/Recipient gender combinations saw an increase in Sensitivity by 8% on average. This supports the claim that Donor/Recipient gender combinations add value in clinical decision making (see Fig. H in Appendix).

Chapter 7: Conclusion

This project aimed to leverage Machine Learning techniques to predict one-year mortality after a heart transplant. Predictive models were constructed to help clinicians better understand underlying patterns in data from Donors and Recipients. This way, correct procedures for treatment may be taken from a medicinal standpoint. This study was continuously validated by clinicians, from Data Preparation to Evaluation. The data was split based on clinical factors such as LVAD Support, Donor/Recipient Gender Combinations, and Time Period of Transplant. Finally, specific features were selected based on the clinician's ability to interpret the created models.

As mentioned previously, performance and interpretability vary inversely. This trade-off is important to understand in deployment. From the analysis done in this work, the Extreme Gradient Boosting model applied to the full dataset with 25 variables is the optimal one. It serves as a "middle-ground" between AUC and number of variables. This model performed better than those from prior studies and uses 25 variables. Although it is not the most interpretable, the model will perform similarly when exposed to future, unseen data.

Essentially, there are two schools of thought. The data scientist wants to optimize model performance, while the clinicians strive for the simplest model to understand. To that end, from a clinical standpoint, the model with 16 variables may be the most optimal. There is a modest discrepancy in AUC when compared to the model with 25 variables, and theoretically, the clinician would trade marginal model performance for interpretability. The system needs to be fed with new data and using 16 parameters rather than 25 is more convenient, easier to track/measure, and simpler to enforce data integrity constraints upon.

Chapter 8: Implications

The last phase in the CRISP-DM Framework is **Deployment** of a selected model. Rarely, creation of the model is not the end of the project as it still needs to be rolled out on a system, maintained, and updated periodically (Wirth and Hipp 7). It is important to continuously validate models built throughout the framework. By doing this, data scientists will be able to gain insight on whether or not models fit the context of the problem.

Although there is a need for the AI in healthcare as highlighted previously, there are a few pertinent issues when it comes to deployment and industry-wide adoption of these systems.

Data Privacy

Data is essential for AI and model training, but some patients may be unwilling to give other entities access to private records. Besides for training, a large data supply is needed for validation and improvement of these models. For widespread deployment, this sensitive data must be shared among numerous institutions. To combat this, these EHR's must be anonymized and patients would need to be fairly informed. The shift to value-based care will support this, and further incentivize organizations to collect and ethically maintain this data for analysis (He et al. 31).

Data Standardization

From a data science standpoint, this is an important aspect to consider. Data standardization refers to the process of transforming data into a common format to be used for analysis. This way, it can be understood regardless of tools and methodologies (He et al. 33). In practice, data is collected in many different ways. It is stored in a variety of formats, databases, and information systems. Although this data may be formatted a certain way in one organization, if it is shared, another organization may not be able to properly interpret this for analysis. With

the complexity and volume of healthcare data in particular, this should occur in the initial phases of model development, even before the CRISP-DM Framework is carried out.

Existing Workforce

There has been significant concern throughout multiple industries of Artificial Intelligence eliminating the need for human workers. Although some jobs have potential to be automated, this will likely limit overall job loss. In healthcare, costs of automation technologies and regulatory and social acceptance are some reasons as to why this may be curbed (Davenport and Kalakota 96).

To overcome these aforementioned challenges, the workforce itself must understand that AI is not here to replace it. Instead, the workforce will be able to leverage it to augment existing workflows and decision making. For an industry-wide implementation, healthcare professionals must develop trust for these systems. Similarly, the onus is on patients to trust institutions to handle their data ethically in hopes of improving their own outcomes.

Chapter 9: Limitations & Future Directions

There are a few limitations this research has faced, the first being the time period of the queried data. The UNOS Data spans from 1990-2016 and does not take into account records after that. With the increased importance of data collection in the space, it is reasonable to assume that future records will follow specific constraints. This entails that the data will contain less errors, yielding a more thorough analysis. The second limitation would be the interpretation of the variables in the full dataset. These are generally subjective since there are over 400, so another set of clinicians could have understood these variables from a different perspective than in this study. Finally, although discussed thoroughly, it was impossible to complete the CRISP-DM Framework. The optimal model was limited to simulation and could not be deployed in a real use case.

For future studies, researchers will be able to leverage "cleaner" data from UNOS as data standards begin to conform. Aside from that, some questions are posed to future data scientist/clinician teams that have not been explored in this study:

Can mortality be predicted independent of time period?

This is particularly significant as this project was limited to one-year mortality. If this limitation was removed, clinicians would be able to understand what may contribute to shorter/longer mortality periods.

How robust will these models be? How will they scale?

Models must adapt to rapid change as new data is collected. Another point to examine would be identifying how these models would scale. This is useful from a data science perspective, and with the emergence of enterprise-wide data and cloud computing, analytic solutions (e.g. Random Forest, Tree Boosting) have potential to scale rather well.

Appendix

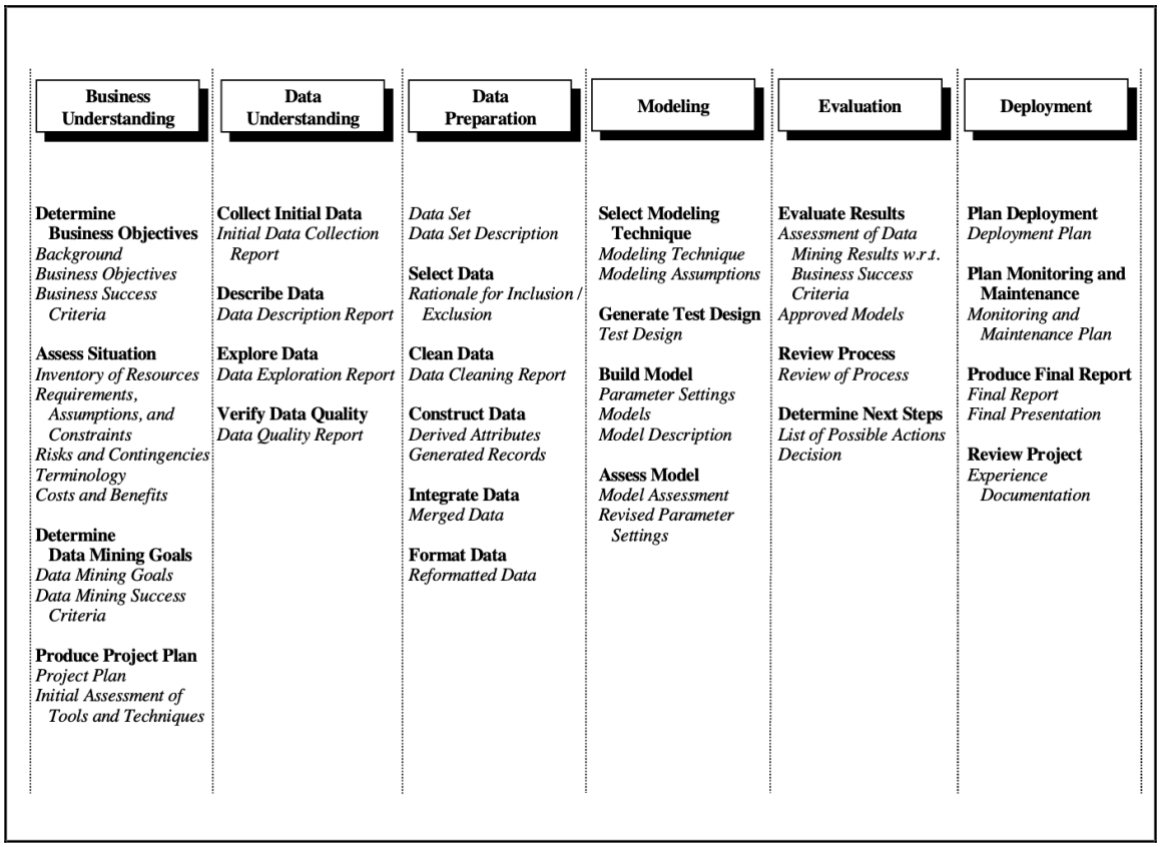

Figure A. Detailed Overview of the CRISP-DM Framework. Wirth, Rüdiger, and Jochen Hipp. *CRISP-DM: Towards a Standard Process Model for Data Mining*. p. 6.

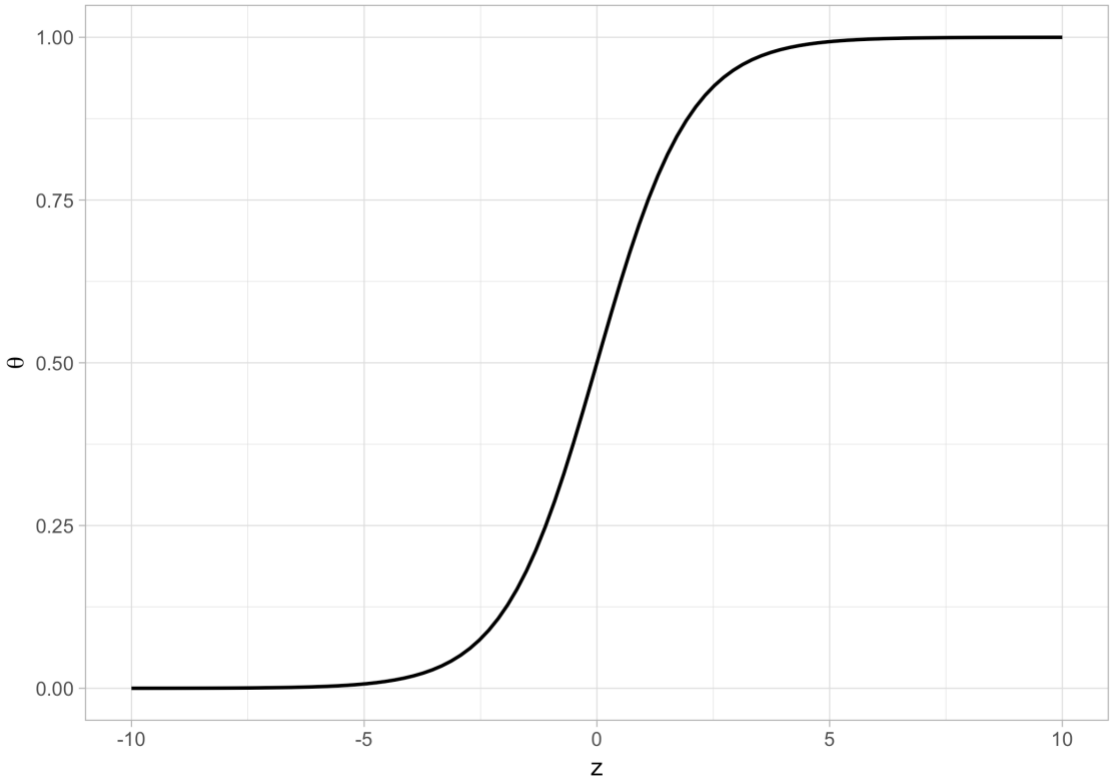

Figure B. Sigmoidal Function in Logistic Regression. Note, the minimum value is 0 and the maximum value is 1. *Logistic Regression Theory for Practitioners - Towards Data Science*. https://towardsdatascience.com/the-data-scientists-field-guide-to-logistic-regression-part-1-intuition-97084b11bd68. Accessed 16 Apr. 2020.

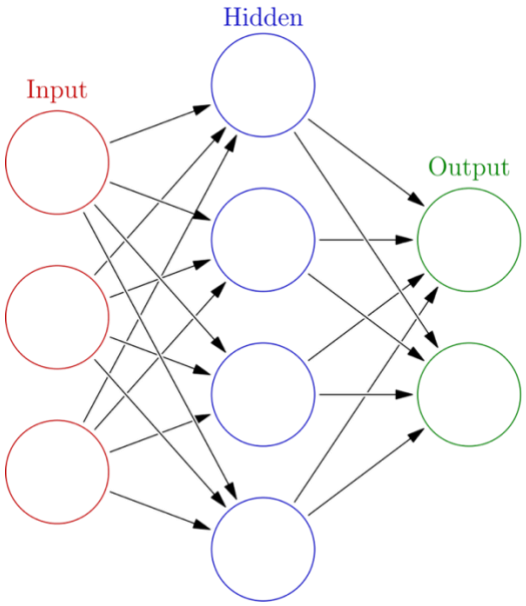

Figure C. Sample Artificial Neural Network. *Management AI: Types Of Machine Learning Systems*. https://www.forbes.com/sites/davidteich/2018/07/06/management-ai-types-of-machine-learning-systems/#390b5ba832fb. Accessed 16 Apr. 2020.

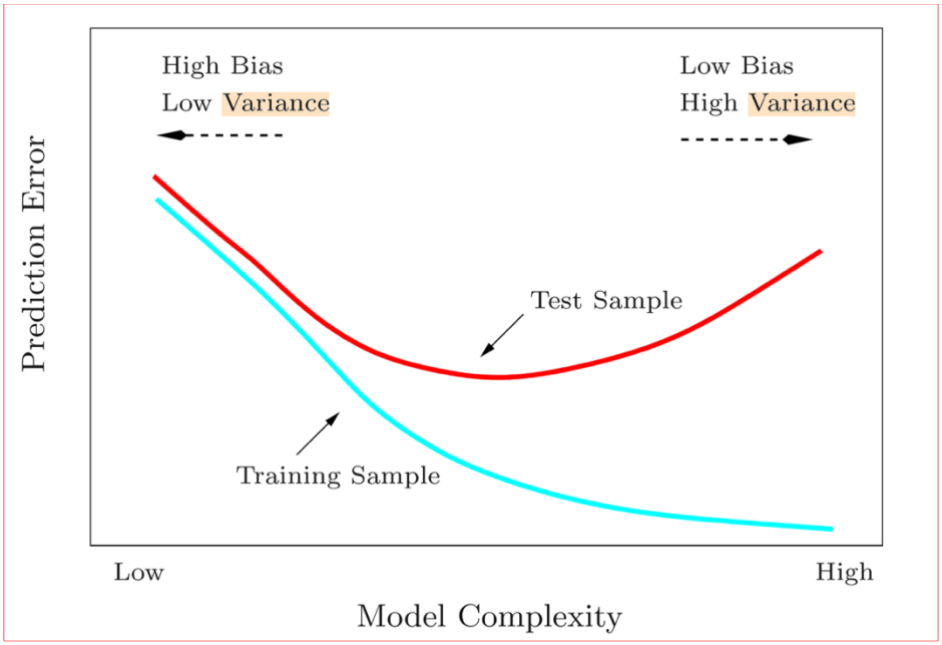

Figure D. Bias-Variance Tradeoff. *Bias and Variance in Machine Learning - Data Driven Investor - Medium.* https://medium.com/datadriveninvestor/bias-and-variance-in-machine-learning-51fdd38d1f86. Accessed 16 Apr. 2020.

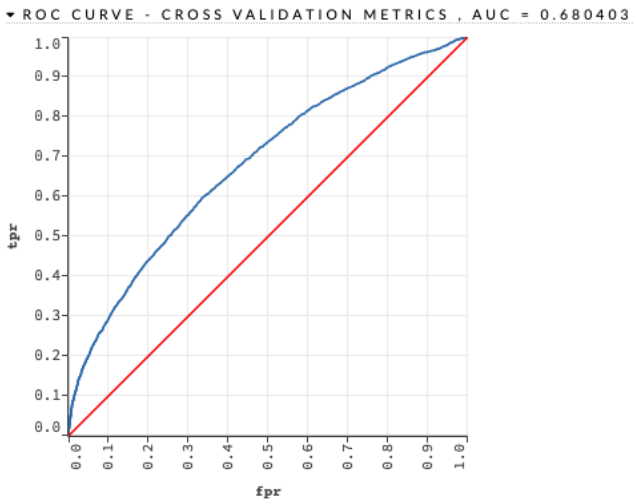

Figure E. ROC Curve from Extreme Gradient Boosting Model Applied to Full Dataset.

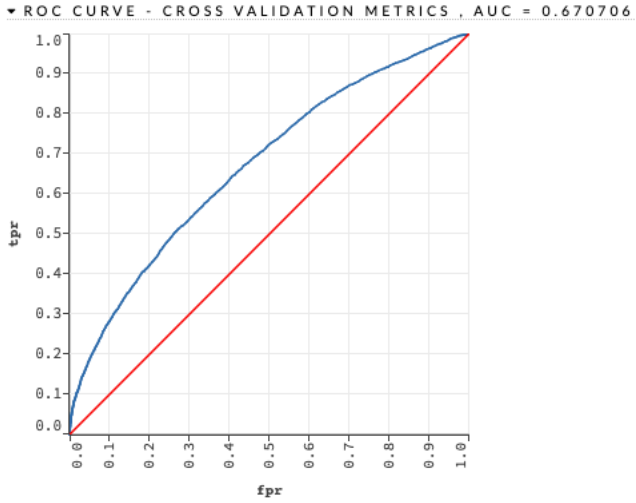

Figure F. ROC Curve from Extreme Gradient Boosting Model Applied to Full Dataset with Significant Variables.

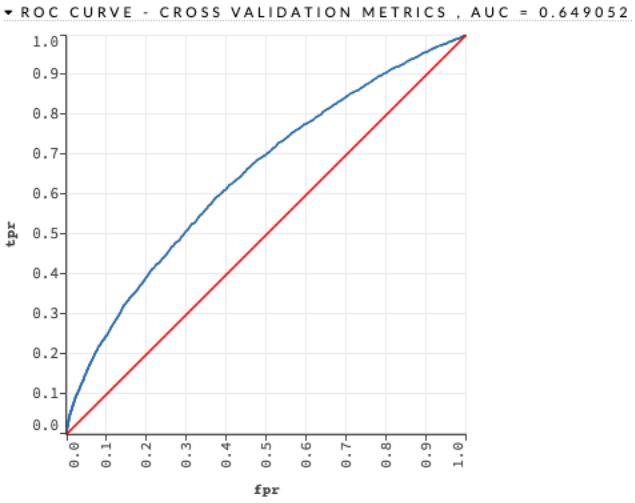

Figure G. ROC Curve from Extreme Gradient Boosting Model Applied to Full Dataset with Clinically Significant Variables.

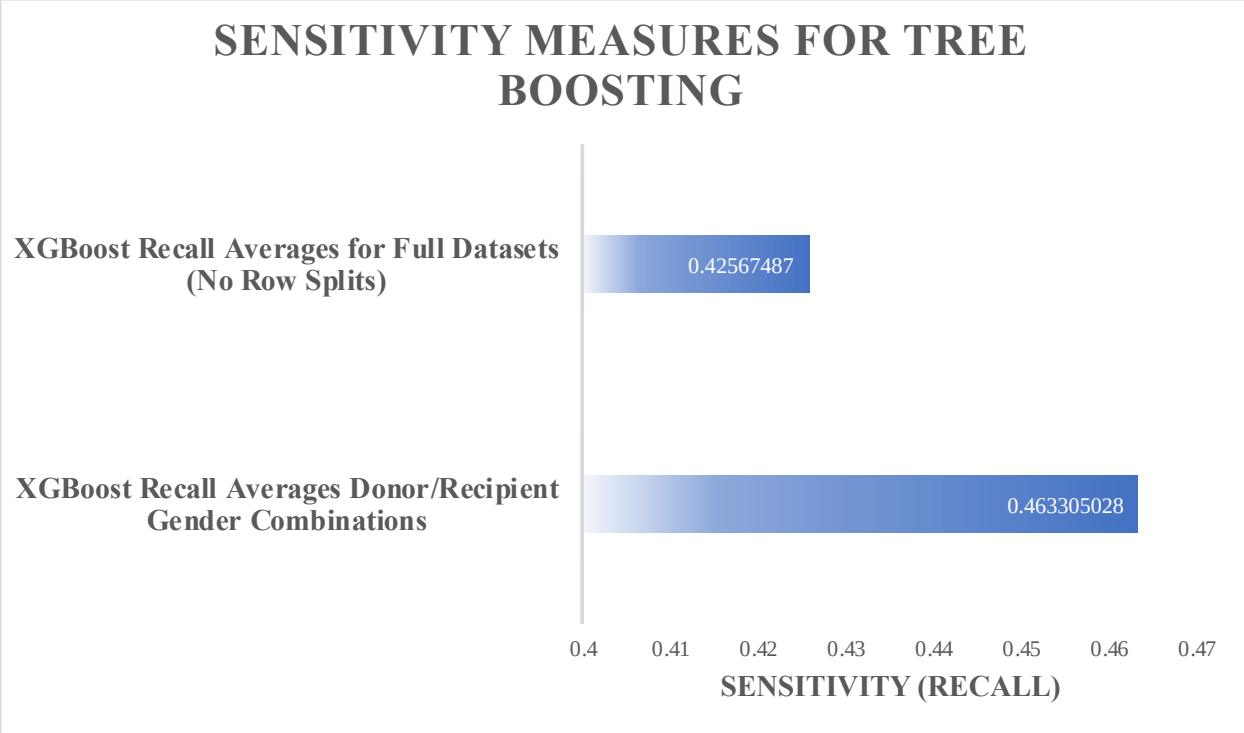

Figure H. Sensitivity Comparisons: Full Datasets and Donor/Recipient Gender Combinations.

References

Amornsamankul, Somkid, et al. “A Comparison of Machine Learning Algorithms and Their Applications.” *International Journal of Simulation: Systems, Science & Technology*, Aug. 2019, pp. 1–17. *DOI.org (Crossref)*, doi:10.5013/IJSSST.a.20.04.08.

Beam, Andrew L., and Isaac S. Kohane. “Big Data and Machine Learning in Health Care.” *JAMA*, vol. 319, no. 13, Apr. 2018, pp. 1317–18. *DOI.org (Crossref)*, doi:10.1001/jama.2017.18391.

Blagus, Rok, and Lara Lusa. “SMOTE for High-Dimensional Class-Imbalanced Data.” *BMC Bioinformatics*, vol. 14, no. 1, Dec. 2013, pp. 1–16. *DOI.org (Crossref)*, doi:10.1186/1471-2105-14-106.

Bradley, Andrew P. “The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms.” *Pattern Recognition*, vol. 30, no. 7, July 1997, pp. 1145–59. *DOI.org (Crossref)*, doi:10.1016/S0031-3203(96)00142-2.

Briscoe, Erica, and Jacob Feldman. “Conceptual Complexity and the Bias/Variance Tradeoff.” *Cognition*, vol. 118, no. 1, Jan. 2011, pp. 2–16. *DOI.org (Crossref)*, doi:10.1016/j.cognition.2010.10.004.

Burrill, Steve. “Health Care Outlook for 2019: Five Trends That Could Impact Health Plans, Hospitals, and Patients.” *Deloitte United States*. *www2.deloitte.com*, https://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/health-care-current-december4-2018.html. Accessed 2 Mar. 2020.

Chen, Tianqi, and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, 2016, pp. 785–94. *arXiv.org*, doi:10.1145/2939672.2939785.

Collier, Matt, et al. *Artificial Intelligence: Healthcare’s New Nervous System*. Industry Outlook Report, Accenture, pp. 1–8.

D, Shashank, et al. “The Internet of Medical Things (IoMT).” *Journal of Pharmaceutical Research*, vol. 16, no. 4, Dec. 2017, p. 290.

Dag, Ali, et al. “Predicting Heart Transplantation Outcomes through Data Analytics.” *Decision Support Systems*, vol. 94, Nov. 2016, pp. 42–52. *DOI.org (Crossref)*, doi:10.1016/j.dss.2016.10.005.

Davenport, Thomas, and Ravi Kalakota. “The Potential for Artificial Intelligence in Healthcare.” *Future Healthcare Journal*, vol. 6, 2019, pp. 94–98.

Fan, Jerome, et al. “Understanding Receiver Operating Characteristic (ROC) Curves.” *CJEM*, vol. 8, no. 01, Jan. 2006, pp. 19–20. *DOI.org (Crossref)*, doi:10.1017/S1481803500013336.

Gedela, Maheedhar, et al. *A Brief Review of Left Ventricular Assist Devices and Their Management*. pp. 19-26.

He, Jianxing, et al. "The Practical Implementation of Artificial Intelligence Technologies in Medicine." *Nature Medicine*, vol. 25, no. 1, Jan. 2019, pp. 30–36. *DOI.org (Crossref)*, doi:10.1038/s41591-018-0307-0.

Heart Failure | National Heart, Lung, and Blood Institute (NHLBI). https://www.nhlbi.nih.gov/health-topics/heart-failure. Accessed 1 Mar. 2020.

Heart Transplant | National Heart, Lung, and Blood Institute (NHLBI). https://www.nhlbi.nih.gov/health-topics/heart-transplant. Accessed 1 Mar. 2020.

Hong, Kimberly N., et al. "Who Is the High-Risk Recipient? Predicting Mortality After Heart Transplant Using Pretransplant Donor and Recipient Risk Factors." *The Annals of Thoracic Surgery*, vol. 92, no. 2, Aug. 2011, pp. 520–27. *DOI.org (Crossref)*, doi:10.1016/j.athoracsur.2011.02.086.

Iyer, Arjun, et al. "Primary Graft Failure after Heart Transplantation." *Journal of Transplantation*, vol. 2011, 2011, pp. 1–9. *DOI.org (Crossref)*, doi:10.1155/2011/175768.

James, Gareth, et al. *An Introduction to Statistical Learning*. Springer New York, 2013. *DOI.org (Crossref)*, doi:10.1007/978-1-4614-7138-7.

Jiang, Fei, et al. "Artificial Intelligence in Healthcare: Past, Present and Future." *Stroke and Vascular Neurology*, vol. 2, no. 4, Dec. 2017, pp. 230–43. *DOI.org (Crossref)*, doi:10.1136/svn-2017-000101.

Johnston, Stephen S., et al. "Using Machine Learning Applied to Real-World Healthcare Data for Predictive Analytics: An Applied Example in Bariatric Surgery." *Value in Health*, vol. 22, no. 5, May 2019, pp. 580–86. *DOI.org (Crossref)*, doi:10.1016/j.jval.2019.01.011.

Kilic, Ahmet, et al. "Donor Selection in Heart Transplantation." *Journal of Thoracic Disease*, vol. 6, no. 8, 2014, pp. 1097–104.

Medved, Dennis, et al. "Improving Prediction of Heart Transplantation Outcome Using Deep Learning Techniques." *Scientific Reports*, vol. 8, no. 1, Feb. 2018, pp. 1–9. *DOI.org (Crossref)*, doi:10.1038/s41598-018-21417-7.

Miller, P. Elliott, et al. "Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights From the UNOS Database." *Journal of Cardiac Failure*, vol. 25, no. 6, June 2019, pp. 479–83. *DOI.org (Crossref)*, doi:10.1016/j.cardfail.2019.01.018.

Mitchell, Rory, and Eibe Frank. "Accelerating the XGBoost Algorithm Using GPU Computing." *PeerJ Computer Science*, vol. 3, July 2017, pp. 1–28. *DOI.org (Crossref)*, doi:10.7717/peerj-cs.127.

NEJM Catalyst. “What Is Value-Based Healthcare?” *Catalyst Carryover*, vol. 3, no. 1, Massachusetts Medical Society, Jan. 2017. *catalyst.nejm.org (Atypon)*, doi:10.1056/CAT.17.0558.

Parreco, Joshua, and Matthew Chatoor. *Comparing Machine Learning Algorithms for Predicting Acute Kidney Injury*. no. 7, July 2019, pp. 725–29.

Rodriguez, J. D., et al. “Sensitivity Analysis of K-Fold Cross Validation in Prediction Error Estimation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, Mar. 2010, pp. 569–75. *DOI.org (Crossref)*, doi:10.1109/TPAMI.2009.187.

Roski, Joachim, et al. “Creating Value In Health Care Through Big Data: Opportunities And Policy Implications.” *Health Affairs*, vol. 33, no. 7, July 2014, pp. 1115–22. *DOI.org (Crossref)*, doi:10.1377/hlthaff.2014.0147.

Sakkis, Georgios, et al. *Stacking Classifiers for Anti-Spam Filtering of e-Mail*. 2001, pp. 1–7.

Sarle, Warren S. *Neural Networks and Statistical Models*. Apr. 1994, pp. 1–12.

Sherman, Rick. *Business Intelligence Guidebook: From Data Integration to Analytics*. Elsevier, Morgan Kaufmann, 2015.

Smith, Grace L., et al. “Worsening Renal Function: What Is a Clinically Meaningful Change in Creatinine during Hospitalization with Heart Failure?” *Journal of Cardiac Failure*, vol. 9, no. 1, Feb. 2003, pp. 13–25. *DOI.org (Crossref)*, doi:10.1054/jcaf.2003.3.

Srivastava, Siddharth, et al. “Deep Learning for Health Informatics: Recent Trends and Future Directions.” *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2017, pp. 1665–70. *DOI.org (Crossref)*, doi:10.1109/ICACCI.2017.8126082.

United Network for Organ Sharing | UNOS | US Organ Transplantation. https://unos.org/. Accessed 1 Mar. 2020.

van Deursen, V. M., et al. “Abnormal Liver Function in Relation to Hemodynamic Profile in Heart Failure Patients.” *Journal of Cardiac Failure*, vol. 16, no. 1, Jan. 2010, pp. 84–90. *DOI.org (Crossref)*, doi:10.1016/j.cardfail.2009.08.002.

Wirth, Rüdiger, and Jochen Hipp. *CRISP-DM: Towards a Standard Process Model for Data Mining*. pp. 1–10.