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

Edwin Kagereki - B00867154

19 November 2021

Contents

Business Understanding

Introduction

Cardiovascular diseases (CVDs), principally ischemic heart disease (IHD) and cardiac stroke, are the leading cause of mortality globally and are often associated with poor survival(Roth 2020).

Acute Coronary Syndrome (ACS) is a term given to diverse presentations related to cardic myopathies of quick onset. Accurate estimation of risk for untoward outcomes after a suspected onset of an ACS may help clinicians chose the type and intensity of therapy. For example, patients predicted to be at higher risk may

receive more aggressive surveillance and/or 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 using multiple linked ICU patients' datasets. Subsequently, the treatment pathway was used to explain the predicted outcome.

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).Although not set in stone, the chances for a patient to have good outcomes 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,""lov*

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

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

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.

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.

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.

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:
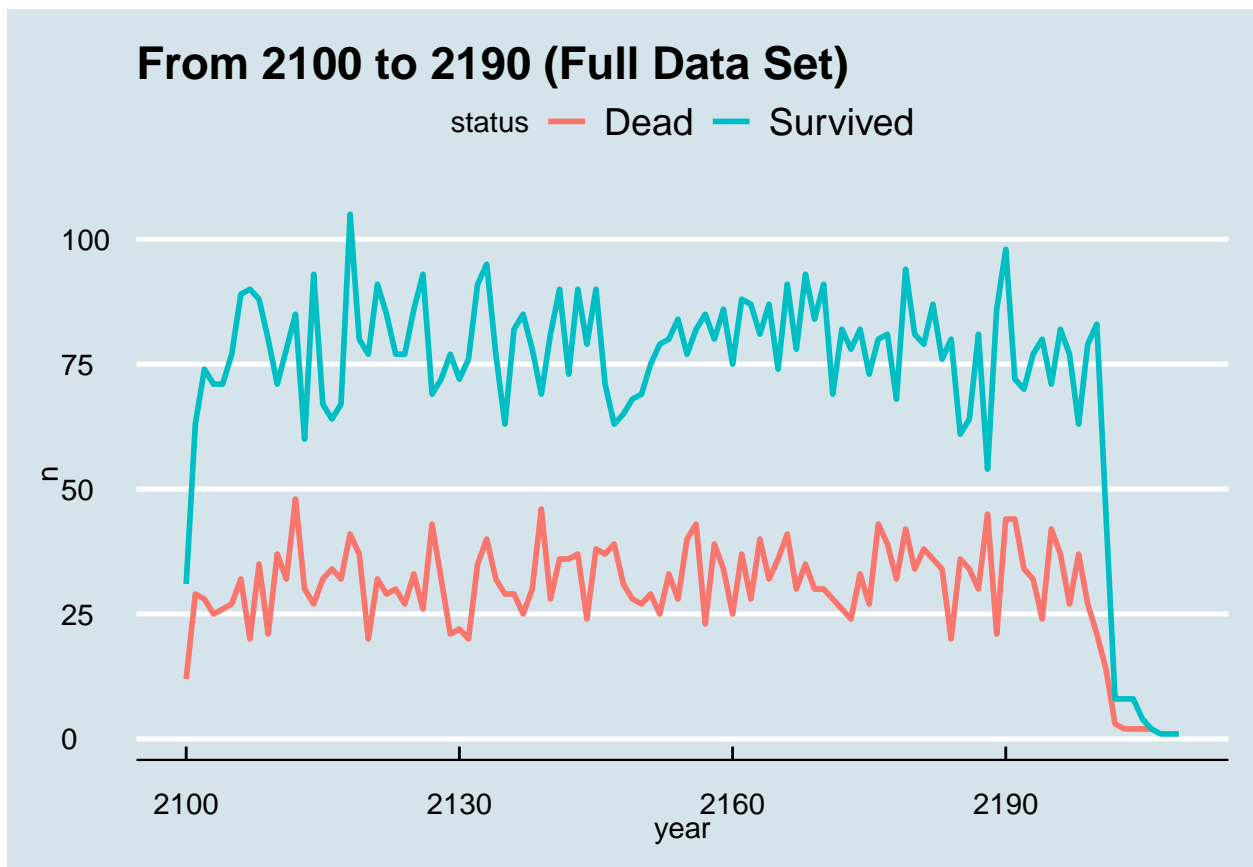
Volumetric analysis of data

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 @Table 2.

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.
- 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

Principal componet analysis was used to reduce the dimensionality of the data.

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.

Construction of event logs

Modeling

Data splitting

For the data splitting strategy, 25% of the hospital admissions were reserved to the test set.

The k-fold cross-validation resampling method was used to create 5 different resamples of the training set which were further split into analysis and assessment sets, producing 5 different performance metrics that were then aggregated. In these resampled datasets, 20% of the hospital admissions were allocated to the validation set and 80% of the hospital stays to the training set.
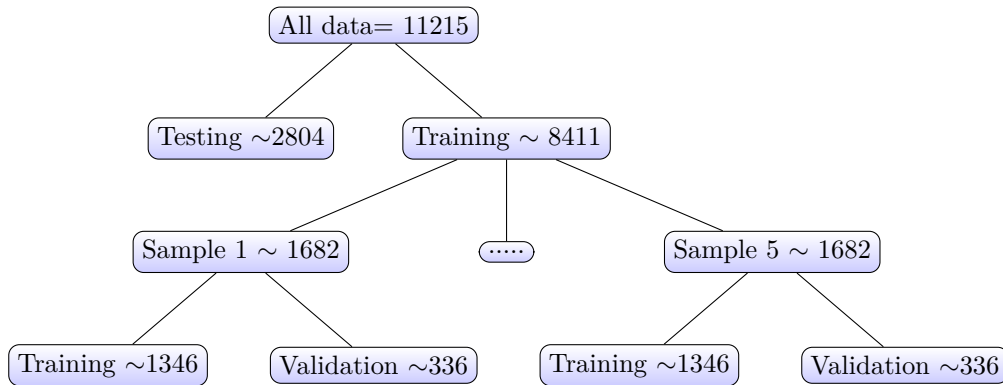

Figure 1: Data splitting algorithm

Model building and tuning

Performance metric and cut-off analysis

In this analysis the F-Beta score was used to identify the model with the highest performance with a score 1.This was meant to:

Increase Precision is a metric that calculates the percentage of correct predictions for the positive class.

Recall calculates the percentage of correct predictions for the positive class out of all positive predictions that could be made.

The selected Beta score of 1 was meant to maximize precision will minimize the false-positive errors, whereas maximizing recall will minimize the false-negative errors.

Finally all the performance metrics of this model were described.

Candidate models   The following candidate models will first be assessed.

- Logistic Regression
- Random Forest,
- XGBoost (extreme gradient boosted trees),
- K-nearest neighbor

Model building and tuning   Grid search algorithm was used to train multiple models simultaneously. We'll also save the validation set predictions (via the call to control_grid()) so that diagnostic information can be available after the model fit. The area under the ROC curve will be used to quantify how well the model performs across a continuum of event thresholds (recall that the event rate—the proportion of stays including children— is very low for these data).

We will use a space-filling design to tune, with 25 candidate models: The random forest is uniformly better across event probability thresholds.

The combination that provides the best performance is the one that you use for your final model. This method is simple to use. You can find the best combination of the values that you provided, and you can run each of the experiments in parallel. However, it's also computationally expensive because so many models are being built. If a hyperparameter isn't important, you might explore different possibilities unnecessarily.
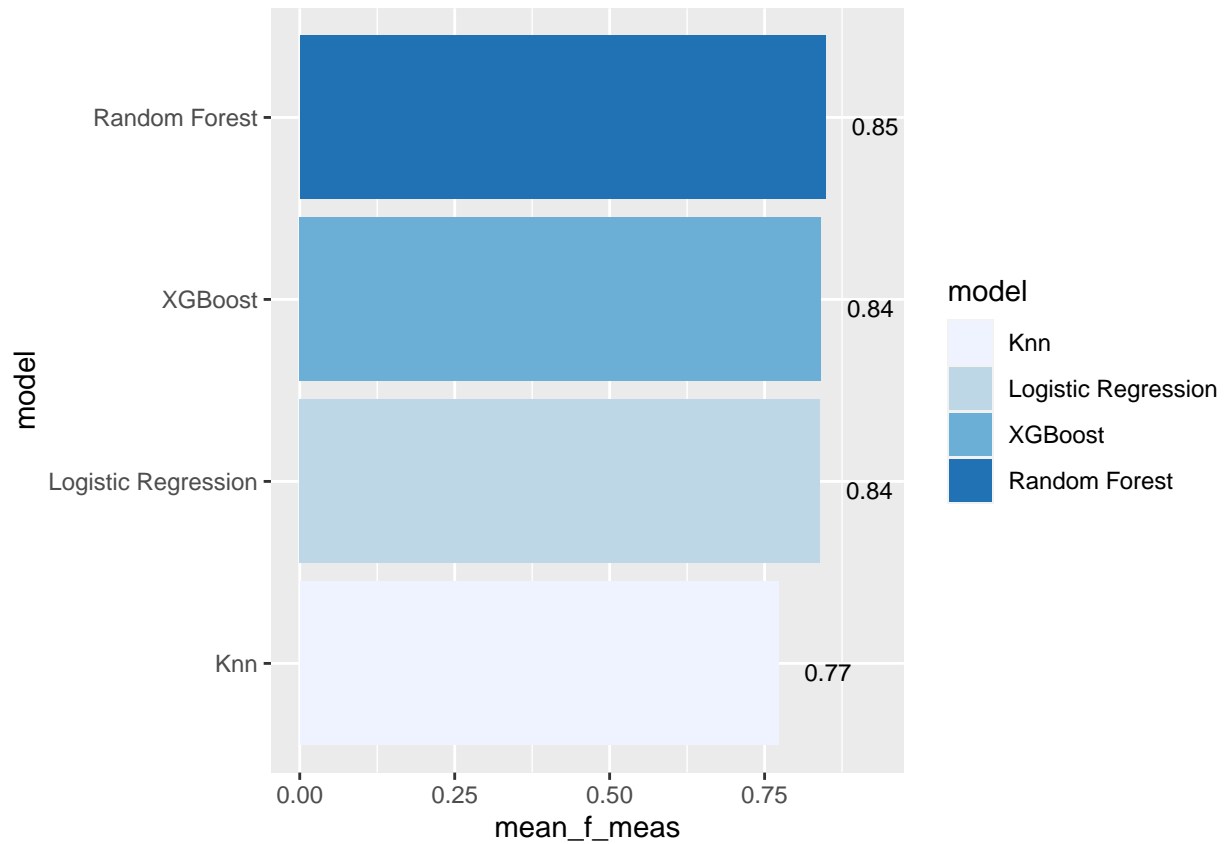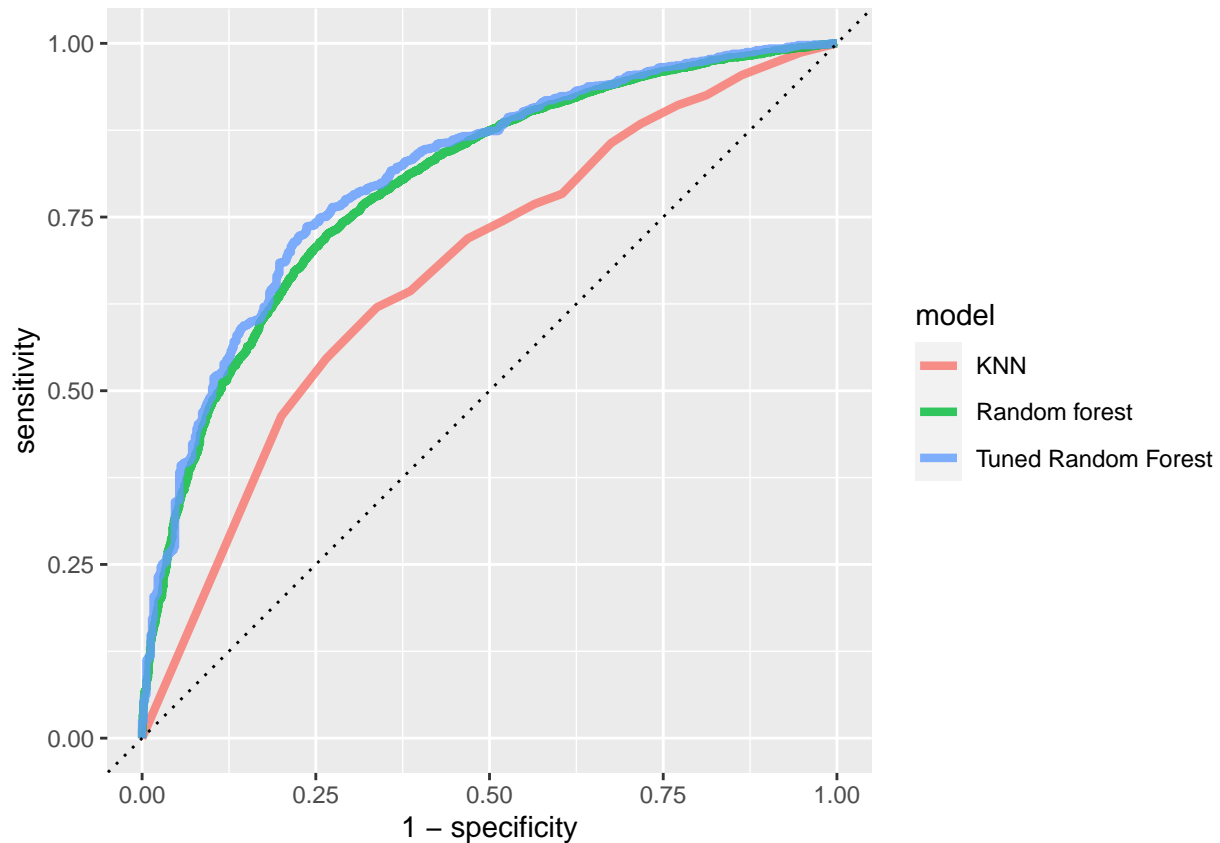
Best model Description

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

Table 2: Attributes and the level of measurement

	variables	types	missing_percent	unique_count	unique_rate
7	AGE	integer	0	84	0.0074900
10	admissionCycle	integer	0	26	0.0023183
12	deadBefore	numeric	0	926	0.0825680
13	dayOfYear	integer	0	366	0.0326349
14	Month	integer	0	12	0.0010700
15	week	integer	0	53	0.0047258
16	weekday	integer	0	7	0.0006242
17	year	integer	0	110	0.0098083
18	hour	integer	0	24	0.0021400
19	FreqGENDER	numeric	0	2	0.0001783
20	FreqADMISSION_TYPE	numeric	0	3	0.0002675
21	FreqADMISSION_LOCATION	numeric	0	6	0.0005350
22	FreqINSURANCE	numeric	0	5	0.0004458
23	FreqRELIGION	numeric	0	20	0.0017833
24	FreqMARITAL_STATUS	numeric	0	8	0.0007133
25	FreqETHNICITY	numeric	0	26	0.0023183
27	FreqLANGUAGE	numeric	0	23	0.0020508

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

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

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

Data attibutes and levels of measurements

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.