Risk Factor Distribution of Cardiovascular Disease in Shanghai, China

by
Yutong Fan

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Business Information System
(Honors Associate)

Presented June 14, 2017
Commencement June 2017

AN ABSTRACT OF THE THESIS OF

Yutong Fan for the degree of Honors Baccalaureate of Science in Business Information System presented on June 14, 2017.
Title: Risk Factor Distribution of Cardiovascular Disease in Shanghai, China

Abstract approved:_____
Bin Zhu

Nowadays, big data and data mining play a more and more important role in numerous industries. Hospital industry is no exception. Hospitals in China have one of the highest patient volumes throughout the world, and they start to take advantage of these data and information, but these valuable data are far away from being well utilized.

In China, the relationship between doctors and patients are intense, and doctors get the blame regardless of whether they are the responsible party. Sometimes, it is the patient's' behavior that causes them to be rehospitalized. This thesis used the decision tree model to mine a dataset provided by Department of Cardiology of a hospital in Shanghai, China and determine the reasons that cause the patients to come back to the hospital.

Key Words: data mining, CART Tree Model
Corresponding e-mail address: yutongfan2016@yahoo.com

Risk Factor Distribution of Cardiovascular Disease in Shanghai, China

by Yutong Fan

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Business Information System
(Honors Associate)

Presented June 14, 2017
Commencement June 2017

Honors Baccalaureate of Science in Business Information System project of Yutong Fan presented on June 14, 2017.

APPROVED:

_____

Bin Zhu, Mentor, College of Business

_____

Zhaohui Wu, Committee Member, College of Business

_____

Shaokun Fan, Committee Member, College of Business

_____

Toni Doolen, Dean, Oregon State University Honors College

I understand that my project will become part of the permanent collection of Oregon State University, Honors College. My signature below authorizes release of my project to any reader upon request.

_____

Yutong Fan, Author

Table of Content

Chapter 1: Introduction

1.1 Summary

Today is the age of big data. Regardless of the industry, data mining is useful for that it enables people to analyze the dataset then identify their inadequacy and come up with best solutions. The healthcare industry is no exception. In this industry, data mining allows people to improve the care for the patients, as well as reduce the spending for the hospitals.

The most important part of data mining is data collection. For this study, we mainly focused on the on the factors that cause rehospitalization of patients with cardiovascular disease in Shanghai. We received the data from the Cardiology department of a hospital located in Shanghai, China. The data include the demographic information, diagnosis, previous diagnosis (if any), and abnormal test index of each patient, which will be further discussed in the Data Description Chapter. In total, we studied 99 patients, among which 49 of them came back at least once after their first hospitalization, and 50 of them did not. The reason is that in China, the concept of big data and data mining is starting to rise but it is not fully adapted in the healthcare industry yet. Therefore, during the data acquisition period, we discovered several deficiencies in the way they run the department, and propose several possible improvements in their patient recording system, as well as the business process in general. These possible improvements will be detailed construed in the Data Acquisition Process Chapter.

There are numerous ways to mine data to support strategic decision-making. Since we do not have a hypothesis in advance, we implemented the idea of unsupervised learning, which is a machine-learning algorithm that finds the underlying model or pattern in the data in order to better understand the data. Of the unsupervised learning algorithm, we practiced the decision tree model, which will be supplemented in the Methodology Chapter. Based on our data, we found a four correlated risk factors with significance and a possible explanation why they hold such an importance.

1.2 Overview of Medical System in China

Since ancient times, doctors and medical personnel are well-respected professions for that they save people's lives and give them second chances. However, in China, it is not true anymore. Deep misunderstandings exist between doctors (or other medical personnel) and patients. Part of the reason is the insufficient coverage of medical insurance. In a vast country like China, it is nearly impossible to cover a generous amount of the medical bills for 1.39 billions of people. At least for now, the Chinese government can only afford to cover the basic medical expenditures. According to the World Health Organization Global Health Expenditure database, China's out-of-pocket health expenditure accounts for 31.99% of total expenditure on health, which is almost three times the proportion of United States' out-of-pocket health expenditures. As of the private expenditure on health, the out-of-pocket health expenditure accounts for 72.35%, whereas United States' percentage is 21.37%.

Human body works mysteriously. There is no guarantee for a cure under any circumstances. However, when patients have to pay a large sum for their medical

treatment, they are under the false impression that they can buy their health. They expect a return for the price they paid, which is their recovery, thus causes the gap of expectations. Sometimes, patients' disobedience of doctors' orders will worsen their conditions, but they still think that the doctors did not treat them well.

Unlike in United States, hospitals in China are mostly public. Public hospitals provide about 90% of medical service, and have the best medical personnel and medical resources. Chinese hospitals are organized based on a 3-tier system. This system evaluates a hospital by its ability to provide medical care, medical education, and medical research. The 3 tiers are Primary, Secondary or Tertiary. Based on the level of service provision, size, medical technology, medical equipment, and management and medical quality, these 3 tiers are further divided into 3 subsidiary levels: A, B and C, with an additional special level - 3AAA, which is reserved for the most specialized hospitals (Li).

The proficiencies of doctors have a relatively huge difference between tiers and levels, comparing to doctors in United States. Therefore, patients tend to go to hospitals with higher rating in order to get the best treatment. As a result, high rating hospitals are overcrowded while resources in low rating hospitals are wasted.

1.3 Overview of Medical Resources in China

China has 1.39 billions of people, which accounts for 18.5% of world population. China's medical industry has been developing rapidly in the past few years thanks to the new policies of healthcare reformation. The development does not have any sign of slowing down; on the contrary, the medical industry has a high potential of growth in the future. According to the latest data from National Bureau of Statistics of China, in 2014,

the total health expenditure of China was 3.5 trillion dollars, which is 4.6 times of the total health expenditure of China in 2004.

The surge of the demand for medical service is caused by the aging of population, the urbanization, the change of lifestyle, and the universal medical insurance system (DELOITTE), however, the supply falls short of the demand. The amount of health institutions in 2014 is only 1.2 times of the amount in 2004. The amount of healthcare personnel in 2014 is only 1.5 times of the amount in 2004. Another problem of the health care system is the humongous population base. United States spends 3.0 trillion dollars on health expenditure in 2014, which is less than that the amount spent in China (3.5 trillion dollars). Nevertheless, when comparing the per capita statistics, United States spends 9402.17 dollars on each individual, which is 22 times of the amount China spends on each citizen.

Chapter 2: Data

2.1 Data Description

The dataset we built can be divided into two groups. The first group includes 49 patients who have been rehospitalized in the past three years. The other group includes 50 patients who have not been rehospitalized in the past three years. The data collected for each patient consist of the following five main categories, which are demographic, diagnosis, previous diagnosis, abnormal laboratory test result (Higher than normal), and abnormal laboratory test result (Lower than normal).

- **Demographic:** Basic demographic information about each patient
- **Diagnosis:** The top 7 most common diagnosis for the most recent hospitalization
- **Previous Diagnosis:** The top 7 most common diagnosis of the previous hospitalization (if any)
- **Abnormal laboratory test result (Higher than normal):** The laboratory test result that is higher than the normal range
- **Abnormal laboratory test result (Lower than normal):** The laboratory test result that is lower than the normal range

Under each main category, there are several subcategories. The complete list can be found in the appendix.

2.2 Data Acquisition Process

The data acquisition process was the most time-consuming part of this project. There are several difficulties encountered along the way, which reflects the immaturity

within their patient data recording system, and at the same time, offers us an opportunity to help the department to better their patient data recording system.

First of all, the Department of Cardiology already has the awareness of the importance of the patient recording system. The department, as well as the hospital in general, has been trying to further the system. However, they still have a long way to go. For starter, the hospital has operated for almost 100 years. The alternation within the system has only been 1 to 2 years. To have a relatively strong system, the department has tremendous amount of work to do in order to digitize their old patient data, especially when considering the large patient population they deal with on the day-to-day basis.

Our original anticipation was to compare rehospitalized patients with patients who were not rehospitalized in a five-year period. Nonetheless, since the hospital had amelioration in their patient data recording system not long ago, it was difficult to gather enough samples. Many patients' data could not be located. Thus, in order to continue this study, we shortened the timespan to three years. Also, when acquiring data, around 70% of the patient data we received were copies of hand-written materials, with sensitive information (such as patient names and current addresses) covered by paper slips.

Secondly, the laboratory test index was not consistent. While entering patient's' laboratory test index, we discovered that the recorded test index were not uniform. For example, for the index *Serum total cholesterol (TC),* only approximately 60% patients have those recorded. There are possibilities that the rest of the sampled patients were tested, and the results indicate that their TC level was normal, but it is impossible for us to know. Therefore, we had to leave TC out as an indicator.

Thirdly, some potentially important data were missing. When we started to review the data sent by the department, we discovered that many demographic information were omitted. To optimize the data, we suggested them to reinforce the data integrity by supplementing the demographic data. Later we found out that this demographic information was quite critical in our model construction. One of the possible significant indicators is the smoking history. According to some similar studies (such as Framingham Heart Study), smoking history is an influential element in cardiovascular disease. However, when doctors admitted their patients, this piece of information was not valued enough to ask every patient. Also, recollecting this piece of information is rather challenging, especially for the patients who are not rehospitalized, due to the tense relationship between doctors and patients in China, which was explained in Chapter 1. Thus, we would recommend the Department of Cardiology to start to ask for and record this information from now on for the sake of further researches.

Chapter 3: Research Methodology

3.1 Data Mining

Data Mining is the process of sorting through a large set of data in order to detect patterns and relationships. The application of data mining techniques to hospital industry is an area that holds promise, but it still needs further development. This chapter outlines the data mining method applied to this study and the reason to pick the selected method.

3.2 Decision Tree Models

The basic model of choice for this study is the decision tree model. Decision trees are used to split a dataset into branch-like segments. These segments generate a model to predict the target value. The target variable is usually the root node. Starting from this node, a one-dimensional tree-like interface is formed. For each node and each split, the object of analysis is usually presented, as well as the distribution of the target in that field (SAS).

Decision Tree Model can be built by asking questions about the aspects of the data. After receiving the answer for one question, a follow-up question is asked until a decision on the category or class is made. The decision tree model is built by organizing the questions asked and the answers received in a hierarchical structure (Kingsford).

Decision tree model is one kind of classification technique, which belongs to unsupervised learning. Unsupervised learning is to detect the underlying structure or pattern of a dataset in order to have a better understanding of the data. In unsupervised learning, there is no correct answer, because it purely depends on algorithms to determine the structure for display.

3.3 Analysis Package Selection

The program used to create the decision tree model is the SPSS Modeler. SPSS (Statistical Product and Service Solutions) is one kind of data mining software owned by IBM. There are several other software that are commonly used in data mining, including Excel, SAS, R, etc., but SPSS is the best choice for that the dataset need not complex analysis, hence, SAS and R are overqualified for that matter. Excel, on the other hand, is better for creating reports and scorecards, which are can only indicate past events instead of future events.

3.4 Types of Decision Tree Models

There are four decision tree models available in the SPSS Modeler, which are CART (Classification and Regression Tree), QUEST (Quick, Unbiased, Efficient Statistical Tree), CHAID (Chi-square Automatic Interaction Detector) and C5.0.

The following table is a simple comparison of the four models:

Model	**Modeling Process**	**Split**	**Target**	**Input Fields**	**Choice of Impurity Function**
CART	Recursive Partition	Binary	Numeric or Categorical	Numeric or Categorical	Gini Index
CHAID	Chi-Square Statistics	Nonbinary	Numeric or Categorical	Numeric or Categorical	No

QUEST	Quadratic Discriminant analysis	Binary	Categorical	Numeric or Categorical	Gini Index
C 5.0	Recursive Partition	Nonbinary	Categorical	Numeric or Categorical	Entropy

Model of Choice:

CART model was selected for the following reasons:

- CART automatically identifies the most significant variables and eliminates non-significant ones. This dataset has a large number of variables, and it is difficult to tell which variables are important, therefore, the algorithm behind the CART model can easily identify the most important variables by itself.
- CART is good at handling the noise in the dataset. This model tends to isolate the outliers in a separate node. Thus, if there are outliers, they will not affect the any values in the tree structure or be ignored.
- Adjustments made to independent variables might change the splitting values, but they will not change the whole tree structure (Timofeev).

Chapter 4: Data Analysis and Result

4.1 Data Adjustment and Selected Variables

Because the sample size is not sufficient, therefore, when the tree model was first built, the tree structure was not stable. Some variables were not emblematic enough to be used as indicators. Therefore, to find the indicative and viable variables, a logistic regression

model was created to detect inadmissible indicators. Also, to keep the important demographic information, categories were combined for certain variables, such as Education and BMI Category. For the *Education* variable, the categories were reduced from (Illiteracy, Elementary school, Middle School, High School, Polytechnic School, Junior College, College, Master) to 2 (Lower than High School and High School and Higher). For the BMI Category, the categories were reduced from 3 (Normal Weight, Overweight, Obese) to 2 (Normal Weight and Overweight).

As a result, the dependent variable and independent variables selected are listed below.

Dependent Variable:

- Hospitalized frequency (Flag, 0 = Hospitalized more than once/1 = Hospitalized only once)

Independent Variables:

- Gender (Flag, Male/Female)
- Education (Flag, Lower Than High School/High School and Higher)
- Marital Status (Categorical, Single/Married/Widowed)
- BMI Category (Flag, Overweight/Normal Weight)
- Family Disease History (Flag, 0 = Yes, patient has a family disease history/1 = No, patient does not have a family history)
- Type II diabetes (Flag, 0 = No/1 = Yes)
- Hypertension (Flag, 0 = No/1 = Yes)
- White blood cell count (H[1]) (Flag, 0 = No/1 = Yes)
- Blood glucose (H) (Flag, 0 = No/1 = Yes)
- Serum direct bilirubin (H) (Flag, 0 = No/1 = Yes)

[1] 'H': Higher than the normal acceptable range

- Blood troponin-I (H) (Flag, 0 = No/1 = Yes)
- Lactate dehydrogenase (H) (Flag, 0 = No/1 = Yes)
- Serum uric acid (H) (Flag, 0 = No/1 = Yes)
- Neutrophilic granulocyte segmented form count (H) (Flag, 0 = No/1 = Yes)
- D-dimer (H) (Flag, 0 = No/1 = Yes)
- Blood urea nitrogen (H) (Flag, 0 = No/1 = Yes)
- Glycated hemoglobin (H) (Flag, 0 = No/1 = Yes)
- L-γ-gamma glutamyltransferase (H) (Flag, 0 = No/1 = Yes)
- Serum creatine kinase (H) (Flag, 0 = No/1 = Yes)
- Oxygen partial pressure (H) (Flag, 0 = No/1 = Yes)
- PT (H) (Flag, 0 = No/1 = Yes)
- Aspartate Aminotransferase (H) (Flag, 0 = No/1 = Yes)
- Bile Acid (H) (Flag, 0 = No/1 = Yes)
- Low density lipoprotein cholesterol (Flag, 0 = No/1 = Yes)
- Total carbon dioxide (Flag, 0 = No/1 = Yes)
- Serum chlorine (Flag, 0 = No/1 = Yes)
- Serum creatinine (Flag, 0 = No/1 = Yes)
- Red Blood Cell Count (L^2) (Flag, 0 = No/1 = Yes)
- Total T3 (L) (Flag, 0 = No/1 = Yes)
- Serum potassium (L) (Flag, 0 = No/1 = Yes)
- Hemoglobin (L) (Flag, 0 = No/1 = Yes)
- Blood platelet count (L) (Flag, 0 = No/1 = Yes)
- High-density lipoprotein cholesterol (L) (Flag, 0 = No/1 = Yes)

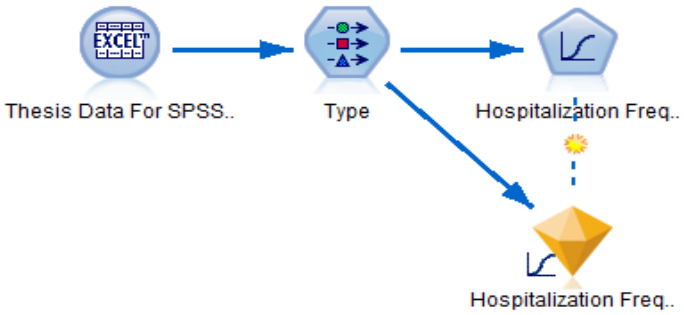
[2] 'L': Lower than the normal acceptable range

4.2 Process and the result provided by the SPSS Modeler

4.2.1 Process

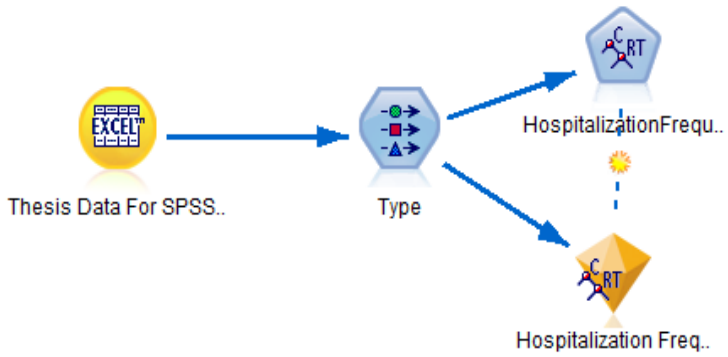
Building the CART tree model in SPSS involves the following steps.

Step 1: Determine which independent variables to use by creating a logistic regression model and eliminating the ones with missing data or do not have significant meanings, also to get the lowest number of patients that are predicted as not coming back but actually came back.



Step 2: Check the aggregate at a cutoff value of 90%.

Step 3: Use the selected dependent and independent variables to create a CART Tree Model.



4.2.2 Results

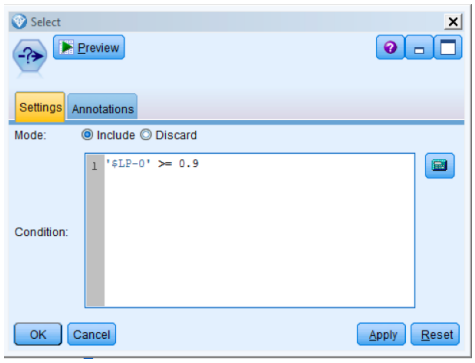
Step 1 result: As shown in the classification table below, 45 out of 49 patients who were predicted to be rehospitalized actually had been rehospitalized. Only 4 out of 49 patients were wrongly predicted. 46 out of 50 patients who were predicted as will not come back to the hospital actually have not come back to the hospital. Only 4 out of 46 patients were wrongly predicted. The total classification correctness is 91.9%. The most important number in this table is the predicted as not coming back (1) but in fact, comes back (0), because this indicates that something abnormal might have happened. It could be the patient did not follow doctor's orders or the doctor might have overlooked some of the symptoms. This figure indicates that this is a good model since the predicted 1 and observed 0 is only 4.

Classification

Observed	Predicted		
	0.0	1.0	Percent Correct
0.0	45	4	91.8%
1.0	4	46	92.0%
Overall Percentage	49.5 %	50.5%	91.9%

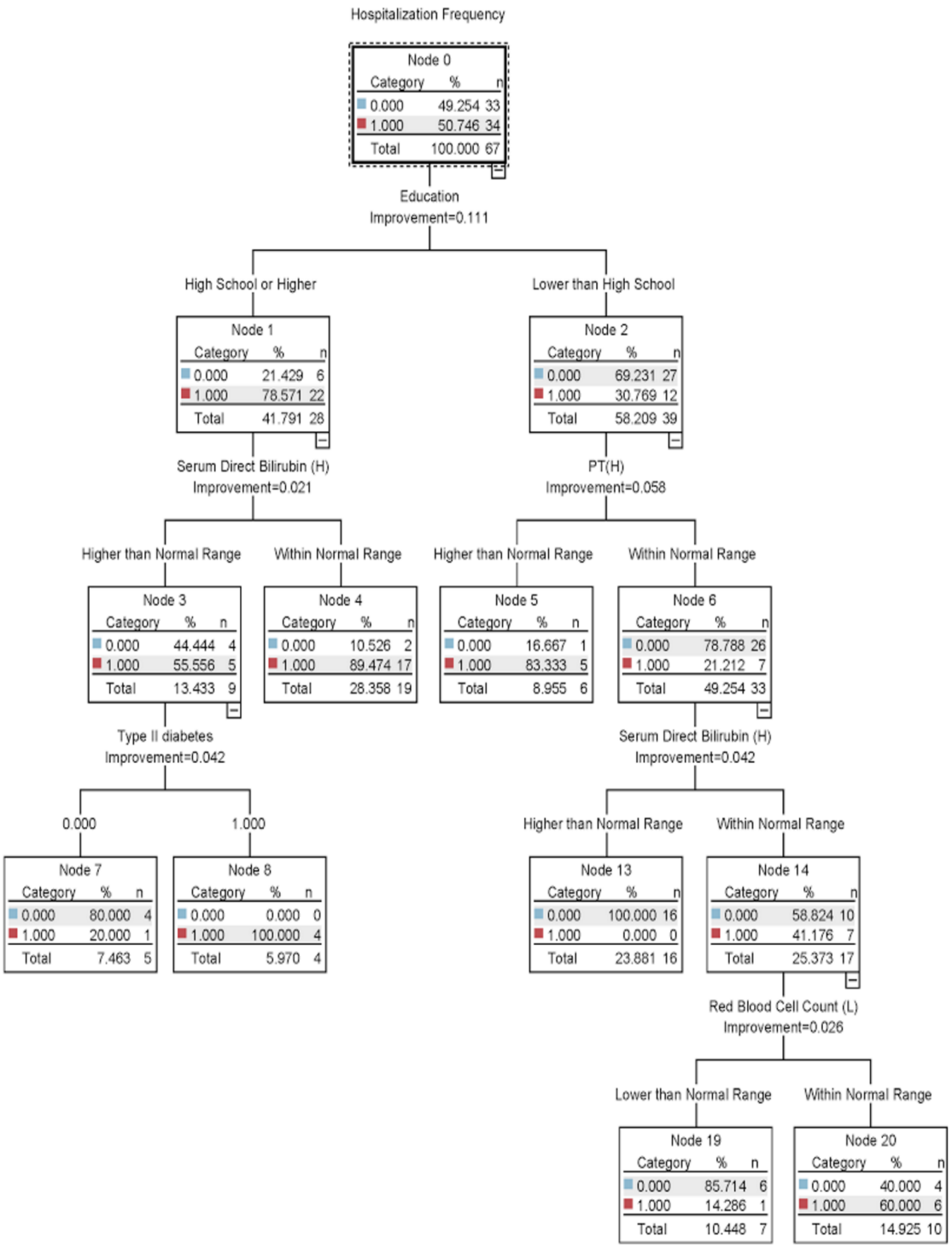
Step 2 Result: If the probability of predicted as will-not-be-rehospitalized patient cutoff at 90%, the patient who is will be rehospitalized is 1.

	Hospitalization Frequency	Record_Count
1	1.000	1
2	0.000	36



Step 3 Result: The following figure shows the CART tree model. In total, the tree model has four tiers. The first indicator is Education. If a patient's education level is *Lower than High School*, he or she has a higher possibility to be rehospitalized. In this example,

69.231% of patients with lower than high school education have been treated more than once while only 21.429 % patients with a *High School or Higher* degree have been hospitalized more than once. The next indicator is PT. PT is a blood test that measures the time it takes for the liquid portion (plasma) of one's blood to clot. It is a measurement of the coagulation function. If PT is low, it means that the blood is clotting faster than normal (MedLinePlus). Among the patients with lower than high school education, if the patient's PT is normal, they have a 78.788% chance to be rehospitalized, and if their PT is higher than the normal range, only 16.67% of chances that they will come back to the hospital. Next important indicator is the Serum direct bilirubin. Among the patients who have a PT value within the normal range, if their serum direct bilirubin is higher than normal, they have 100% chance of coming back, while if it is within the normal range, they only have a less than 60% chance of coming back. The last tier is the red blood cell count. When the red blood cell count is lower than the normal range, there is 85.714% of chance that the patient will be rehospitalized whereas if the value is within the normal range, the patient only has a 40% chance to revisit the hospital.


Hospitalization Frequency
Node 0
Category % n
0.000 49.254 33
1.000 50.746 34
Total 100.000 67
Education
Improvement=0.111
High School or Higher
Lower than High School
Node 1
Category % n
0.000 21.429 6
1.000 78.571 22
Total 41.791 28
Node 2
Category % n
0.000 69.231 27
1.000 30.769 12
Total 58.209 39
Serum Direct Bilirubin (H)
Improvement=0.021
PT(H)
Improvement=0.058
Higher than Normal Range
Within Normal Range
Higher than Normal Range
Within Normal Range
Node 3
Category % n
0.000 44.444 4
1.000 55.556 5
Total 13.433 9
Node 4
Category % n
0.000 10.526 2
1.000 89.474 17
Total 28.358 19
Node 5
Category % n
0.000 16.667 1
1.000 83.333 5
Total 8.955 6
Node 6
Category % n
0.000 78.788 26
1.000 21.212 7
Total 49.254 33
Type II diabetes
Improvement=0.042
Serum Direct Bilirubin (H)
Improvement=0.042
0.000
1.000
Higher than Normal Range
Within Normal Range
Node 7
Category % n
0.000 80.000 4
1.000 20.000 1
Total 7.463 5
Node 8
Category % n
0.000 0.000 0
1.000 100.000 4
Total 5.970 4
Node 13
Category % n
0.000 100.000 16
1.000 0.000 0
Total 23.881 16
Node 14
Category % n
0.000 58.824 10
1.000 41.176 7
Total 25.373 17
Red Blood Cell Count (L)
Improvement=0.026
Lower than Normal Range
Within Normal Range
Node 19
Category % n
0.000 85.714 6
1.000 14.286 1
Total 10.448 7
Node 20
Category % n
0.000 40.000 4
1.000 60.000 6
Total 14.925 10

Chapter 5 Conclusion

5.1 General Conclusion

Numerous researches have already shown that there are significant relations between PT, serum direct bilirubin, and the risk of coronary artery disease. The relationship between PT and coronary artery disease is that when PT value increases, the cardiac function reduces, therefore in theory, should bring the patient back to the hospital. However, our result says otherwise. We ascribe it to behavioral reasons. When a patient has a high PT value (thus less cardiac function), they tend to have arrhythmia and myocardial infarction (Wang), which would make them feel uncomfortable and pay more attention to their medical condition. Especially when the patient is relatively more well educated, they would probably take the doctor more seriously. Therefore, it is not something that can be controlled by the doctor.

There are several hypothesis of the impact of serum direct bilirubin towards coronary artery disease, but the specific mechanism is still unknown. Research shows that the relationship between bilirubin and coronary artery disease is a U relationship, the concentration is too low or too high both hurt the body (Lian).

One possible way to interpret this result is that when patient is not well educated, chances are that they would not strictly follow doctor's orders regarding to their eating, drinking, and smoking habit, and other lifestyle related problem. These bad habit could cause the shorten of Prothrombin time and increase the platelet adhesion, which could worsen the vassal clogging. Based on this result, the Department of Cardiology could increase the frequency of follow-up visit, in order to have a closer monitor of the lifestyle of the patient. Patients who are rehospitalized unexpectedly might cause them to direct their

anger to the doctors and blame them for that. By recognizing the risk factors that could lead them back to the hospital can prepare both the doctor and the patient in advance, and also allows the doctors to customize the follow-up visit.

5.2 Business Contribution

In recent years, more and more hospitals and other related industry start to utilize data mining to get to significant medical findings. The implementation of data mining tools does not necessarily require user to have an in-depth medical information. By using the right variables, the hidden relationship and connections will reveal themselves. In addition, since this study is regionalized, it allows professionals to compare the differences between different regions and even different countries. For example, in the Framingham study, obesity is a huge indicator of cardiovascular disease, but since in China, not many people are obese, this risk factor lost its significance.

5.3 Directions for Further Research

This study sampled 100 patients within three years, which is not considerably conclusive. If the department can better their patient recording system and have a bigger sample pool, the result could be more significant. For further research, one could collect a larger dataset that includes more participants in both groups to reduce data bias and generate more valuable results.

Also, one possible important variable is the smoke history. According to the Framingham Heart Study, cigarette smoking increases the risk of heart disease. Based on the dataset received, however, smoking history could not be specified. For further research, we suggest to include smoking history as a variable.

Bibliography:

Chen, Y. (2015, January 27). Prothrombin time (PT). Retrieved June 13, 2017, from https://medlineplus.gov/ency/article/003652.htm

China Statistics Press. (2015). China Statistical Yearbook. Retrieved June 13, 2017, from http://www.stats.gov.cn/tjsj/ndsj/2015/indexeh.htm

Deloitte. (2015, May 12). China Healthcare Provider Market. Retrieved June 13, 2017, from https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/life-sciences-health-care/deloitte-cn-lshc-china-healthcare-provider-market-zh-150512.pdf

Excellence in Medical Affair Recent trends in China Medical Affairs and Outcomes from China Medical Affairs Survey 2016 (Rep.). (2016). Deloitte. Retrieved May 22, 2017.

Fang, H., & Wang, D. (2007, July 18). 血清胆红素与冠心病的关系 (The Relationship between Serum Bilirubin and Coronary Artery Disease). Retrieved June 13, 2017, from http://journal.9med.net/html/qikan/lcyx/lcjyzz/200512346/lz/20080831112635953_268617.html

Health Expenditures. (2017, January 20). Retrieved June 13, 2017, from https://www.cdc.gov/nchs/fastats/health-expenditures.htm

Kingsford, C., & Salzberg, S. L. (2008, September). What are decision trees? Retrieved June 13, 2017, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701298/

Li, X., Huang, J., & Zhang, H. (2008, September 20). An analysis of hospital preparedness capacity for public health emergency in four regions of China: Beijing, Shandong, Guangxi, and Hainan. Retrieved June 13, 2017, from https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-8-319#IDAXJ1JFB

Lian, Y., & Zhang, P. (2016). 胆红素与冠心病关系的研究进展. Retrieved June 13, 2017, from http://e-zhxxgbzz.yiigle.com/CN116031201601/862654.jhtml

Out-of-pocket health expenditure (% of private expenditure on health). (n.d.). Retrieved June 13, 2017, from http://data.worldbank.org/indicator/SH.XPD.OOPC.ZS?view=map

SAS. (n.d.). Decision Trees— What Are They? . Retrieved June 13, 2017, from http://support.sas.com/publishing/pubcat/chaps/57587.pdf

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications* (Unpublished master's thesis). Humboldt University.

Wang, M. (n.d.). *冠心病患者凝血机制的改变及其临床意义* (7th ed., Vol. 33). Modern Preventive Medicine. Retrieved June 13, 2017.

Appendices:

Appendix 1: Data Collected

I. Demographic:

   A. Gender

   B. Age

   C. Height

   D. Weight

   E. Level of Education

   F. Marital status

   G. Heart disease family history

   H. Frequency of hospitalization

   I. Home zip code

II. Diagnosis:

   A. Coronary Atherosclerosis

   B. Type II diabetes

   C. Hypertension

   D. Atrial Fibrillation

   E. Anemia

   F. Cardiac Insufficiency

   G. Renal Insufficiency

III. Previous Diagnosis (If any):

A. Coronary Atherosclerosis

B. Type II diabetes

C. Hypertension

D. Atrial Fibrillation

E. Anemia

F. Cardiac Insufficiency

G. Renal Insufficiency

IV. Abnormal Laboratory Test Result (Higher than normal):

A. White blood cell count (WBC)

B. Blood glucose

C. Serum direct bilirubin (DBIL)

D. Blood troponin-I (TnI)

E. Lactate dehydrogenase (LDH)

F. Serum uric acid (SUA)

G. Neutrophilic granulocyte segmented form count

H. D-dimer

I. Blood urea nitrogen （BUN)

J. Glycated hemoglobin (GHB)

K. L-γ-gamma glutamyltransferase (GGT)

L. Serum creatine kinase (CK)

M. Oxygen partial pressure

N. PT

O. Aspartate Aminotransferase (AST)

P. Bile Acid

Q. 2h plasma glucose (2hPG)

R. Apolipoprotein-E (APOE)

S. Low density lipoprotein cholesterol (LDL-C)

T. Total carbon dioxide (TCO2)

U. Serum chlorine

V. Serum creatinine (Scr)

W. BNP

V. Abnormal Laboratory Test Result (Lower than normal):

A. High-density lipoprotein cholesterol （HDL-C)

B. Apo lipoprotein A1 （ApoA1）

C. Red blood cell count （RBC）

D. Alanine aminotransferase (ALT)

E. Total T3

F. Serum potassium

G. Hemoglobin (Hb)

H. Hematocrit (HCT)

I. Blood platelet count (BPC)

J. Serum total cholesterol (TC)

K. Determination of Free T3 (FT3)

L. Mean corpuscular hemoglobin concentration (MCHC)

M. Fibrinogen (FBG)

N. Apo lipoprotein B (ApoB)

Appendix 2: A Comparison between China and America on Health Expenditures in 2014:

Statistics	China	United States
Total Health Expenditure	3.5 trillion dollars	3.0 trillion dollars
Total Health Expenditure (per Capita)	419.73 dollars	9402.54 dollars
Total Health Expenditure (% of GDP)	5.55%	17.14%
Out-of-pocket health expenditure (% of total expenditure on health)	31.99%	11.05%
Out-of-pocket health expenditure (% of private expenditure on health)	72.35%	21.37%