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

Edwin Kagereki - B00867154

16 October 2021

Contents

Business Understanding

Business Objectives

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; Roth 2020) This project aims at predicting the mortality of patients admitted in ICU with suspected cardiovascular events using data from patients who stayed in critical care units.

Background

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

Organization The project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived because the project did not impact clinical care and all protected health information was deidentified.

Problem area

- Identify the problem area (e.g., marketing, customer care, business development, etc.)
- Describe the problem in general terms
- Check the current status of the project (e.g., Check if it is already clear within the business unit that a data mining project is to be performed, or whether data mining needs to be promoted as a key technology in the business)
- Clarify prerequisites of the project (e.g., What is the motivation of the project? Does the business already use data mining?)
- If necessary, prepare presentations and present data mining to the business
- Identify target groups for the project result (e.g., Are we expected to deliver a report for top management or an operational system to be used by naive end users?)
- Identify the users' needs and expectations

Current solution

- Describe any solution currently used to address the problem
- Describe the advantages and disadvantages of the current solution and the level to which it is accepted by the users

Business Objectives

Describe the customer's primary objective, from a business perspective. In addition to the primary business objective, there are typically a large number of related business questions that the customer would like to address. For example, the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor, while a secondary business objective might be to determine whether lower fees affect only one particular segment of customers.

- Informally describe the problem to be solved
- Specify all business questions as precisely as possible
- Specify any other business requirements (e.g., the business does not want to lose any customers)
- Specify expected benefits in business terms
- **Beware of setting unattainable goals — make them as realistic as possible.**

Business Success Criteria

Describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific and readily measurable, such as reduction of customer churn to a certain level, or general and subjective, such as "give useful insights into the relationships." In the latter case, be sure to indicate who would make the subjective judgment.

- Specify business success criteria (e.g., Improve response rate in a mailing campaign by 10 percent and sign-up rate by 20 percent)
- Identify who assesses the success criteria

Each of the success criteria should relate to at least one of the specified business objectives.

Before starting the situation assessment, you might analyze previous experiences of this problem — either internally, using CRISP-DM, or externally, using pre-packaged solutions.

Assess Situation

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and in developing the project plan.

Inventory of Resources

List the resources available to the project, including personnel (business and data experts, technical support, data mining experts), data (fixed extracts, access to live warehoused or operational data), computing resources (hardware platforms), and software (data mining tools, other relevant software).

Hardware Resources

- Identify the base hardware
- Establish the availability of the base hardware for the data mining project
- Check if the hardware maintenance schedule conflicts with the availability of the hardware for the data mining project
- Identify the hardware available for the data mining tool to be used (if the tool is known at this stage)

Sources of Data and Knowledge

MIMIC III Database MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with x patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

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 (DUA). Adherence to the terms of the DUA is paramount.

LOIC tables The Loin

American Heart Association® guidelines for CPR and ECC (2020) This document provided know

Personnel Sources

- Identify project sponsor (if different from internal sponsor as in Organization)
- Identify system administrator, database administrator, and technical support staff for further questions
- Identify market analysts, data mining experts, and statisticians, and check their availability
- Check availability of domain experts for later phases

Remember that the project may need technical staff at odd times throughout the project, for example during data transformation.

Requirements, Assumptions, and Constraints

List all requirements of the project, including schedule of completion, comprehensibility, and quality of results and security, as well as legal issues. As part of this output, make sure that you are allowed to use the data.

List the assumptions made by the project. These may be assumptions about the data, which can be verified during data mining, but may also include non-verifiable assumptions related to the project. It is particularly important to list the latter if they will affect the validity of the results.

List the constraints made on the project. These constraints might involve lack of resources to carry out some of the tasks in the project in the time required, or there may be legal or ethical constraints on the use of the data or the solution needed to carry out the data mining task.

Requirements

- Specify target group profile
- Capture all requirements on scheduling
- Capture requirements on comprehensibility, accuracy, deploy ability, maintainability, and repeatability of the data mining project and the resulting model(s)
- Capture requirements on security, legal restrictions, privacy, reporting, and project schedule

Assumptions

- Clarify all assumptions (including implicit ones) and make them explicit (e.g., to address the business question, a minimum number of customers with age above 50 is necessary)
- List assumptions on data quality (e.g., accuracy, availability)
- List assumptions on external factors (e.g., economic issues, competitive products, technical advances)
- Clarify assumptions that lead to any of the estimates (e.g., the price of a specific tool is assumed to be lower than $1,000)
- List all assumptions regarding whether it is necessary to understand and describe or explain the model (e.g., how should the model and results be presented to senior management/sponsor)

The list of assumptions also includes assumptions at the beginning of the project, i.e., what the starting point of the project has been.

Constraints

- Check general constraints (e.g., legal issues, budget, timescales, and resources)
- Check access rights to data sources (e.g., access restrictions, password required)
- Check technical accessibility of data (operating systems, data management system, file or database format)
- Check whether relevant knowledge is accessible
- Check budget constraints (fixed costs, implementation costs, etc.)

Risks and Contingencies

List the risks, that is, the events that might occur, impacting schedule, cost, or result. List the corresponding contingency plans: what action will be taken to avoid or minimize the impact or recover from the occurrence

of the foreseen risks.

Identify Risks

- Identify business risks (e.g., competitor comes up with better results first)
- Identify organizational risks (e.g., department requesting project doesn't have funding for the project)
- Identify financial risks (e.g., further funding depends on initial data mining results)
- Identify technical risks
- Identify risks that depend on data and data sources (e.g., poor quality and coverage)

Develop Contingency Plans

- Determine conditions under which each risk may occur
- Develop contingency plans

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

Costs and Benefits

Prepare a cost-benefit analysis for the project, comparing the costs of the project with the potential benefits to the business if it is successful.

- Estimate costs for data collection
- Estimate costs of developing and implementing a solution
- Identify benefits (e.g., improved customer satisfaction, ROI, and increase in revenue)
- Estimate operating costs

The comparison should be as specific as possible, as this enables a better business case to be made.

Remember to identify hidden costs, such as repeated data extraction and preparation, changes in workflows, and time required for training.

Determine 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

Describe the intended outputs of the project that enable the achievement of the business objectives. Note that these are normally technical outputs.

- Translate the business questions to data mining goals (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).
- Specify data mining problem type (e.g., classification, description, prediction, and clustering). For more details about data mining problem types, see Appendix 2.

It may be wise to re-define the problem. For example, modeling product retention rather than customer retention when targeting customer retention delivers results too late to affect the outcome.

Data Mining Success Criteria

Define the criteria for a successful outcome to the project in technical terms, for example a certain level of predictive accuracy or a propensity-to-purchase profile with a given degree of "lift." As with business success criteria, it may be necessary to describe these in subjective terms, in which case the person or persons making the subjective judgment should be identified.

- Specify criteria for model assessment (e.g., model accuracy, performance and complexity)
- Define benchmarks for evaluation criteria
- Specify criteria which address subjective assessment criteria (e.g., model explain ability and data and marketing insight provided by the model)

Remember that the data mining success criteria are different than the business success criteria defined earlier.

Remember it is wise to plan for deployment from the start of the project.

Produce Project Plan

Describe the intended plan for achieving the data mining goals and thereby achieving the business goals.

Project Plan

List the stages to be executed in the project, together with their duration, resources required, inputs, outputs, and dependencies. Wherever possible, make explicit the large-scale iterations in the data mining process— for example, repetitions of the modeling and evaluation phases. As part of the project plan, it is also important to analyze dependencies between time schedule and risks. Mark results of these analyses explicitly in the project plan, ideally with actions and recommendations for actions if the risks are manifested.

Although this is the only task in which the project plan is directly named, it nevertheless should be consulted continually and reviewed throughout the project. The project plan should be consulted at minimum whenever a new task is started or a further iteration of a task or activity is begun.

- Define the initial process plan and discuss the feasibility with all involved personnel
- Combine all identified goals and selected techniques in a coherent procedure that solves the business questions and meets the business success criteria
- Estimate the effort and resources needed to achieve and deploy the solution. (It is useful to consider other people's experience when estimating timescales for data mining projects. For example, it is often postulated that 50-70 percent of the time and effort in a data mining project is used in the Data Preparation Phase and 20-30 percent in the Data Understanding Phase, while only 10-20 percent is spent in each of the Modeling, Evaluation, and Business Understanding Phases and 5-10 percent in the Deployment Phase.)
- Identify critical steps
- Mark decision points
- Mark review points
- Identify major iterations

Initial Assessment of Tools and Techniques

At the end of the first phase, the project team performs an initial assessment of tools and techniques. Here, it is important to select a data mining tool that supports various methods for different stages of the process, since the selection of tools and techniques may influence the entire project.

- Create a list of selection criteria for tools and techniques (or use an existing one if available)
- Choose potential tools and techniques
- Evaluate appropriateness of techniques
- Review and prioritize applicable techniques according to the evaluation of alternative solutions

Data Understanding

Collect Initial Data

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.

Initial Data Collection Report

Describe all the various data used for the project, and include any selection requirements for more detailed data. The data collection report should also define whether some attributes are relatively more important than others.

Remember that any assessment of data quality should be made not just of the individual data sources but also of any data that results from merging data sources. Because of inconsistencies between the sources, merged data may present problems that do not exist in the individual data sources.

Data requirements planning

- Plan which information is needed (e.g., only for given attributes, or specific additional information)
- Check if all the information needed (to solve the data mining goals) is actually available

Selection criteria

- Specify selection criteria (e.g., Which attributes are necessary for the specified data mining goals? Which attributes have been identified as being irrelevant? How many attributes can we handle with the chosen techniques?)
- Select tables/files of interest
- Select data within a table/file
- Think about how long a history one should use (e.g., even if 18 months of data are available, only 12 months may be needed for the exercise)

Be aware that data collected from different sources may give rise to quality problems when merged (e.g., address files merged with a customer database may show inconsistencies of format, invalidity of data, etc.).

Insertion of data

- If the data contain free text entries, do we need to encode them for modeling or do we want to group specific entries?
- How can missing attributes be acquired?
- How can we best extract the data?

Remember that some knowledge about the data may be available from non-electronic sources (e.g., from people, printed text, etc.).

Remember that it may be necessary to preprocess the data (time-series data, weighted averages, etc.).

Describe Data

Examine the "gross" properties of the acquired data and report on the results.

Data Description Report

Describe the data that has been acquired, including the format of the data, the quantity of the data (e.g., the number of records and fields within each table), the identities of the fields, and any other surface features that have been discovered.

Volumetric analysis of data

- Identify data and method of capture
- Access data sources
- Use statistical analyses if appropriate
- Report tables and their relations
- Check data volume, number of multiples, complexity
- Note if the data contain free text entries

Attribute types and values

- Check accessibility and availability of attributes
- Check attribute types (numeric, symbolic, taxonomy, etc.)
- Check attribute value ranges
- Analyze attribute correlations
- Understand the meaning of each attribute and attribute value in business terms
- For each attribute, compute basic statistics (e.g., compute distribution, average, max, min, standard deviation, variance, mode, skewness, etc.)
- Analyze basic statistics and relate the results to their meaning in business terms
- Decide if the attribute is relevant for the specific data mining goal
- Determine if the attribute meaning is used consistently
- Interview domain experts to obtain their opinion of attribute relevance
- Decide if it is necessary to balance the data (based on the modeling techniques to be used)

Keys

- Analyze key relationships
- Check amount of overlaps of key attribute values across tables

Review assumptions/goals

- Update list of assumptions, if necessary

Explore Data

This task tackles the data mining questions that can be addressed using querying, visualization, and reporting techniques. These analyses may directly address the data mining goals. However, they may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed before further analysis can occur.

Data Exploration Report

Describe the results of this task, including first findings or initial hypotheses and their impact on the remainder of the project. The report may also include graphs and plots that indicate data characteristics or point to interesting data subsets worthy of further examination.

Data exploration

- Analyze properties of interesting attributes in detail (e.g., basic statistics, interesting sub-populations)
- Identify characteristics of sub-populations

Form suppositions for future analysis

- Consider and evaluate information and findings in the data descriptions report
- Form a hypothesis and identify actions
- Transform the hypothesis into a data mining goal, if possible
- Clarify data mining goals or make them more precise. A "blind" search is not necessarily useless, but a more directed search toward business objectives is preferable.
- Perform basic analysis to verify the hypothesis

Verify Data Quality

Examine the quality of the data, addressing questions such as: Is the data complete (does it cover all the cases required)? Is it correct or does it contain errors? If there are errors, how common are they? Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they?

Data Quality Report

List the results of the data quality verification; if there are quality problems, list possible solutions.

- Identify special values and catalog their meaning

Review keys, attributes

- Check coverage (e.g., whether all possible values are represented)
- Check keys
- Verify that the meanings of attributes and contained values fit together
- Identify missing attributes and blank fields
- Establish the meaning of missing data
- Check for attributes with different values that have similar meanings (e.g., low fat, diet)
- Check spelling and format of values (e.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter)
- Check for deviations, and decide whether a deviation is "noise" or may indicate an interesting phenomenon
- Check for plausibility of values, (e.g., all fields having the same or nearly the same values)

Review any attributes that give answers that conflict with common sense (e.g., teenagers with high income levels).

Use visualization plots, histograms, etc. to reveal inconsistencies in the data.

Data quality in flat files

- If data are stored in flat files, check which delimiter is used and whether it is used consistently within all attributes
- If data are stored in flat files, check the number of fields in each record to see if they coincide

Noise and inconsistencies between sources

- Check consistencies and redundancies between different sources
- Plan for dealing with noise
- Detect the type of noise and which attributes are affected

Remember that it may be necessary to exclude some data since they do not exhibit either positive or negative behavior (e.g., to check on customers' loan behavior, exclude all those who have never borrowed, do not finance a home mortgage, those whose mortgage is nearing maturity, etc.).

Review whether assumptions are valid or not, given the current information on data and business knowledge.

Data Preparation

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

Dataset description: This is the description of the dataset(s) used for the modeling or for the major analysis work of the project.

Select Data

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.

Rationale for Inclusion/Exclusion

List the data to be used/excluded and the reasons for these decisions.

- Collect appropriate additional data (from different sources—in-house as well as externally)
- Perform significance and correlation tests to decide if fields should be included
- Reconsider Data Selection Criteria in light of experiences of data quality and data exploration (i.e., may wish include/exclude other sets of data)
- Reconsider Data Selection Criteria in light of experience of modeling (i.e., model assessment may show that other datasets are needed)
- Select different data subsets (e.g., different attributes, only data which meet certain conditions)
- Consider the use of sampling techniques (e.g., A quick solution may involve splitting test and training datasets or reducing the size of the test dataset, if the tool cannot handle the full dataset. It may also be useful to have weighted samples to give different importance to different attributes or different values of the same attribute.)
- Document the rationale for inclusion/exclusion
- Check available techniques for sampling data

Based on Data Selection Criteria, decide if one or more attributes are more important than others and weight the attributes accordingly. Decide, based on the context (i.e., application, tool, etc.), how to handle the weighting.

Clean Data

Raise the data quality to the level required by the selected analysis techniques. This may involve the selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling.

Data Cleaning Report

Describe the decisions and actions that were taken to address the data quality problems reported during the Verify Data Quality Task. If the data are to be used in the data mining exercise, the report should address outstanding data quality issues and what possible effect this could have on the results.

- Reconsider how to deal with any observed type of noise
- Correct, remove, or ignore noise

- Decide how to deal with special values and their meaning. The area of special values can give rise to many strange results and should be carefully examined. Examples of special values could arise through taking results of a survey where some questions were not asked or not answered. This might result in a value of 99 for unknown data. For example, 99 for marital status or political affiliation. Special values could also arise when data is truncated—e.g., 00 for 100-year-old people or all cars with 100,000 km on the odometer.
- Reconsider Data Selection Criteria in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data).

Remember that some fields may be irrelevant to the data mining goals and, therefore, noise in those fields has no significance. However, if noise is ignored for these reasons, it should be fully documented as the circumstances may change later.

Construct Data

This task includes constructive data preparation operations such as the production of derived attributes, complete new records, or transformed values for existing attributes.

- Check available construction mechanisms with the list of tools suggested for the project
- Decide whether it is best to perform the construction inside the tool or outside (i.e., which is more efficient, exact, repeatable)
- Reconsider Data Selection Criteria Data Selection Criteria in light of experiences of data construction (i.e., you may wish include/exclude other sets of data)

Derived Attributes

Derived attributes are new attributes that are constructed from one or more existing attributes in the same record. An example might be: area = length * width.

Why should we need to construct derived attributes during the course of a data mining investigation? It should not be thought that only data from databases or other sources should be used in constructing a model. Derived attributes might be constructed because:

- Background knowledge convinces us that some fact is important and ought to be represented although we have no attribute currently to represent it
- The modeling algorithm in use handles only certain types of data—for example we are using linear regression and we suspect that there are certain non-linearities that will be not be included in the model
- The outcome of the modeling phase suggests that certain facts are not being covered

Derived attributes

- Decide if any attribute should be normalized (e.g., when using a clustering algorithm with age and income, in certain currencies, the income will dominate)
- Consider adding new information on the relevant importance of attributes by adding new attributes (for example, attribute weights, weighted normalization)
- How can missing attributes be constructed or imputed? [Decide type of construction (e.g., aggregate, average, induction).]
- Add new attributes to the accessed data

Before adding Derived Attributes, try to determine if and how they ease the model process or facilitate the modeling algorithm. Perhaps "income per person" is a better/easier attribute to use than "income per household." Do not derive attributes simply to reduce the number of input attributes.

Another type of derived attribute is the single-attribute transformation, usually performed to fit the needs of the modeling tools.

Single-attribute transformations

- Specify necessary transformation steps in terms of available transformation facilities (for example, change a binning of a numeric attribute)
- Perform transformation steps

Transformations may be necessary to change ranges to symbolic fields (e.g., ages to age ranges) or symbolic fields ("definitely yes," "yes," "don't know," "no") to numeric values. Modeling tools or algorithms often require them.

Generated Records

Generated records are completely new records, which add new knowledge or represent new data that is not otherwise represented (e.g., having segmented the data, it may be useful to generate a record to represent the prototypical member of each segment for further processing).

Check for available techniques if needed (e.g., mechanisms to construct prototypes for each segment of segmented data).

Integrate Data

These are methods for combining information from multiple tables or other information sources to create new records or values.

Merged Data

Merging tables refers to joining together two or more tables that have different information about the same objects. At this stage, it may also be advisable to generate new records. It may also be recommended to generate aggregate values.

Aggregation refers to operations where new values are computed by summarizing information from multiple records and/or tables.

- Check if integration facilities are able to integrate the input sources as required
- Integrate sources and store results
- Reconsider Data Selection Criteria in light of experiences of data integration (i.e., you may wish to include/exclude other sets of data)

Remember that some knowledge may be contained in non-electronic format.

Format Data

Formatting transformations refers primarily to syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool.

Reformatted Data

Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Rearranging attributes Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

Reordering records It might be important to change the order of the records in the dataset. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute.

Reformatted within-value

- These are purely syntactic changes made to satisfy the requirements of the specific modeling tool
- Reconsider Data Selection Criteria in light of experiences of data cleaning (i.e., you may wish to include/exclude other sets of data)

Dataset

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

Dataset Description

This is the description of the dataset(s) used for the modeling or for the major analysis work of the project.

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

References

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.