UNIVERSITY OF SOUTHERN CALIFORNIA,
VITERBI SCHOOL OF ENGINEERING

Final Project Report

PHOENIX LENS

INF560-Data Informatics Professional Practicum

PROJECT BY-
Bhavesh Motwani
Nikita Gupta
Priyambada Jain
Sovan Rath

DATE-
04/25/2017

Table of Contents

1.INTRODUCTION

1.1 TEAM MEMBERS

NAME	ROLE	CONTACT	EMAIL
BHAVESH MOTWANI	Project Lead	3235409767	bmotwani@usc.edu
NIKITA GUPTA	GUI Developer	2134216756	nikitag@usc.edu
PRIYAMBADA JAIN	Technical Lead	2134214495	priyambj@usc.edu
SOVAN RATH	Project Manager	2134217096	sovanrat@usc.edu

1.2 PROJECT DESCRIPTION

Phoenix Lens is a data driven predictive model been developed to aid USC Keck School of Medicine during surgical procedure by predicting the requirement of blood during the process of surgery thereby reducing the wastage of blood and other resources.

Phoenix lens is a data-bound, multidisciplinary approach developed to fulfill the following objectives:

1. Predict the requirement of allogeneic blood during the process of surgery.
2. Assisting the surgeon to limit the allogeneic blood requirement during the surgery.
3. To improvise the efficiency of decision - “Is blood transfusion required?”.
4. To reduce the cost incurred during the surgery.
5. To ease the process of obtaining blood before surgery.

1.3 MOTIVATION

A blood transfusion is a safe, common procedure in which blood is given to the patient through an intravenous (IV) line in one of their blood vessels. Blood transfusions are done to replace blood lost during surgery or due to a serious injury.The amount of external blood that will be required for successful surgery is a preliminary measure that plays a pivotal role.The real motivation behind the system was to predict the amount of blood that will be required before surgery so as to avoid or minimize the wastage of blood incurred by the keck school of medicine.

1.4 PROJECT INFO

Project Title-	PHOENIX LENS
Date Started-	01/21/2017
Date Completed-	04/24/2017
Sponsor/Champion-	KECK SCHOOL OF MEDICINE

2.EXECUTIVE SUMMARY

2.1 Problem Statement

Predict the requirement of allogeneic blood during the process of surgery. Assisting the surgeon to limit the allogeneic blood requirement during the surgery. We will be delivering a web GUI (Graphical User Interface) where the system will predict PRBC_Ordered and Red blood Cells used based on the following input parameters:

- Age
- Surgical Procedure
- Surgical Specialty
- Patient Type
- SN_BM_Pre_OP_INR
- SN_BM_Pre_OP_Platelet_Count
- Results Before Surgery (Hemoglobin)

2.2 PROJECT SCOPE

Process Boundaries - Gives the practitioners,an extra perspective based on more accurate statistical methods and doesn't limit their observation or findings.

Subprocesses:

1. Predicting the units of blood required during surgery.

2.3 MILESTONES

During the project design we followed the Six-Sigma management techniques that comprises of DMAIC cycle. DMAIC cycle comprises of the following phases along with the deliverables :

i) Design Phase - Project Charter (Definition), Voice of Customer (VOC).
ii) Measure Phase - SIPOC, Process Map And Value Stream Map
iii) Analyse Phase - Root Cause Analysis(Fishbone Diagram)
iv) Improve Phase - Design,Development And Testing
v) Control Phase - Control And Monitoring Plan

The final deliverable is a package that contains a web based GUI along with a generalized prediction algorithm at the back end.A well -developed application in the form web application to help doctor make proper estimations of the no. of units of blood required for surgery based
upon recommendation system prediction.

2.4 TEST SCENARIO

Challenges -
Prediction - provides at most 0.9 units of deviation with data available prior to surgery.
Theoretical Acceptance - ~ 1.0 units (As the data grows it will reduce over time.)

2.5 CONCLUSION

There is a certain statistical relation between the various features in the given data set.
We have identified and predicted one such feature Red blood cells using the following features.

1) Age
2) Surgical Procedure
3) Surgical Specialty
4) Patient Type
5) SN_BM_Pre_OP_INR
6) SN_BM_Pre_OP_Platelet_Count
7) Results Before Surgery (Hemoglobin)

Which is then used to predict- SN_BM_PRBC_Ordered.

This would help the medical practitioner predict an approximate amount of blood that will be used during the surgery and reduces wastage.

2.6 PROJECT RECOMMENDATIONS

Prior to obtaining the data,it was recommended to de-identify the records in order to avoid any conflicts of privacy. The recommendations were to use machine learning as it is an emerging branch that could solve seemingly unsolvable problems.

2.7 ACTIONS TAKEN

Some of the actions taken to tackle the current problem were:

1. All the available data related to surgeries was collected from the Keck School Of Medicine.
2. The data was cleaned to avoid any inconsistencies.
3. The data was visualized to observe relations and patterns between features.
4. Finally,the data was passed to an analytical engine to get the desired output.

2.8 BENEFITS

A few of the benefits of our project are-
It would serve as an aid to the medical practitioners helping them better analyse their intuition and knowledge.It would give a more accurate answer to the amount of blood to be ordered and hence reduce wastage.Ultimately, it would reduce the cost incurred to the Keck School and the patient by reducing wastage.

3. LEAN SIGMA SIX PROJECT PHASES

3.1 DEFINE PHASE

During the Define phase of a Six Sigma DMAIC (Define, Measure, Analyze, Improve, Control) project, the team was responsible for clarifying the purpose and scope of the project, for getting a basic understanding of the process to be improved, and for determining the customers' perceptions and expectations for quality.

The problem faced by the Keck School of medicine is the increasing cost of surgeries due to the use of allogeneic blood and the wastage or lack of availability of blood during surgeries.

Blood transfusions are done to replace blood lost during surgery or due to a serious injury.The amount of external blood that will be required for successful surgery is a preliminary measure that plays a pivotal role.

Our team proposed an introduction of data driven predictive model(Pheonix Lens) which would be estimating use of autologous blood whenever possible and predict the approximate units of blood required for the surgery.During this design phase, we analysed the VOC(Voice of consumer) and presented the [project](#) [charter](#) to address the concern. Also during the monthly presentation with the Stakeholder we presented the project charter to them in order to verify that it addresses all the desired concerns.

3.2 MEASURE PHASE

Keck School of Medicine provided us the process map,which described in depth of patient blood ordering. The process that is being followed is as follows:

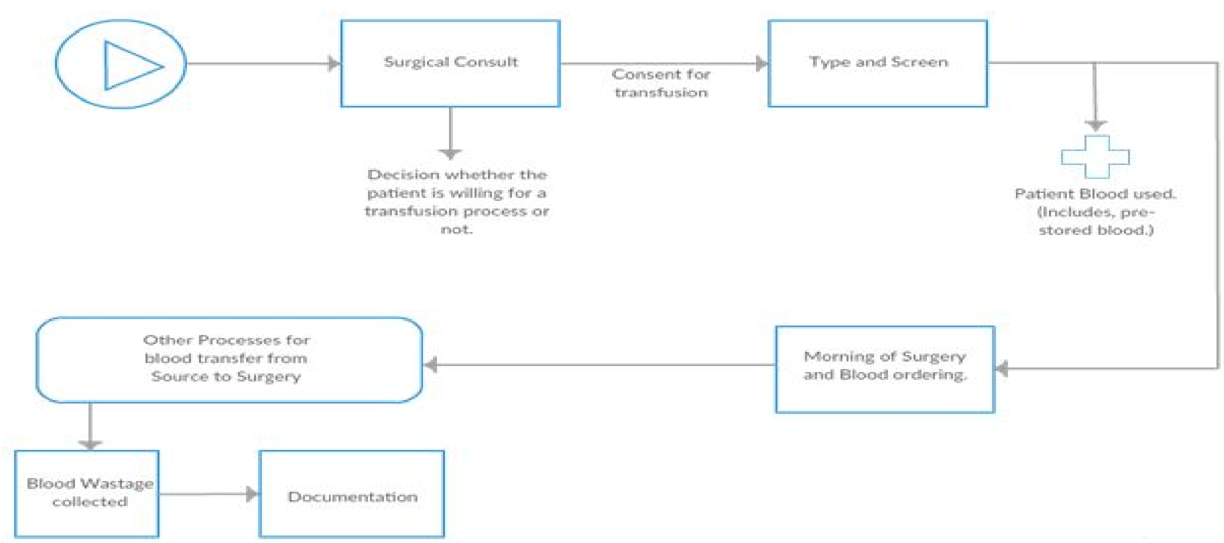


Figure 1

Based on the inputs provided by our customers at the Keck school of medicine, we analyse that the use of predictive system before the actual ordering of the blood to validate the estimated quantity of blood. The following process map introduces Phoenix Lens in the patient blood ordering system.

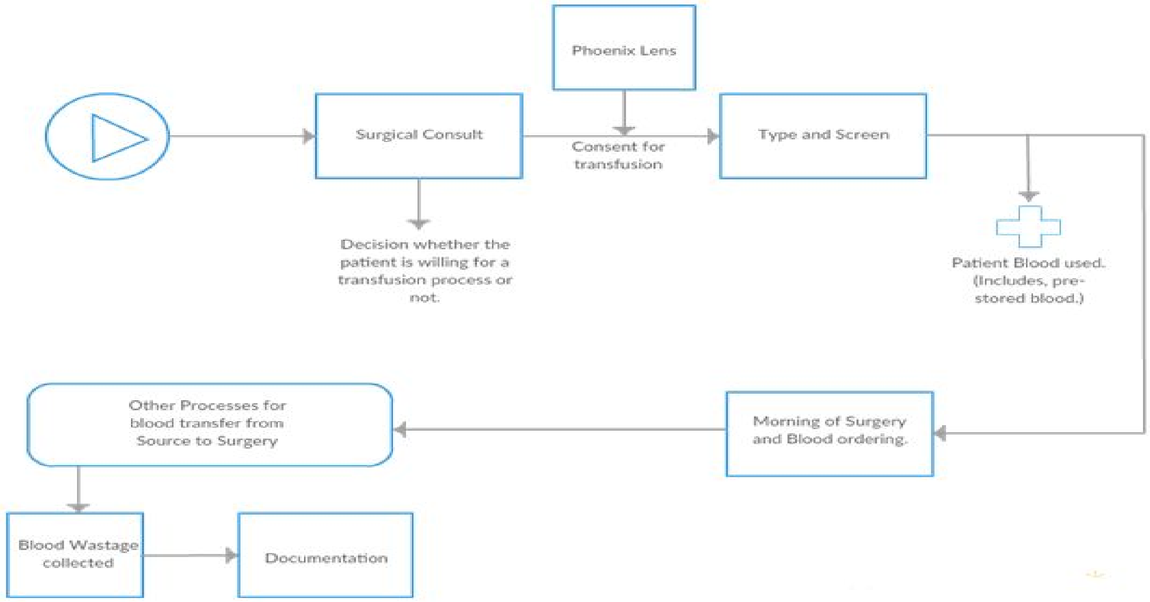


Figure 2

We analyzed the process further using SIPOC diagram. A SIPOC diagram is a tool used to identify all relevant elements of a process improvement project before work begins. It helps define a complex project and is typically employed at the Measure phase of the Six Sigma .

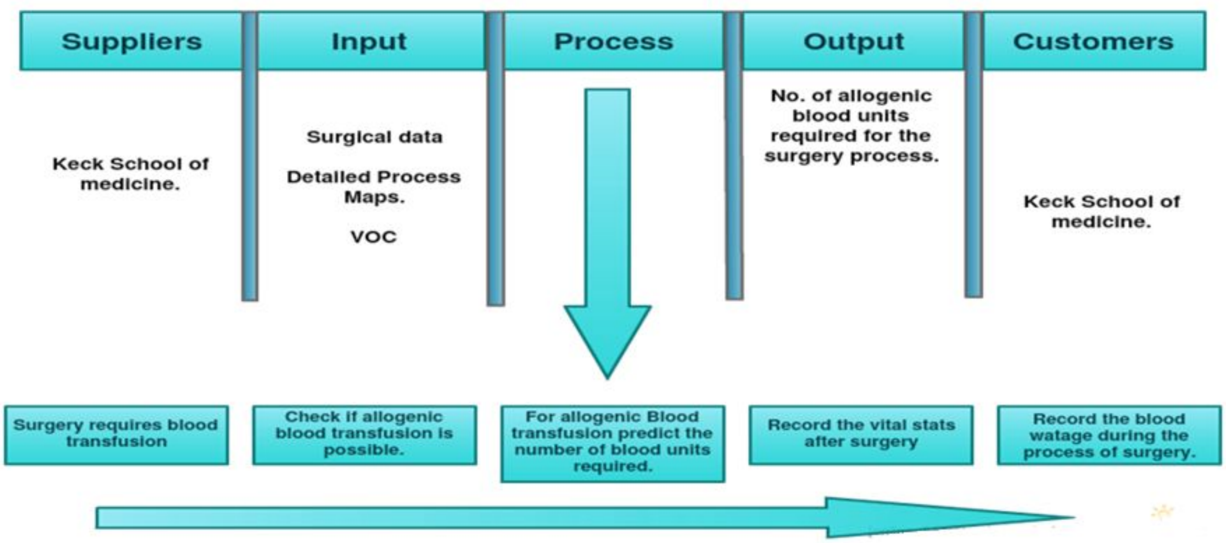


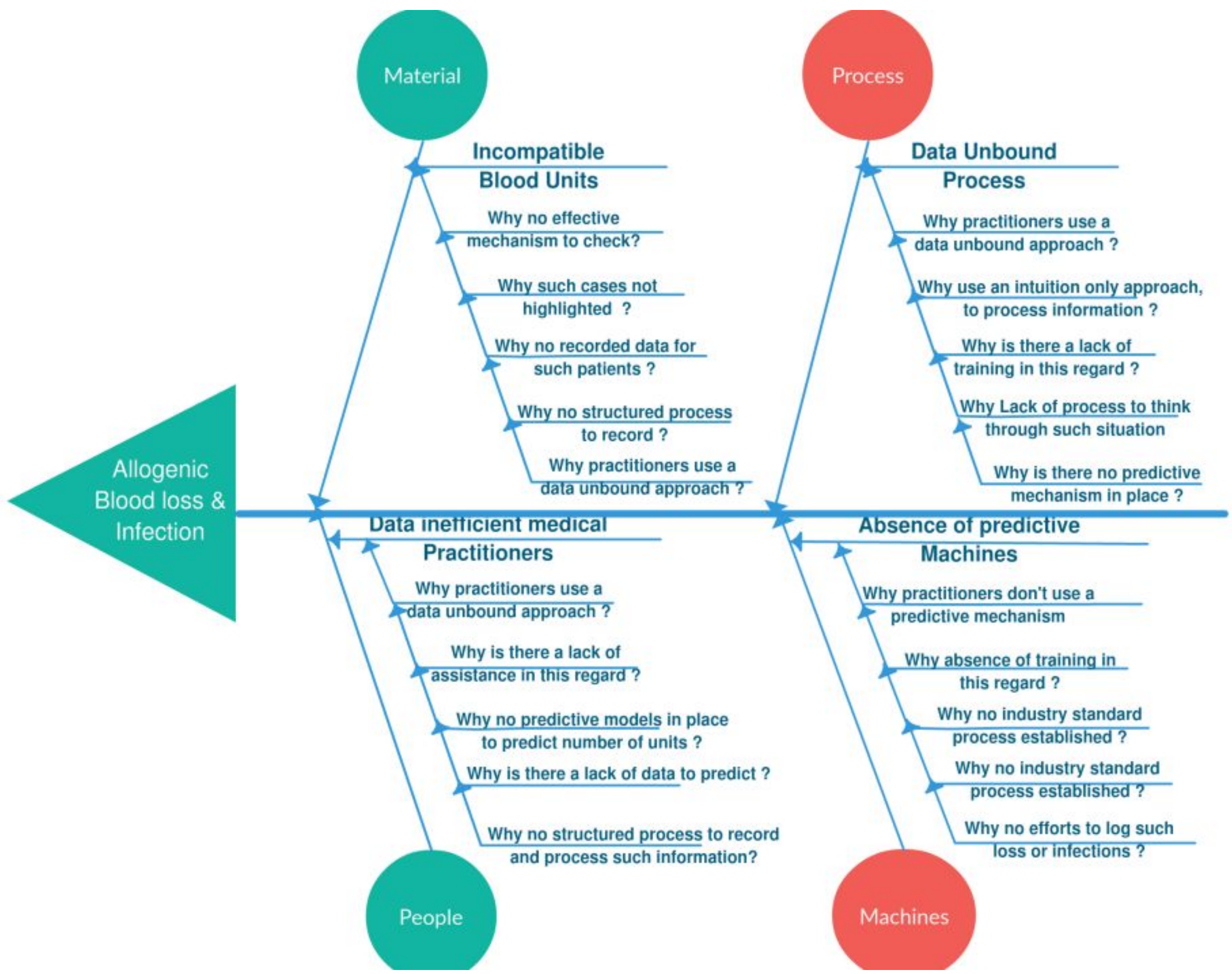
Figure 3

It would be an efficient way to decide on the process of allogeneic blood transfusion by taking a data based decision where we look at the past records of blood transfusion cases and try to predict the number of blood units that will be required.Based on the available parameter before the surgery, we will try to predict the approximate number of blood units required for that particular surgery. Also the model will predict the usage of the blood units during the surgery (SN_BM_Red_Blood_Cells).

Data Collection Planning and Execution - We developed a R script to clean the data and make it fit for the model construction.

1) Converted the string columns to float type
2) Converted the Surgical Procedure to category and assigned the cat codes.
3) Converted the Surgical Speciality to category and assigned the cat codes.
4) Removed the logical and statistical outliers.

3.3 ANALYSE PHASE



Fish Bone for the root cause analysis

Figure 4

We analysed the data and by using the fishbone diagram came into a conclusion that a predictive mechanism is necessary to assist the medical practitioners before the ordering of blood units. We list down all the potential solutions that have been given a fair thought. We also present the solution taken after careful consideration of data.

Potential Solutions

1.Predict all the components of the blood separately. (Red Blood Cells, Plasma and Crycinophilates.)
2. Predict only the Red_Blood_cells.
3. Predict the Pre_RBC.
4. Predict the estimated blood loss (EBL) column for the data set.
5. Predict the blood loss by using (Pre_RBC - Red_Blood_Cells)

These possible solutions have their own merits and challenges.

The first solution doesn't agree with the VOC. The solution predicts the components after the surgical process hence shall be of least usage to the customer. Further, we deduced that only RBC has noteworthy correlation with PRE_RBC which is the value needed to be calculated before the surgery. Therefore, in the light of these evidences the idea was dropped.
The second solution to predict RBC is good however this data is found during or after the course of surgery. Predicting this wouldn't reduce the blood loss as blood has been ordered previously. Predicting PRE_RBC is critical towards reducing the loss as the ordering process depends on it. Although predicting the EBL would provide the practitioners with the number of blood units lost however our VOC states that EBL is unreliable and the record is filled by the staff without the usage of any proper measure. The final solution, although brilliant doesn't have a good correlation with the data and hence is not chosen.

Finalized Solution

Our solution believes that the PRE_RBC prediction is critical towards the blood ordering process. This column in our dataset can point towards an allogeneic or an autologous transfusion. However the absolute correlation of this column with other fields available before the surgery is within the range of (0.1 - 0.3). Therefore this column needs the support of a column RBC which is available after the surgical process and has a positive correlation of 0.6.

Our chosen solution based on our analysis is to predict the PRE_RBC in a two step process. First, to predict the number of blood units (RBC) used during the process of surgery. Further based on our prediction, predict the number of blood units required before the surgery (PRE_RBC).

Charts that helped us come up with the solutions.

1. Correlation heat maps.
2. Process maps & SIPOC

3. VOC (Customer Questionnaire.)
4. Fishbone diagram

Formula

Y0 = SN - BN - RBC
Inputs - Age
Surgical Procedure
Surgical Speciality
Patient Type
INR
Platelets
Hemoglobin

Y1 = SN -BN - PRE_RBC
Inputs - Age
Surgical Procedure
Surgical Speciality
Patient Type
INR
Platelets
Hemoglobin
SN - BN - RBC

SWOT

Strength -

Agrees with the VOC.
Data available before surgery.

Weakness -

Poor correlation with test and other data available before surgery

Opportunities -

Any change in the prediction mechanism will assist the practitioners.
As the storing and record keeping mechanism becomes stable, and more records are available we shall give better results.

Threats -

Poor prediction results would reduce the relevance of the product and thereby decrease the usage of predictive mechanism towards determining the number of blood units required before surgery.

Pilot prescribed

The solution should be available as a simple web app attached to the data source. After the records for the test are available, the practitioners should be able to predict the number of blood units. This will assist them during their process of analysis before the blood ordering.

3.4 IMPROVE PHASE

The phoenix lens is developed using the following learning model architecture.

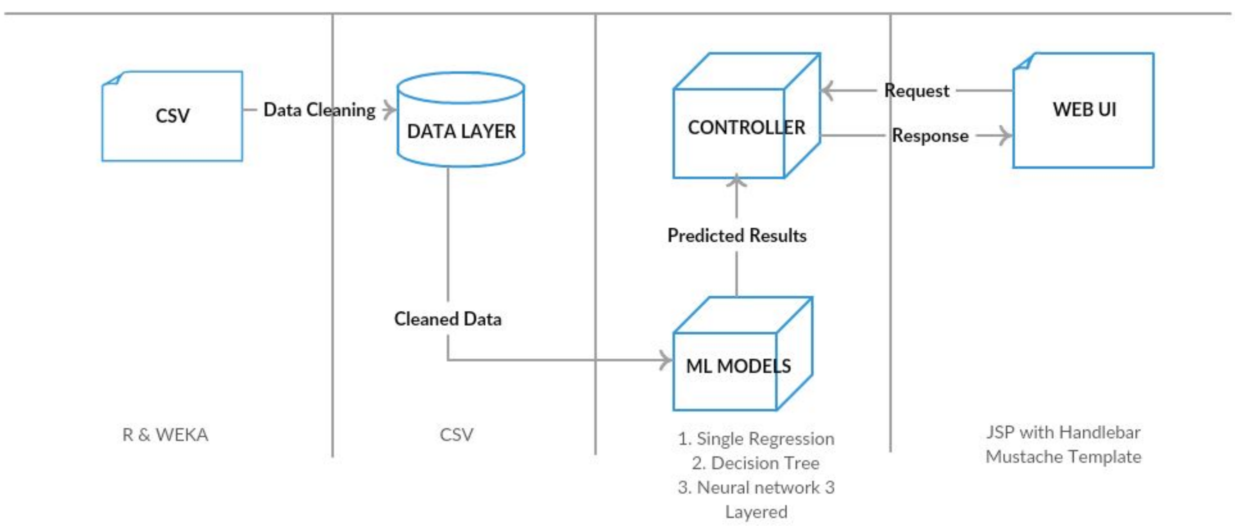


Figure 5

Followed by analyzing the risks that can be encountered during our model design. We followed the risk management analysis where we analyzed the risks and also provided the corresponding mitigation.

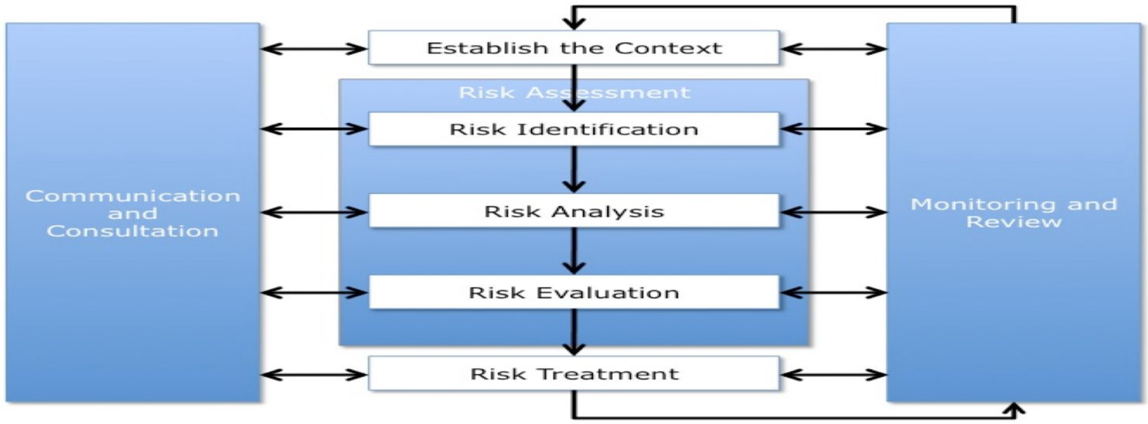


Figure 6

This helped in process of identification, assessment, and prioritization of risks.

TASKS	RISK IDENTIFICATION	TYPE OF RISK	RPN	RISK CALCULATIONS			CONSEQUENCES	MITIGATION
				Likelihood	*Impact*	*Risk Rating*		
1. Analysis Data	Missing Values	Technical	High	(2) Moderate	(2) Moderate	(3) Moderate	Decrease in Efficiency	Preprocessing of Data before analysis.
2. Analysis Data	Incorrect values	Technical	High	(2) Medium	(1) Major	(3) Medium	Error rate increases in Predictive System.	Double check the values before use in live system.
3. Analysis Phase	Less Computational Space	Technical	High	(2) Medium	(1) Major	(3) Medium	Memory out of bound not able to compute the results.	Move the system to cloud server.
4. During Surgery	Power Failure	Technical	High	(3) Unlikely	(1) Major	(3) Medium	Failure of the predictive system.	Ensure power backup in initial stages only.

Who conducted the Risk Assessment? **Completed by:** Phoenix Lens **Signature:** **Date:** 02/18/2017	**Who approved the Risk Assessment?** **Approved by:** Keck School of Medicine **Signature:** **Date:** 03/09/2017

Figure 7

A **work breakdown structure** (**WBS**) is a key project deliverable that organizes the team's work into manageable sections. Following is the Pheonix Lens WBS where we granulated the model and distributed the tasks among the team to begin with development.

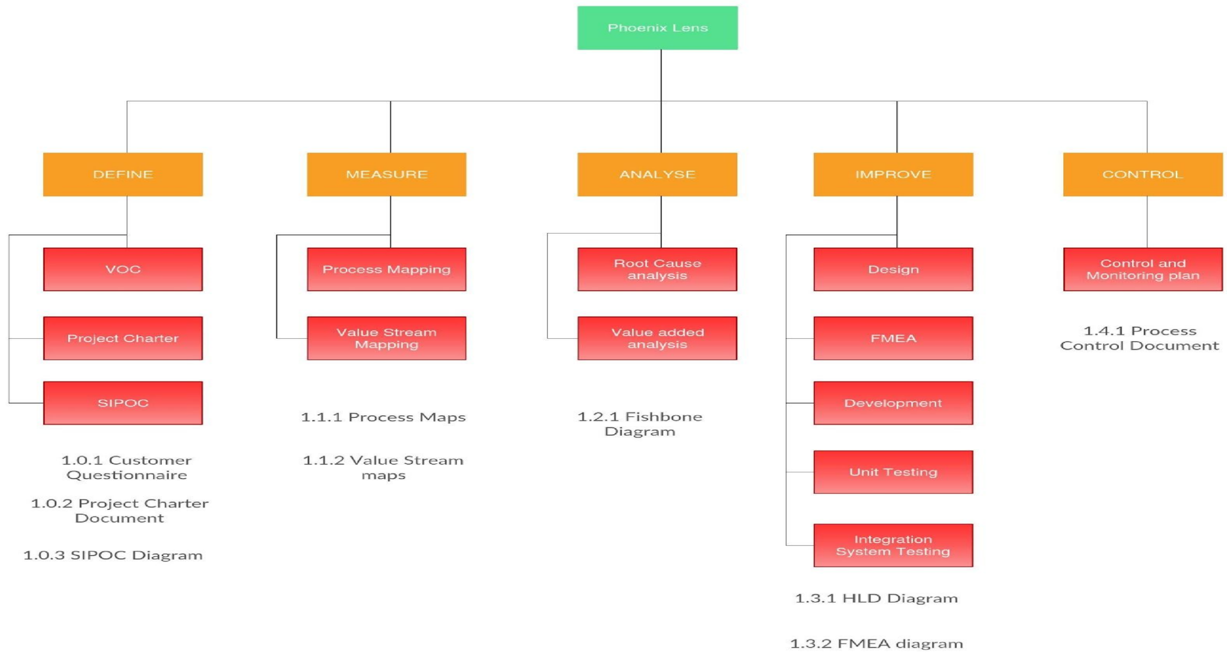


Figure 8

The improve phase involved the following steps-

Design Development and Testing

- Data Pre-processing: Cleaning and modeling of the data.

- Data Visualization and Feature Engineering: This is one of the pivotal section where we decided the independent variables (based on process and correlation).

- Model Construction: Deployed various model as explained later and based on the performance matrix selected the best out of them.

- GUI Interface: Now we designed the GUI interface where the surgeon can enter the input parameters and the model will predict the required PRBC to be ordered.

- Integration: Once the interface was done our team progressed with the model connectivity with the GUI.

- Testing: Based on the dataset provided by the Keck School of Medicine we examined the model for the predictions.

3.5 CONTROL PHASE

As the development is continuously iterative process, we will provide a feature through which the user can provide us feedback for scope of improvement.

We proposed to control the process by using a process control plan (PCP) to monitor the predictions made by the system Phoenix lens. This should be used towards checking whether our system is functioning within the limits of LSL and HSL.This would include the improvements to the system based on the feedbacks provided by the stakeholders.

4.RESULTS AND SYSTEM IMPLEMENTATION

4.1 System Implementation :

During the system implementation we have utilized the following technology stack :

1) R - data cleaning and Data Visualization
2) Weka - Data Visualization
3) Python - pandas - data frames for processing
4) Python - scikit - machine learning library and model evaluation
5) Graphical User Interface : Bootstrap, HTML, CSS, JavaScript, JQuery, Python,Flask.

Feature Engineering :

First things first for the system to be developed the foremost requirement is the data set followed by understanding the significance of variable.

Phoenix Lens is based on the dataset provided by the USC Keck School Of Medicine.

Description of the dataset : The data set contained the following fields:

- Masked FIN
- SN-BM-Pre-OP INR(international normalized ratio)
- SN-BM-Pre-OP Platelet Count
- SN - BM - PRBC Ordered
- Masked SurgCaseID
- SURG_PROCEDURE
- SURGICAL_SPECIALTY
- Surgeon Hash Name
- PATIENT_TYPE
- Allogeneic Blood Transfusion

- SN - BM - Red Blood Cells
- SN - BM - Fresh Frozen Plasma
- SN - BM - Platelets
- SN - BM - Cryoprecipitate
- ResultsBeforeSurgery
- ResultAfterSurgery
- EBL: Estimated blood loss

Correlation Analysis :

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.
A positive **correlation** indicates the extent to which those variables increase or decrease in parallel.

A negative **correlation** indicates the extent to which one variable increases as the other decreases.

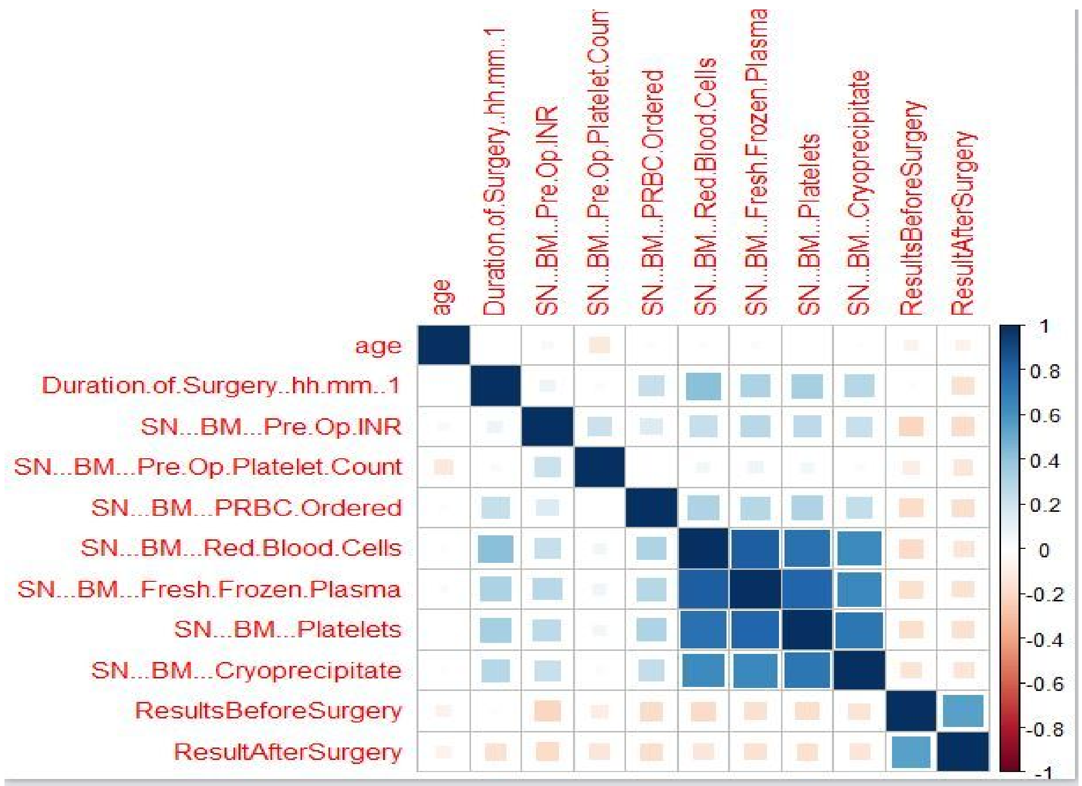


Figure 9

Feature Selection:

Based on the analysis done using heatmap as described in Figure 9 we came up with the following features as independent features for our model.

Independent Variables/Inputs:

1) Age,
2) Surgical_Procedure,
3) Surgical_Specaility,
4) SN_BM_Pre_OP_INR,
5) SN_BM_Platelet_Count,
6) ResultsBeforeSurgery(Hemoglobin before surgery)

Dependent Variable/Output:

SN_BM_PRBC_Ordered

Machine learning model:

1) ***Linear Regression Model:***

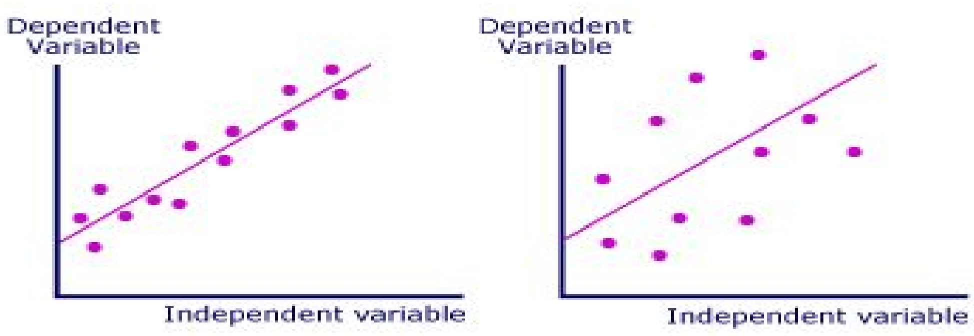
Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Linear regression is an attractive model because the representation is so simple.The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be: **y = B0 + B1*x**

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).



Graph A Graph B

Inputs : Age, Surgical_Prod, Surgical_Specaility, SN_BM_Pre_OP_INR, SN_BM_Platelet_Count, ResultBeforeSurgery.

Output : SN_BM_PRBC_Ordered

Approach :

1. Based on the parameters that are available with the surgeon before hand i.e prior to the start of the surgery.

2. Calculated the correlation of these attributes with the dependent variable (SN_BM_PRBC_Ordered).

3. As Surgical_Speciality and Surgical_Prod determine the SN_BM_PRBC_Ordered extensively, our main focus was to concentrate the model around these parameters and generate different models around them for better results

4. Surgical_Specaility are of 29 types, each has it's own model that we used based on the specific instance.

2) Random Forest Regression Model:

Decision trees involve the greedy selection of the best split point from the dataset at each step. This algorithm makes decision trees susceptible to high variance if they are not pruned. This high variance can be harnessed and reduced by creating multiple trees with different samples of the training dataset (different views of the problem) and combining their predictions. This approach is called bootstrap aggregation or bagging for short.

A limitation of bagging is that the same greedy algorithm is used to create each tree, meaning that it is likely that the same or very similar split points will be chosen in each tree making the different trees very similar (trees will be correlated). This, in turn, makes their predictions similar, mitigating the variance originally sought.

We can force the decision trees to be different by limiting the features (rows) that the greedy algorithm can evaluate at each split point when creating the tree. This is called the Random Forest algorithm. Like bagging, multiple samples of the training dataset are taken and a different tree trained on each. The difference is that at each point a split is made in the data and added to the tree, only a fixed subset of attributes can be considered.

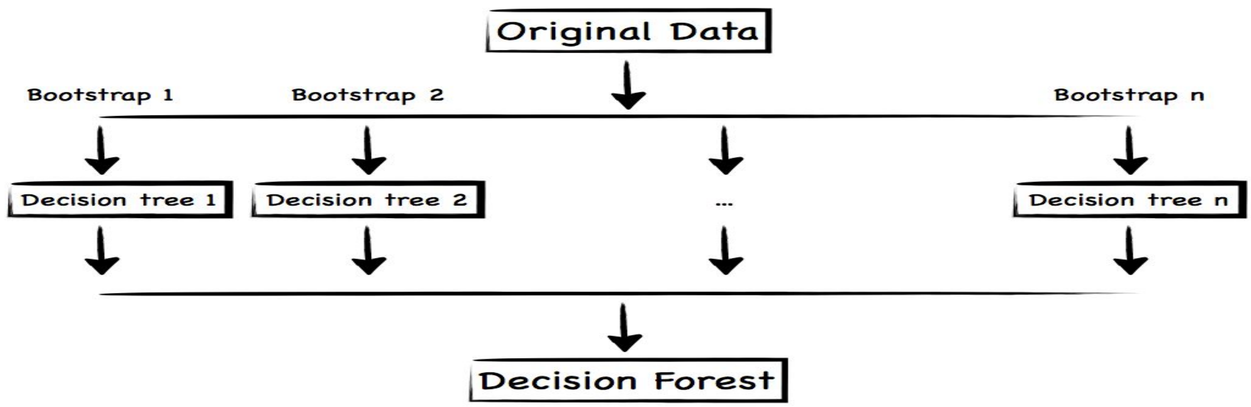


Figure 10

Inputs : Age, Surgical_Prod, Surgical_Specaility, Patient Type,SN_BM_Pre_OP_INR, SN_BM_Platelet_Count, ResultBeforeSurgery.

Output : SN_BM_PRBC_Ordered

Approach : Based on the parameters that are available with the surgeon before hand i.e prior the start of the surgery we created a pipeline of two models where the Red Blood Cells are predicted using the inputs followed by the prediction of PRBC_Ordered.

4.2 Performance Measurement:

In statistics, the mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by :

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|\,.$$

As the name suggests, the mean absolute error is an average of the absolute errors |ei| = |yi - xi| , where yi is the prediction and xi the true value.

Results :

We tried the following machine learning algorithms and compared the results based on MAE(Mean Absolute Error).

1) **Linear Regression**

Linear Regression Model Mean Absolute Error : 1.8777

2) **KNN Regression**

K Nearest Neighbour Mean Absolute Error : 1.42865

3) **Decision Tree Algorithm**

Decision Tree Algorithm (Single tree) : 1.2484

4) **Random Forest Model**

Random Forest Model Mean Absolute Error : 0.779642

As indicated from the results we decided to use **Random Forest Model** as it provided the least MAE error which is indicative of a good result and fair accuracy for the predictions.

Test Scenario and Acceptance Criteria:

Challenges - Prediction - provides at most 0.9 units of deviation with data available prior to surgery.

Theoretical Acceptance - ~ 1.0 units (As the data grows it will reduce over the time.)

Current Acceptance - 0.78 units. of deviation for PRBC_Ordered.

4.3 Graphical User Interface:

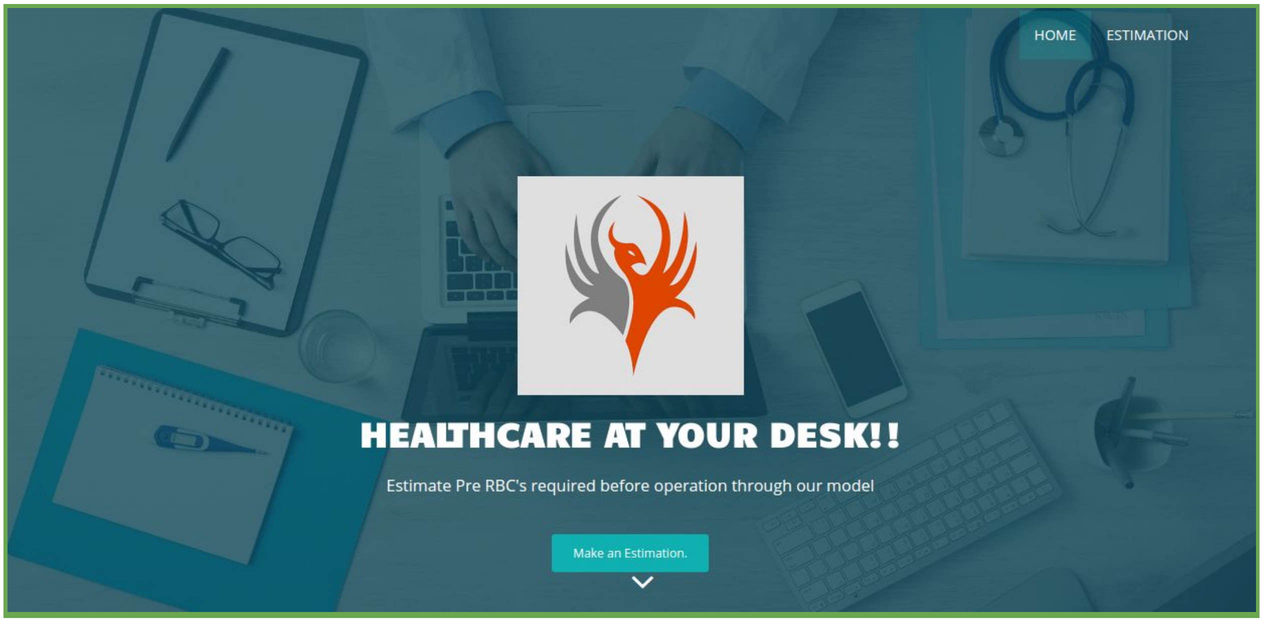


Figure 11

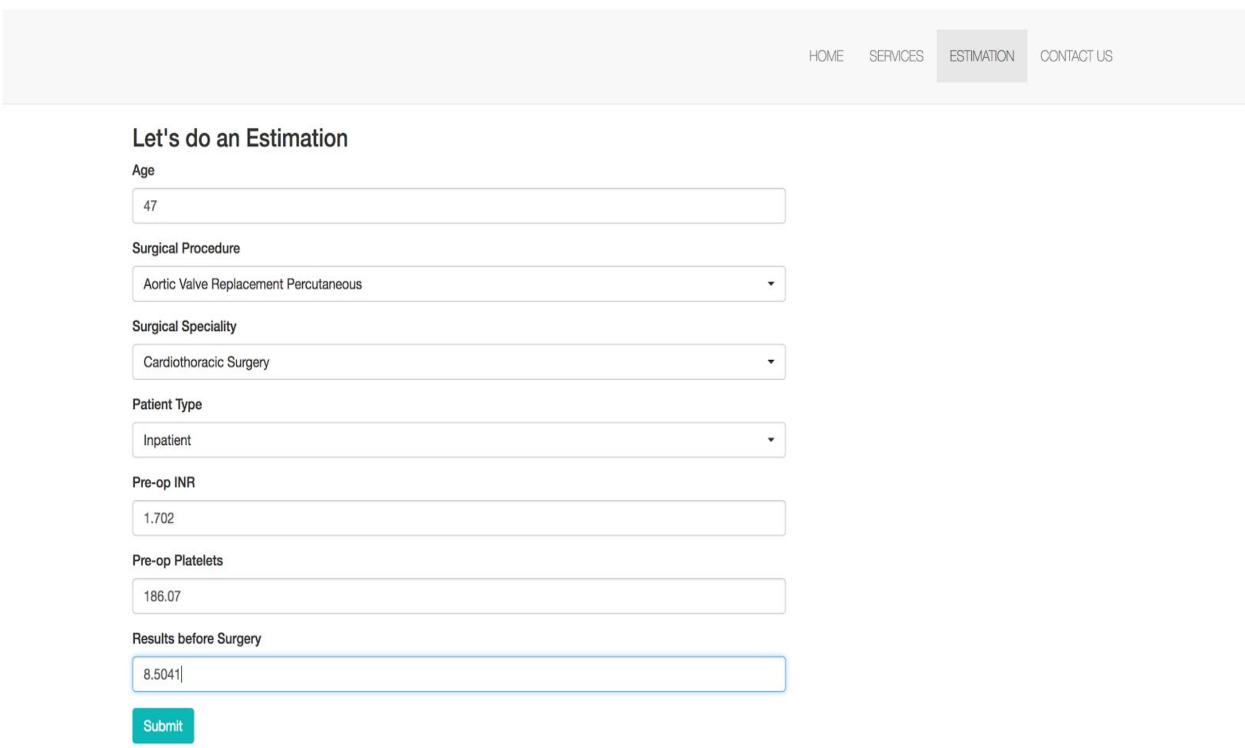


Figure 12

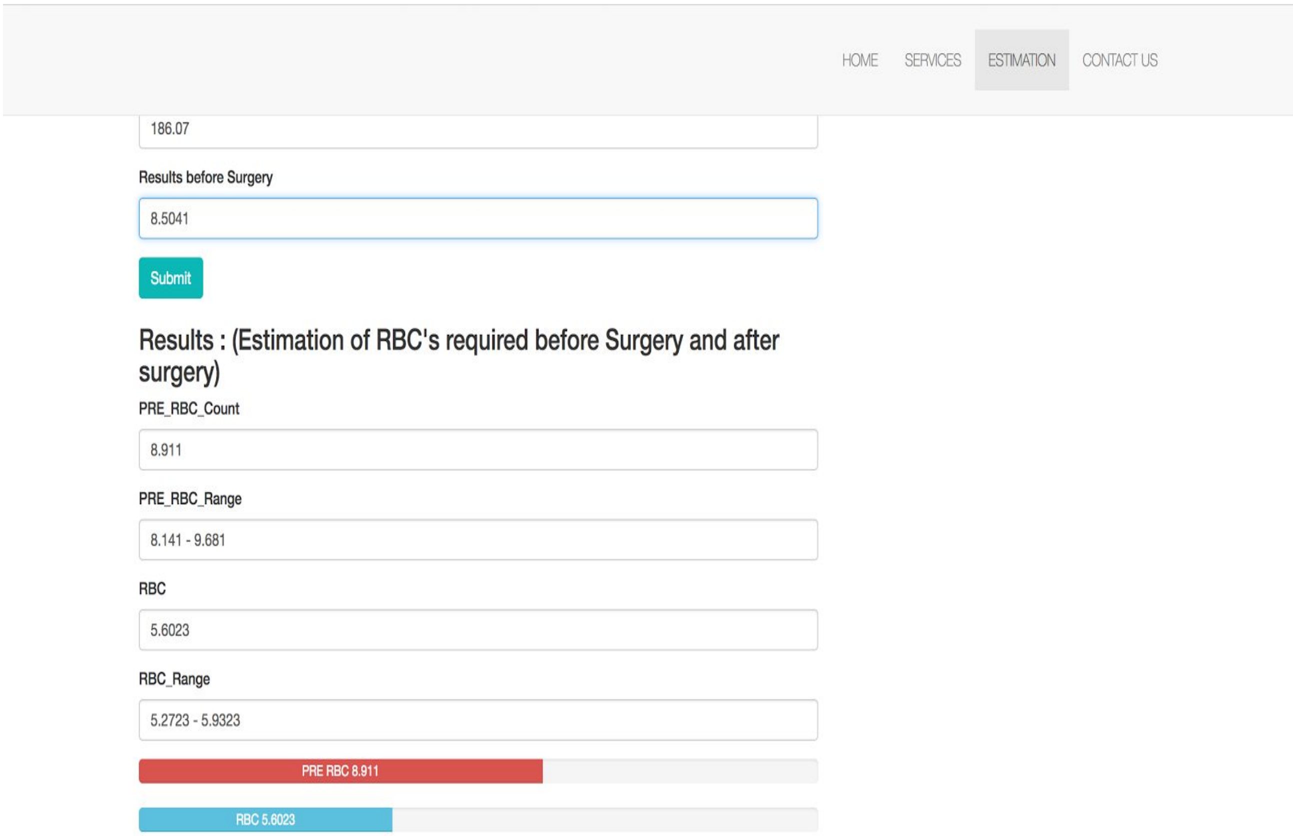

HOME
SERVICES
ESTIMATION
CONTACT US
186.07
Results before Surgery
8.5041
Submit
Results : (Estimation of RBC's required before Surgery and after surgery)
PRE_RBC_Count
8.911
PRE_RBC_Range
8.141 - 9.681
RBC
5.6023
RBC_Range
5.2723 - 5.9323
PRE RBC 8.911
RBC 5.6023

Figure 13

5. References

- ❖ Project Charter
  https://drive.google.com/a/usc.edu/file/d/0B-rnx_dKiVzuMXZpdkN3YllITm8/view?usp=sharing
- ❖ FishBone Diagram
  https://drive.google.com/open?id=0B-rnx_dKiVzuVmk3YlA2LWFXc2M
- ❖ Gantt Chart
  https://drive.google.com/open?id=0B23it1bceRVJVk5PUmhlTFctZ2c
  https://drive.google.com/open?id=0B23it1bceRVJVzNlS0RZalBfd3M
- ❖ Technical Report
  https://drive.google.com/open?id=0B-rnx_dKiVzuWWh2dl9vSTFBMm8
- ❖ Process Map for Patient Blood Ordering
  https://blackboard.usc.edu/bbcswebdav/pid-4605998-dt-content-rid-13344286_2/courses/20171_inf_560_32429/Process%20Map.pdf
- ❖ Measure Phase Analysis
  https://drive.google.com/open?id=0B-rnx_dKiVzuX0ttZ0NXNkg4dGc
- ❖ Analyse Phase
  https://drive.google.com/open?id=0B-rnx_dKiVzuVmk3YlA2LWFXc2M
- ❖ Improve Phase
  https://drive.google.com/open?id=1NcfXMNVY7RqX5up4ybtthrrVbevkuM27IVaEnUpokpc
- ❖ Linear Regression
  http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- ❖ Random Forest
  http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- ❖ MAE
  http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html