Explainable artificial intelligence model to predict mortality in patients presenting with acute coronary syndromes

Edwin Kagereki - B00867154

08 November 2021

Contents

Business Understanding

Introduction

Cardiovascular diseases (CVDs), principally ischemic heart disease (IHD) and cardiac stroke, are the leading cause of global mortality. In addition to their increasing global prevalence, they are often associated with poor survival(Roth 2020).

Acute Coronary Syndrome (ACS) is a term given to a continuum of CVDs ranging from ST-segment elevation myocardial infarction (STEMI) to non–ST-segment elevation myocardial infarction (NSTEMI) and unstable angina. Accurate estimation of risk for untoward outcomes after a suspected onset of an ACS may help clinicians chose the type and intensity of therapy. Such predictions may be helpful as patients predicted to be at higher risk may receive more aggressive surveillance and/or earlier treatment, while patients predicted to be at lower risk may be managed less aggressively.

Problem statement The establishment of prognosis model for patients with suspected ACS is important in critical care medicine.Numerous risk-prediction models for differing outcomes exist for the different types of ACS.

These models however have some limitations. First, most models have been developed from large randomized clinical trial populations in which the generalizability to risk prediction in the average clinician's experience is questionable(Eagle KA Lim MJ 2004).

Second, given the dynamic nature of the treatment environment, predicting future behavior while the treatment is underway may help the clinicians make decisions proactively.

This project aimed at developing a risk-prediction tool for ACS, focusing on clinical end point of all-cause mortality. The motivation was to determine the utility of machine learning at the population level using multiple linked ICU patients' datasets.

Project Objectives

This project developed a tool for application in the decision-making environment. This was done in two steps:

1. A binary classification model was developed. This model used:

- Patient demographics
- Interventions within the golden hour (laboratory, medication, procedures and microbiology studies).

The concept of the golden hour refers to the vital period by which a patient with a suspected cardiovascular event should be receiving definitive treatment to prevent death or irreparable damage to the heart.Although not set in stone, the chances to save a patient are usually high if substantive medical attention is given within an hour of the cardiac event(Johnson 2016). In this study the golden hour cut-off was 60 minutes after initial contact with the hospital. This included all interventions that were done prior to admission.

2. Process mining was used to explain the outcome of the patient based on the care pathway followed. Two concepts were applied:

- Process discovery - Processes followed by the two classes will be described.
- Conformance checking - The care pathway followed by both patients will be checked for conformance with the American Heart Association(AHA) guidelines for CPR and ECC (ACLS 2020). To run the process mining the timestamped interventions (laboratory, medication, procedures) were used.

Project Plan

Sources of Data and Knowledge

MIMIC III Database MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with 46,520 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside , laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (including post-hospital discharge).

Although de-identified, the datasets described herein contain detailed information regarding the clinical care of patients, and as such it must be treated with appropriate care and respect.

Researchers seeking to use the database must:

1. Become a credentialed user on PhysioNet. This involves completion of a training course in human subjects research.
2. Sign the data use agreement. Adherence to the terms of the DUA is paramount.

LOIC tables The Loin(Institute 2021) data tables will be used to enrich the laboratory dataset.

American Heart Association guidelines for CPR and ECC This AHA guideline(ACLS 2020) will be used as the gold standard for ACS workflow. The workflows for the patients in this project will be assessed for cornformity with this workflow.

Terminology

1. **Acute Cardiac syndrome:** Acute coronary syndrome (ACS) refers to a spectrum of clinical presentations ranging from those for ST-segment elevation myocardial infarction (STEMI) to presentations found in non–ST-segment elevation myocardial infarction (NSTEMI) or in unstable angina. It is almost always associated with rupture of an atherosclerotic plaque and partial or complete thrombosis of the infarct-related artery. Candidates of acute cardiac syndrome were identified using the *DIAGNOSIS* in the *ADMISSIONS* table which provides a preliminary. This column was is a free text diagnosis for the patient on hospital admission. The diagnosis was assigned by the admitting clinician and did use a systematic ontology. Candidate cased were identified by using the key words commonly used in the diagnosis of acute coronary syndrome and the related differential diagnosis. These were:

"stemi,""acute coronary syndrome,""angina,""tachycardia,""aortic aneurysm,""pericardi,""ortic dissection,""coronary artery dissection,""cardiomyopathy,""heart failure,""mitral valve disease,""mitral stenosis,""coronary artery disease,""chf,""congestive heart failure,""heart failure,""telemetry,""myocardial infaction,""cardiac arrest,""myocardial infarction,""aortic stenosis,""st elevated,""pericardial effusion," "cardiomyopathy,""cath lab,""tamponade,""tamponede"

2. Angiotensin-converting enzyme (ACE) inhibitors are medications that help relax the veins and arteries to lower blood pressure. ACE inhibitors prevent an enzyme in the body from producing angiotensin II, a substance that narrows blood vessels. The following terms were used to identify ACE's from the list of *DRUG* column of the *PRESCRIPTIONS TABLE* table:

"benazepril," "captopril," "enalapril,""enalaprilat,""fosinopril," "lisinopril," "moexipril," "perindopril," "quinapril," "ramipril,""trandolapril"

3. Beta blockers (beta-adrenergic blocking agents)

Medications that reduce blood pressure. Beta blockers work by blocking the effects of the hormone epinephrine, also known as adrenaline.The following terms were used to identify Beta blocker from the list of *DRUG* column of the *PRESCRIPTIONS TABLE* table:

"acebutolol,""atenolol,""betaxolol,""bisoprolol,""carteolol,""carvedilol,""labetalol,""metoprolol,""nadolol,""nebivolol," "penbutolol,""pindolol,""propanolol,""sotalol,""timolol"

4. Glycoprotein IIb/IIIa inhibitors

These drugs are frequently used during percutaneous coronary intervention (angioplasty with or without intracoronary stent placement). They work by preventing platelet aggregation and thrombus formation.

The following terms were used to identify Glycoprotein IIb/IIIa inhibitors's from the list of *DRUG* column of the *PRESCRIPTIONS TABLE* table:

“abciximab,”“eptifibatide,”“tirofiban,”“roxifiban,”“orbofiban”

5. P2Y12 inhibitors *“clopidogrel,”“prasugrel,”“ticlopidine,”“ticagrelor”*
6. HMGCoA *“altoprev,”“amlodipine,”“atorvastatin,”“caduet,”“crestor,”“ezallor,”“fluvastatin,”“lescol,”“lipitor,”“livalo,”“lo*
7. A glossary of data mining terminology, illustrated with examples relevant to the business problem in question

- Check prior availability of glossaries; otherwise begin to draft glossaries
- Talk to domain experts to understand their terminology
- Become familiar with the business terminology

Success Criteria

The success of this project was to generate forward-looking, predictive insights to improve the management of the care pathway in patients suspected to have ACS by:

- Successfully predicting the ICU mortality outcome of the patient with suspected ACS based on the patient demographics and the interventions given within the first one hour.
- Identify patients in which undesirable events will likely be observed in the based on the interventions.

Inventory of resources

Software For this project the following software will be used:

1. PostgreSQL
2. R
3. Python

Computing resources The analysis will be done on a Windows desktop and a Linux server. Github repository was used for the CI/CD pipeline.

Requirements, Assumptions, and Constraints

For this analysis, the all the medical records were not analysed.

It is also assumed that:

- The patients in this population were only treated in this hospital, therefore mortality are only captured in this hospital.
- All the pre-hospitalization interventions were captured.

Risks and Contingencies

- Although the process mining will give a better predictive description of the patient outcomes, alternative surrogate modeling methods like decision tree maybe used.

Data Mining Goals

A business goal states objectives in business terminology; a data mining goal states project objectives in technical terms. For example, the business goal might be, “Increase catalog sales to existing customers,” while a data mining goal might be, “Predict how many widgets a customer will buy, given their purchases over the past three years, relevant demographic information, and the price of the item.”

Data Mining Goals

1. Build a binary classification machine learning model to predict all-cause mortality of patients based on demographics and interventions within first hour of suspected ACS event.
2. Compare the conformity of care path in patients who died and patients who survived with the with the ACLS care path.
3. Assess and report any work flow variants, and differences between patients who died and those who survived.

Data Mining Success Criteria

After training the binary classifier, evaluation measures will be used to assess the performance of the model. The predictive performance of the classifier will be assessed by calculating the number of correctly identified class patients (true positives), the number of correctly recognized patients that are not member of the class (true negatives), the number of the patients that are wrongly recognized (false positives) and the number of the examples that were not identified (false negatives). By using these measures, a confusion matrix will be constituted.

		Ground Truth Values	
		Positive	Negative
Predicted Values	Positive	true positive (tp)	false positive (fp)
	Negative	false negative (fn)	true negative (tn)

The following measures will be calculated from this table:

- Accuracy.
- Precision.
- Recall.
- Specificity.

To benchmark the classification model performance, the results ranging from accuracy of (70% to 90%) as reported in larger, though different models used in the prediction of mortality in CVDs will be used(Sherazi et al. 2020).

Data Understanding

Data Access

Having met the criterian and gained access, I also learned about and calculated various severity scores for each patient which I will talk about later on in the article. When allowed access to the MIMIC-III database, it is suggested that you transfer all of this information into a RDMS (relational database management system) and Physionet has tutorials on how to transfer the database into a local instance of the PostgreSQL RDMS which I followed. After connecting to the PostgreSQL database, I was able to easily make SQL queries and connect my database to many helpful tools such as pgAdmin4 which provides a GUI (graphical user interface) for the database.

Although the database includes 26 tables, only the following tables will be included in the analysis:

- **ADMISSIONS:** Contains information regarding a patient's admission to the hospital. Information available includes timing information for admission and discharge, demographic information, the source of the admission, and so on. Record of 58,976 unique admissions.
- **PATIENTS:**Defines each patient in the database, i.e. defines a single patient. There are 46,520 patients recorded.
- **SERVICES:**Lists services that a patient was admitted/transferred under.This table contains 73,343 entries.

- **DIAGNOSIS_ICD**Identify type of data sources (online sources, experts, written documentation, etc.)
- **MICROBIOLOGYEVENTS:**Contains microbiology information, including cultures acquired and associated sensitivities.There are 631,726 rows in this table.
- **PRESCRIPTIONS:**Contains medication related order entries, i.e. prescriptions.This table contains 4,156,450 rows.
- **PROCEDUREEVENTS_MV:**Contains procedures for patients. This table has 258,066 rows.
- **D_ITEMS:** Definition table for all 12,487 items in the ICU databases.
- **D_LABITEMS:** Definition table for 753 laboratory measurements.

Selection criteria

For this analysis only the patients admitted with suspected acute cardiac syndrome were included. In this subset, patients who died within the first hour of treatment were also excluded.

Derived variables of data

The following derived variables were calculated:

1. Age - The age was computed by subtracting the *DOB* from the *ADMITTIME*. Any figure above 300 was adjusted by subtracting 211, since any age above 300 was ages over 89 had been shifted such that the patient age appears to be 300 in the database.

2. Splitting of Datetime Features - The following features were extracted from teh *ADMITTIME* feature:

- Day of the year
- Week of the year
- Month
- Year
- Hour of day

3. The Length of stay: This was calculated from the previous admissions.

4. Admission cycle: If the patient had multiple admissions, what was the admission cycle in this case.

Data exploration

The MIMIC III dataset contained 20399 43.85 female and % and 26121 56.15 male patients. The cumulative incidence of suspected ACS was 24.11%. The report is summarized in Table 1 below.

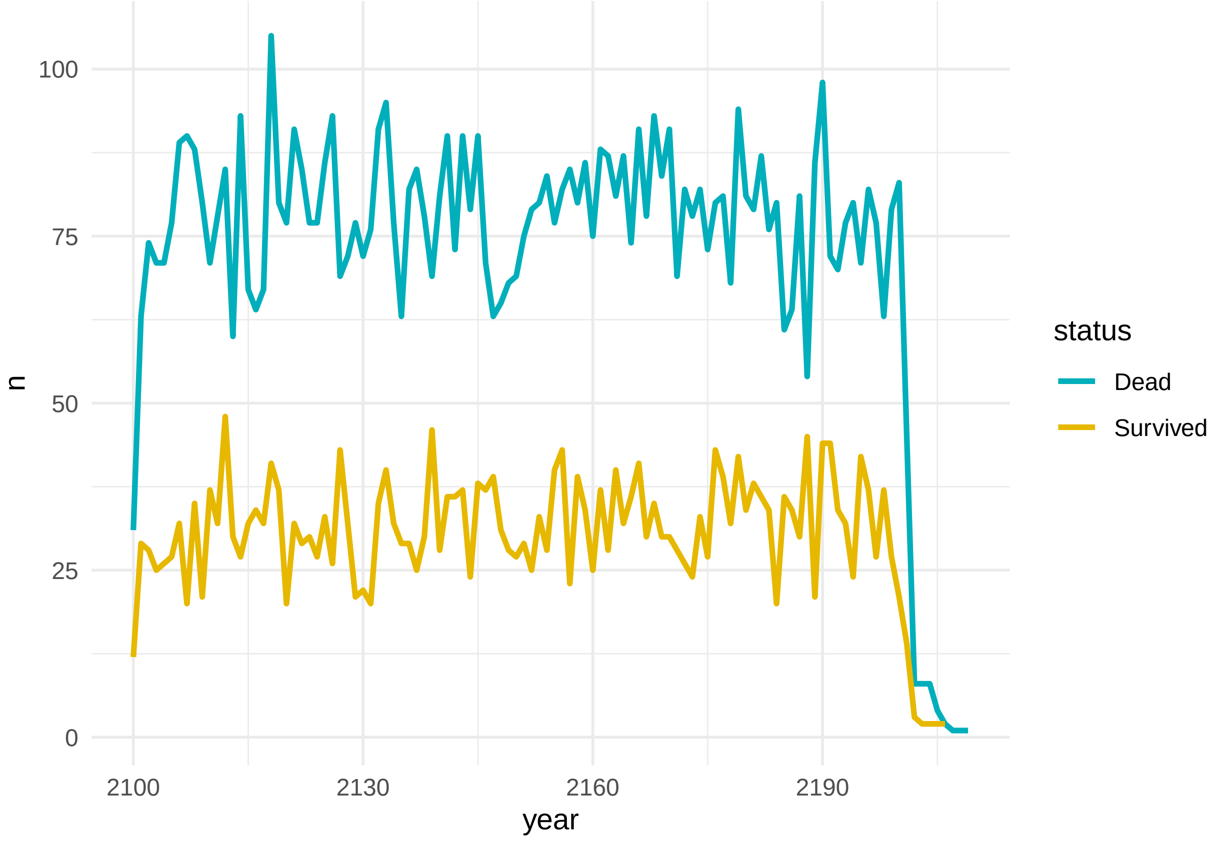
Table 2: **Patient Characteristics**

Characteristic	**Deceased**, N = 3,237	**Survived**, N = 7,978	**p-value**
GENDER, n (%)			<0.001
F	1,423 (44%)	2,841 (36%)	
M	1,814 (56%)	5,137 (64%)	
AGE, Mean (SD)	74.90 (12.16)	66.26 (13.53)	<0.001
INSURANCE, n (%)			<0.001
Private	465 (14%)	2,833 (36%)	
Public	2,762 (85%)	5,105 (64%)	
Self Pay	10 (0.3%)	40 (0.5%)	
MARITAL_STATUS, n (%)			<0.001
Living alone	1,455 (45%)	3,096 (39%)	
Living with Partner	1,533 (47%)	4,623 (58%)	
UNKNOWN	249 (7.7%)	259 (3.2%)	
Ethinicity, n (%)			<0.001
ASIAN	42 (1.3%)	147 (1.8%)	

Characteristic	**Deceased**, N = 3,237	**Survived**, N = 7,978	**p-value**
BLACK	186 (5.7%)	593 (7.4%)	
CAUCASIAN	2,224 (69%)	5,733 (72%)	
HISPANIC	53 (1.6%)	249 (3.1%)	
OTHER	47 (1.5%)	186 (2.3%)	
UNKNOWN	685 (21%)	1,070 (13%)	

These data was subset using a total of unique subset of the words as shown below.

In addition the patients were seen over a span of time ranging from 2100-07-09 to 2209-07-14. This temporal distribution is shown in the chart below:



Volumetric analysis of data

The total dataset was estimated to be more than 100GBs of data. The selected subset of data tables was about 3.47 GBs.

Attribute types and values

And see Table @ref(dtypes).

Assumptions/Limitations

- The patients in this population were only treated in this hospital, therefore all important events like hospitalization and mortality are only captured in this hospital.
- All the pre-hospitalization interventions were captured.

Table 3: Attributes and the level of measurement

variables	types	missing_count	missing_percent	unique_count	unique_rate
X	integer	0	0.00000	11215	1.0000000
SUBJECT_ID	integer	0	0.00000	10289	0.9174320
HADM_ID	integer	0	0.00000	11215	1.0000000
ADMITTIME	character	0	0.00000	11169	0.9958984
DISCHTIME	character	0	0.00000	11203	0.9989300
DEATHTIME	character	10258	91.46679	958	0.0854213
ADMISSION_TYPE	character	0	0.00000	3	0.0002675
ADMISSION_LOCATION	character	0	0.00000	6	0.0005350
DISCHARGE_LOCATION	character	0	0.00000	15	0.0013375
INSURANCE	character	0	0.00000	5	0.0004458
LANGUAGE	character	0	0.00000	47	0.0041908
RELIGION	character	0	0.00000	20	0.0017833
MARITAL_STATUS	character	0	0.00000	8	0.0007133
ETHNICITY	character	0	0.00000	35	0.0031208
EDREGTIME	character	0	0.00000	4281	0.3817209
EDOUTTIME	character	0	0.00000	4281	0.3817209
DIAGNOSIS	character	0	0.00000	3022	0.2694605
HOSPITAL_EXPIRE_FLAG	integer	0	0.00000	2	0.0001783
HAS_CHARTEVENTS_DATA	integer	0	0.00000	2	0.0001783
GENDER	character	0	0.00000	2	0.0001783
DOB	character	0	0.00000	9210	0.8212216
EXPIRE_FLAG	integer	0	0.00000	2	0.0001783
AGE	integer	0	0.00000	84	0.0074900
LOS2	numeric	0	0.00000	8253	0.7358894
Period	integer	0	0.00000	12	0.0010700
endGoldenHour	character	0	0.00000	11169	0.9958984
admissionCycle	integer	0	0.00000	26	0.0023183
nAdmissions	integer	0	0.00000	24	0.0021400
deadBefore	numeric	0	0.00000	926	0.0825680
DIAGNOSIS2	character	0	0.00000	3022	0.2694605
dayOfYear	integer	0	0.00000	366	0.0326349
Month	integer	0	0.00000	12	0.0010700
week	integer	0	0.00000	53	0.0047258
weekday	integer	0	0.00000	7	0.0006242
year	integer	0	0.00000	110	0.0098083
hour	integer	0	0.00000	24	0.0021400
DIAGNOSIS3	character	0	0.00000	2655	0.2367365

- The unit of analysis is the admission

Data Preparation

Data quality was assessed dimensions of completeness, uniqueness, validity, accuracy and consistency. Any issue picked was addressed in the data cleaning step below.

Data Cleaning

1. There were no duplicates.
2. Completeness:

- Any treatment offered before admission had missing admission ID(*ADMISSION.HADM_ID*). This was inferred from the patient id (*ADMISSION.SUBJECT_ID*) and treatment period.
- The language(*ADMISSION.LANGUAGE*) and ethnicity (*ADMISSION.ETHNICITY*). These were replaced with "UNKNOWN"

3. Consistency:

- The diagnosis(*ADMISSION.DIAGNOSIS*) provides a preliminary, free text diagnosis for the patient on hospital admission as assigned by the admitting clinician and does not use a systematic ontology. Data cleaning was done by removing unnecessary text and ensuring common acronyms and abbreviations referred to the same diagnosis.
- The Loinc table was used to identify the actual test.

Categorical variable encoding

1. Categorical variables with high cardinality: diagnoses, medicine, procedures.

- This target encoding/hash encoding was used.

2. Categorical variables with less cardinality but need to preserve the variance: gender,

- Frequency encoding was used

Dimension reduction

The final dataset contained x columns.

Dimension reduction was done using

https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29

Final Dataset

Finally the chosen dataset factoring in 90% of the variance was factored in. This was used for the final analysis.

Splitting of machine learning dataset

These are the dataset(s) produced by the data preparation phase, used for modeling or for the major analysis work of the project.

Construction of event logs

Modeling

Select Modeling Techniques

As the first step in modeling, select the actual initial modeling technique. If multiple techniques are to be applied, perform this task separately for each technique.

Remember that not all tools and techniques are applicable to each and every task. For certain problems, only some techniques are appropriate (See Appendix 2, where techniques appropriate for certain data mining problem types are discussed in more detail). “Political requirements” and other constraints further limit the choices available to the data mining engineer. It may be that only one tool or technique is available to solve the problem at hand—and that the tool may not be absolutely the best, from a technical standpoint.

Modeling Technique

Record the actual modeling technique that is used.

Decide on appropriate technique for exercise, bearing in mind the tool selected.

Modeling Assumptions

Many modeling techniques make specific assumptions about the data.

- Define any built-in assumptions made by the technique about the data (e.g., quality, format, distribution)
- Compare these assumptions with those in the Data Description Report
- Make sure that these assumptions hold and go back to the Data Preparation Phase, if necessary

Generate Test Design

Prior to building a model, it is necessary to define a procedure to test the model’s quality and validity. For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore, the test design specifies that the dataset should be separated into training and test sets. The model is built on the training set and its quality estimated on the test set.

Test Design

Describe the intended plan for training, testing, and evaluating the models. A primary component of the plan is to decide how to divide the available dataset into training data, test data, and validation test sets.

- Check existing test designs for each data mining goal separately
- Decide on necessary steps (number of iterations, number of folds, etc.)
- Prepare data required for test

Build Model

Run the modeling tool on the prepared dataset to create one or more models.

Parameter Settings

With any modeling tool, there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice.

- Set initial parameters
- Document reasons for choosing those values

Models

Run the modeling tool on the prepared dataset to create one or more models.

- Run the selected technique on the input dataset to produce the model
- Post-process data mining results (e.g., edit rules, display trees)

Model Description

Describe the resulting model and assess its expected accuracy, robustness, and possible shortcomings. Report on the interpretation of the models and any difficulties encountered.

- Describe any characteristics of the current model that may be useful for the future
- Record parameter settings used to produce the model
- Give a detailed description of the model and any special features
- For rule-based models, list the rules produced, plus any assessment of per-rule or overall model accuracy and coverage
- For opaque models, list any technical information about the model (such as neural network topology) and any behavioral descriptions produced by the modeling process (such as accuracy or sensitivity)
- Describe the model's behavior and interpretation
- State conclusions regarding patterns in the data (if any); sometimes the model reveals important facts about the data without a separate assessment process (e.g., that the output or conclusion is duplicated in one of the inputs)

Assess Model

The model should now be assessed to ensure that it meets the data mining success criteria and passes the desired test criteria. This is a purely technical assessment based on the outcome of the modeling tasks.

Model Assessment

Summarize results of this task, list qualities of generated models (e.g., in terms of accuracy), and rank their quality in relation to each other.

- Evaluate results with respect to evaluation criteria
- Test result according to a test strategy (e.g.: Train and Test, Cross-validation, bootstrapping, etc.)
- Compare evaluation results and interpretation
- Create ranking of results with respect to success and evaluation criteria
- Select best models
- Interpret results in business terms (as far as possible at this stage)
- Get comments on models by domain or data experts
- Check plausibility of model
- Check effect on data mining goal
- Check model against given knowledge base to see if the discovered information is novel and useful
- Check reliability of result
- Analyze potential for deployment of each result
- If there is a verbal description of the generated model (e.g., via rules), assess the rules: Are they logical, are they feasible, are there too many or too few, do they offend common sense?
- Assess results
- Get insights into why a certain modeling technique and certain parameter settings lead to good/bad results

"Lift Tables" and "Gain Tables" can be constructed to determine how well the model is predicting.

Revised Parameter Settings

According to the model assessment, revise parameter settings and tune them for the next run in the Build Model task. Iterate model building and assessment until you find the best model.

Adjust parameters to produce better models.

Evaluation

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives, and seeks to determine if there is some business reason why this model is deficient. It compares results with the evaluation criteria defined at the start of the project.

A good way of defining the total outputs of a data mining project is to use the equation:

$$RESULTS = MODELS + FINDINGS$$

In this equation, we are defining that the total output of the data mining project is not just the models (although they are, of course, important) but also the findings, which we define as anything (apart from the model) that is important in meeting the objectives of the business or important in leading to new questions, lines of approach, or side effects (e.g., data quality problems uncovered by the data mining exercise). Note: Although the model is directly connected to the business questions, the findings need not be related to any questions or objectives, as long as they are important to the initiator of the project.

Evaluate Results

This step assesses the degree to which the model meets the business objectives, and seeks to determine if there is some business reason why this model is deficient. Another option is to test the model(s) on test applications in the real application, if time and budget constraints permit.

Moreover, evaluation also assesses other generated data mining results. Data mining results cover models that are related to the original business objectives and all other findings. Some are related to the original business objectives while others might unveil additional challenges, information, or hints for future directions.

Assessment of Data Mining Results w.r.t. Business Success Criteria

Summarize assessment results in terms of business success criteria, including a final statement related to whether the project already meets the initial business objectives.

- Understand the data mining results
- Interpret the results in terms of the application
- Check effect on for data mining goal
- Check the data mining result against the given knowledge base to see if the discovered information is novel and useful
- Evaluate and assess results with respect to business success criteria (i.e., has the project achieved the original Business Objectives)
- Compare evaluation results and interpretation
- Rank results with respect to business success criteria
- Check effect of result on initial application goal
- Determine if there are new business objectives to be addressed later in the project, or in new projects
- State recommendations for future data mining projects

Approved Models

After accessing models with respect to business success criteria, select and approve the generated models that meet the selected criteria.

Review Process

At this point, the resulting model appears to be satisfactory and appears to satisfy business needs. It is now appropriate to make a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked. At this stage of the data mining exercise, the Process Review takes the form of a Quality Assurance Review.

Review of Process

Summarize the process review and list activities that have been missed and/or should be repeated.

- Provide an overview of the data mining process used
- Analyze the data mining process. For each stage of the process ask:
  - Was it necessary?
  - Was it executed optimally?
  - In what ways could it be improved?
- Identify failures
- Identify misleading steps
- Identify possible alternative actions and/or unexpected paths in the process
- Review data mining results with respect to business success criteria

Determine Next Steps

Based on the assessment results and the process review, the project team decides how to proceed. Decisions to be made include whether to finish this project and move on to deployment, to initiate further iterations, or to set up new data mining projects.

List of Possible Actions

List possible further actions along with the reasons for and against each option.

- Analyze the potential for deployment of each result
- Estimate potential for improvement of current process
- Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available)
- Recommend alternative continuations
- Refine process plan

Decision

Describe the decisions made, along with the rationale for them.

- Rank the possible actions
- Select one of the possible actions
- Document reasons for the choice

Deployment

Plan Deployment

This task starts with the evaluation results and concludes with a strategy for deployment of the data mining result(s) into the business.

Deployment Plan

Summarize the deployment strategy, including necessary steps and how to perform them.

- Summarize deployable results

- Develop and evaluate alternative plans for deployment
- Decide for each distinct knowledge or information result
- Determine how knowledge or information will be propagated to users
- Decide how the use of the result will be monitored and its benefits measured (where applicable)
- Decide for each deployable model or software result
- Establish how the model or software result will be deployed within the organization's systems
- Determine how its use will be monitored and its benefits measured (where applicable)
- Identify possible problems during deployment (pitfalls to be avoided)

Plan Monitoring and Maintenance

Monitoring and maintenance are important issues if the data mining results become part of the day-to-day business and its environment. A careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. In order to monitor the deployment of the data mining result(s), the project needs a detailed plan for monitoring and maintenance. This plan takes into account the specific type of deployment.

Monitoring and Maintenance Plan

Summarize monitoring and maintenance strategy, including necessary steps and how to perform them.

- Check for dynamic aspects (i.e., what things could change in the environment?)
- Decide how accuracy will be monitored
- Determine when the data mining result or model should not be used any more. Identify criteria (validity, threshold of accuracy, new data, change in the application domain, etc.), and what should happen if the model or result could no longer be used. (update model, set up new data mining project, etc.).
- Will the business objectives of the use of the model change over time? Fully document the initial problem the model was attempting to solve.
- Develop monitoring and maintenance plan.

Produce Final Report

At the end of the project, the project team writes up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experience, or a final presentation of the data mining result(s).

Final Report

At the end of the project, there will be at least one final report in which all the threads are brought together. As well as identifying the results obtained, the report should also describe the process, show which costs have been incurred, define any deviations from the original plan, describe implementation plans, and make any recommendations for future work. The actual detailed content of the report depends very much on the intended audience.

- Identify what reports are needed (slide presentation, management summary, detailed findings, explanation of models, etc.)
- Analyze how well initial data mining goals have been met
- Identify target groups for report
- Outline structure and contents of report(s)
- Select findings to be included in the reports
- Write a report

Final Presentation

As well as a final report, it may be necessary to make a final presentation to summarize the project – maybe to the management sponsor, for example. The presentation normally contains a subset of the information contained in the final report, structured in a different way.

- Decide on target group for the final presentation and determine if they will already have received the final report
- Select which items from the final report should be included in final presentation

Review Project

Assess what went right and what went wrong, what was done well, and what needs to be improved.

Experience Documentation

Summarize important experience gained during the project. For example, pitfalls, misleading approaches, or tips for selecting the best-suited data mining techniques in similar situations could be part of this documentation. In ideal projects, experience documentation also covers any reports that have been written by individual project members during the project.

- Interview all significant people involved in the project and ask them about their experience during the project
- If end users in the business work with the data mining result(s), interview them: Are they satisfied? What could have been done better? Do they need additional support?
- Summarize feedback and write the experience documentation
- Analyze the process (things that worked well, mistakes made, lessons learned, etc.)
- Document the specific data mining process (How can the results and the experience of applying the model be fed back into the process?)
- Generalize from the details to make the experience useful for future projects

Appendix

```r
con <- DBI::dbConnect(RPostgreSQL::PostgreSQL(),
  host = "AWS end point",
  user = "eKagereki",
  password = rstudioapi::askForPassword("Database password")
)

data <- tbl(con, "ADMISSIONS")
```

References

ACLS. 2020. "Acute Coronary Syndromes Algorithm - ACLS Version Control: This Document Follows 2020 American Heart Association Guidelines for CPR and ECC. American Heart Association Guidelines Are Updated Every Five Years." https://www.acls.net/images/algo-acs.pdf.

Eagle KA Lim MJ, et al., Dabbous OH. 2004. "A Validated Prediction Model for All Forms of Acute Coronary Syndrome: Estimating the Risk of 6-Month Postdischarge Death in an International Registry." *Jama*. https://doi.org/doi:10.1001/jama.291.22.2727.

Institute, Regenstrief. 2021. "LOINC." *LOINC*. https://loinc.org/downloads/.

Johnson, et al, A. 2016. "MIMIC-III Clinical Database." https://doi.org/https://doi.org/10.13026/C2XW26.

Roth, et al, Gregory A. 2020. "Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update from the GBD 2019 Study." *Journal of the American College of Cardiology* 76 (25): 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010.