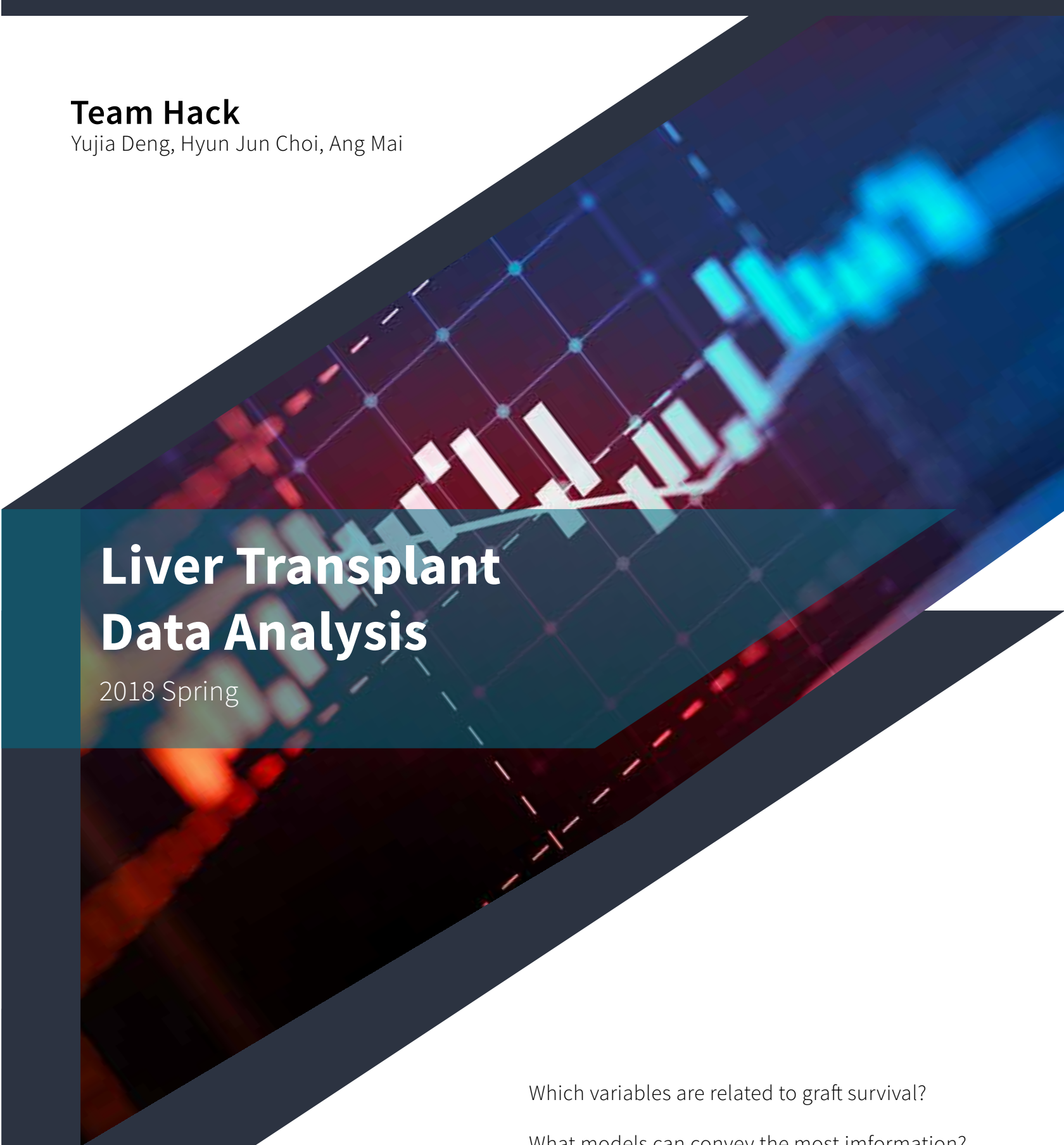
Team Hack
Yujia Deng, Hyun Jun Choi, Ang Mai



Liver Transplant Data Analysis

2018 Spring

Which variables are related to graft survival?

What models can convey the most imformation?

What prediction can we get?

Team Members: Personal Information

Name: Yujia Deng, Hyun Jun Choi, Ang Mai
Address: University of Southern California, Los Angeles, CA 90084
Email: yujiaden@usc.edu, choi797@usc.edu, angmai@usc.edu

Project Information

Project Title: Predict Graft Survival after OLT using Pre-Transplant Features
Date Started: 1/11/2018
Date Completed: 5/3/2018
Project Sponsor/Champion: Professor Ana Farzindar

Executive Summary

This project aims at the problem identifying features associated with graft survival after OLT.

In of the scope of the project includes data obtaining and studying, data analysis and deliverable creation and presentation to stakeholders. Out of the scope is data collection and data publication.

Project milestones were set and achieved each month, with the first phase of defining project goals, customer deliverables and other project details by using project charter and SIPOC diagram ending in early March, second phase of measuring performance by using process mapping and start developing statistical analysis ending in early April, third phase of analyzing causes of bad performing or overfitting models and comparing models ending in mid April, fourth phase of improving model performance by using other methods and developing other potential solutions for bad performing models ending in late April, and the fifth phase of controlling project performance by finalizing reports and PowerPoint presentation to stakeholders and finishing end product ending in early May.

The end product is a website containing information about data exploration and prediction, and it is tested multiple times and is only accepted when all functions work properly including contents built with Tableau and R. Analysis of the data lead to the conclusion that there are certain variables that are significantly associated with graft

survival while some are not, and the recommendation for patients/doctors to use the prediction tool for graft survival after a donor-recipient pair has been formed.

Tasks completed during the projects include obtaining and studying the dataset and background information, analyzing data consisting of data cleaning, using different approaches for analysis, evaluating performance of models, and transforming results into easy-to-understand messages through presentation and website. Classification models were trained and tested to measure performance while regression and survival analysis models were checked using performance measurement such as R^2, AIC and C-Index. After the comparison of the information conveyed by different models, survival analysis is selected for building the prediction tool. With the tool, patients/doctors will benefit from being able to predict graft survival for each new patient at all future time frame up to about 4000 days and therefore can make more informed decisions about the surgery accordingly.

Lean Six Sigma Project

Define Phase

Cost of Poor Quality Statement

According to the data from American Liver Foundation, each year in U.S, around 6,000 recipients have been benefited from liver transplant.[1] The number of liver transplants has been increasing annually, however, in contrast with the rapid increase of the liver demand, there is a severe shortage of liver donors and success of the surgery is not guaranteed. 8,082 liver transplants were performed in 2017 and 13,869 patients are still waiting for a liver as of Jan 2018, and graft failure accounts for about 20% of death in one month after the surgery from 2010 to 2015.[2]

Therefore, it is extremely important to undergo careful evaluations before deciding the operation of transplant to avoid futility and waste of resources for an unlikely successful surgery. The opportunity of the project is to explore the possibility of applying data analysis to help doctors and patients make more informed decision prior to orthotopic liver transplant.

Customer Satisfaction (Voice of the Customer)

A poor quality end product impacts customer's usage and therefore the value of the project. A website containing data exploration and prediction tools, the end product should be aesthetically viewable, conveniently accessible, and easily navigated.

Customers should also be granted authorizations to access the private contents on the website.

The voice of customer is tested during the demonstration of the end product and validated when the customer understands how to navigate within the website, explore data, and use the prediction tool without any issues.

Tools Application

Python and R are used in building models. During the process, we have learned how to build support vector machine (SVM) and neural network (NN) models for classification in Python, use subset selection methods in R and develop cox regression model in R. Tableau is used to build data visualization and we have gained knowledge in using the tool. We also learned how to use Shiny in R to build online applications that can be used for data exploration and prediction. Our learning experience also includes building a website using Google Sites with functions of embedding contents from Tableau and Shiny apps.

Measure Phase

Process Mapping/Process Visualization

Different levels of process maps have been developed during the project. A high level process map describes supplier, inputs, process, outputs, and customers and is abbreviated as SIPOC. A more detailed process map is the common process map, outlining the main tasks in the project. Detailed process map is the most detailed, including both main processes and their sub-processes. Last but not least, a functional process map is created to easily visualize people responsible for each process and the corresponding function. Please see the Appendix A for the process maps.

The Vital Few

This project aims at the problem identifying features associated with graft survival after OLT. The vital few covariates contributing to the response graft survival are selected using different methods, including boosting decision tree, subset selection and purposeful selection based significance and model performance. More details are outlined in the “Result and System Implementation” section.

Data Collection Planning and Execution

Data collection is out of scope and the data is provided by the Department of Surgery at Keck Medicine of USC. The dataset contains 84603 observations of 159 variables. The first step in data preprocessing is to remove invalid observations, whose graft survival time is larger than patient survival time, and useless variables, post-transplant features. Then the process includes checking missing values and determining imputing methods. If a categorical variable has an originally defined unknown category, missing values are assigned to the category; if not, then proportion of missing values is checked and if larger than 0.05, an unknown category is created for the missing values, if less than or equal to 0.05, missing values are assigned the mode.[3] For continuous variables, predictive mean matching is used in implementing the missing values. Data preprocessing lead to a dataset of 42169 observations of 145 variables useful for regression and survival analysis in addition to graft status (GSTATUS) and graft survival time (GTIME).

For classification, new datasets are subsetted from the dataset after imputation based on graft failure time. Then, censored data including the observations whose graft status is alive but graft survival time is less than the time of interest are removed. The following step is to create a new indicator variable of the time of interest. For example, when creating a dataset to analyze graft survival within one month, observations whose graft status is alive and graft survival time is less than 30 days are removed, then an indicator variable G30 is created and assigned value 1 if the graft status is 1 and assigned value 0 if the graft status is 0. As we are interested in analyzing graft status in 1 month, 2 months, 3 months, 6 months, 9 months, and 1 year, 6 subsetted datasets are created each with a new indicator variable.

Measurement System Analysis

Measurement system analysis is performed to analyze the stability, bias, linearity, repeatability, and/or reproducibility of a measurement system.[4] In the project, the measurement system measures the performance of models developed for data analysis, and the measurement system analysis focus on measuring the stability of model performance. Reference model performance comes from current studies, which is about 66.2% accuracy of classifying graft status after the surgery.[5]

Measurement system analysis looks at the multiple models implemented and records the test accuracy for each models, revealing that most of the models achieve reasonable accuracy below the reference line.

Tools Application

Flowchart tools are used in the process of creating process maps, and Google Slide is used to recreate the process maps for them to be presentable. Performance metrics are defined based on literatures and current studies in the field, including AUC-ROC (measures classification accuracy) for classification models, R^2 (measures amount of variation in the dependent variable accounted by the independent variables) and AIC (measures the amount of information lost) for regression models and C-Index (analogous to AUC-ROC) for cox regression models. Knowledge in accessing the performance of machine learning and statistical models are gained throughout the process.

Analysis Phase

Charts that Help Analyze Data Collected From the Measure Phase

In order to identify the variables related to graft survival after OLT, in addition to focusing on the short-term graft survival, it is useful to gain a understanding of long-term graft survival rate, whose causes are analyzed using fishbone diagram as shown in Appendix B.

Additional Charts used by Lean Six Sigma

In addition to fishbone diagram, a work breakdown structure (WBS) is created to effectively organizes the team's work into manageable tasks, as shown in Appendix B.

Y=f(X) Formula

In the project of identifying important variables associated with graft survival, the outcome of potential solutions is list of variables that are selected as important to graft survival, the inputs to achieve the outcome is all available useful features from the dataset, and the inputs are put into different models to achieve the outcome.

In classification, boosted decision tree is used to select top features to be used in SVM and NN models. The process can be formulated as
$Selected\ features = Boosted\ Decision\ Tree\ (Available\ useful\ variables)$.

In regression, backward selection is used to select a linear regression model. The process starts with $GTIME = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$ and ends when an

optimal model is achieved with lowest AIC score. The process can be formulated as
$X's\ in\ selected\ model = Backward\ selection\ (Available\ useful\ variables)$

In cox regression, variables are selected in a purposeful mechanism. Cox regression model can be formulated as $h(t \mid X) = h(t \mid X = 0)exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$ and the feature selection process can be formulated as
$X's\ in\ final\ cox\ regression\ model = Purposeful\ selection\ (Available\ useful\ variables)$

Strengths, Weaknesses, Opportunities, and Threats

Strengths of our team includes statistical backgrounds of the teammates as two of us major in Statistics. We are also good at making aesthetic visualizations using various tools and aim at creating neat, visually enjoyable and powerful presentations.

With a background focusing on data analytics and statistics, our team is not strong at coding and integrating codes written in different platforms including R and Python. We also should avoid switching goals in the middle of the project as new ideas appear but instead perform more research at the early stage to finalize a goal that we will work on throughout the project.

As the need of liver is increasing consistently, there are numerous opportunities to perform data analysis and use the results to help doctor and patients make better decisions of the surgery, as well as help patients survive longer after the surgery, and ideally reduce liver-related diseases and therefore reduce the need of liver transplant.

However, environmental degradation heavily influences people's health. The advance of medicine and technology might help cure many diseases nowadays, but not all diseases are curable. The bad habits that people have in their lives also posed threats to their health, making it harder to reduce the need of certain medications and surgeries, including liver transplant.

Root Causes

The fishbone diagram is used to analyze long-term low graft survival rate. To analyze short-term graft survival, we plot the graft survival rate in 1 month, 2 months, 3 months, 6 months, 9 months, and 1 year and also explore the data by gender, age group, and blood type in Appendix C. Graft seems to be able to survive at least 80% of the time under 1 year and the trend seems smooth and there are indeed differences of people having different biological characteristics, so the reason for decreasing graft survival might relate to the degradation of the graft functions as time passes and the biological factors.

Correlation

By examining the scatter plots between graft survival time and each feature, there seem not to be any obvious correlation between graft survival time and any individual feature. Correlation plots of continuous variables are also created, and no hidden correlation is identified other than repetitive measures such as weight at transplant and weight at registration. Please see Appendix C for the correlation plot.

Sources of Variation

Model performance varies as we perform cross-validation and/or tune models by adding or removing variables used in the analysis. Classification models' performance metrics such as area under the receiver operating characteristic curve (AUC-ROC), true positive rate, false positive rate, positive predictive value, F1-score, R^2, AIC and C-Index all vary as the number of variables included in multivariate models changes.

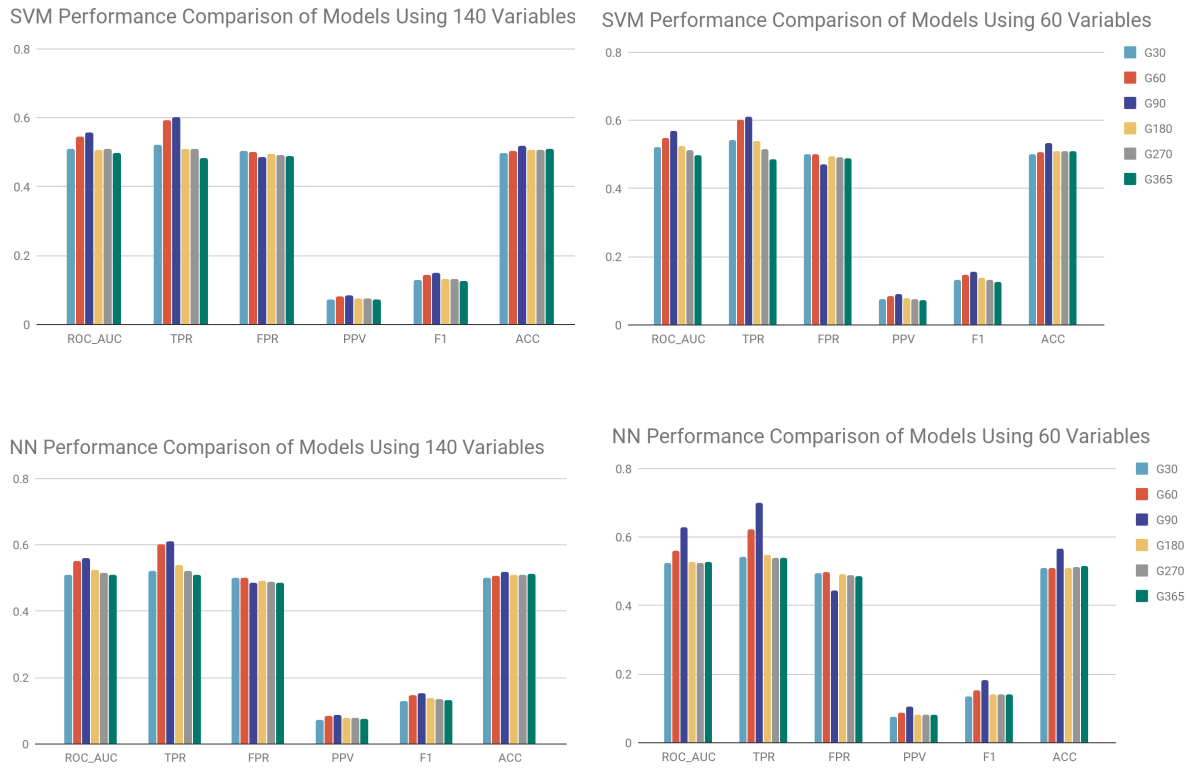


Figure 1. Graphs showing performance metrics of classification models. Variation is visible as number of variables used changes.

Potential Solutions

To identify important features related to graft survival, one potential way is to use tree-based methods to select the features. Boosted decision tree is implemented to select the top features for predicting graft survival in six different time frames under one year. Among the six selections, many donor-related variables appear multiple times to be on top of the list. Subset selection is another potential solution to identify important variables in linear regression model attempting to predict graft survival time. Purposeful selection can also be a solution to find out features related to graft survival in cox regression.

Improve Phase

Alternative Solutions Considered

For linear regression models, an alternative selection method can be separating the data by age group so different analyses are performed for MELD or PELD score.

To improve the selection methods for cox regression, univariate models are fit and significant variables are selected to be included in a multivariate model.

Another solution is in addition to the above steps, the multivariate models is further tuned by eliminating variables creating errors in the model and adding variables of interest such as blood type, age, gender, and state of residency.

Please see Appendix D for the variables selected in classification, linear regression and cox regression after improvement.

Recommended Solution(s)

After comparing the models based on their model performance and amount of information conveyed, the cox regression model is selected as the recommended solution to build a prediction tool for graft survival. Cox regression can predict the probability of experiencing graft failure at time t+1 assuming the graft has survived until time t. One concern is that there are time-varying variables included in the data set such as the number of previous transplant, without adjusting the model to incorporate the time-varying effect, the model's accuracy might not be as high. Overall, the model has accuracy 0.583.

Ways to Pilot the Recommended Solution(s)

The solution is tested by inputting new values of the variables used in the model and check the prediction. Predicted probability of graft survival should not be higher than 1 and lower than 0. Reasonable values should be close to the survival rate from data exploration, which shows that the average is about 0.89.

Project Plan

A WBS is created specifically for the recommended solution.

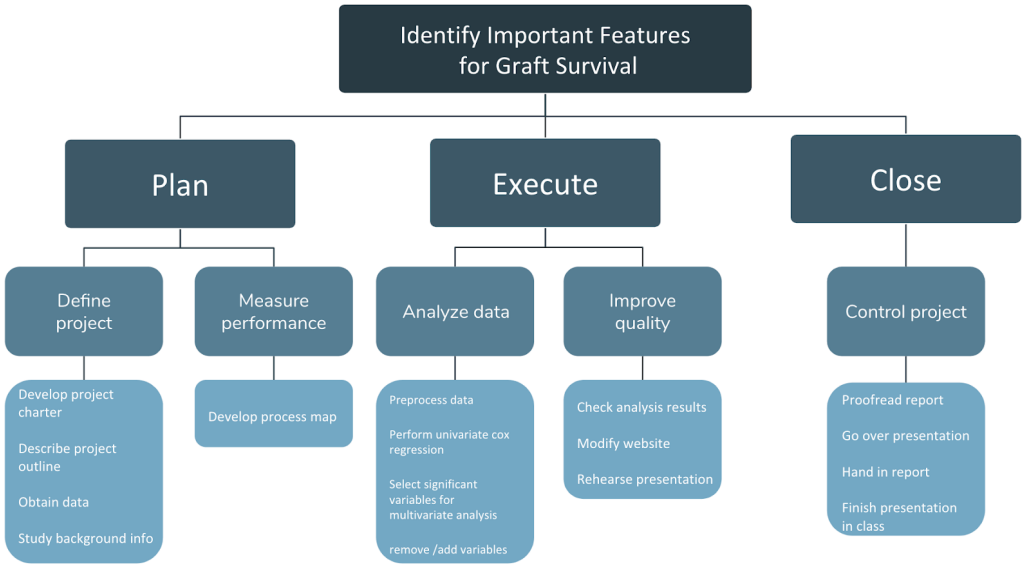


Figure 2. Work Breakdown Structure for developing prediction tool using cox regression.

Control Phase

The project is controlled by weekly meetings with the teaching assistant at the latter half of the semester and regular weekly or bi-weekly team meetings throughout the semester. Day-to-day processes are controlled by live communication among teammates, updating tasks done or difficulties encountered, and the issues would be discussed in the next team meeting. The Gantt chart is also used as a control tool to keep up with the schedule as shown in Appendix B.

One challenge in implementing control is coming up with new ideas and still following the Gantt Chart. Therefore, we need to adjust the Gantt Chart to extend the

define and analysis processes, but also need to make sure we still meet the deadline of closing the project.

After closing the project, we will communicate with the customers to add authorized users as desired. We will also maintain the website and the contents on the site. If the customers have any questions, we will also be available through email posted on the website.

Result and System Implementation

Machine Learning Approaches

Classification

1. Neural Network (NN)

NN is a supervised machine learning algorithm that learns a function by training a function on a dataset. With a set of features and an output, the model can be trained as as classification model. Between input and output, there are hidden layers in the model.

Two of the parameters in NN is epoch, number of forward and backward passes of all training examples, and batch size, the number of training examples in one forward/backward pass.[6] Different values of epoch and batch size are checked and the optimal test performance and computing time are reached at epoch 300 and batch size 200.

For each of the 6 subsetted data sets, boosted decision tree is implemented to select top features that will be later used in the NN classification model. 20% of data are separated as test set. Models are trained and validated on the other 80% of the data using 10-fold cross validation, and then tested on the test set. Data are also oversampled to fix the unbalanced class problem, as proportion of observations with alive graft is much higher than observations with dead graft. Performance metrics are outputted for comparison. On average, NN is able to reach about 50% of AUC-ROC.

More details regarding NN can be found in Appendix E.

2. Support Vector Machine (SVM)

SVM is a supervised machine learning classifier. Given a set of labeled features and an output, the algorithm outputs an optimal hyperplane that categorizes new observations.[7]

The kernel function is used to study the type of relationships in data, and needs to be chosen with consideration as the data is not linearly separable.[8] Another parameter is gamma, which indicates how far the influence of a single training observation can reach.[9] The radial basis function is used as the Kernel function and 10 the gamma.

20% of data are separated as test set. Models are trained and validated on the other 80% of the data using 10-fold cross validation, and then tested on the test set. Data are also oversampled to fix the unbalanced class problem, as proportion of observations with alive graft is much higher than observations with dead graft. Performance metrics are outputted for comparison. SVM is performed on each of the 6 subsetted data sets, and reaches AUC-ROC at about 50%.

More details regarding NN can be found in Appendix F.

3. Comparison

Overall, both models achieve AUC-ROC about 0.53 on average predicting graft status over different time frames up to 1 year after the surgery. Compared to SVM, NNs seems to perform better overall in G90. In G90 prediction case, the AUC of NN is 0.62 and that of SVM is 0.56.

Linear Regression

A linear regression is modeled as $GTIME = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$,

$k = 145$ as there are 145 features available for analysis after data cleaning. Backward selection is used, taking the full model with all features as the start, and remove one variable each time to achieve optimal AIC.

The parameters β_i are estimated using ordinary least square, which is minimizing the sum of squares of the differences between the observed GTIME and the predicted values by the linear function. The coefficient estimates of β_i represent the magnitude of association between the corresponding X and GTIME while all the other variables are constant.

The baseline model has AIC 570831.1 and the selected model has AIC 569712.5. R^2 is 0.027. New values of the independent variables can be inputted and a predicted GTIME will be outputted.

Linear regression model assumes linearity, normality, and homoscedasticity (equal variances) and the assumptions are checked using residual plots and the linearity and homoscedasticity assumptions are violated as shown in Appendix G.

Survival Analysis

Survival analysis is performed using cox regression, which is modeled as $h(t \mid X) = h(t \mid X = 0)exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$. $h(t \mid X)$ is the hazard function given a set of features X, which can be understood as the probability of an individual experiencing graft failure at time t+1 if the person has survived until time t. In the final model, $k = 17$.

The parameters β_i are estimated using partial maximum likelihood. The exponential of the coefficient estimate represents the effect of the corresponding feature has on the ratio of two hazard function while all the other variables are constant. For example, for continuous variable, the hazard ratio between people whose $X_i = x + 1$ and people whose $X_i = x$ is exp(β_i) , meaning that people whose $X_i = x + 1$ is exp(β_i) times more likely than people whose $X_i = x$ to experience graft failure while all other variables are the same; for categorical variables, the hazard ratio between people of category j and people of baseline model l is exp($\beta_j - \beta_l$), meaning that people of category j is exp($\beta_j - \beta_l$) times more likely than people of baseline model l to experience graft failure while all other variables are the same.[10]

The model achieves C-index 0.583. New values of the independent variables can be inputted and the model will output a predicted hazard function.

Cox regression assumes proportional hazard and linearity. The assumptions are tested in the cph.zh function in R using chi-square test. Significant test (p-value<0.05) indicates non-linearity and except for one level in the state of residency variable everything else is non-significant, so the model assumptions can be said to be met as shown in Appendix H.

System Implementation

After evaluating the performance, assumptions and amount of information conveyed by different models, survival analysis is selected for building the prediction tool as the cox regression seems to meet the assumptions and is able to predict the hazard function of time in addition to merely the graft survival time in linear regression or the graft survival status in classification.

The selected cox regression model is incorporated into a Shiny app written in R. Users will be able to input new values of the independent variables in the model in the app. A graph of the hazard function will be plotted with time on the x-axis and survival probability on the y-axis, and survival probability at times of interest include 1 month, 2 months, 3 months, 6 months, 9 months and 1 year will be shown as shown in Appendix H.

Prototype and Demo

A prototype of the end product is a website where users can navigate for data analysis tool and prediction tool. Tableau and Shiny in R are used to develop contents on the site as well.

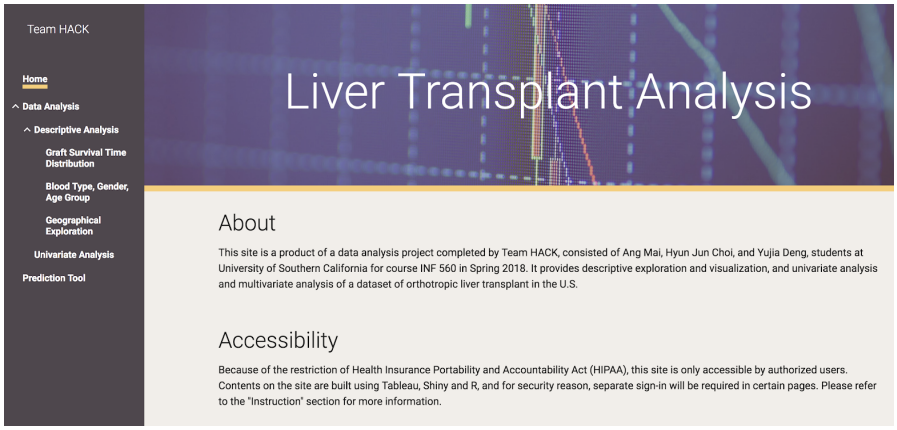


Figure 3. Page of Website presenting analysis results and tools

To ensure security of the information, the website is only accessible by authorized users. To view the contents on built with Tableau and Shiny, additional authorizations need to be granted. To fully access the site, there are 3 authorizations needed to be granted.

1. Access to the site
2. Access to the Tableau contents in the site
3. Access to certain Shiny contents in the site

Customers can communicate with us for adding authorized users.

The menu is located on the left. The home page includes information about the website including instructions on accessing the contents and contact information. Under “Data Analysis”, users can explore the data in the “Descriptive Analysis”

section the distribution of graft survival time, graft survival statistics by blood type, age group, and gender, and geographical exploration. To assess the relationship between graft survival time and any other variables, users can log in to "Univariate Analysis" to check out the scatter plot, Kaplan-Meier curves, and survival curve outputted by univariate cox regression between GTIME and one of the independent features by selecting GTIME as the "Y Feature" and the desired independent feature as the "X Feature". Users can also choose to visualize the correlation between any two variables by selecting the desired "Y Feature" and "X Feature". In the prediction tool, users input values of the variables on the left and on the right the app will output a survival curve and underneath survival probability at different times of interest.

Appendix A

Different Process Maps

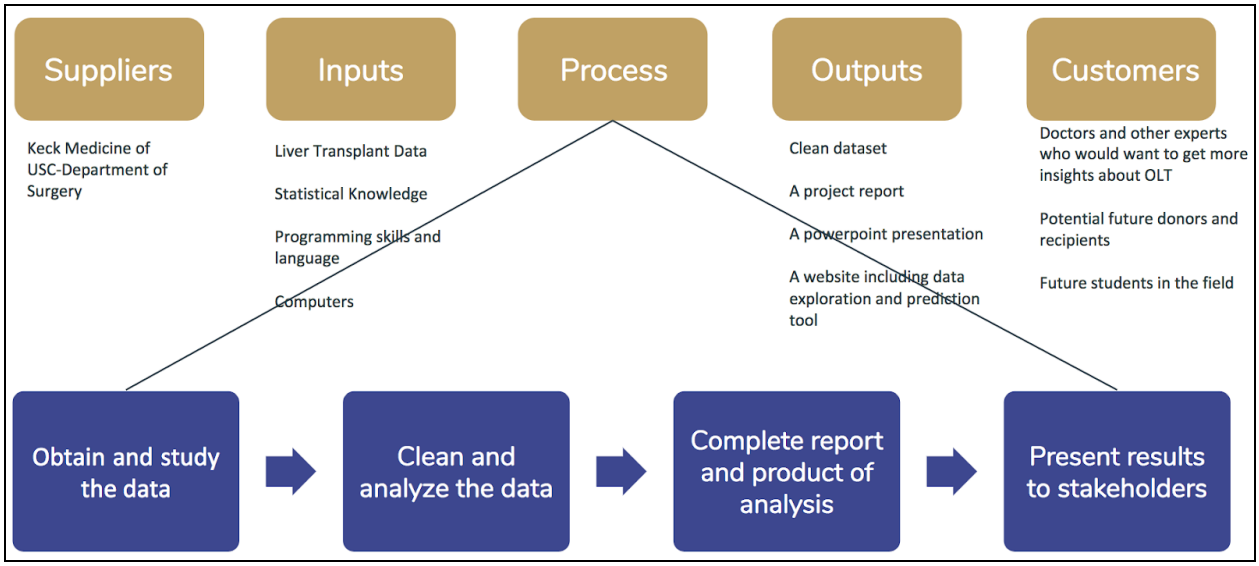


Figure A1. SIPOC.

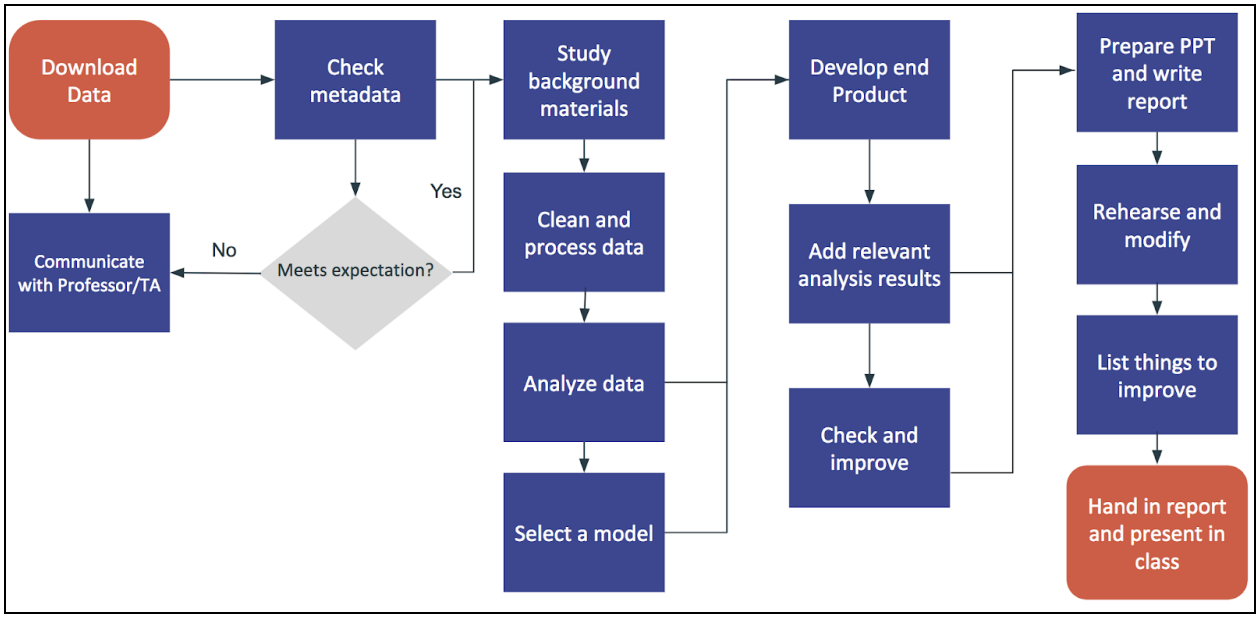


Figure A2. Common Process Map.

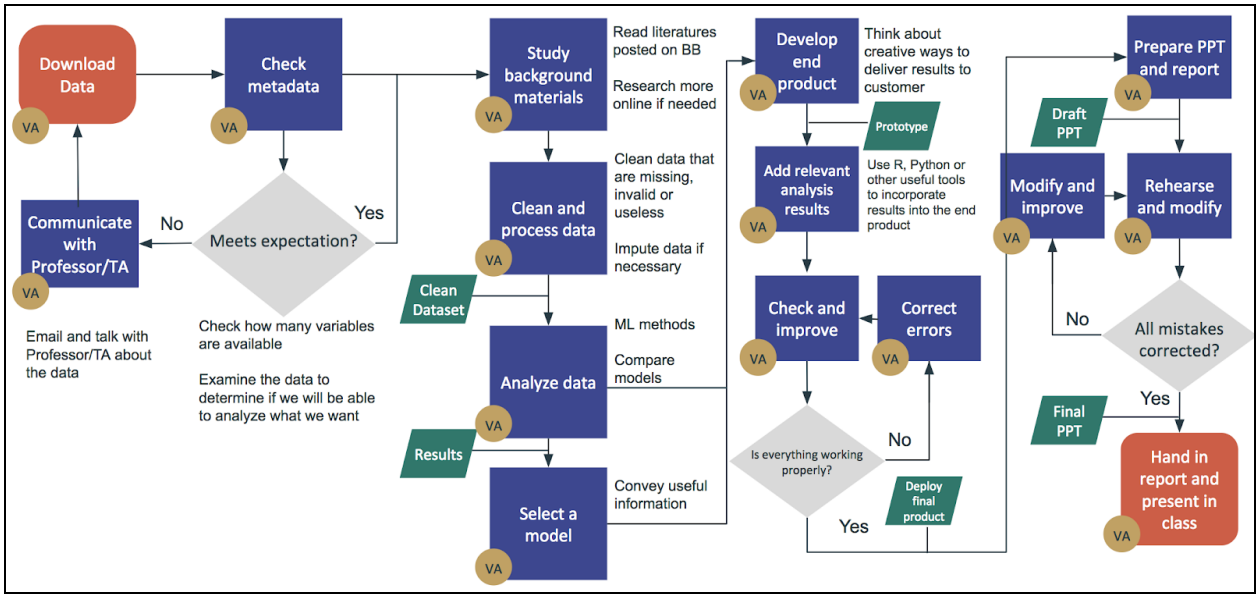


Figure A3. Detailed Process Map.

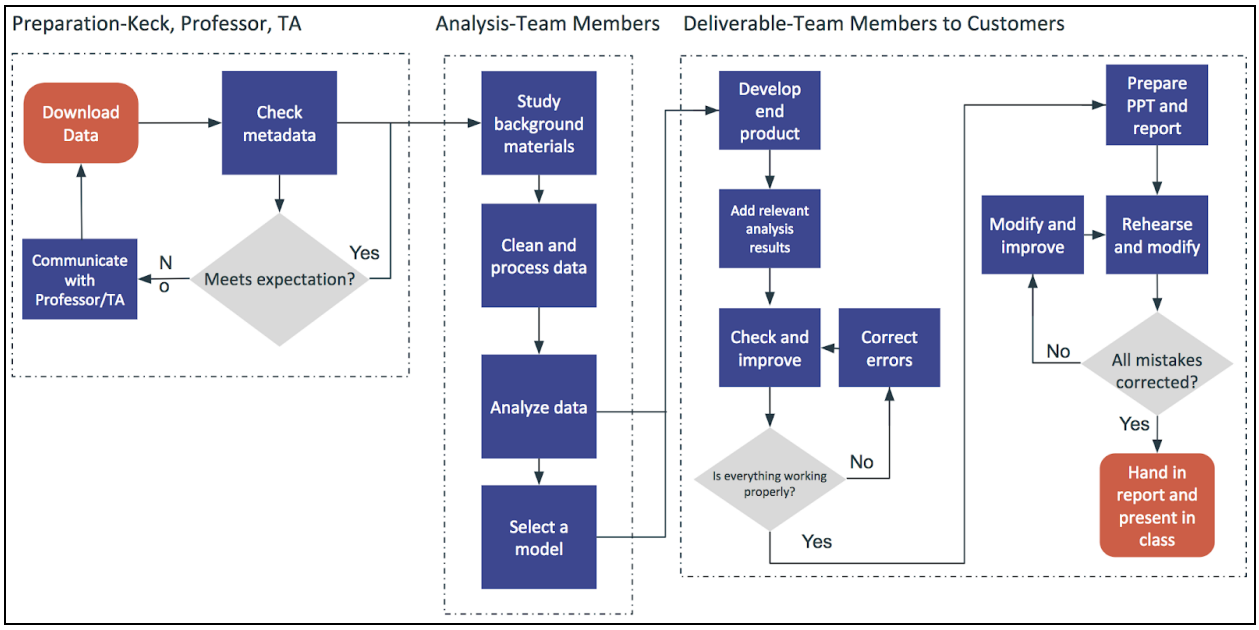


Figure A4. Functional Process Map.

Appendix B

Other Charts/Graphs used in Lean Six Sigma

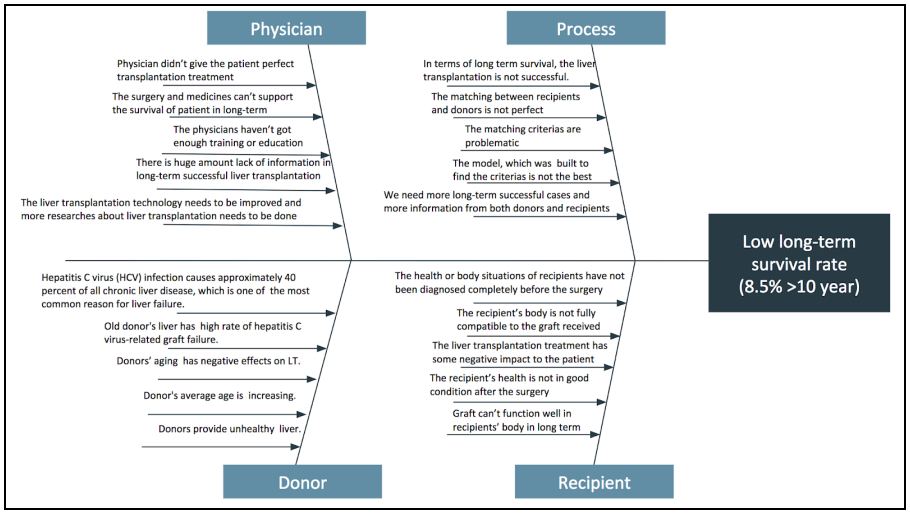


Figure B1. Fishbone Diagram of analyzing long-term graft survival rate.

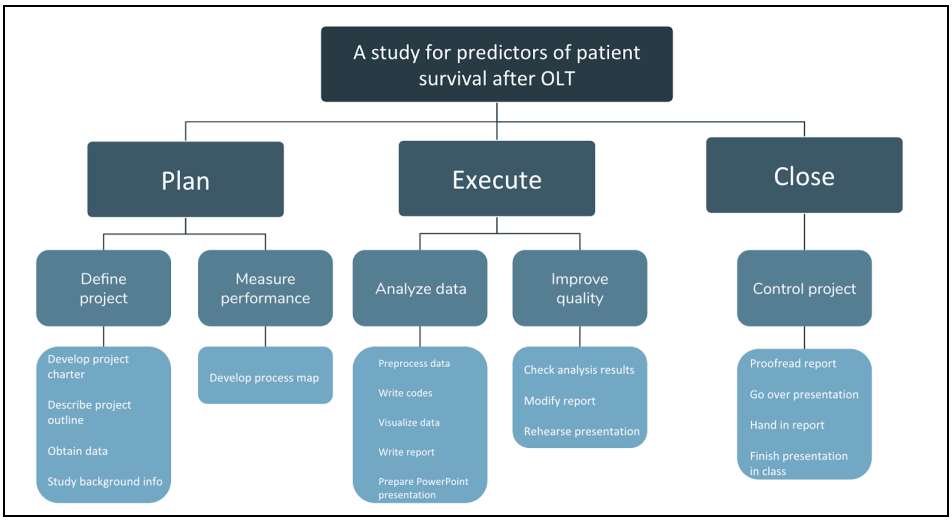


Figure B2. Originally developed WBS for the tasks of the project.

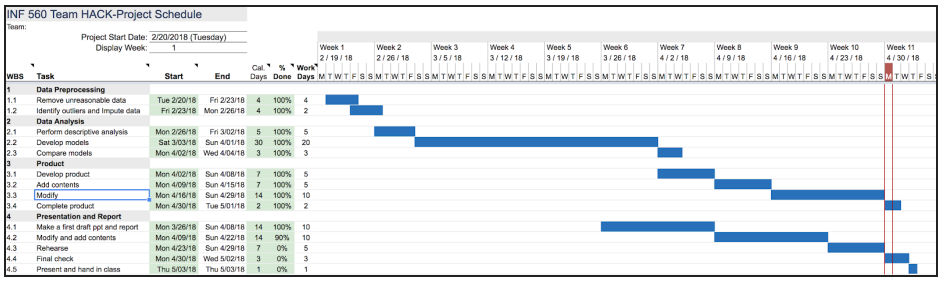


Figure B3. Most recently updated Gantt Chart.

Appendix C

Graphs used in Data Exploration

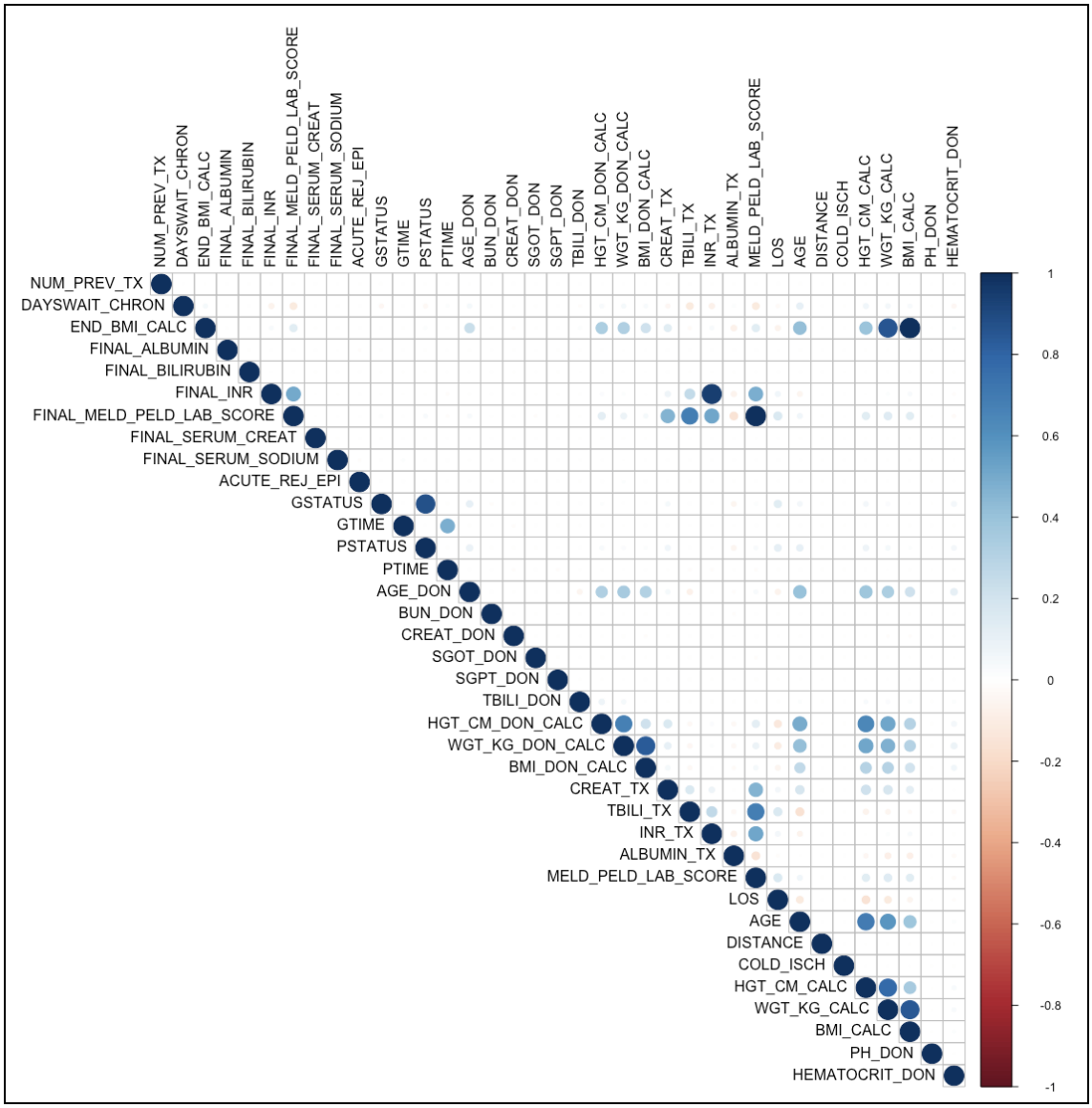


Figure C1. Correlation matrix among continuous variables.

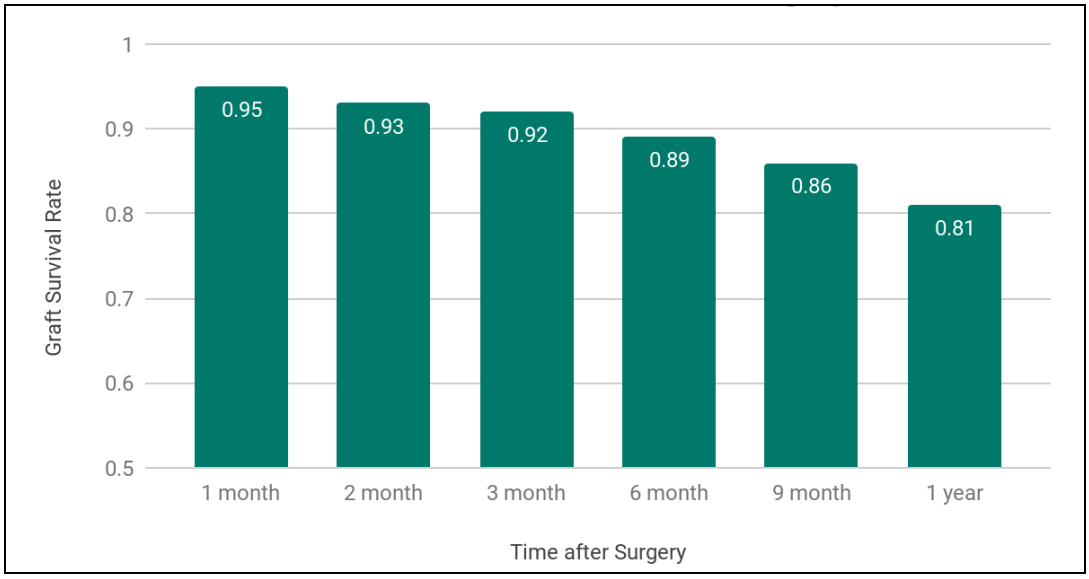


Figure C2. Graft survival rate at different times after surgery.

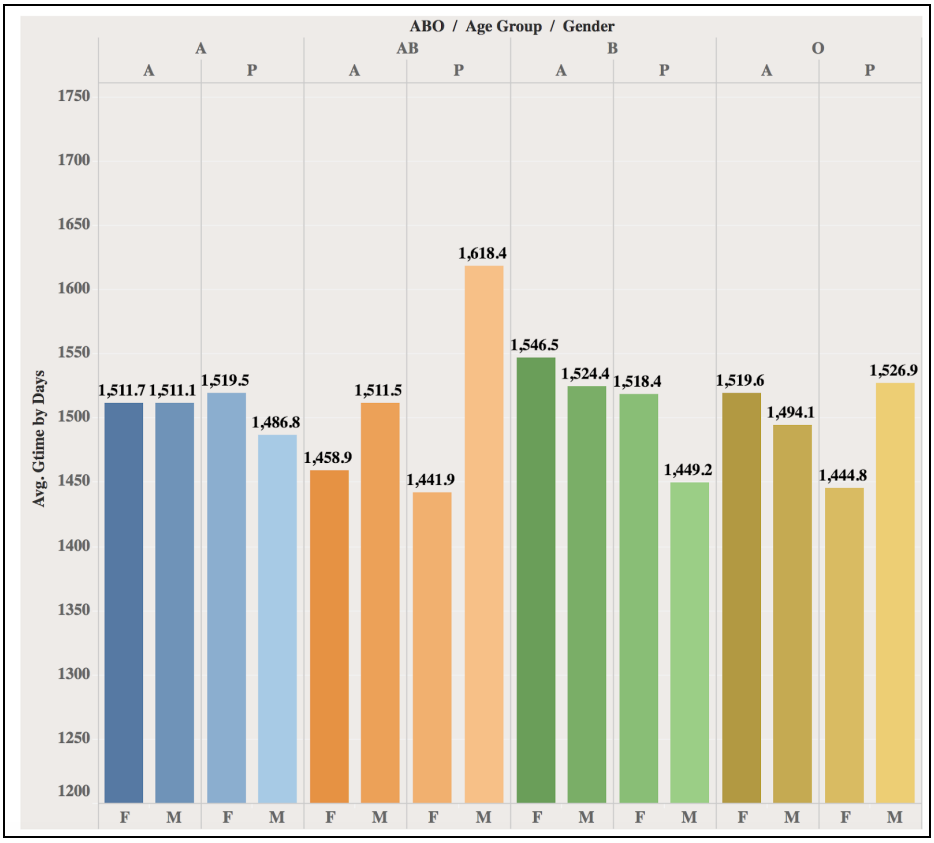


Figure C3. Average graft survival time by ABO/Age group/Gender of recipient.

Appendix D

Variables Used in Different Models

Top 30 Variables from Boosted Decision Tree

G30	G60	G90	G180	G270	G365
ASCITES_TX	AGE_DON	AGE_DON	AGE_DON	AGE_DON	PREV_TX
TATTOOS	CREAT_TX	TBILI_TX	AGE	AGE	AGE_DON
WORK_INCOME_T	PERM_STATE_TRR	HEMATOCRIT_DOI	TBILI_TX	TBILI_TX	DDAVP_DON
PREV_TX	HGT_CM_CALC	HGT_CM_CALC	DAYSWAIT_CHRO	DAYSWAIT_CHRO	PERM_STATE_TRR
PULM_INF_DON	BMI_DON_CALC	PERM_STATE_TRR	PERM_STATE_TRR	PERM_STATE_TRR	DAYSWAIT_CHRON
CMV_DON	HEMATOCRIT_DOI	WGT_KG_DON_CA	PREV_TX	PREV_TX	PULM_INF_DON
GENDER_DON	TBILI_TX	DAYSWAIT_CHRO	ALBUMIN_TX	ALBUMIN_TX	HEMATOCRIT_DON
CMV_IGM	DAYSWAIT_CHRO	PREV_TX	HEMATOCRIT_DOI	HEMATOCRIT_DOI	EXC_EVER
GENDER	SGPT_DON	BMI_DON_CALC	CREAT_TX	CREAT_TX	CMV_IGM
HCV_SEROSTATUS	ALBUMIN_TX	AGE	INR_TX	INR_TX	FINAL_ASCITES
DDAVP_DON	AGE	CREAT_TX	HGT_CM_CALC	HGT_CM_CALC	GENDER_DON
PT_DIURETICS_DO	PREV_TX	ALBUMIN_TX	WGT_KG_DON_CA	WGT_KG_DON_CA	NON_HRT_DON
HIST_HYPERTENS	BUN_DON	BUN_DON	BMI_DON_CALC	BMI_DON_CALC	GENDER
HIST_OTH_DRUG_	COLD_ISCH	FINAL_BILIRUBIN	BUN_DON	BUN_DON	DA1
ARGININE_DON	FINAL_SERUM_CR	SGPT_DON	PERM_STATE	PERM_STATE	HIST_HYPERTENS_DON
INOTROP_SUPPOR	TBILI_DON	TBILI_DON	END_BMI_CALC	END_BMI_CALC	EXC_CASE
PREV_AB_SURG_T	WGT_KG_DON_CA	END_BMI_CALC	DB2	DB2	ETHCAT
PT_T4_DON	FINAL_BILIRUBIN	INR_TX	COLD_ISCH	COLD_ISCH	PREV_AB_SURG_TRR
ABO_DON	END_BMI_CALC	DDR2	FINAL_BILIRUBIN	FINAL_BILIRUBIN	CMV_DON
LI_BIOPSY	INR_TX	SGOT_DON	FINAL_SERUM_CR	FINAL_SERUM_CR	HIST_OTH_DRUG_DON
FINAL_ASCITES	DDR2	COLD_ISCH	TBILI_DON	TBILI_DON	PRI_PAYMENT_TRR
PROTEIN_URINE	END_STAT	FINAL_INR	END_STAT	END_STAT	HBV_CORE
CMV_IGG	DA1	INSULIN_DON	ETHCAT	ETHCAT	PT_DIURETICS_DON
HIST_CIG_DON	FINAL_INR	DIAG	HOME_STATE_DO	HOME_STATE_DO	CREAT_TX
EXC_EVER	INSULIN_DON	DISTANCE	MELD_PELD_LAB	MELD_PELD_LAB	ASCITES_TX
AGE_DON	HGT_CM_DON_CA	ETHCAT	FINAL_ALBUMIN	FINAL_ALBUMIN	AGE
INSULIN_DON	DISTANCE	CREAT_DON	CREAT_DON	CREAT_DON	HCV_SEROSTATUS
ABO	CDC_RISK_HIV_DO	PERM_STATE	HGT_CM_DON_CA	HGT_CM_DON_CA	PT_T4_DON
EXC_CASE	ETHCAT	FINAL_MELD_PELI	DDR2	DDR2	CLIN_INFECT_DON
FINAL_ENCEPH	DB2	DB2	SGPT_DON	SGPT_DON	ABO_DON

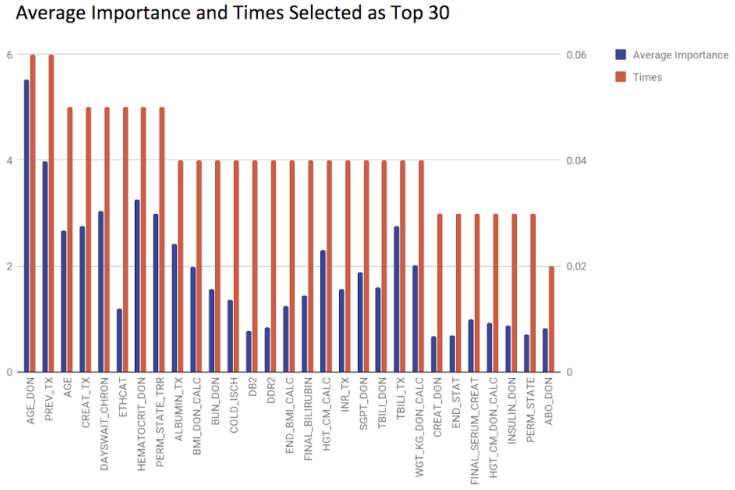


Figure D1. Variables selected by boosted decision tree for classification (left); average importance score and the number of times selected for each of the top 30 variables (right).

Selected Variables in Linear Regression	Estimate
Blood Type	
A	
A1	-4.00
A1B	41.76
A2	359.98
A2B	226.55
AB	-31.09
B	-1.24
O	-12.26
Gender	
Male	1.28
Female	
Education	
1-None	
2-Grade school (0-8)	-0.84
3-High school (9-12)	-0.89
4-Attended college/technical school	-0.01 .
5-Associate/Bachelor degree	-0.93
6-Post-college graduate degree	-0.89
996-Less than 5 years	-0.85
998-Unknown	-0.92
Diabetes	
1-No diabetes	
2-Type 1	0.66 *
3-Type 2	-0.11
4-Type other	-0.29
5-Type unknown	0.74
998-Status unknown	-0.45
Ethnicity	
1-White	
2-Black	-4.07

4-Hispanic	-3.63
5-Asian	-1.57
6-Amer Ind/Alaska Native	6.66
7-Native Hawaiian/other Pacific Islander	90.34
9-Multiracial	40.42
BMI	-1.40 *
Number of Previous Transplantation	9.62
Days on liver waiting list	-0.0077
Type of Exception Relative to HCC	
HBL	
HCC	282.20
Non-HCC	285.40
Region of Transplantation [11]	
1-Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Eastern Vermont	
2-Delaware, District of Columbia, Maryland, New Jersey, Pennsylvania, West Virginia, Northern Virginia	6.76 24.34
3-Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, Puerto Rico	1.63
4-Oklahoma, Texas	42.24 .
5-Arizona, California, Nevada, New Mexico, Utah	42.45
6-Alaska, Hawaii, Idaho, Montana, Oregon, Washington	50.00 .
7-Illinois, Minnesota, North Dakota, South Dakota, Wisconsin	-2.79
8-Colorado, Iowa, Kansas, Missouri, Nebraska, Wyoming	13.50
9-New York, Western Vermont	17.56
10-Indiana, Michigan, Ohio	23.16
11-Kentucky, North Carolina, South Carolina, Tennessee, Virginia	
State of Residency at Registration	
AK-Alaska	
AL-Alabama	-74.89
AR-Arkansas	-28.05
AS-American Samoa	-149.40
AZ-Arizona	-146.60
CA-California	-112.00
CO-Colorado	-150.90
CT-Connecticut	-68.85
DC-District of Columbia	-33.03
DE-Delaware	85.24
FL-Florida	-97.46
GA-Georgia	-78.92
GU-Guam	-410.00
HI-Hawaii	-88.07
IA-Iowa	-74.26
ID-Idaho	29.85
IL-Illinois	-76.82
IN-Indiana	-162.70
KS-Kansas	-52.96
KY-Kentucky	-102.2
LA-Louisiana	-60.70
MA-Massachusetts	-104.4
MD-Maryland	-84.50
ME-Maine	43.96
MI-Michigan	-129.7
MN-Minnesota	-116.3
MO-Missouri	-88.85
MS-Mississippi	-196.20
MT-Montana	-62.58
NC-North Carolina	-147.50
ND-North Dakota	-229.20
NE-Nebraska	-45.54
NH-New Hampshire	-78.09
NJ-New Jersey	-66.04
NM-New Mexico	-18.41
NV-Nevada	-118.90
NY-New York	-98.82
OH-Ohio	-83.79

OK-Oklahoma	-73.99
OR-Oregon	-109.40
PA-Pennsylvania	-93.32
PR-Puerto Rico	-120.10
RI-Rhode Island	-128.00
SC-South Carolina	-115.10
SD-South Dakota	-42.67
TN-Tennessee	-90.50
TX-Texas	-78.23
UT-Utah	-102.20
VA-Virginia	-120.90
VI-Virgin Islands	-586.80
VT-Vermont	-124.40
WA-Washington	-84.01
WI-Wisconsin	-81.49
WV-West Virginia	-115.50
WY-Wyoming	113.00
ZZ-Unknown	-167.10
Ascites	
1-Absent	
2-Slight	-3.24
3-Moderate	8.57
4-Unknown	15.3
Bilirubin	-0.14
Dialysis prior week	
Yes	-8.49
No	
Encephy	
1-None	
2-1 to 2	15.56 .
3- 3 to 4	16.50
4-Unknown	53.58 *
INR	-2.81
Serum Creatinine	3.34
MELD or PELD Score	-0.81 .

*Figure D2. Variables selected by backward selection for linear regression. The symbol next to the coefficient estimate indicates the p-value of testing whether the estimate is zero. If the p-value is small, the test is significant and the estimate is said to not equal to zero and therefore there is significant relationship between the corresponding variable and GTIME (0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1).*

Selected Variables in Cox Regression	Exponential of Coefficient Estimate
Recipient Age	1.00 ***
Recipient Gender	
Male	0.094
Female	
Recipient Blood Type	
A	
A1	0.98
A1B	0.56
A2	1.11
A2B	-0.000027
AB	1.027
B	-0.95 .
O	1.04 *
Recipient BMI	0.93
Recipient MELD/PELD Score	1.001
Previous Transplant	
Yes	1.545 ***
No	
State of Residency Where Transplant will Take Place	
AK-Alaska	

AL-Alabama	0.60
AR-Arkansas	0.60
AS-American Samoa	0.000055
AZ-Arizona	0.62
CA-California	0.57
CO-Colorado	0.59
CT-Connecticut	0.62
DC-District of Columbia	0.66
DE-Delaware	0.63
FL-Florida	0.67
GA-Georgia	0.59
GU-Guam	0.98
HI-Hawaii	0.44
IA-Iowa	0.61
ID-Idaho	0.51
IL-Illinois	0.69
IN-Indiana	0.82
KS-Kansas	0.67
KY-Kentucky	0.74
LA-Louisiana	0.66
MA-Massachusetts	0.62
MD-Maryland	0.71
ME-Maine	0.85
MI-Michigan	0.75
MN-Minnesota	0.61
MO-Missouri	0.65
MP-Northern Mariana Islands	0.000029
MS-Mississippi	0.56 .
MT-Montana	0.56
NC-North Carolina	0.61
ND-North Dakota	0.46 .
NE-Nebraska	0.74
NH-New Hampshire	0.58 .
NJ-New Jersey	0.64
NM-New Mexico	0.56 .
NV-Nevada	0.81
NY-New York	0.76
OH-Ohio	0.67
OK-Oklahoma	0.70
OR-Oregon	0.57 .
PA-Pennsylvania	0.75
PR-Puerto Rico	0.63
RI-Rhode Island	0.85
SC-South Carolina	0.72
SD-South Dakota	0.67
TN-Tennessee	0.57 .
TX-Texas	0.63
UT-Utah	0.55 .
VA-Virginia	0.62
VI-Virgin Islands	1.03
VT-Vermont	0.61
WA-Washington	0.44 **
WI-Wisconsin	0.65
WV-West Virginia	0.58 .
WY-Wyoming	0.94
ZZ-Unknown	1.00
Days on Waiting List	-0.998 ***
Recipient Ethnicity	
1-White	
2-Black	1.05 .
4-Hispanic	1.09 **
5-Asian	1.08 .
6-Amer Ind/Alaska Native	0.87
7-Native Hawaiian/other Pacific Islander	0.89
9-Multiracial	1.10

Donor Age	1.009 ***
Donor Gender Male Female	 0.98
Donor Blood Type A A1 A1B A2 A2B AB B O	 1.06 0.56 1.11 0.000027 1.03 0.95 . 1.04 *
Donor BMI	0.78 ***
Donor Type Deceased Alive	 0.91 .
Donor Hematocrit	1.01 ***
Donor Less Than 7 Days Old at Time of Donation Yes No	 7.88 **

*Figure D2. Variables used in cox regression. The symbol next to the coefficient estimate indicates the p-value of testing whether the estimate is zero. If the p-value is small, the test is significant and the estimate is said to not equal to zero and therefore there is significant relationship between the corresponding variable and graft survival (0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1).*

Appendix E

More Details about Neural Network

a. The Target Function

The target feature is GSTATUS, which is 0 or 1.

The input value $S^{(l)}$ of hidden layer can be obtained from equation: $S^{(l)} = W^{(l)}X^{(l-1)} + b^{(l)}, \quad 1 \le l \le L$

The output of sigmoid function can be calculated from equation: $X^{(l)} = sigmoid(S^{(l)}) = \frac{1}{1+e^{-S^{(l)}}}, \quad 1 \le l \le L$

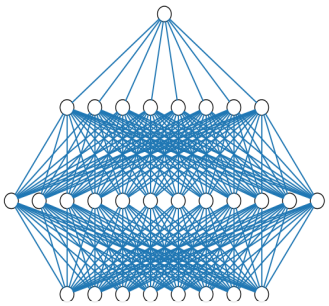


Fig.E1

If the loss function is standard least square error, the error of output can be calculated as $E(X^{(L)}) = \frac{1}{2}(X^{(L)} - y)^2$

Derivative of error for output layer is $\delta^{(L)} = (X^{(L)} - y)(1 - X^{(L)})X^{(L)}$

And for hidden layer, the derivative of error is

$$\delta_i^{(l-1)} = x_i^{(l-1)}(1 - x_i^{(l-1)}) \sum_{j=0}^{d^{(l)}} \delta_j^{(l)} W_{i,j}^{(l)}, \quad 1 \le l \le L - 1$$

Then update the weights and bias for output layer and hidden layer respectively:

$$W_{i,j}^{(l)} := W_{i,j}^{(l)} - \alpha \delta_j^{(l)} x_i^{(l-1)}, \quad 1 \le l \le L$$

$$b^{(l)} = b^{(l)} - \eta \delta^{(l)}, \quad 1 \le l \le L$$

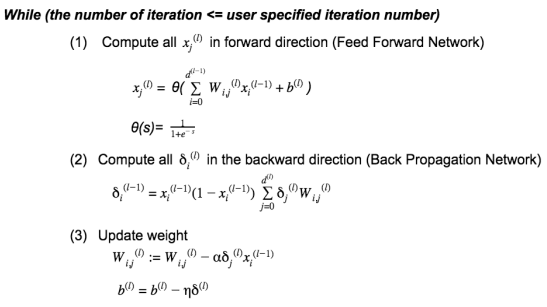
b. Representation

The input features of NN model can be any subsets of 145 features. For example, if we use 8 features among 145 features, the input layer structure is same to input layer structure in Figure E1, where the model has 8 input nodes.

c. System Structure

We use 60 variables as inputs, so we have one input layer-60 input features, one hidden layer-12 neurons (activation function of hidden layer-sigmoid function), one hidden layer-9 neurons (activation function of hidden layer-sigmoid function), one hidden layer-8 neurons (activation function of hidden layer-sigmoid function), one hidden layer-8 neurons (activation function of hidden layer-sigmoid function), one hidden layer-7 neurons (activation function of hidden layer-sigmoid function) and one output layer (activation function of output layer-sigmoid function. In this structure, there are 200 unknown variables. Between input layer and first hidden layer, there are 96 unknown variables. Between first hidden layer and second hidden layer, there are 96 variables. From the second layer to output layer, there are 8 variables.

d. The Learning Algorithm



After finishing data preprocessing, we set input features and target output. With the pseudo code of NN Model.1, if we use 60 features, dimension of each input data is 60. After computing $x^{(1)}$ in forward direction, we have 12-dimension input data, $x^{(2)}$. This calculation uses the first part of pseudocode. Then if we compute $x^{(2)}$ in forward direction again, we can get $x^{(3)}$, which is 9 dimensions. Finally, after computing $x^{(6)}$ in forward direction, we can get one-dimension output value of this model. From the error, we need to calculate δ from output layer to input layer. When we update weights, it should be updated from input layer to output layer. To make this learning process efficient, we use ADAM optimizer. We train and test model with 100 batch sizes and 100 numbers of epoch. The learning rate of ADAM is 0.01.[12]

e. Improvements and Modifications

To improve our NN model, we have considered different structures of NN, activation function, numbers of epoch, batch size and optimizers. It was examined and the result shows that epoch number 300, batch size 200, and ADAM optimizer 0.01 give best model accuracy.

Appendix F

More Details about Support Vector Machine

a. The Target Function

The target feature is graft survival status, labeled with 6 categories in 1 month, 2 months, 3 months, 6 months, 9 months, and 1 year. The input features of model are selected from boosted decision tree. The output of SVM model is 0 or 1 in each graft survival label.

If we use kernel function ψ((a, b))=(a, b, $a^2 + b^2$), the i-th row and j-th column of Q matrix is $Y_j Y_j K(X_i, X_j)$, where X_i means each values of each variable and 20-dimension coordinates of each data point in our model. Y_i represents the label of each graft status, valuing 1 or 0. Because the kernel function is

$$\psi((a, b))=(a, b, a^2 + b^2),\ K(X_i, X_j)\ is\ X_i^T X_j + (X_i^T X_i)(X_j^T X_j).$$

$$Max_{\alpha} \sum_{n=1}^{N} \alpha_n - 1/2\alpha^T Q\alpha \quad (1)$$

$$Such\ that\ for\ all\ n,\ \alpha_n \geq 0,\ \sum_{n=1}^{N} \alpha_n y_n = 0.$$

After performing the steps above, we can get W by using α by $W = \sum_{n=1}^{N} \alpha_n y_n z_i^T$, where z_i is the nonlinear transformation of X_i. In this case, W is 1 by 3 vector and z is 3 by 1 vector, because if x is (a, b), the nonlinear transformation of x, ψ((a, b)), is (a, b, $a^2 + b^2$), which is in z space. Finally, we can obtain get b from the equation (2).

$$b^{(*)} = (Max_{i:y^i=-1} W^{*T} X^i + Min_{i:y^i=1} W^{*T} X^i) * - 1/2 \quad (2)$$

b. Representation

The input features of SVM model can be subsets of 145 features. For example, if we only use 2 features among 145 features, each data points and the SVM model can be plotted like the Figure F1.

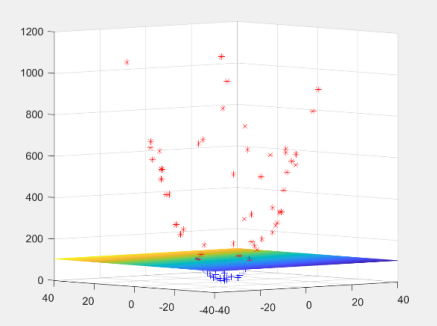


Fig.F1

c. System Structure

For the SVM model, we have one input 20-dimension input features. Each dimension presents each data field which was selected by boosted decision tree from original data set. The kernel is Radial basis function. The gamma value is 10. [13]

Appendix G

Assessing Assumptions of Linear Regression

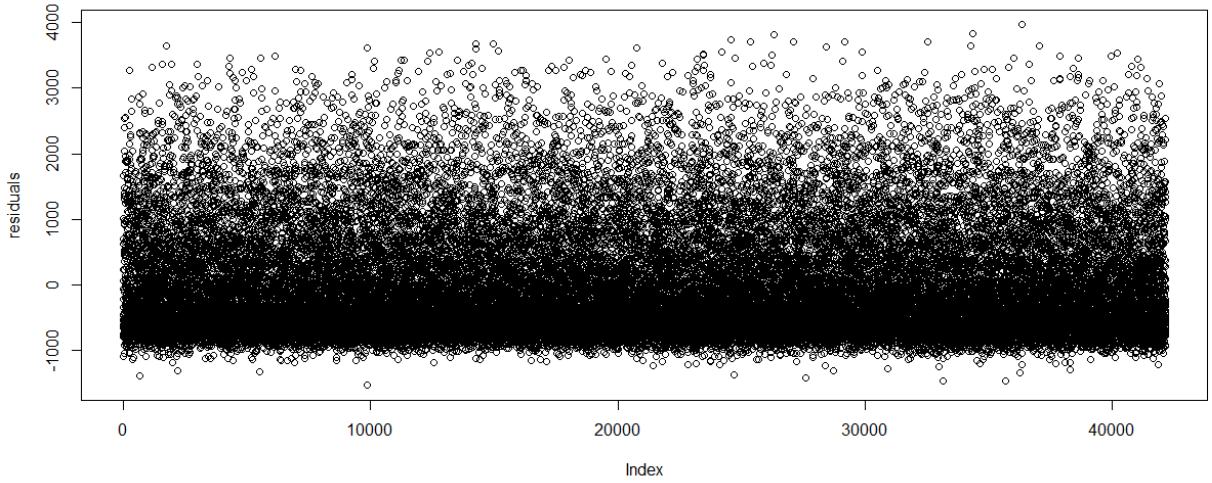


Figure G1. Residual plot of linear regression model, showing violation of linearity and homoscedasticity. Plot meeting the assumptions should show equal variance above and below 0 across all values on the x-axis.

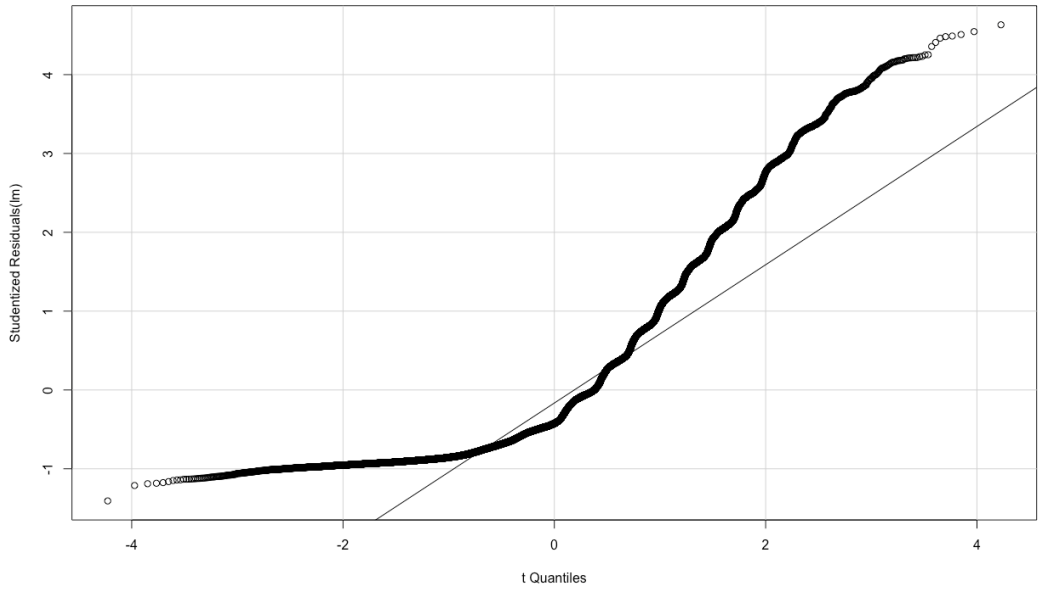


Figure G2. Q-Q plot of linear regression model, showing violation of normality. Plot meeting the assumption should show scatter point closely aligning with the solid line.

Appendix H

Assessing Assumption of Cox Regression

	rho	chisq	p
DAYSWAIT_CHRON	-0.003863	1.91e-01	0.6622
ETHCAT2	-0.001923	4.51e-02	0.8319
ETHCAT4	-0.007273	6.47e-01	0.4212
ETHCAT5	-0.015381	2.90e+00	0.0888
ETHCAT6	-0.002995	1.10e-01	0.7404
ETHCAT7	0.009425	1.08e+00	0.2978
ETHCAT9	0.001390	2.36e-02	0.8779
PERM_STATE_TRRAL	-0.015168	2.80e+00	0.0942
PERM_STATE_TRRAR	-0.010812	1.42e+00	0.2329
PERM_STATE_TRRAS	0.006827	5.26e-07	0.9994
PERM_STATE_TRRAZ	-0.011547	1.62e+00	0.2026
PERM_STATE_TRRCA	-0.012657	1.95e+00	0.1625
PERM_STATE_TRRCO	-0.013598	2.25e+00	0.1336
PERM_STATE_TRRCT	-0.019552	4.65e+00	0.0310
PERM_STATE_TRRDC	-0.008289	8.37e-01	0.3602
PERM_STATE_TRRDE	-0.008855	9.55e-01	0.3285
PERM_STATE_TRRFL	-0.012019	1.76e+00	0.1848
PERM_STATE_TRRGA	-0.010657	1.38e+00	0.2397
PERM_STATE_TRRGU	0.000605	4.44e-03	0.9469
PERM_STATE_TRRHI	-0.007588	7.01e-01	0.4024
PERM_STATE_TRRIA	-0.011525	1.62e+00	0.2037
PERM_STATE_TRRID	-0.013007	2.06e+00	0.1512
PERM_STATE_TRRIL	-0.013533	2.23e+00	0.1354
PERM_STATE_TRRIN	-0.014942	2.72e+00	0.0992
PERM_STATE_TRRKS	-0.011527	1.62e+00	0.2033
PERM_STATE_TRRKY	-0.015185	2.81e+00	0.0938
PERM_STATE_TRRLA	-0.010234	1.27e+00	0.2588
PERM_STATE_TRRMA	-0.011422	1.59e+00	0.2075
PERM_STATE_TRRMD	-0.014629	2.61e+00	0.1065
PERM_STATE_TRRME	-0.006716	5.49e-01	0.4586
PERM_STATE_TRRMI	-0.011815	1.70e+00	0.1924
PERM_STATE_TRRMN	-0.010197	1.27e+00	0.2605
PERM_STATE_TRRMO	-0.010771	1.41e+00	0.2347
PERM_STATE_TRRMP	0.001776	4.64e-08	0.9998
PERM_STATE_TRRMS	-0.015833	3.05e+00	0.0806
PERM_STATE_TRRMT	-0.005343	3.47e-01	0.5556
PERM_STATE_TRRNC	-0.010697	1.39e+00	0.2379
PERM_STATE_TRRND	-0.009902	1.19e+00	0.2747
PERM_STATE_TRRNE	-0.012854	2.01e+00	0.1561
PERM_STATE_TRRNH	-0.007908	7.61e-01	0.3829
PERM_STATE_TRRNJ	-0.016013	3.12e+00	0.0772
PERM_STATE_TRRNM	-0.016154	3.18e+00	0.0747
PERM_STATE_TRRNV	-0.010169	1.26e+00	0.2618
PERM_STATE_TRRNY	-0.012568	1.92e+00	0.1655
PERM_STATE_TRROH	-0.015188	2.81e+00	0.0937
PERM_STATE_TRROK	-0.011462	1.60e+00	0.2060
PERM_STATE_TRROR	-0.015255	2.83e+00	0.0923
PERM_STATE_TRRPA	-0.014260	2.48e+00	0.1156
PERM_STATE_TRRPR	-0.013414	2.19e+00	0.1390
PERM_STATE_TRRRI	-0.012243	1.83e+00	0.1766
PERM_STATE_TRRSC	-0.008350	8.49e-01	0.3569
PERM_STATE_TRRSD	-0.017063	3.55e+00	0.0597
PERM_STATE_TRRTN	-0.014603	2.60e+00	0.1071
PERM_STATE_TRRTX	-0.013012	2.06e+00	0.1511
PERM_STATE_TRRUT	-0.012973	2.05e+00	0.1523
PERM_STATE_TRRVA	-0.009618	1.13e+00	0.2886
PERM_STATE_TRRVI	-0.004013	1.97e-01	0.6573
PERM_STATE_TRRVT	-0.007564	6.97e-01	0.4039
PERM_STATE_TRRWA	-0.016239	3.21e+00	0.0731
PERM_STATE_TRRWI	-0.012717	1.97e+00	0.1605
PERM_STATE_TRRWV	-0.016313	3.24e+00	0.0718
PERM_STATE_TRRWY	-0.005876	4.20e-01	0.5169
PERM_STATE_TRRZZ	-0.012684	1.96e+00	0.1616

	rho	chisq	p
PREV_TXY	-0.010780	1.42e+00	0.2338
AGE_DON	-0.022633	6.58e+00	0.0103
DON_TYL	-0.001313	2.11e-02	0.8846
MELD_PELD_LAB_SCORE	0.003240	1.27e-01	0.7217
AGE	0.013024	2.13e+00	0.1443
SHARE_TY4	0.001226	1.84e-02	0.8922
SHARE_TY5	-0.014445	2.55e+00	0.1104
SHARE_TY6	0.008317	8.41e-01	0.3592
HEMATOCRIT_DON	0.019109	4.49e+00	0.0340
LT_ONE_WEEK_DONY	0.005388	3.54e-01	0.5519
GENDERM	0.005046	3.11e-01	0.5770
GENDER_DONM	-0.002008	4.92e-02	0.8244
ABOA1	-0.006115	4.58e-01	0.4988
ABOA1B	0.012411	1.87e+00	0.1715
ABOA2	0.007915	7.64e-01	0.3822
ABOA2B	0.086941	2.36e-06	0.9988
ABOAB	-0.019885	4.82e+00	0.0281
ABOB	-0.011691	1.67e+00	0.1966
ABOO	-0.006079	4.50e-01	0.5023
ABO_DONA1	0.006042	4.46e-01	0.5044
ABO_DONA1B	0.021296	5.54e+00	0.0186
ABO_DONA2	0.000870	9.23e-03	0.9234
ABO_DONA2B	0.009145	1.02e+00	0.3123
ABO_DONAB	0.007946	7.70e-01	0.3801
ABO_DONB	-0.010304	1.30e+00	0.2551
ABO_DONO	-0.001614	3.18e-02	0.8584
BMI_CALC	0.006652	5.60e-01	0.4542
BMI_DON_CALC	0.003898	1.84e-01	0.6677
GLOBAL	NA	9.58e+01	0.3438

Figure H1. Outputs of testing proportional hazard assumption. P value less than 0.05 indicates violation of the assumption.

Reference

1. “Liver Disease Information.” *American Liver Foundation*, www.liverfoundation.org/.

2. Kim, W. R., et al. “OPTN/SRTR 2012 Annual Data Report: Liver.” American Journal of Transplantation, vol. 14, no. S1, 2014, pp. 69–96., doi:10.1111/ajt.12581.

3. Masconi, Katya L, et al. “Reporting and Handling of Missing Data in Predictive Research for Prevalent Undiagnosed Type 2 Diabetes Mellitus: a Systematic Review.” *EPMA Journal*, vol. 6, no. 1, Nov. 2015, doi:10.1186/s13167-015-0028-0.

4. “Measurement System Analysis (MSA).” MoreSteam Lean Six Sigma Training and Technology, www.moresteam.com/toolbox/measurement-system-analysis.cfm.

5. Schrem, Harald, et al. “Value and Limitations of the BAR-Score for Donor Allocation in Liver Transplantation.” Langenbeck's Archives of Surgery, vol. 399, no. 8, 2014, pp. 1011–1019., doi:10.1007/s00423-014-1247-x.

6. “Epoch vs Iteration When Training Neural Networks.” Terminology - Epoch vs Iteration When Training Neural Networks - Stack Overflow, stackoverflow.com/questions/4752626/epoch-vs-iteration-when-training-neural-networks?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa.

7. Patel, Savan. “Chapter 2 : SVM (Support Vector Machine) - Theory – Machine Learning 101 – Medium.” Medium, Machine Learning 101, 3 May 2017, medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72.

8. “Kernel Method.” *Wikipedia*, Wikimedia Foundation, 27 Apr. 2018, en.wikipedia.org/wiki/Kernel_method.

9. “RBF SVM Parameters.” RBF SVM Parameters - Scikit-Learn 0.19.1 Documentation, scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html.

10. Chen, Zehua. “Interpretation of proportional hazards regression models.” ST3242 Notes. https://www.stat.nus.edu.sg/~stachenz/ST3242Notes4.pdf

11. “Organ Procurement and Transplantation Network.” *Regions - OPTN*, optn.transplant.hrsa.gov/members/regions/.

12. Goodfellow, Ian, et al. *Deep Learning*. Aaron Courville Online Book, 2017.

13. Ng, Andrew. “CS229 Lecture notes.” https://see.stanford.edu/materials/aimlcs229/cs229-notes3.pdf