CARLSON
SCHOOL OF MANAGEMENT
UNIVERSITY OF MINNESOTA

PREDICTIVE ANALYTICS

MSBA 6420

PNEUMONIA & MRSA PATIENT CLASSIFICATION AND RISK PREDICTION

Lavanya Basava Raju, Ananya Mishra, Xiurong Lin, Ryan Borowicz

December 5, 2016

Contents

1.0 Business Understanding

Hospital-acquired infections (HAI) are associated with increased attributable morbidity, mortality, prolonged hospitalization, and economic costs. Although there are many decade academic study and preventive policy deployment, HAI is still a major threaten to patient safety in the United States (US). As one of the five major costly infections, ventilator-associated pneumonia cost $40,144 per incident annually with Monte Carlo simulation, which contributed to 31% of overall costs of these five major infections. The current incidence of ventilator-associated pneumonia (VAP) ranges from two to 16 episodes for 1000 ventilator-days, with an attributable mortality of 3-17%.

Considering the health harm and costs of HAI, researchers have been focusing on the evaluation of cost and identify patients at high risk for HAI during stays at the hospital. Since it is not easy to access the hospital's classified data, many studies were only able to work on the macro cost estimation using the public database, such as National Healthcare Safety Network(NHSN), to help providers to justify investing in prevention. Zimlichman et al. (2013) modeled the variation of main infection types within a large population of patients and used simulation to extrapolate totals for the US health care system. Also, Lee et al. (2012) used a quasi-experimental design with interrupted time series with comparison series to examine changes in trends of two health care–associated infections that were targeted by the CMS policy.

Simultaneously, some studies focus on HAI high risk prediction at the micro level. Goodman et al. (2016) had tried to develop a user-friendly decision tree to predict which organisms are extended-spectrum beta-lactamase(ESBL) producing bacteria to guide appropriate antibiotic therapy. Its results suggested that a clinical decision tree can be used to estimate a bacteremia patient`s likelihood of infection with ESBL-producing bacteria. The decision tree`s positive and negative predictive values were 90.8% and 91.9%, respectively. The medical scoring systems are widely used to predict risk of morbidity or mortality and to evaluate outcomes in patients with certain illness. The value of such scoring systems is to provide a simple predictive tool with certain relevant factors for clinical use. Since up until 2011, there exists no such a scoring system for HAI, Chang et al. (2011) had used 7 different HAI devices as variables and 1000 cases to develop a scoring system to predict HAI that was derived from Logistic Regression (LR) and validated by Artificial Neural Networks(ANN)simultaneously.

With these tangible and intangible costs for patients and hospitals in mind, we set out to devise an approach to solving the problem of HAI's through building a predictive model that balanced the costs of preventative medication against the costs of post-incident treatment. The model should be able to predict the likelihood that a patient would be susceptible to obtaining an infection, and provide a clear method to ensure that front-line staff members have access to this information to implement in their daily routines. We will discuss the details of the model as well as the deployment approach in the sections below.

2.0 Data Understanding & Preparation

2.1. Data Understanding

The Informs Data Mining Contest provided the following datasets for use in the competition which we utilized as the basis for our analysis. They provided separate datasets for training and test, where the test datasets were stripped of rows containing diagnosis codes related to pneumonia, which is one of our target labels. Because of this we ended up splitting the training set into training and test sets in building our models.

The four datasets are summarized in the table below:

	Patient Conditions	Patient Demographics	Hospital Events	Medications
Records	213,018	68,619	38,956	256,672
Variables	4	18	56	75

TABLE 1 - DATASET SUMMARY

Patient Conditions Dataset:

This dataset had 213,018 records and 4 variables, and provides a listing of all patient conditions by patient id and year. There are 43,151 unique patients with 619 different conditions.

Patient Demographics Dataset:

This dataset had 68,619 records and 18 variables, and represents patient demographic information such as age, sex, gender, race, education level, income, and marriage status. It includes 52,280 unique patients across a variety of demographic traits.

Hospital Events Dataset:

The Hospital Events dataset included 38,956 patient visits by 11,846 patients. It includes information on patient hospital visits such as the primary and secondary conditions diagnosed on the visit, whether it was related to an emergency room visit or procedure, as well as the doctor and facility costs and payment information related to the visit. These patients had 554 different conditions.

Medications Dataset:

The Medications dataset included 256,672 patient medication records. Of these, there were 33,863 patients taking 5,631 different medications. The dataset includes information on the medication name, type of pharmacy obtained in, the associated diagnosis code, the pharmacy class, and the cost of the medication.

2.1.1 Feature Reduction

Prior to beginning the exploratory analysis, each of the datasets was individually reviewed to remove extraneous variables that did not make sense in the context of the business problem. There were also a variety of variables included in the original datasets that the contest mentioned to remove. The following variables were removed from the datasets prior to analysis:

Hospital Events:

- Removed due to contest instruction – EVENTRN, FF1PTYPE, VARPSU, VARSTR, PERWT03F
- Removed due to lack of definition or duplication - DUID (duplicative of DUPERSID), FFEEIDX (lack of definition), IMPFLAG (duplicative of NUMNIGHX)
- Removed due to level of detail - Cost metrics – In the original dataset there were 16 separate variables related to the different types of payment methods used to pay for the doctor and facility charges, such as whether it was paid through insurance, private pay, Medicaid, etc. We chose to use the summary metrics for total payments and charges related to the visits, rather than each of the individual amounts, as we did not believe these to have an impact on the likelihood of a patient obtaining MRSA or pneumonia.
    - Fcmedicaid, Fcmedicare, Fcotherfed, Fcotherinsur, Fcotherpub, Fcotherpv, Fcprivateins, Fcselfpaid, Fcstate, Fctricare, Fcveterans, Fcworkcomp, Mdmedicaid, Mdmedicare, Mdotherfed, Mdotherinsur, Mdotherpub, Mdotherpv, Mdprivateins, Mdselfpaid, Mdstate, Mdtricare, Mdveterans, Mdworkcomp

Medications:

- Removed due to contest instruction - RXCCC1X, RXCCC2X, RXCCC3X
- Removed due to level of detail – RXNAME/RXNDC (aggregated data at rxclass level), cost metrics (rxotherfed, rxotherinsur, rxotherpub, rxotherpv, rxprivateins, rxselfpaid, rxstate, rxtricare, rxveterans, rxworkcomp), rx class sub-levels (TC1S1, TC1S1_1, TC1S1_2, TC1S2, TC1S2_1, TC1S3, TC1S3_1, TC2S1, TC2S1_1, TC2S1_2, TC2S2, TC3S1, TC3S1_1)
- Removed due to lack of definition – CLMOMFLG, INPCFLG
- Removed due to duplication – DUID, PID, values captured by PHARTP1 (PHARTP2 – PHART9)
- Removed due to incorrect data – GBO, RXFORM, RXFRMUNT, RXHHNAME, RXQUANTY, RXRECIDX, RXSTRENG, RXSTRUNT
- Removed due to lack of relevance - LINKIDX, PCIMPFLG, PREGCAT, PURCHRD, RXFLG, VARPSU, VARSTR

2.1.2 Missing Values Treatment

After removing the extraneous variables, we then looked at handling the missing values. Within the datasets, the contest provided the following nomenclature for these types of values, specified as:

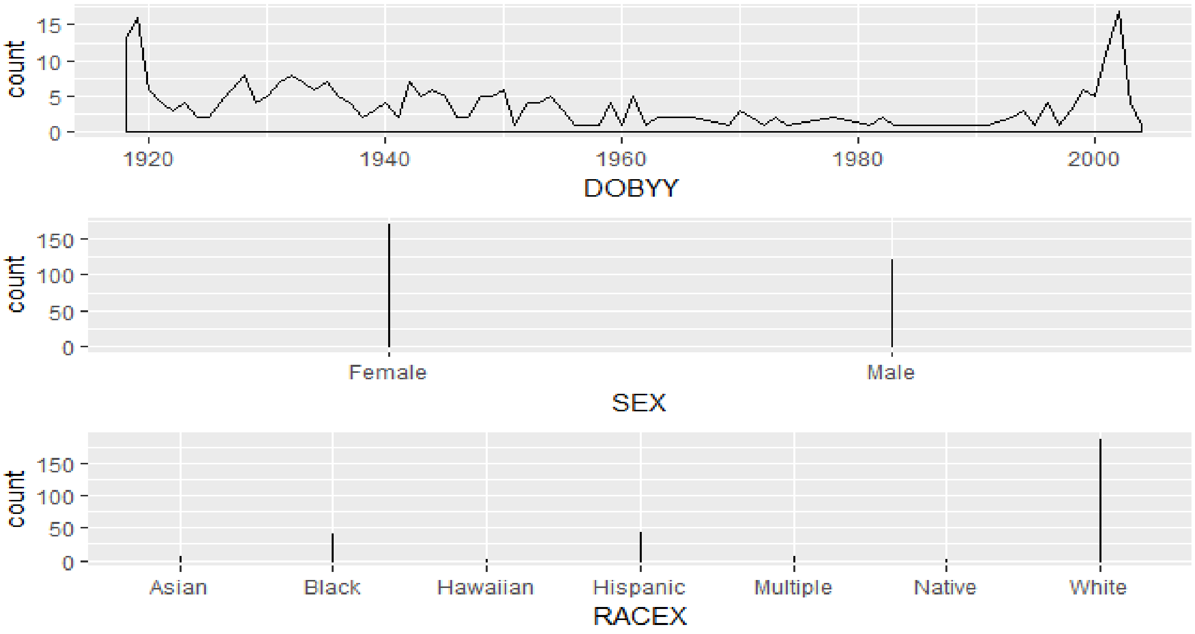
- -9 - Not ascertained – classified as “ND” in our analysis
- -8 - Unknown – classified as “NA” in our analysis
- -7 - Refused to Answer - classified as “ND” in our analysis
- -1 - Inapplicable - classified as “NA” in our analysis

These values are used in various contexts so it would not be correct to make the blanket assumption to list them all as “NA” values. As such, we approached it by converting the -1 and -8 values to “NA” values, and classifying the -9 and -7 values into their own category of “ND” for not determined. It should be noted that the interpretation of these values was left to our assumptions, as they do not always make sense in the data. For example, there are ICD diagnosis codes with values of -1 and -9. By classifying the -1 values as “NA” we are essentially excluding these results from our analysis, while the -9 values will be considered as a separate category. There were also numerous instances that could be related to data-entry issues or due to combining various datasets. An example of this is all rows where there is an emergency room event id, but the associated emergency room flag of 1 or 2 for yes or no is populated with one of the above-mentioned values (-9, -8, -7, -1). In addition to these values, the data also included period values “.”, which we imported as NA’s.

2.1.3 Data Visualization

Over the course of the analysis, several visualizations were performed in an iterative fashion to view the data for overall patterns as well as the before and after results of various transformations. The visuals below describe our target variable to give a general sense of the type of patient who has acquired one of these conditions in the past:

<u>Demographic Characteristics:</u>



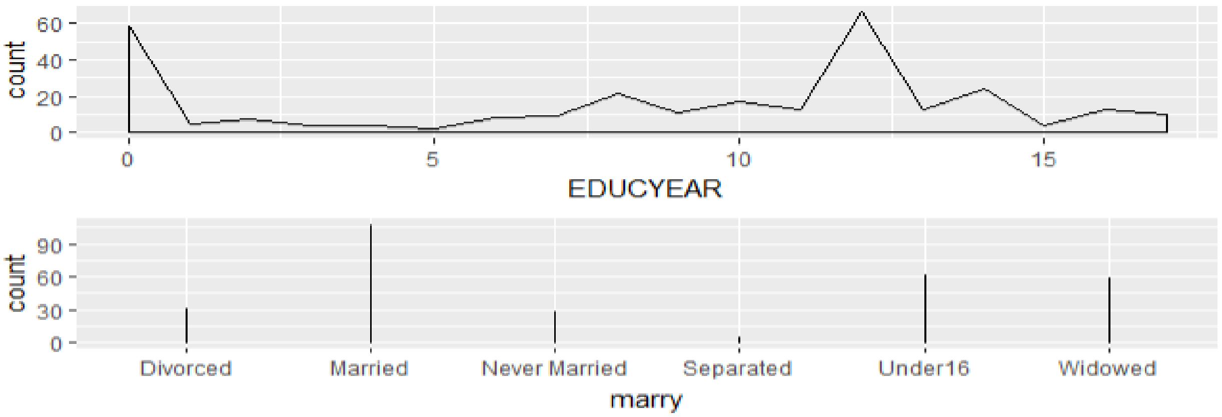


Figure 1 - Target patient demographics

Hospital & Medications Information:

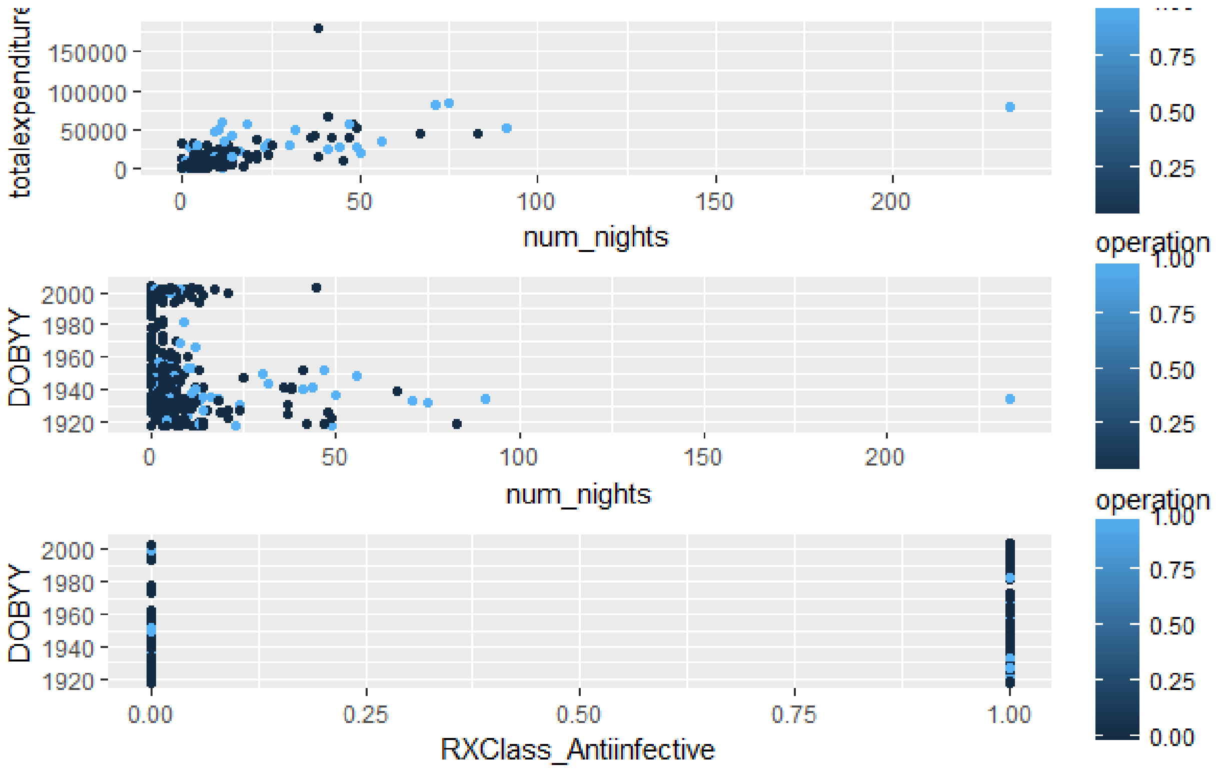


Figure 2 - bivariate analysis

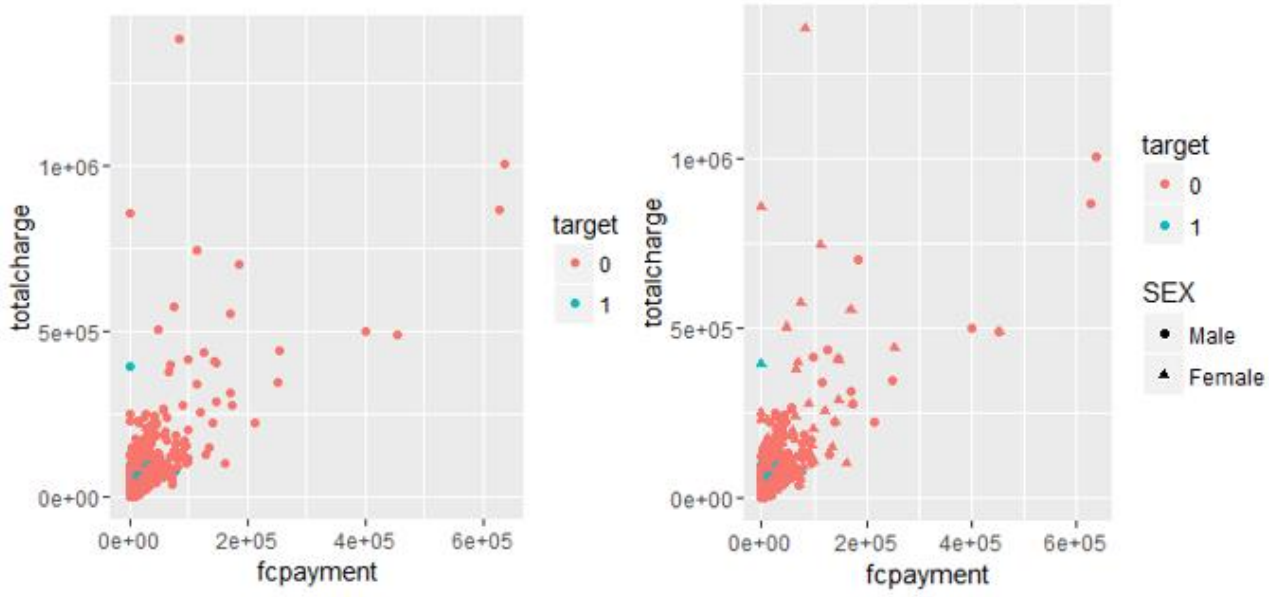


FIGURE 3 - TOTAL CHARGE - TARGET VS. NON-TARGET

2.2. Data Preparation

2.2.1 Individual Dataset Transformations

Patient Conditions Dataset:

After removing the DUID, the Conditions dataset consists of the following variables:

Variable	Description	Values
DUPERSID	Patient Identifier	Numerous
ICD9CODX	Patient Condition Identifier (ICD9 Code)	Numerous (grouped by ICD class)
Year	Year diagnosed	2003, 2004

The following transformations were performed:

- Converting the year variable to a factor
- Removing the records where the IC9CODX is "NA"
- Removing duplicate records

Patient Demographics Dataset:

This dataset consisted of 18 variables describing different demographic characteristics of patients. There were several columns classifying the race of a patient (RACEX - race, RACEAX – race - Asian, RACEBX – race - black, RACEWX – race - white, RACETHNX - race - other, HISPANX - Hispanic, HISPCAT – Hispanic Type) that we combined into one variable (RACE). The Poverty variable looked to be a grouping of patients by poverty status, but was excluded from our analysis because it didn’t correlate with the Income variable, as people with high incomes were classified as “Poor” or “Near Poor” poverty levels.

Variable	Description	Values
DUPERSID	Patient Identifier	Various
DOB	Year of Birth	Various
SEX	Gender	1 = Male, 2 = Female
RACE	Ethnicity	1 = White, 2 = Black, 3 = Native, 4 = Asian, 5 = Hawaiian, 6 = Multiple
MARRY	Marriage Status	1 = Married, 2 = Widowed, 3 = Divorced
EDUCYEAR	Education Year	0 = Never attended school, 1-12 = grades 1 – 12, 13+ = years college
Income	Annual Income	Various

Hospital Events Dataset:

The following transformations were performed on the hospital dataset:

- Removed EMERROOM variable due to duplication of Emergency ID field
- Converted the charge and payment amounts to $0 when the value was showing as -1
- The ICD codes were classified based on the International Statistical Classification of Diseases and Related Health Problems standard. Refer to appendix for the full classification values.

Due to the inconsistencies in the data several assumptions were made, including:

- Assuming an Emergency Room identifier of -1 was equivalent to an NA value
- For cases where there were event id’s but the associated emergency room flag was set to “No”, we counted these as false alarms not requiring an emergency room stay, and changed the number of imputed nights to 0 for these records.
- Used the actual number of nights’ field when the value was positive, and the imputed nights field when the actual value was negative (i.e. – 1)
- For the variables specific to emergency room visits we considered the meaning of the “NA” values (-1, -7, -8, -9) to all be the same as blank values.
- By transforming the ICD diagnosis code variable to a binomial, we are saying that the primary diagnosis is the same as the secondary diagnosis in predicting our target variable

Variable	Description	Values
DUPERSID	Patient Identifier	

Year	Year of data collection	2003, 2004
Icd1x	Primary Diagnosis	Grouped into the following categories: 001-139-infectious diseases, 140-239-neoplasms, 240-279-endocrine_nutritional_metabolic, 280-289-blood diseases, 290-319-mental disorders, 320-359-nervous system, 360-389-sense organs, 390-459-circulatory system, 460-519-respiratory system, 520-579-digestive system, 580-629-genitourinary system, 630-679-pregnancy_complications, 680-709-skin diseases, 710-739-musculoskeletal system, 740-759-congenital anomalies, 760-779-perinatal period, 780-799-ill-defined conditions, 800-999-injury and poisoning, E-External Causes, V-health service contact
Icd2x-4x	Secondary Diagnosis	Same as Icd1x
ANYOPER	Any operations or surgeries performed	1 = Yes, 2 = No
Pro1x, Pro2x	Primary/Secondary Procedure	Various
DSCHPMED	Medicines prescribed at discharge	1 = Yes, 2 = No
ERHEVIDX	Event ID for corresponding emergency room visit	Unique Hospital Visit Identifier
RSNINHOS	Reason entered hospital	1 = Surgery, 2 = Treatment, 3 = Diagnostic, 4 = Baby birth, 5 = Baby pre-birth, 91 = Other
SPECCOND	Hospital stay related to condition	1 = Yes, 2 = No
NUMNIGHT	Number of nights stayed at provider	0 - 40
NUMNIGHX	Number of nights in hospital-edited/imputed	0 - 40
Fcpayment	Facility payment	
Fctotal	Total facility cost	
Mdpayment	Physician reimbursement	
Mdtotal	Total physician cost	
Totalcharge	Total charges	
Totalexpenditure	Total reimbursement	

Medications Dataset:

The decision was made to classify individual prescriptions by their medication class rather than individual values due to the large number of individual RX values and the need to include each of these as predictor variables in the model.

Due to the inconsistencies in the data several assumptions were made, including:

- Assuming the diabetes flag variable represented a medicine prescribed for diabetes
- The RXICD code values were not found to have a link back to the hospital dataset so there was no way to link a medication with its associated hospital visit if related

Variable	Description	Values
DUPERSID	Patient Identifier	
Year	Year of data collection	2003, 2004
RXNAME	Medication Name (imputed)	Removed individual values – used medication class to classify these values
TC1	Medication Class	1 = Anti-infective, 20 = Antineoplastics, 28 = Biologicals, 40 = Cardiovascular agents, 57 = Central nervous system agents, 81 = Coagulation modifiers, 87 = Gastrointestinal agents, 97 = Hormones, 105 = miscellaneous agents, 115 = Nutritional products, 122 = Respiratory agents, 133 = Topical agents, 218 = Alternative medicines, 242 = Psychotherapeutic agents, 254 = Immonologic agents, 331 = Radiologic agents, 358 = Metabolic agents
PHARTP1	Pharmacy Type	1 = mail order, 2 = in another store, 3 = in HMO/Clinic/Hospital, 4 = drug store, 5 = online
DIABFLG	RX Insulin or Diabetic Equipment/Supply	1 = Yes, 2 = No
Rxexpenditure	Medication expenditure	
rxpayment	Amount paid total	

Merging the Datasets:

After analyzing the individual datasets, we wanted to identify common identifiers that could be used to join the different datasets. What we found was that there was not a common unifying strand between the four datasets, but rather they seemed to be datasets obtained from separate sources that didn't share a common language. The patient identifier was common to all of the datasets, but the patients were different in each of the datasets. As a result of this, we needed to identify an approach to merge the datasets. We decided to use the hospital dataset as the base dataset since it provided the most relevant information to the problem, and is where we determined our target value from based on the ICD diagnosis code. From there we performed the following steps:

- **Joined the hospital dataset to the demographics dataset** - We joined the demographics dataset to this based on the combination of patient id and year.

- **Moved the conditions data to patient id level and created a temporary table to add conditions as predictor variables at the ICD class level** - The conditions table included multiple diagnosis codes for most patients, and would create duplicate records in our dataset if joined directly. To address this problem, we created a temporary table of unique patients and transformed the conditions into predictor variables that could be used in building the model. This added a condition count variable to the dataset summing the total number of conditions for each patient, as well as individual (0,1) dummy variables for each of the associated ICD diagnosis codes associated for that patient according to the ICD class.

- **Moved the medications data to patient id level and created a temporary table to add medications as predictor variables at the medication class level -** The Medications dataset created a similar problem in that it had multiple records per patient for each of their associated medications. To incorporate at the patient id level, we transformed these examples into predictor variables, with one variable per medication aggregated at the pharmacy class level.

Target Variable Creation:

After merging the datasets, we created our target variable based on the following business logic:

- ICD diagnosis codes between 480 & 486 relate to pneumonia cases
- ICD diagnosis code of V09 relates to MRSA cases

We made sure to exclude these values from the conditions dataset prior to merging to ensure we weren’t incorporating the target variable as part of one our model predictors.

3.0 Modeling

3.1. Overview

The purpose of the analysis is to classify infected patients based on their demographics, hospital information, conditions, and medication history. Supervised classification algorithms are used in the analysis to classify these patients. The following techniques were tried, tested, and compared to choose the best classification model:

1. Decision Tree
2. kNN – k-Nearest Neighbors
3. Logistic Regression
4. Naïve Bayes
5. Neural Networks
6. SVM – Support Vector Machine
7. Ensemble Modeling

In each of the techniques, the goal was to build a cost sensitive model that would minimize the average misclassification costs per patient, per the cost matrix in section 3.2.4 below.

3.2. Pre-Processing Steps

The data is at patient-year level with patient demographics, which include DOB, number of years of education, race, gender, marital status and income; hospital information, which includes information on the payment, admission to the emergency room, number of nights stayed, reason for being admitted in the emergency room and the procedures underwent; conditions history, which includes all the conditions the patient has or had in the past; and medications history, which include the medications and pharmacy types where the patient bought the medicine.

3.2.1 Balancing Dataset

This data had an infected to non-infected patient records ratio as 291:12944. As the data is unbalanced, we used the "Sample" functionality to get a sample of the dataset that allows for a better classification. After using this functionality with 0.1 for non-infected patient data and 1.0 for infected patient data, the ratio of infected to non-infected patient records changes to 291:674.

3.2.2 Data Transformation

After balancing, we used "Numerical to Binomial" functionality on all the attributes that were converted to dummy variables to hold ones or zeros such as medication types, procedures, and pharmacy types; and used "Normalize" functionality to normalize the numerical attributes such as DOB, income, and payment details.

3.2.3 Dimension Reduction

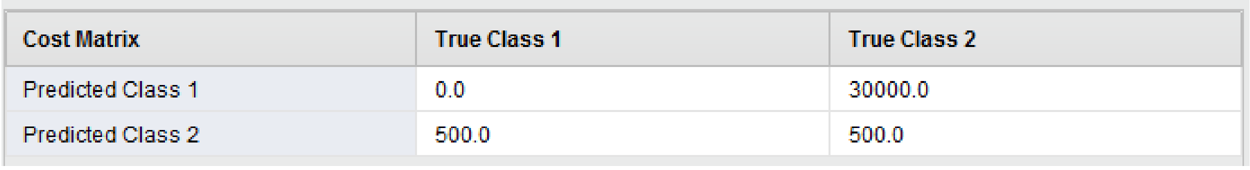
"PCA/Principal Component Analysis" functionality was used to reduce dimensions from procedure and condition data. There were 74 procedures and 23 condition types in the dataset. A patient could have undergone one or multiple procedures and had one or multiple conditions. PCA functionality reduced these 97 variables to give 11 principal components that explain most of the variance in the data.

“Join” functionality was used to combine the results of PCA with the rest of data at the patient-year level.

3.2.4 Validation and Misclassification Cost

“Cross Validation” functionality is used to validate the model and “Find Threshold”, “Apply Threshold” and “Performance Cost” are used to provide a cost matrix and tune the models.

Below is the cost matrix used for all the models, where class 1 is Non-Infected and class 2 is Infected.



Cost Matrix	True Class 1	True Class 2
Predicted Class 1	0.0	30000.0
Predicted Class 2	500.0	500.0

Another cost matrix with 500 replaced with 1 and 30000 replaced with 10000 was also used.

3.2.5 Feature Selection

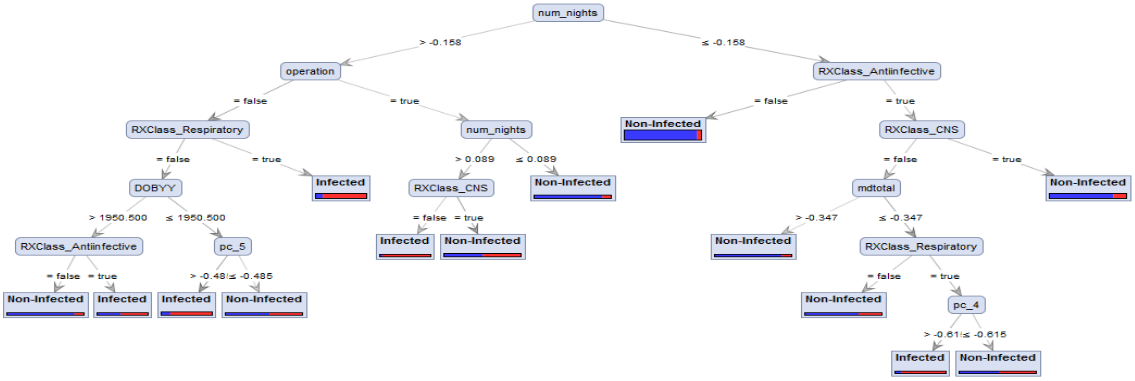
“Forward Selection” and “Optimize Parameters” functionalities where used to determine attributes and parameters that give the best models. The snapshots of the overall process are provided in the appendix.

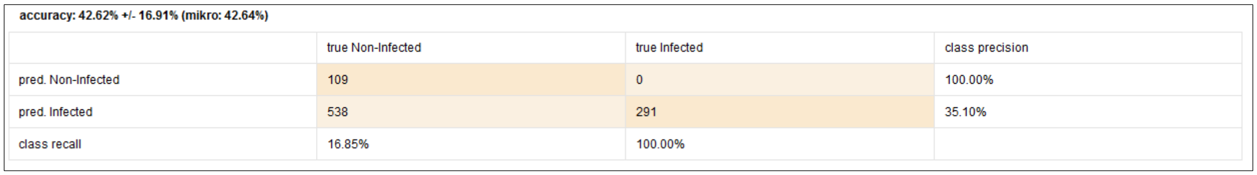
3.3. Models

3.3.1 Decision Tree

Decision trees have a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node. A tree can be learned by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same values as the target variable, or when splitting no longer adds value to the predictions.

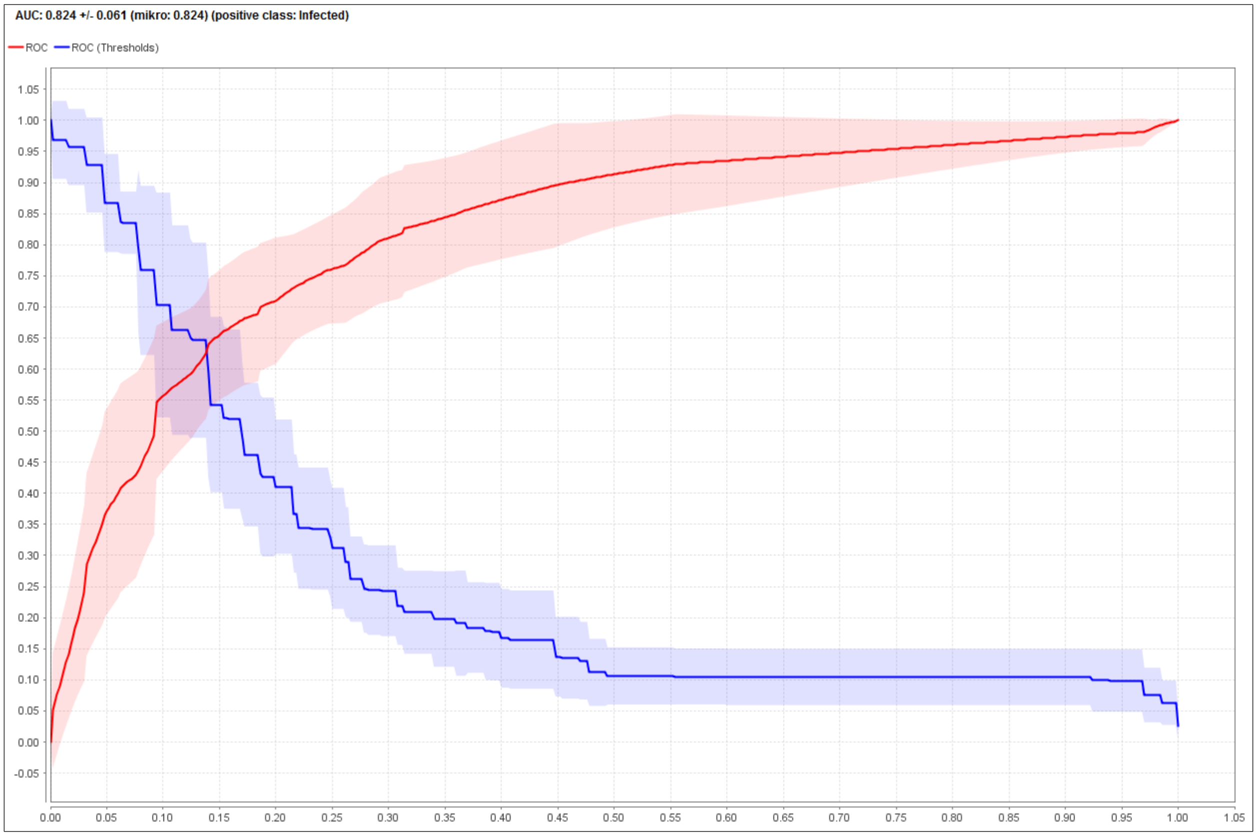
Below is the tree and the outputs from the model:





accuracy: 42.62% +/- 16.91% (mikro: 42.64%)

	true Non-Infected	true Infected	class precision
pred. Non-Infected	109	0	100.00%
pred. Infected	538	291	35.10%
class recall	16.85%	100.00%	



AUC: 0.824 +/- 0.061 (mikro: 0.824) (positive class: Infected)

Parameters

Decision Tree – Applied post pruning and pre-pruning with number of pre-pruning alternatives as 3, criterion as “information gain”, maximal depth of tree as 20, confidence as 0.05, minimal gain as 0.04, minimal leaf size as 15, and minimal size for split as 15

From the decision tree results we can see that, of the 291 infected patients,

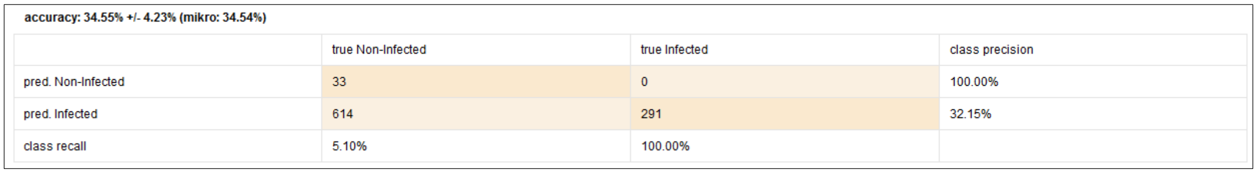
1. 211 stayed at the hospital for a longer duration, of which, 39 underwent an operation and 95 had conditions involved with the respiratory system, 59 were older patients and 16 had taken medication for infections in the past
2. Among the remaining 82 patients, 22 had taken medications for CNS in the past, 44 had taken medication for infections in the past and the rest 16 had not taken any sort of medication related to infections

From the model, the value of recall is 100% which shows that all patients likely to be infected are rightly classified. However, the misclassification error outcome of this model indicates that about 83% of patients who are not infected would have to take preventive measures.

The snapshot of the tree structure with rules is placed in the appendix.

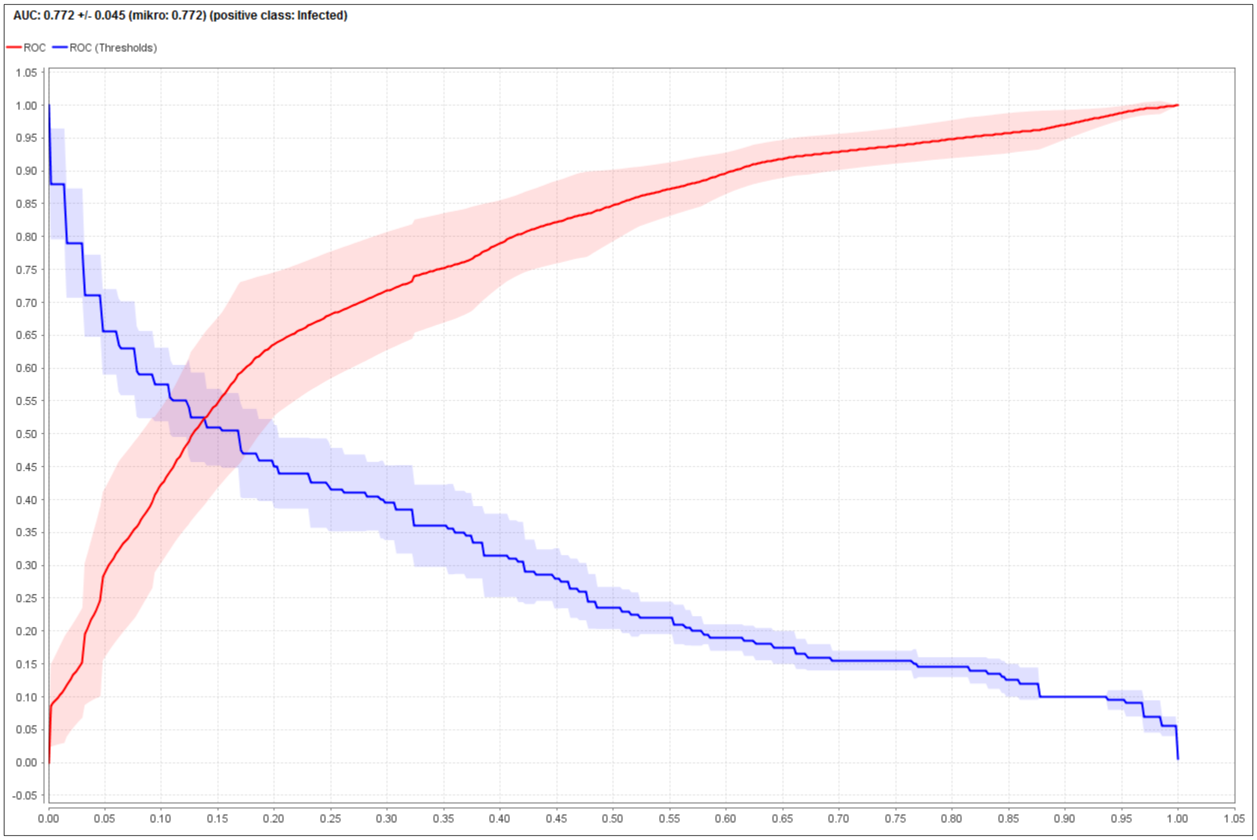
3.3.2 kNN

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure, such as distance functions. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors, measured by a distance function. Below are the outputs from the model:



accuracy: 34.55% +/- 4.23% (mikro: 34.54%)

	true Non-Infected	true Infected	class precision
pred. Non-Infected	33	0	100.00%
pred. Infected	614	291	32.15%
class recall	5.10%	100.00%	



Parameters –

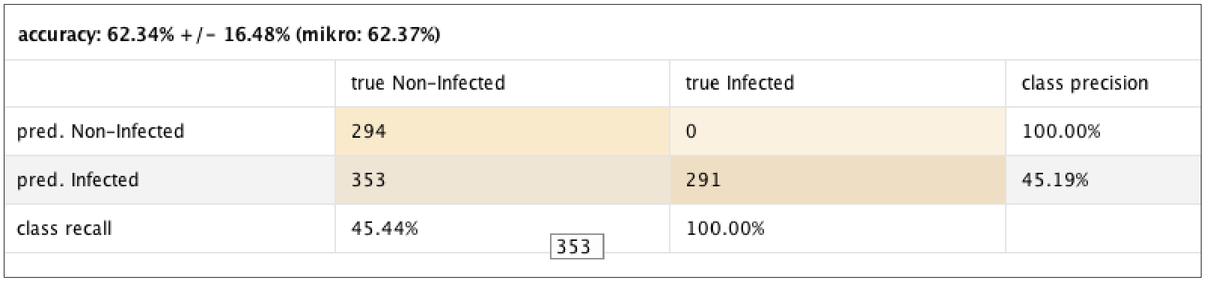
kNN – Used 20 for value of k, “Mixed measures” for measure type, and “Mixed Euclidean distance” for the measure

From the model, the value of recall is 100% which shows that all patients likely to be infected while at hospital are rightly classified. However, the misclassification error outcome of this model indicates that about 95% of patients who are not infected would have to take preventive measures.

3.3.3 Logistic Regression

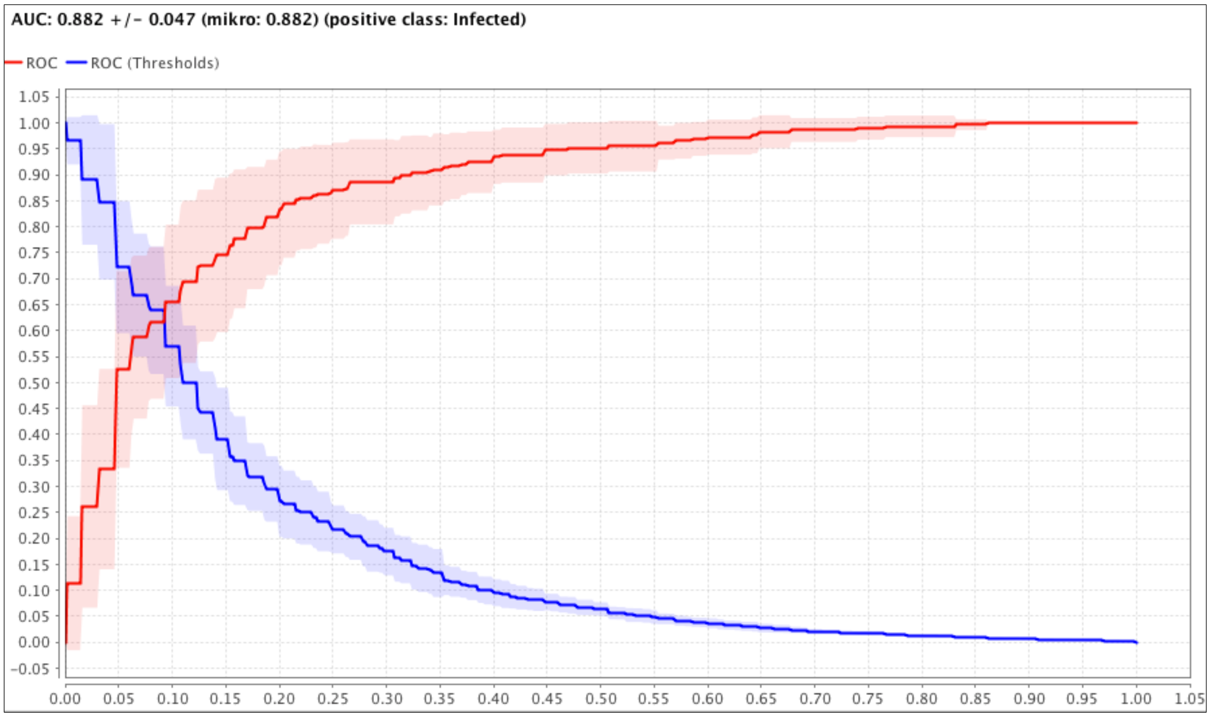
Logistic regression is one of the most commonly-used statistical techniques. It is used with data where there is a binary success-failure outcome or response variable, or where the outcome takes the form of

a binomial proportion. Like linear regression, where the algorithm estimates the relationship between predictor variables and an outcome variable; logistic regression estimates the probability that the outcome variable assumes a certain value, rather than estimating the value itself.
Below are the outputs from the model:



accuracy: 62.34% +/- 16.48% (mikro: 62.37%)

	true Non-Infected	true Infected	class precision
pred. Non-Infected	294	0	100.00%
pred. Infected	353	291	45.19%
class recall	45.44%	100.00%	



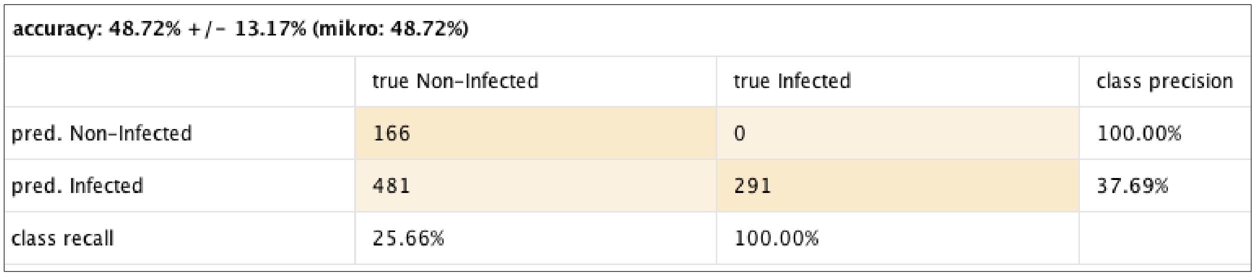
AUC: 0.882 +/- 0.047 (mikro: 0.882) (positive class: Infected)

From the model, the value of recall is 100% which shows that all patients likely to be infected while at hospital are rightly classified. However, the misclassification error outcome of this model indicates that about 54% of patients who are not infected would have to take preventive measures.

3.3.4 Naïve Bayes

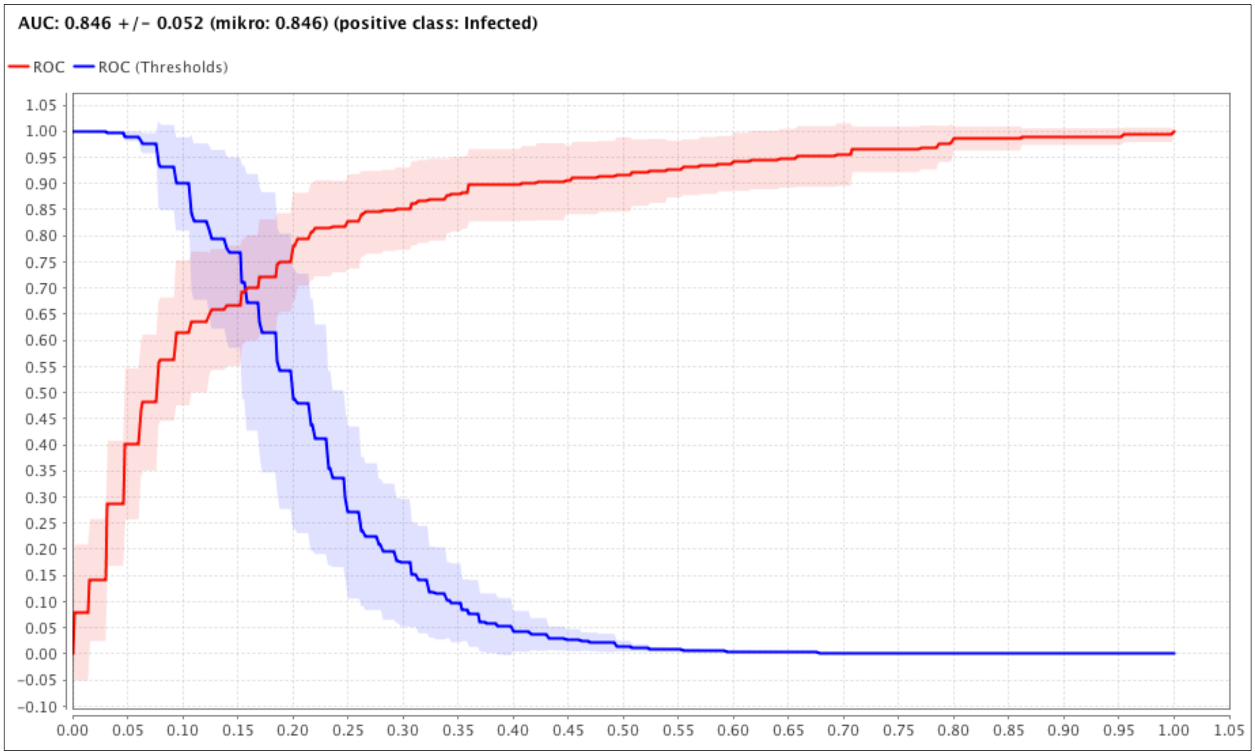
Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong naive independence assumptions between the features. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Below are the outputs from the model:



accuracy: 48.72% +/- 13.17% (mikro: 48.72%)

	true Non-Infected	true Infected	class precision
pred. Non-Infected	166	0	100.00%
pred. Infected	481	291	37.69%
class recall	25.66%	100.00%	



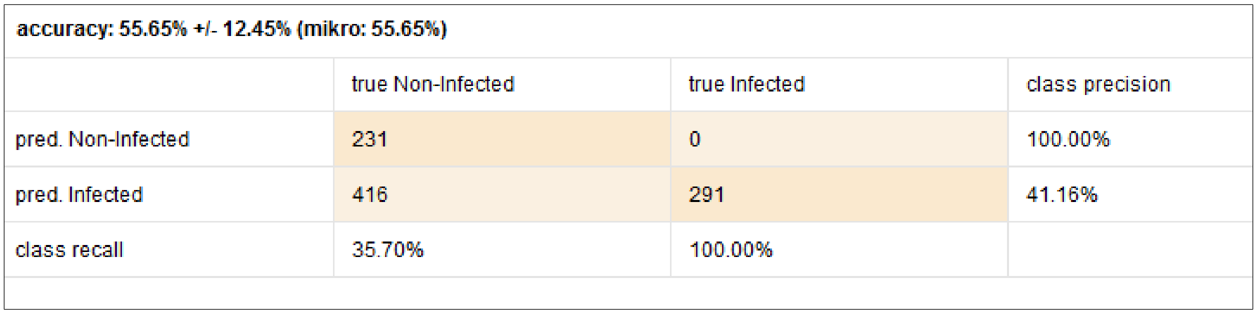
AUC: 0.846 +/- 0.052 (mikro: 0.846) (positive class: Infected)

From the model, the value of recall is 100% which shows that all patients likely to be infected while at hospital are rightly classified. However, the misclassification error outcome of this model indicates that about 74% of patients who are not infected would have to take preventive measures.

3.3.5 Neural Networks

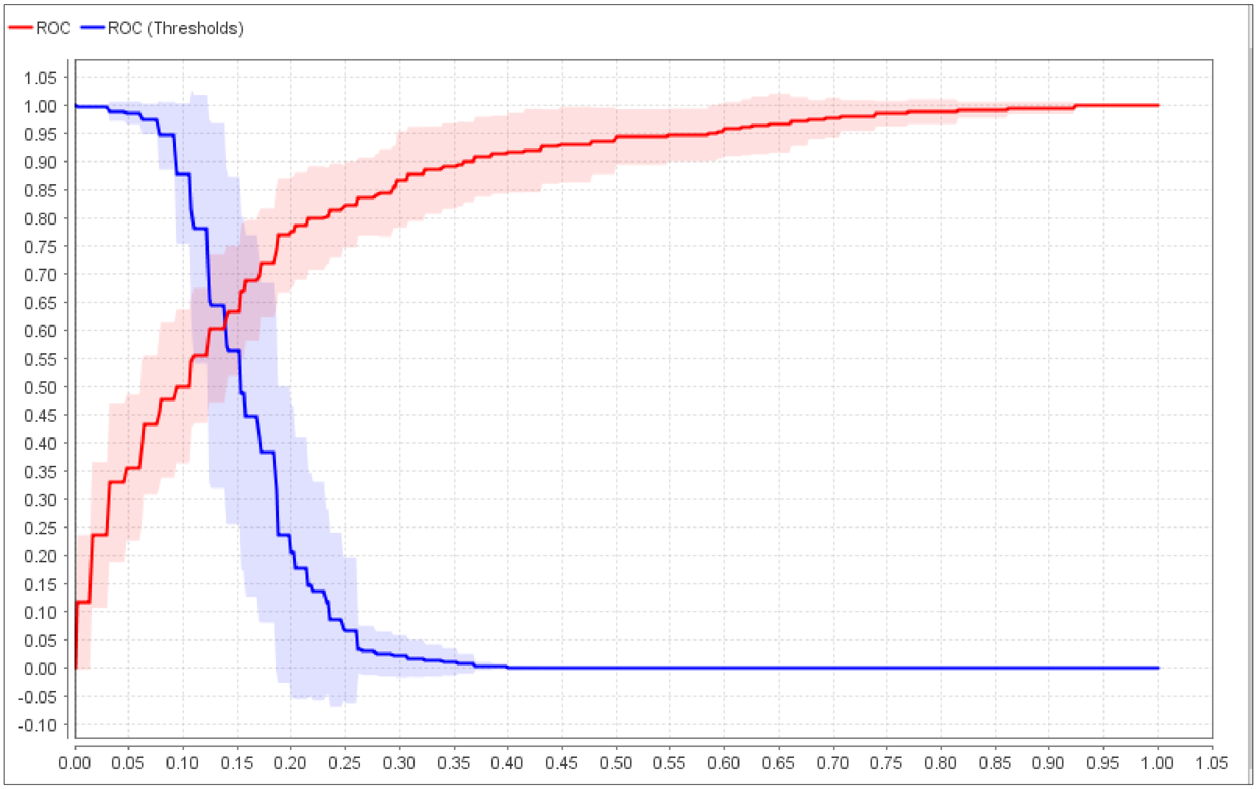
Neural networks are a specific type of deep learning algorithm used in prediction and classification problems. As neural networks work on interval data, we skipped the step to convert variables to binomial, and considered only numerical attributes. Apart from this, everything in the pre-processing step remains as-is.

Below are the outputs from the model:



accuracy: 55.65% +/- 12.45% (mikro: 55.65%)

	true Non-Infected	true Infected	class precision
pred. Non-Infected	231	0	100.00%
pred. Infected	416	291	41.16%
class recall	35.70%	100.00%	



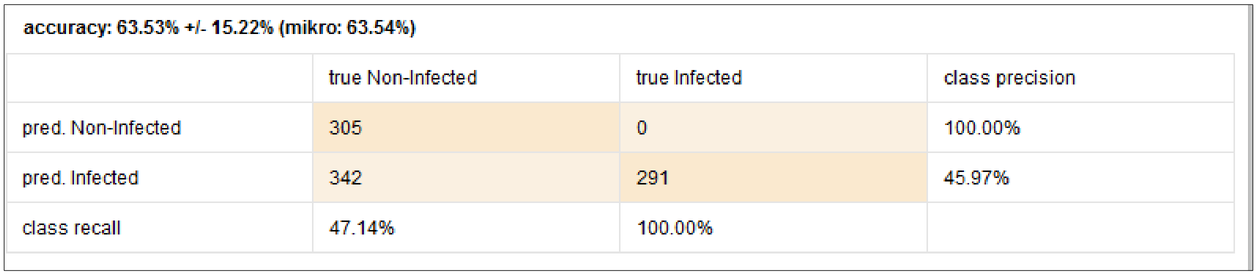
Parameters
Training cycle is 500, learning rate=0.3 and momentum=0.2. We used just 1 hidden layer for the model.

From the model, the value of recall is 100% which shows that all patients likely to be infected while at hospital are rightly classified. However, the misclassification error outcome of this model indicates that about 64% of patients who are not infected would have to take preventive measures.

3.3.6 Support Vector Machine (SVM)

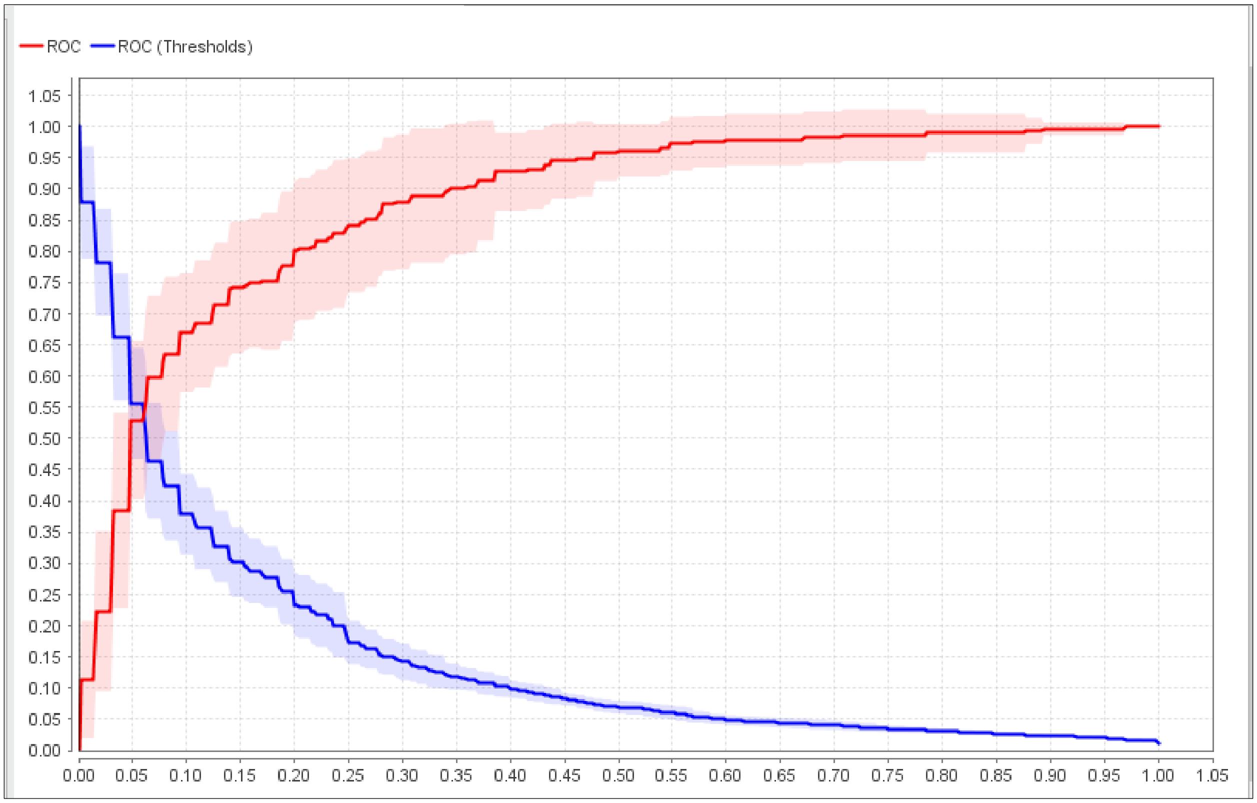
Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression, and outlier detection. As this is a classification problem with lot of attributes, SVM could be a good approach to deal with the problem. As SVM works on interval data, we omitted the step to convert variables to binomial and considered only numerical attributes.

Below is the output for the model:



accuracy: 63.53% +/- 15.22% (mikro: 63.54%)

	true Non-Infected	true Infected	class precision
pred. Non-Infected	305	0	100.00%
pred. Infected	342	291	45.97%
class recall	47.14%	100.00%	



Parameters –
Kernel type=dot and kernel cache=200

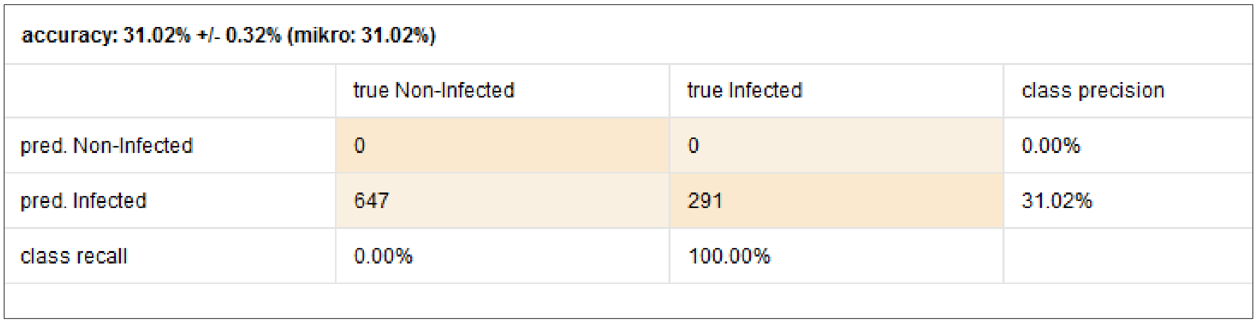
From the model, the value of recall is 100% which shows that all patients likely to be infected while at hospital are rightly classified. However, the misclassification error outcome of this model indicates that about 53% of patients who are not infected would have to take preventive measures.

3.3.7 Ensemble Modeling

Ensemble modeling or metamodeling is the process of running two or more related but different analytical models, and then synthesizing the results into a single score or spread to improve the accuracy of predictive analytics and data mining applications. Here the data is trained on a collection of models like SVM, Decision Tree, Neural Net etc.

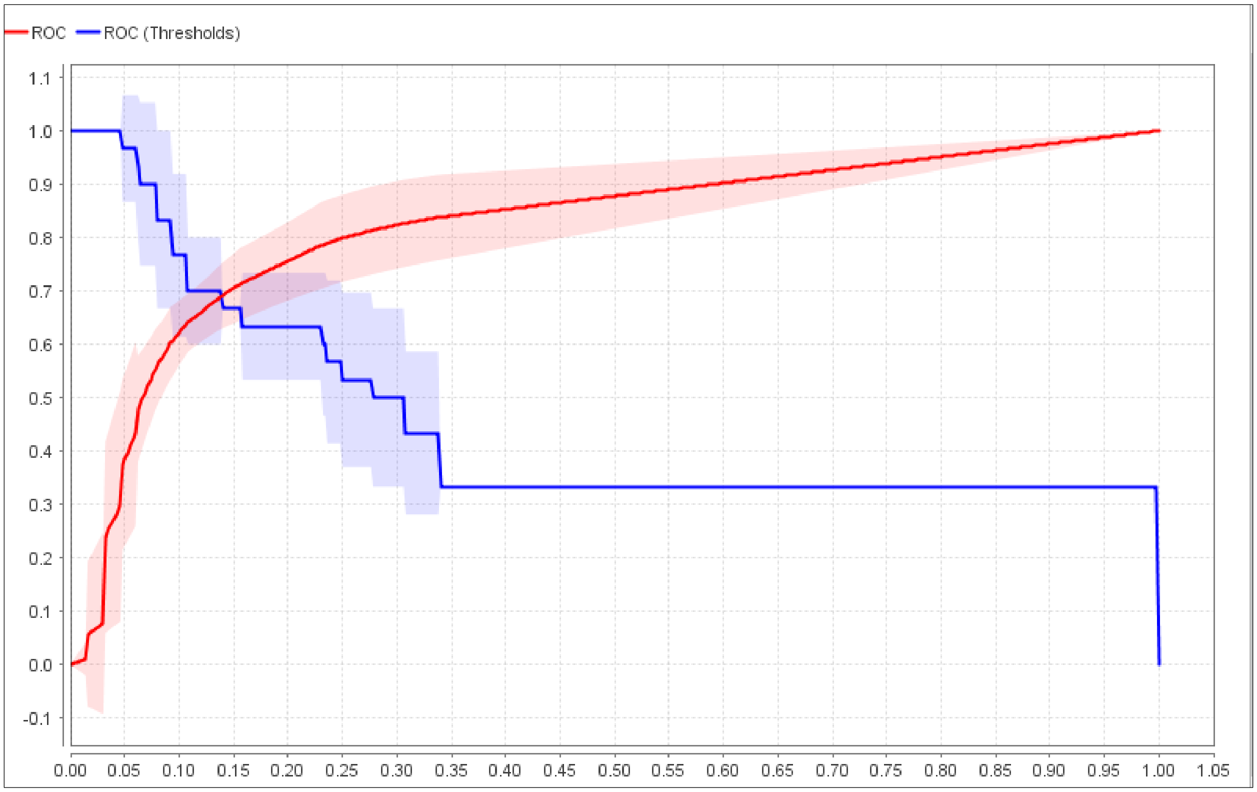
We used naïve bayes, neural networks and decision tree in our ensemble.

Below are the outputs of the model:



accuracy: 31.02% +/- 0.32% (mikro: 31.02%)

	true Non-Infected	true Infected	class precision
pred. Non-Infected	0	0	0.00%
pred. Infected	647	291	31.02%
class recall	0.00%	100.00%	



Though the model predicts all the patients infected rightly, it indicates that every non-infected patient should be given preventive measure.

4.0 Model Evaluation

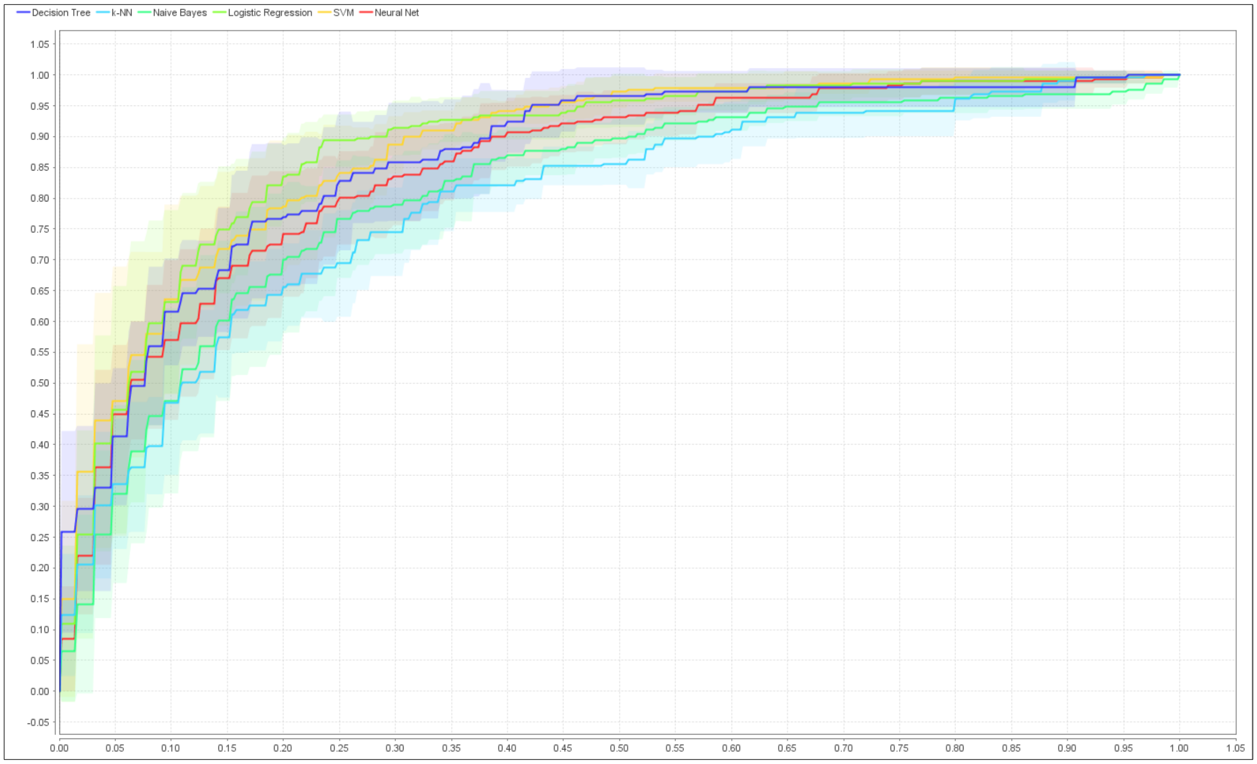
As the primary purpose of the analysis was to classify the patients who would be likely to pick up the infection while at the hospital, the main modeling goal was to get a recall of 100%, which means all the infected patients are rightly classified. Post this, we checked for the misclassification cost and AUC. Misclassification can also be interpreted as the percentage of patients who are non-infected and require preventive measure. AUC is an abbreviation for area under the curve. It is used in classification analysis to determine which of the used models predicts the classes best. It is true positive rates plotted against false positive rates. The closer AUC for a model comes to 1, the better it is. So, models with higher AUCs are preferred over those with lower AUCs.

In comparing the results from different models, Logistic Regression and Support Vector Machine (SVM) seem to provide best results, followed by Neural Networks, Naïve Bayes, Decision Tree and k-Nearest Neighbors (kNN).

Find below the comparison of metrics across different models and the ROC curve comparing the models:

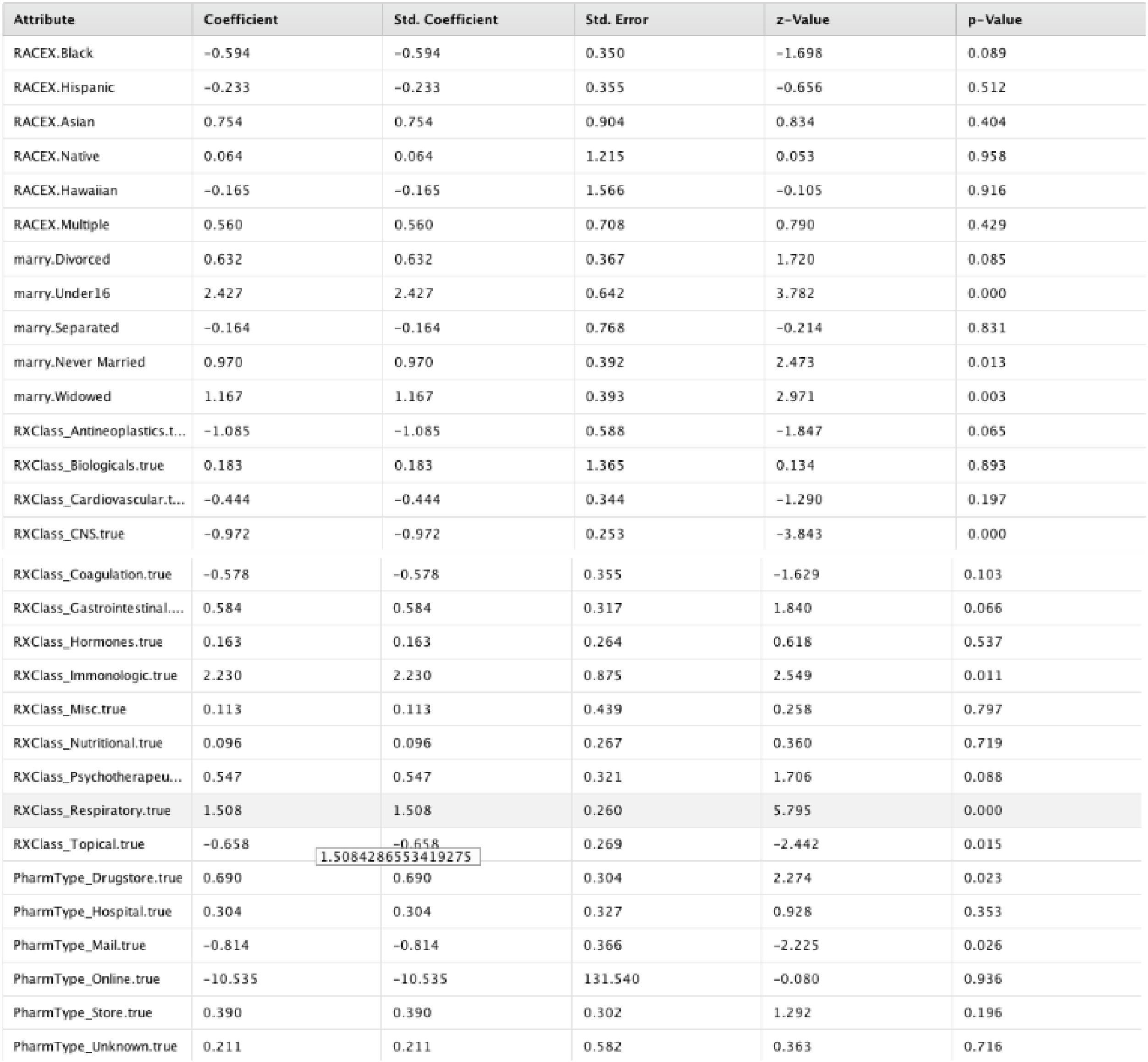
(positive class – Infected, negative class – Non-Infected)
(Recall – Classifying infected patients as infected – which is the primary goal of our analysis)

	Accuracy %	Recall %	AUC	F-measure %	Misclassification Cost	Non-Infected Patients requiring preventive measure
Decision Tree	42.62	100	0.824	53.19	442.004	83 %
kNN	34.55	100	0.772	48.72	482.372	95 %
Logistic Regression	62.34	100	0.882	64.23	343.400	54 %
Naïve Bayes	48.72	100	0.846	55.50	411.502	74 %
Neural Networks	55.65	100	0.852	59.16	376.859	64 %
SVM	63.53	100	0.877	64.50	337.48	53 %

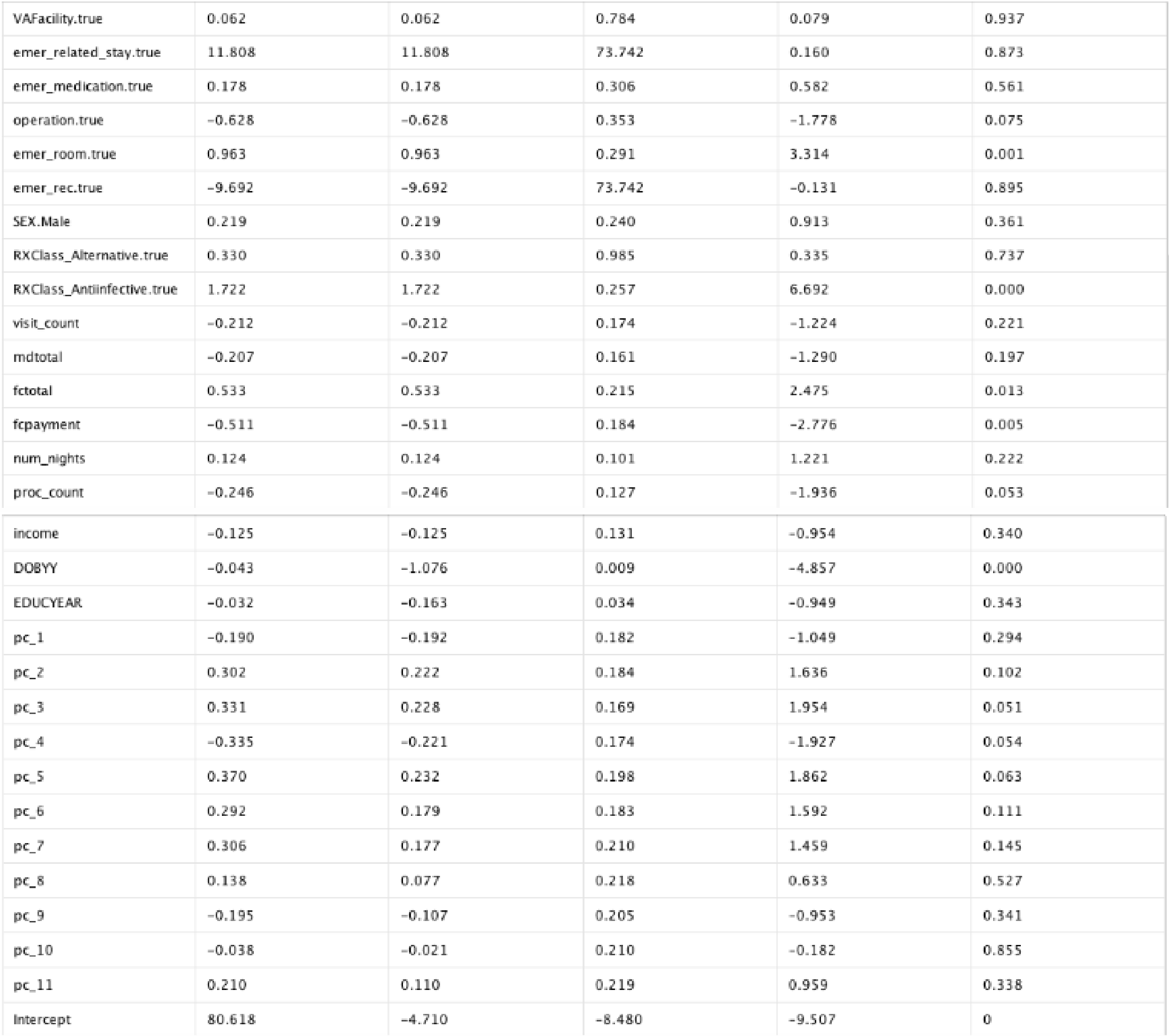


SVM gave slightly better misclassification cost compared to the logistic regression model. But we picked the logistic regression for the better interpretability of the patient attributes.

<u>Logistic Regression Output:</u>



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
RACEX.Black	-0.594	-0.594	0.350	-1.698	0.089
RACEX.Hispanic	-0.233	-0.233	0.355	-0.656	0.512
RACEX.Asian	0.754	0.754	0.904	0.834	0.404
RACEX.Native	0.064	0.064	1.215	0.053	0.958
RACEX.Hawaiian	-0.165	-0.165	1.566	-0.105	0.916
RACEX.Multiple	0.560	0.560	0.708	0.790	0.429
marry.Divorced	0.632	0.632	0.367	1.720	0.085
marry.Under16	2.427	2.427	0.642	3.782	0.000
marry.Separated	-0.164	-0.164	0.768	-0.214	0.831
marry.Never Married	0.970	0.970	0.392	2.473	0.013
marry.Widowed	1.167	1.167	0.393	2.971	0.003
RXClass_Antineoplastics.t...	-1.085	-1.085	0.588	-1.847	0.065
RXClass_Biologicals.true	0.183	0.183	1.365	0.134	0.893
RXClass_Cardiovascular.t...	-0.444	-0.444	0.344	-1.290	0.197
RXClass_CNS.true	-0.972	-0.972	0.253	-3.843	0.000
RXClass_Coagulation.true	-0.578	-0.578	0.355	-1.629	0.103
RXClass_Gastrointestinal....	0.584	0.584	0.317	1.840	0.066
RXClass_Hormones.true	0.163	0.163	0.264	0.618	0.537
RXClass_Immonologic.true	2.230	2.230	0.875	2.549	0.011
RXClass_Misc.true	0.113	0.113	0.439	0.258	0.797
RXClass_Nutritional.true	0.096	0.096	0.267	0.360	0.719
RXClass_Psychotherapeu...	0.547	0.547	0.321	1.706	0.088
RXClass_Respiratory.true	1.508	1.508	0.260	5.795	0.000
RXClass_Topical.true	-0.658	-0.658	0.269	-2.442	0.015
PharmType_Drugstore.true	0.690	0.690	0.304	2.274	0.023
PharmType_Hospital.true	0.304	0.304	0.327	0.928	0.353
PharmType_Mail.true	-0.814	-0.814	0.366	-2.225	0.026
PharmType_Online.true	-10.535	-10.535	131.540	-0.080	0.936
PharmType_Store.true	0.390	0.390	0.302	1.292	0.196
PharmType_Unknown.true	0.211	0.211	0.582	0.363	0.716



VAFacility.true	0.062	0.062	0.784	0.079	0.937
emer_related_stay.true	11.808	11.808	73.742	0.160	0.873
emer_medication.true	0.178	0.178	0.306	0.582	0.561
operation.true	−0.628	−0.628	0.353	−1.778	0.075
emer_room.true	0.963	0.963	0.291	3.314	0.001
emer_rec.true	−9.692	−9.692	73.742	−0.131	0.895
SEX.Male	0.219	0.219	0.240	0.913	0.361
RXClass_Alternative.true	0.330	0.330	0.985	0.335	0.737
RXClass_Antiinfective.true	1.722	1.722	0.257	6.692	0.000
visit_count	−0.212	−0.212	0.174	−1.224	0.221
mdtotal	−0.207	−0.207	0.161	−1.290	0.197
fctotal	0.533	0.533	0.215	2.475	0.013
fcpayment	−0.511	−0.511	0.184	−2.776	0.005
num_nights	0.124	0.124	0.101	1.221	0.222
proc_count	−0.246	−0.246	0.127	−1.936	0.053
income	−0.125	−0.125	0.131	−0.954	0.340
DOBYY	−0.043	−1.076	0.009	−4.857	0.000
EDUCYEAR	−0.032	−0.163	0.034	−0.949	0.343
pc_1	−0.190	−0.192	0.182	−1.049	0.294
pc_2	0.302	0.222	0.184	1.636	0.102
pc_3	0.331	0.228	0.169	1.954	0.051
pc_4	−0.335	−0.221	0.174	−1.927	0.054
pc_5	0.370	0.232	0.198	1.862	0.063
pc_6	0.292	0.179	0.183	1.592	0.111
pc_7	0.306	0.177	0.210	1.459	0.145
pc_8	0.138	0.077	0.218	0.633	0.527
pc_9	−0.195	−0.107	0.205	−0.953	0.341
pc_10	−0.038	−0.021	0.210	−0.182	0.855
pc_11	0.210	0.110	0.219	0.959	0.338
Intercept	80.618	−4.710	−8.480	−9.507	0

The most significant patient attributes that come from the model are age, marital status, race, number of procedures that the patient underwent, total charges (actual cost to the hospital) and payment (actual payment made to the hospital) of the patient stay, type of medication the patient received, if the patient underwent an operation, if the patient was admitted in the emergency room and the type of pharmacy from where the patient picked the medicine.

- Older patients more likely they are to get infected
- Single patients are more likely to be infected that their married counterparts
- White patients are more likely than the patients from the other race to be infected
- Higher the number of procedures the patient underwent at the hospital more are the chances of the patient getting infected
- Antineoplastics, CNS, Coagulation, Gastrointestinal, Immunologic, Topical, Respiratory, Psychotherapeutic are the medications that the patients infected with MSRA and Pneumonia also took
- Higher the charges required to be payed to the hospital, more are the chances that the patient will be infected
- The patients being admitted to the emergency room have a higher chance of getting infected

- The type of pharmacy where the patients picked the medicine – especially by mail and drugstore influences them getting infected

Cost saving calculations for Logistic Regression and SVM:

Since we have the data for patient costs we can calculate the savings provided by our model. The total cost of our 291 target patients for the years 2003 & 2004 was $8,576,761.12, or $29,473.41 per patient. If we had used our logistic regression model to classify all patients during these years, we would have had the following costs:

- 5,882 patients predicted and actual non-infected = $0
- 7,062 patients predicted infected and actual non-infected = cost of preventative treatment (~$500) = 7,062 * $500 = $3,531,000
- 291 patients predicted infected and actual non-infected = cost of preventative treatment (~$500) = 291 * $500 = $145,500
- Total Cost Savings = 8,576,761.12 – 3,531,000 – 145,500 = **$4,900,261.12**

Using the SVM model we would obtain the following results:

- 6,102 patients predicted and actual non-infected = $0
- 6,842 patients predicted infected and actual non-infected = cost of preventative treatment (~$500) = 6,842 * $500 = $3,421,000
- 291 patients predicted infected and actual non-infected = cost of preventative treatment (~$500) = 291 * $500 = $145,500
- Total Cost Savings = 8,576,761.12 – 3,421,000 – 145,500 = **$5,010,261.12**

These costs do not incorporate all the intangible costs associated with having a patient take the preventative medicine when they didn't actually need it including side effects, potential waiting time and annoyance factor, etc.

5.0 Deployment

5.1. Approach to Implement the Model in a Live Setting

The deployment of this model will incorporate several different steps and groups within the organization. Because of the sensitivity of the patient data we were not given information regarding which hospital the data came from, so we can't glean too much insight around the existing processes at the hospital. For this deployment, we'll use the example of a medium to large sized hospital that is not well versed in analytics and is currently in the process of implementing an Electronic Health Record (EHR) system for their patient records. The two primary use cases for our model are existing and new patients, which will impact the information we can use in the predictions. For example, for new patients we may not have all of the information regarding historical condition and medication information, as well as associated costs of care at previous facilities. We provide a number of recommendations targeted at improving the data quality and integrity of the systems below, but for deployment purposes we will assume none of these will have taken place and the model will be implemented as-is. This will work fine for existing patients as a one-time effort to update the patient's medical record with a new field that will capture the patient's likelihood of obtaining pneumonia or MRSA. Updating this value would require integrating the multiple disparate sources of data from each system. For our analysis we had flat files that could be used as the inputs. In a production setting, we would need to create these files from the source systems using SQL queries or some other tool. We would then need to perform the data cleansing activities in R before running the model in RapidMiner. The results of the model would need to be incorporated in a SQL update script to modify the values for the new field on the patient record. Ideally, a data pipeline would be created that automated these steps so the procedure could be run in a nightly batch process. R is a free open source tool, but would require having a data/business analyst with the necessary skills to run. RapidMiner would come at a cost, and depending on the hospital’s openness to taking this on, the model may need to be re-written in a free open-source program such as R or Python. On the front-end side of the operation, the staff including patient intake, nurses, and doctors would be made aware of the increased pneumonia and MRSA risk through the patient’s health record, and could take necessary precautions including providing preventative medication such as Zyvox and Vancomycin.

IT Infrastructure & Data Collection Procedures:

Based on the state of the data inconsistency and disparate data sources, we are making the assumption that the hospital is currently operating on multiple systems, and does not have a unified source for obtaining all of the information needed to create our predictive model. The hospital could run our model without making any changes on the provided data, but there are a number of steps they could take to improve data quality and consistency that would lead to more meaningful results. The first recommendation for the hospital would be to invest in an EHR system that will house all of the information in one place. This may include a data warehousing solution as well that would provide a better foundation for analysis. Looking back at our predictor variables across each dataset, we have:

- Hospital – Visit Count, payment information (mdtotal, fctotal, fcpayment, totalcharges, totalexpenditure)
- Medications - RX Class, Pharmacy Type
- Demographics – Date of Birth, Sex, Race, Education Year, Marriage Status, Income
- Conditions - Condition Class, Condition Count

Of this information, the majority should be stored or able to be calculated from the electronic patient record database. Depending on the level of integration of the hospital systems obtaining the payment information and medication information may prove more difficult. It is likely that these would be three disparate systems, with the payment information being stored in an ERP, the patient records being stored in an EHR system, and the medication information being stored in a separate pharmacy system. For this analysis, it is alright that they remain separate, but there has to be a unifying thread between the systems. One of the biggest issues with our analysis was that we couldn’t connect the hospital visits with the medication information or the conditions data. As a result, we could really only say that a certain patient went to the hospital and took these medications on a certain year, and not that the medications were taken as a result of the conditions identified during the hospital visit. Because of this there is the potential for data leakage in our results, with the symptoms of our target variable condition leaking into our predictor variables. To address this, the hospital visit identifier should be added in the pharmacy system. Additional steps the hospital could take to improve data quality include:

- Overall

  - Create a method to track whether a hospital visit is related to an HAI
  - Provide a consistent approach to handling NA values in the database that makes sense in the business context. The current method of using -7, -8, -9, and -1 values is highly confusing and easily leads to misinterpretation.
  - Historical condition and medication history is an important predictor of a patient's susceptibility to obtaining an HAI. Obtaining this information for new patients that transfer hospitals would be very useful for the models, and could be classified in the hospital database differently than records
- Hospital:
  - Data around the specific types of devices/methods used during the hospital visit would be useful in predicting HAI, as certain types of procedures or methods have been linked in the past with higher rates of HAI (i.e. arterial lines, CVS's, draining tubes, etc.).
  - Incorporate additional information on other predisposing factors contributing to HAI, such as smoking habits and diabetes.
- Medications:
  - Create a method to link the medications data with the hospital visits data.
- Demographics:
  - The analysts should be careful in using this data in the model as several of the variables can be misleading and open to user's specifying incorrect values (income, education level, poverty level). As such, there should be a process to spot-check these before entering in the database to remove any extreme outliers.

Adjusting the Model for the New Patient Use Case:

The summary section provided the general approach for implementing the model. This procedure will work for existing patients whose medical and medication history the hospital has access to. For new patients, the hospital will not have all the information we used as predictors in our model. Because of this a second model was created to incorporate only information the hospital would have for new patients, including demographic information and an initial condition diagnosis. For this use case, depending upon the time available to make the decision on preventative treatment measures, the pipelined process could be run to update the new patient record with a value for the new field indicator. If there is not time to complete this process, the doctor would need to make a spot decision based upon his prior knowledge and new understanding of the underlying factors the model has shown to be predictive of patients likely to obtain pneumonia or MSRA.

5.2. Legal / Patient Compliance

There is another consideration that can't be overlooked in the context of this analysis, that being obtaining patient permission to use the preventative medication as well as the various demographic information characteristics utilized in our models. For the demographic information, we are dependent upon the patient providing accurate results and being willing to provide all the information, although the more relevant features such as age are mandatory. Things such as income and education level are not highly predictive so they may be excluded in the analysis. Our cost-sensitive model placed a very high cost on false negatives, which equates to predicting a patient will not become infected when they become infected. By doing this, we are decreasing the false negatives at the cost of increased false positives, which in this case is giving preventive medication to patients who will not end up getting the condition. We identified the costs of the preventative medicine within our model, but there is no certainty that the preventive medications do not have side-effects or that the patient will even agree to take it. These factors add to our potential cost, but are not in a quantitative manner that would easily facilitate incorporation in the model.

6.0 References

Eyal Zimlichman, MD, MSc1,2; Daniel Henderson, MD, MPH1; Orly Tamir, PhD, MSc, MHA1; et al.(2013) Health Care–Associated Infections, A Meta-analysis of Costs and Financial Impact on the US Health Care System. JAMA intern Med. 22: 2039–2046.

P.J. Jenksa, M. Laurentb, S. McQuarryc, R. Watkinsb et al.(2014) Clinical and economic burden of surgical site infection (SSI) and predicted financial consequences of elimination of SSI from an English hospital. Journal of Hospital Infection.86(1):24-33.

Ying-Jui Chang,Min-Li Yeh,Yu-Chuan Li,Chien-Yeh Hsu ,Chao-Cheng Lin,Meng-Shiuan Hsu,Wen-Ta Chiu.(2011) Predicting Hospital-Acquired Infections by Scoring System with Simple Parameters. PLOS|One

Katherine E. Goodman, Justin Lessler, Sara E. Cosgrove, Anthony D. Harris, Ebbing Lautenbach, Jennifer H. Han, Aaron M. Milstone, Colin J. Massey, and Pranita D. Tamma.(2016) A Clinical Decision Tree to Predict Whether a Bacteremic Patient Is Infected With an Extended-Spectrum β-Lactamase–Producing Organism. Clinical Infectious Diseases. 63(7):896-903.

François Barbier, Antoine Andremont, Michel Wolff, Lila Bouadma (2013) Hospital-acquired pneumonia and ventilator-associated pneumonia: recent advances in epidemiology and management. Current Opinion in Pulmonary Medicine. 19(3):216-228.

Sheng WH, Wang JT, Lu DC, Chie WC, Chen YC, et al. (2005) Comparative impact of hospital-acquired infections on medical costs, length of hospital stay and outcome between community hospitals and medical centres. J Hosp Infect 59: 205–214.

Mahieu LM, Buitenweg N, Beutels P, De Dooy JJ (2001) Additional hospital stay and charges due to hospital-acquired infections in a neonatal intensive care unit. J Hosp Infect 47: 223–229.

Haley RW, Culver DH, White JW, Morgan WM, Emori TG (1985) The nationwide nosocomial infection rate. A new need for vital statistics. Am J Epidemiol 121: 159–167.

7.0 Appendix

7.1. Team Member Contributions

* Refer to Team-Member Contribution Evaluation Forms

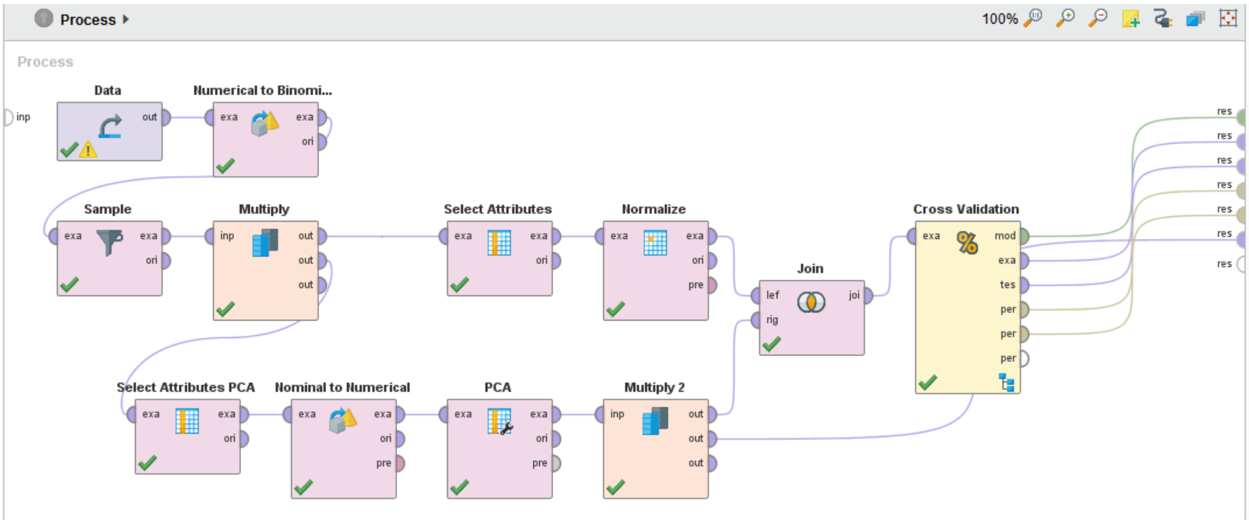
7.2. Data Preparation Source Codes

We used R for data preparation and cleansing activities. Find the detailed codes in the below embedded file:

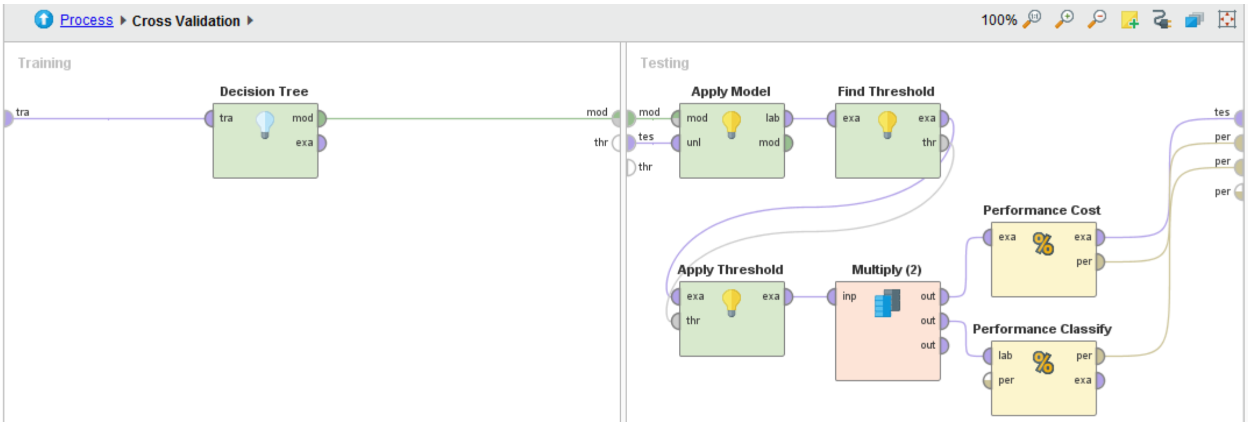
MSBA6320-PA-Projec
t-Codes.Rmd

7.3. Modeling Source Files

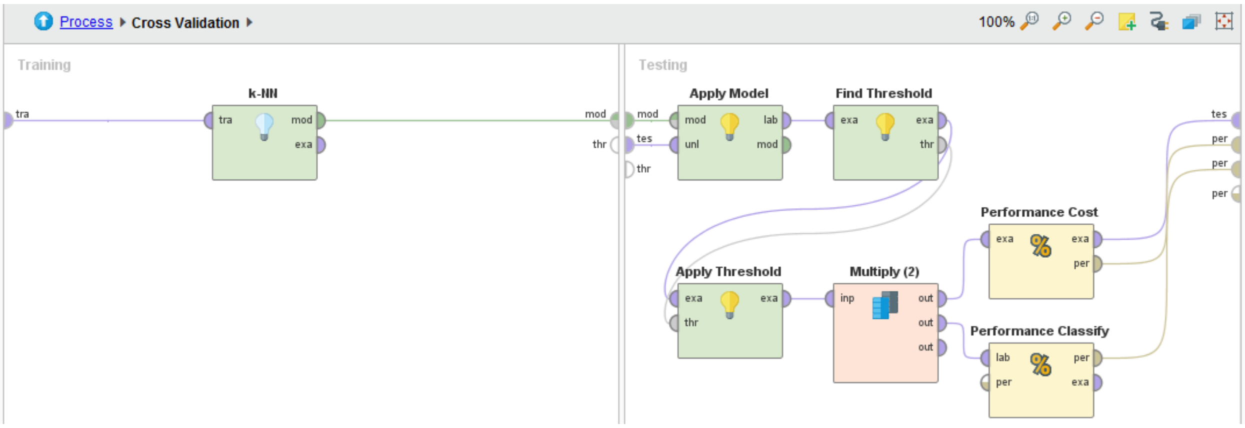
7.3.1 Pre-Processing



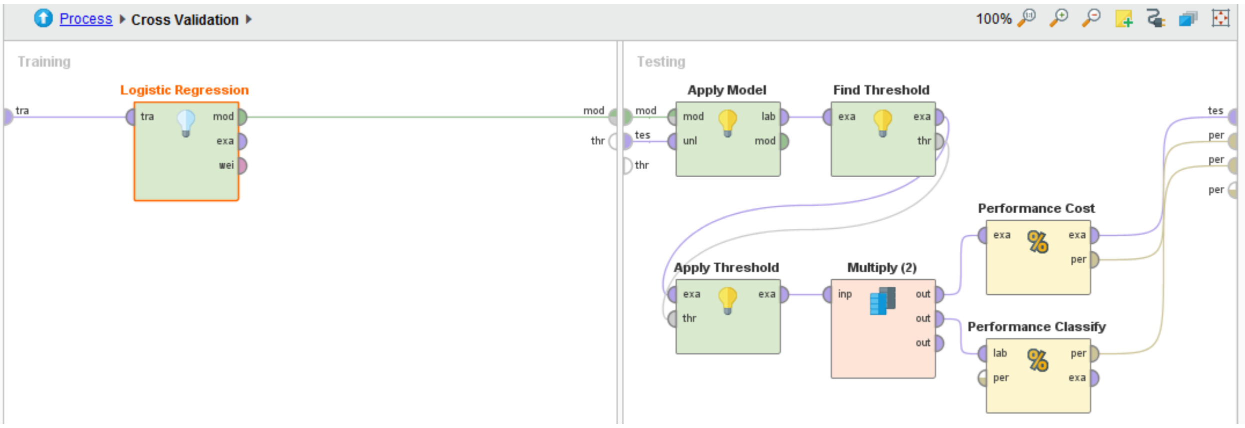
7.3.2 Decision Tree



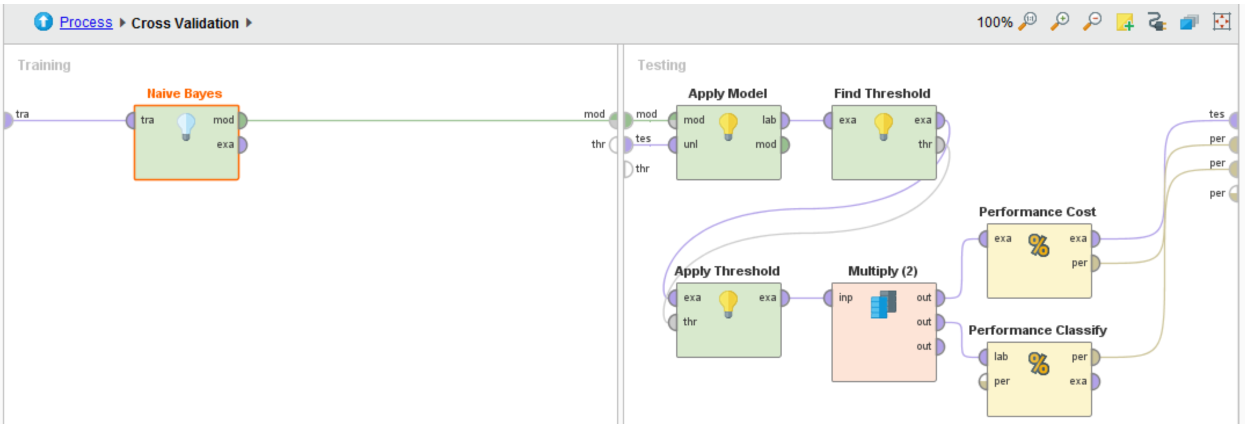
7.3.3 kNN



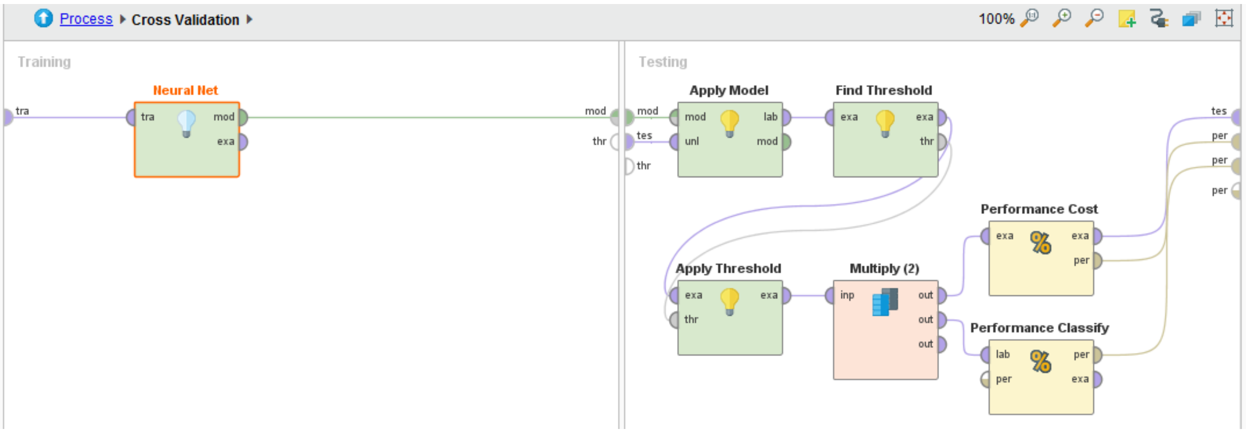
7.3.4 Logistic Regression



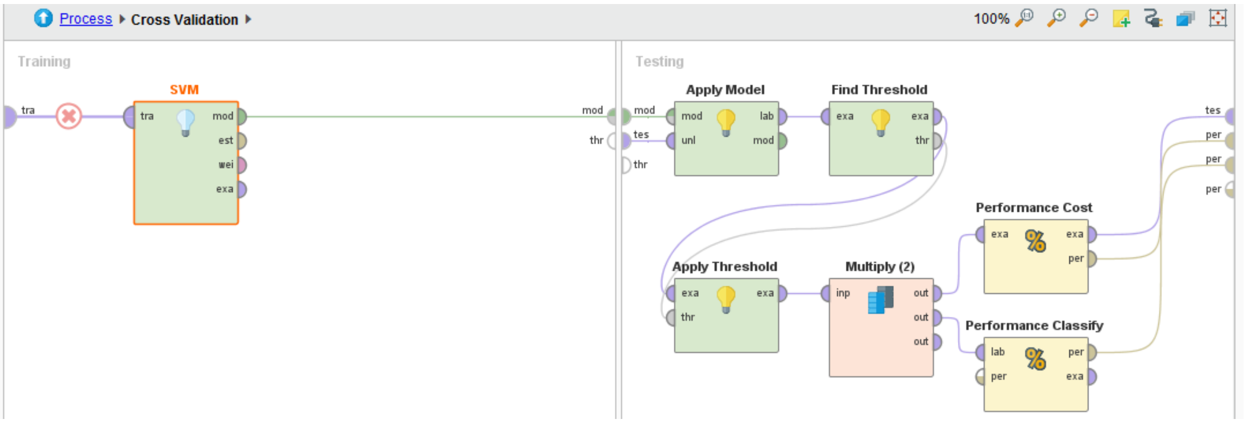
7.3.5 Naïve Bayes



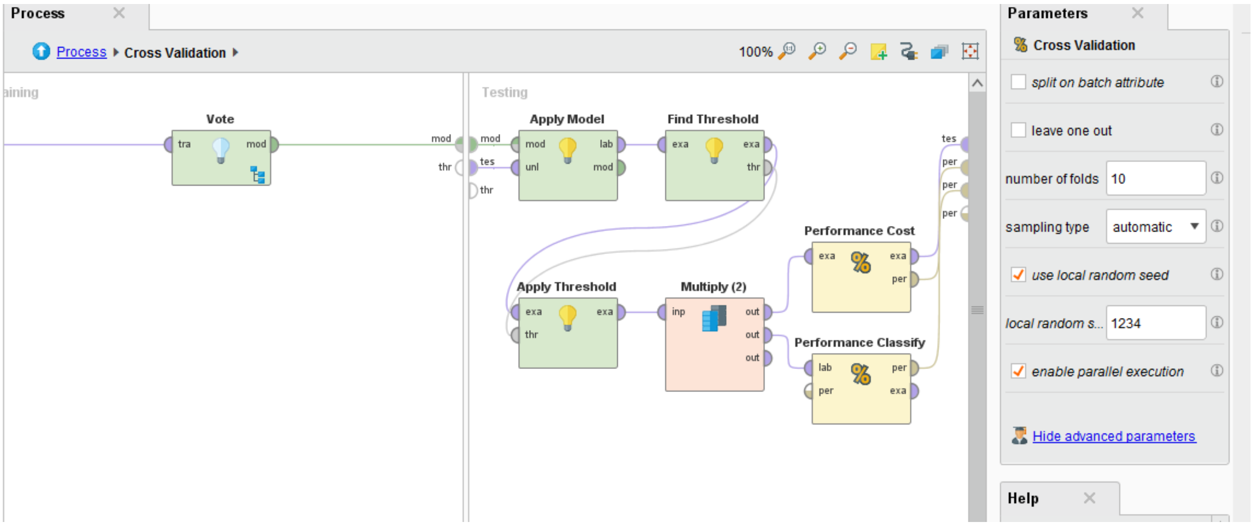
7.3.6 Neural Networks



7.3.7 Support Vector Machines



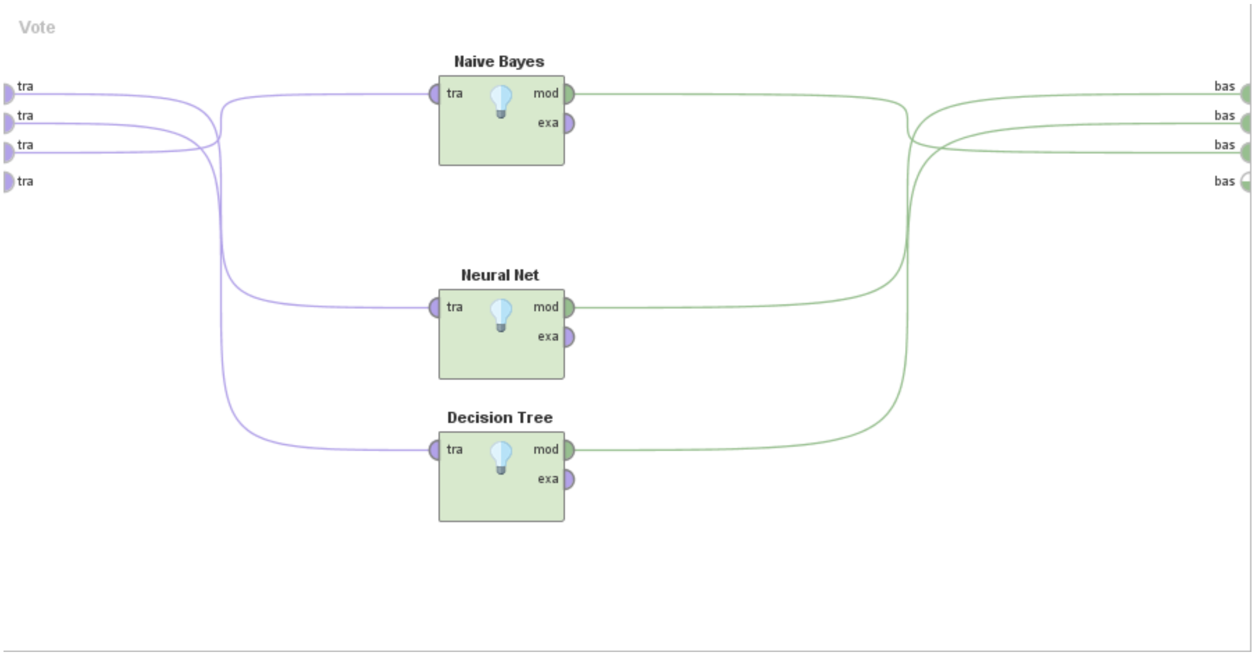
7.3.8 Ensemble



Parameters

Cross Validation

- ☐ split on batch attribute
- ☐ leave one out
- number of folds: 10
- sampling type: automatic
- ☑ use local random seed
- local random s...: 1234
- ☑ enable parallel execution


Vote
Naive Bayes
Neural Net
Decision Tree
tra
mod
exa
bas

Extra Credit – Explanatory Case Study

University of Minnesota Health Solicits Student Body Assistance in EHR Database & Predictive Modeling Exercise

As the Chief Technology Officer at University of Minnesota Health, Mark Johnson knew that his job was on the line. Across the industry, hospitals and health maintenance organizations (HMO's) were implementing new technology and touting the benefits of their electronic health record (EHR) systems and data science teams. Mark had been struggling for almost a year to garner support within the university to provide more resources toward Minnesota Health's own effort at integrating their medical records and pharmacy systems. The project had initially been highly regarded with a lot of positive momentum, but due to a lack of early results and budget overruns the project had lost steam. This served as a constant source of controversy in the weekly executive meetings, with the CEO and CFO wondering if they were wasting money on the venture, and debating internally about disbanding the project entirely.

Seeing the end in sight, Mark knew that drastic action was required. He had previously suggested the idea of involving university students in the project as a potential cost-saving measure since he could not obtain the needed resources through other channels. He was concerned that the executive management team would view this as an act of desperation, putting into question his own leadership abilities. The next board meeting was coming up in just four days, and he knew he couldn't raise the topic again without having fully vetted the plan.

Mark knew exactly the person to call to discuss the idea with. He dialed up Azhara Khalil of the University of Minnesota Graduate Career Center to discuss the possibility of using Master of Science in Business Analytics (MSBA) students to assist in the project. He knew these students would have the requisite business and technical skills to take on a project of this magnitude with such a fast turn-around time. After discussing on the phone, Mark and Azhara were able to coordinate a time to meet with a group of select students the following day. After expressing his gratitude and hanging up, he immediately began coordinating with his team to gather the necessary background materials for the students to hit the ground running.

UNIVERSITY OF MINNESOTA

Driven to Discover℠

University of Minnesota MSBA Program

The University of Minnesota's MSBA program had built an outstanding reputation during its short existence dating to 2014. Students who enrolled in this program dedicated themselves to 45 rigorous credits completed in a one-year program, and experienced 100% job placement with companies across the United States upon graduation. The student body included 84 students from a variety of educational and cultural backgrounds. This created an exciting environment full of new ideas, but could often lead to challenges in managing the internal dynamics of project teams. Despite these challenges, the students had performed well together in similar situations during their program.

First Meeting with the Team

After another sleepless night, Mark grabbed his coffee the next day and set off on his trip to the West Bank side of campus to meet with the MSBA students. Azhara had mentioned there would be a group of four to five students participating in the project, but had not provided any other details. Upon meeting the group, introductions were made. The group consisted of the following members:

- Sandeep, a 26-year-old male international student from India
- Tina, a 30-year-old Chinese American woman who had been working in Minnesota for the last two years
- Tom, a 28-year-old male student who was a native Minnesotan
- Shruthi, a 25-year-old female international student from India

After meeting the team, Mark explained the problem he was facing and some of the details around the University's current EHR implementation project. The students had recently completed a class in Database Management so were intimately aware of the types of challenges these projects can encounter. Being aspiring data scientists, the group first wanted to look at the data.

Having come prepared, Mark pulled out a flash drive containing the details of the project and provided the documents he had compiled the previous day that explained all the work completed on the project to this point. He had been told by Azhara that the students were self-starters so he assumed they would want to go ahead and get started on creating an approach to complete the work. He provided the students a completion deadline of three days to allow him to present the findings at the weekly executive meeting, thanked everybody for their time, and wished them luck before exiting the room.

The students were stunned, each looking around the room at one another in surprise. Had the CTO really just left them with a three-day deadline on a massive project with no clear direction?

The Students Get to Work

With little time and no plan in place, the students knew they needed to get started immediately. They began working through the files, and came to a consensus on the four primary sets of information that would need to be accounted for in their proposal. These datasets included patient conditions, patient demographics, hospital events, and patient medication information. The group modeled out the conceptual diagram for the system on the white board to get a better visual understanding of the information flow.

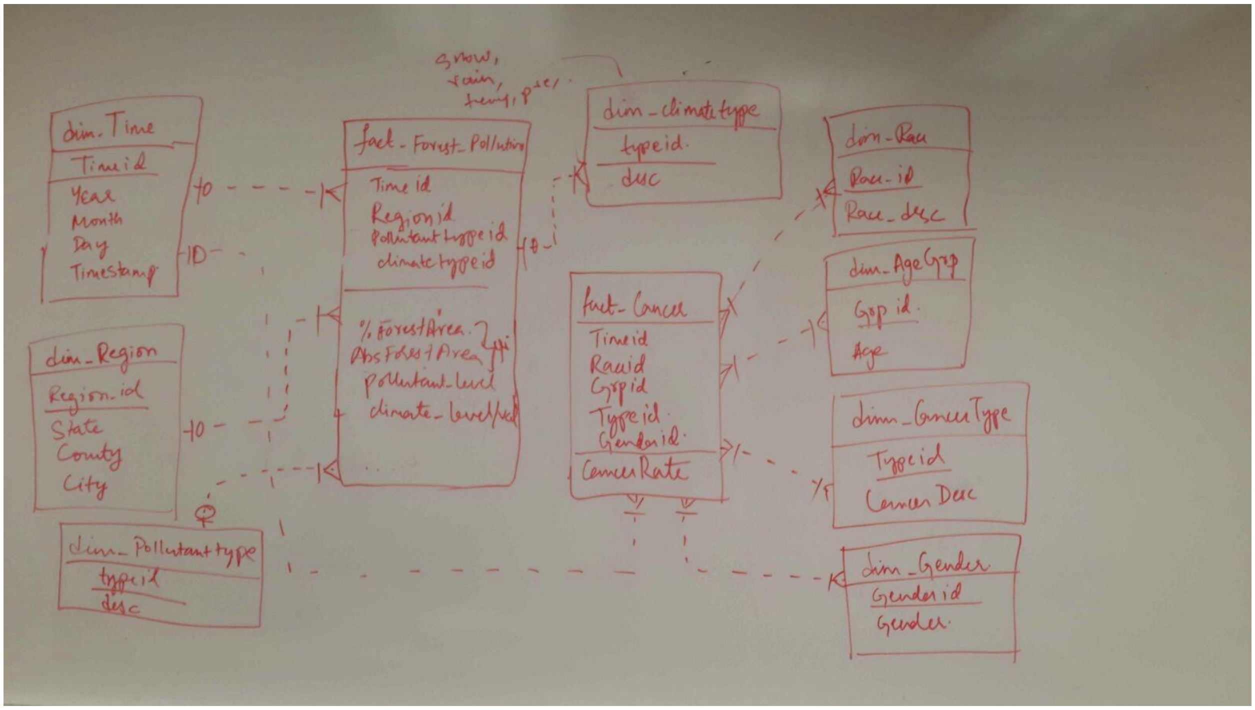


Figure 4 - Conceptual Database Diagram

While reviewing the diagram the team couldn't help but feel something was missing. Despite the seeming interconnectedness of the datasets, there didn't seem to be a way to combine them into a fashion suitable for analysis. Having already been discussing for six hours they decided to break for the day and come back the following morning to continue.

A New Challenge is Added to the Project Scope

The next day the students opened their inbox to an email from Mark, requesting the team to continue their current work, but also switch gears and focus on a new issue that he believed to be a pressing item needed for the executive meeting. Mark wanted the group to come up with an approach to predict hospital patients that were more likely to obtain a hospital acquired infection (HAI) such as MSRA and pneumonia. He mentioned briefly that this would save the university a lot of money, but didn't clarify beyond that. The students had yet to explain to Mark the issues they were having with the existing project, and were now faced with an entirely new problem for which they had very limited domain knowledge. The group knew they were in a precarious situation, but also didn't feel they had much of a choice in the manner. Despite the circumstances the team worked through the night to try and come up with an idea to address the problem. The discussions seemed to go in circles, however, as the group did not handle ambiguity well and were still getting used to each other as classmates. They went to sleep at 4 am without a plan in place.

The Team Hits an Impasse with the Deadline Looming

With the deadline looming the next day the team was stuck. They could not come to an agreement on the approach to take, and knew they were running out of time. They decided to solicit feedback from a trusted advisor at the school. They prepared a list of questions they were hoping for advice on to move forward with the project to meet the deadline for Mark's meeting:

1. How can we better manage the internal dynamics of the team to complete the project on time?
2. What is the best approach for researching and learning about unfamiliar domains to assist in a data science exercise?
3. What approaches can be used to integrate datasets that can’t be directly linked, and what are the potential issues that could result?

They believed the answers to these questions would provide the framework to complete the analysis and move forward with the project, but were concerned about the scope becoming unmanageable for their team.