Prediction of in-hospital mortality for patients suffering from coronary atherosclerosis using electronic health records

(Shayantan Banerjee, Research scholar, Indian Institute of Technology Madras)

Introduction

India has witnessed an unprecedented increase in the occurrence of heart diseases and stroke in the past 25 years. The percentage of all reported cases has almost doubled during the last two decades. Reliable and detailed estimates of the burden of different cardiovascular diseases requires a time-consuming and robust statistical analysis of the risk factors associated with the disease. With the advent of Electronic Health Records, it has become increasingly easier to conduct exploratory data analysis on this data, build predictive and interpretable models to estimate the burden of different cardiovascular diseases. The results of such an analysis can result in lower readmission rates, reduced burden on hospital resources and reduced mortality.

Objectives

A predictive model will be very helpful in detecting high-risk patients. After a patient is admitted and the blood tests and clinical data (like Insurance, Admission type and Gender) are collected, we will build predictive models to classify survivability of patients within 30 days of admission. This is the most critical period for surviving coronary atherosclerosis.
We considered MIMIC III (Medical Information Mart for Intensive Care III) free hospital database. This database contains de-identified data from over 50,000 patients who were admitted to Beth Israel Deaconess Medical Center in Boston, Massachusetts from 2001 to 2012. In order to get access to the data for this project, you will need to request access at this link.
The main objective of this study was to build five different predictive models using demographic data, treatment data, lab measurements reported within 24 hours of admission, a combined model using all of the above and free text clinical notes. A comparative analysis of the results would then elucidate the best set of features that gives the most predictive power.

Analysis

Data exploration: We used the MIMIC III (Medical Information Mart for Intensive Care III) free hospital database. This database contains de-identified data from over 50,000 patients who were admitted to Beth Israel Deaconess Medical Center in Boston, Massachusetts from 2001 to 2012. In order to get access to the data for this project, you will need to request access at this link https://mimic.physionet.org/gettingstarted/access/).
The following tables were considered for the analysis: ADMISSIONS table, DRGCODES table, PATIENTS table, LABEVENTS table, NOTEEVENTS table and the DIAGNOSIS table for our purposes. A snapshot and description of each of the tables is shown below.

- ADMISSIONS table: this contains the patient admission details like sample id, hospital admission id, admission type, admission location, insurance type, etc.

	ROW_ID	SUBJECT_ID	HADM_ID	ADMITTIME	DISCHTIME	DEATHTIME	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	INSURANCE
0	21	22	165315	2196-04-09 12:26:00	2196-04-10 15:54:00	NaN	EMERGENCY	EMERGENCY ROOM ADMIT	DISC-TRAN CANCER/CHLDRN H	Private
1	22	23	152223	2153-09-03 07:15:00	2153-09-08 19:10:00	NaN	ELECTIVE	PHYS REFERRAL/NORMAL DELI	HOME HEALTH CARE	Medicare
2	23	23	124321	2157-10-18 19:34:00	2157-10-25 14:00:00	NaN	EMERGENCY	TRANSFER FROM HOSP/EXTRAM	HOME HEALTH CARE	Medicare

- PATIENTS table: contain patient details like DOB, hospital stay id etc

	ROW_ID	SUBJECT_ID	GENDER	DOB	DOD	DOD_HOSP	DOD_SSN	EXPIRE_FLAG
0	234	249	F	2075-03-13	NaN	NaN	NaN	0
1	235	250	F	2164-12-27	2188-11-22 00:00:00	2188-11-22 00:00:00	NaN	1
2	236	251	M	2090-03-15	NaN	NaN	NaN	0

- LABEVENTS table: Lab measurements associated with the patients along with the date and time.

	ROW_ID	SUBJECT_ID	HADM_ID	ITEMID	CHARTTIME	VALUE	VALUENUM	VALUEUOM	FLAG
0	281	3	NaN	50820	2101-10-12 16:07:00	7.39	7.39	units	NaN
1	282	3	NaN	50800	2101-10-12 18:17:00	ART	NaN	NaN	NaN
2	283	3	NaN	50802	2101-10-12 18:17:00	-1	-1.00	mEq/L	NaN

- NOTEEVENTS table: This table contains the free text clinical notes entered by the physician.

	ROW_ID	SUBJECT_ID	HADM_ID	CHARTDATE	CHARTTIME	STORETIME	CATEGORY	DESCRIPTION	CGID	ISERROR	TEXT
0	174	22532	167853.0	2151-08-04	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2151-7-16**] Dischar...
1	175	13702	107527.0	2118-06-14	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2118-6-2**] Discharg...
2	176	13702	167118.0	2119-05-25	NaN	NaN	Discharge summary	Report	NaN	NaN	Admission Date: [**2119-5-4**] D...

- DRGCODES table: This table contains the Disease Related Group codes to classify patients into different groups where each group has the same hospital resource use.

	ROW_ID	SUBJECT_ID	HADM_ID	DRG_TYPE	DRG_CODE	DESCRIPTION	DRG_SEVERITY	DRG_MORTALITY
0	342	2491	144486	HCFA	28	TRAUMATIC STUPOR & COMA, COMA <1 HR AGE >17 WI...	NaN	NaN
1	343	24958	162910	HCFA	110	MAJOR CARDIOVASCULAR PROCEDURES WITH COMPLICAT...	NaN	NaN
2	344	18325	153751	HCFA	390	NEONATE WITH OTHER SIGNIFICANT PROBLEMS	NaN	NaN

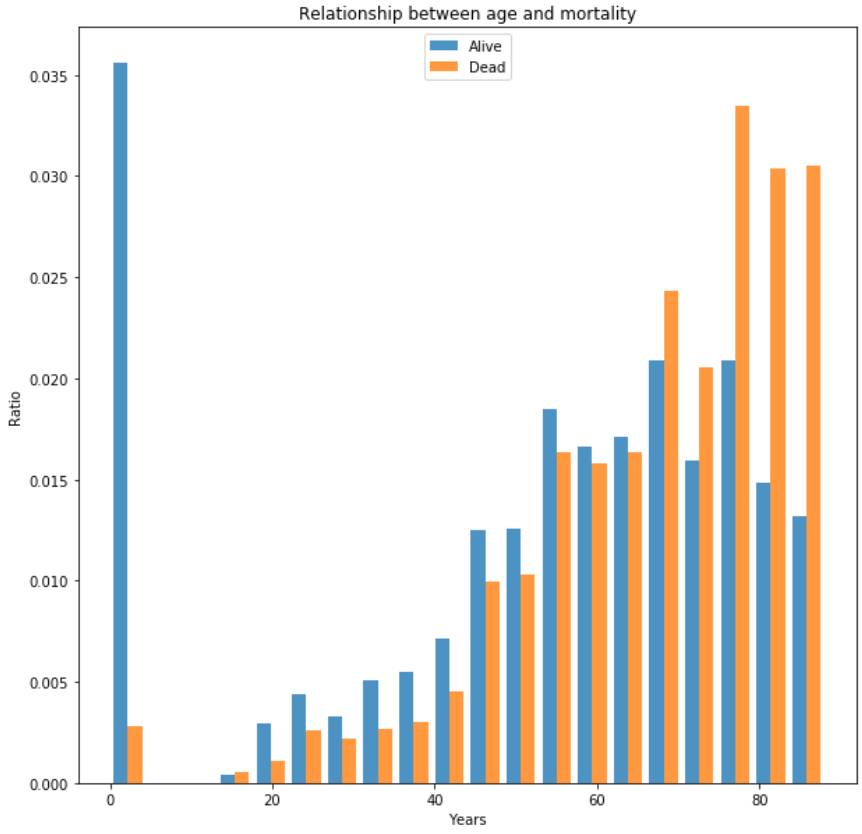
- DIAGNOSIS table: This table contains the diagnosis codes for different patients. A particular patients can have multiple hospital admissions and consequently multiple diagnosis codes.
  We selected only those patients that had a diagnosis of coronary atherosclerosis or and ICD9 code of 41401 reported on the database.

	ROW_ID	SUBJECT_ID	HADM_ID	SEQ_NUM	ICD9_CODE
0	1297	109	172335	1.0	40301
1	1298	109	172335	2.0	486
2	1299	109	172335	3.0	58281

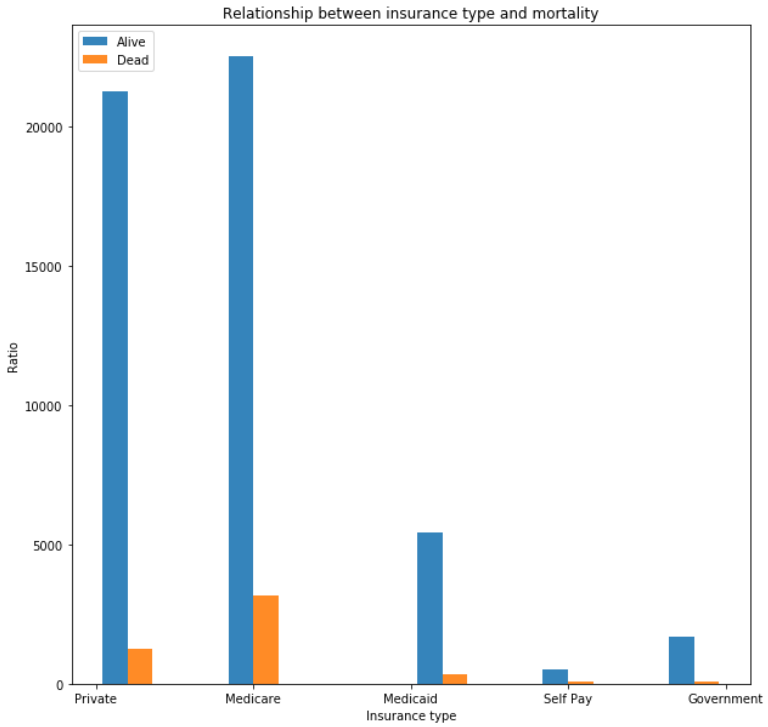
Exploratory data analysis

We considered only discharge notes of patients for further analysis. Mortality period was calculated by subtracting deathtime from admittime. The labels for the dataset were ***0 (those patients who will survive after 30 days) or 1 (those patients who will die within 30 days).*** The risk of a patient dying within 30 days of admission is the dependent variable that we are trying to predict.

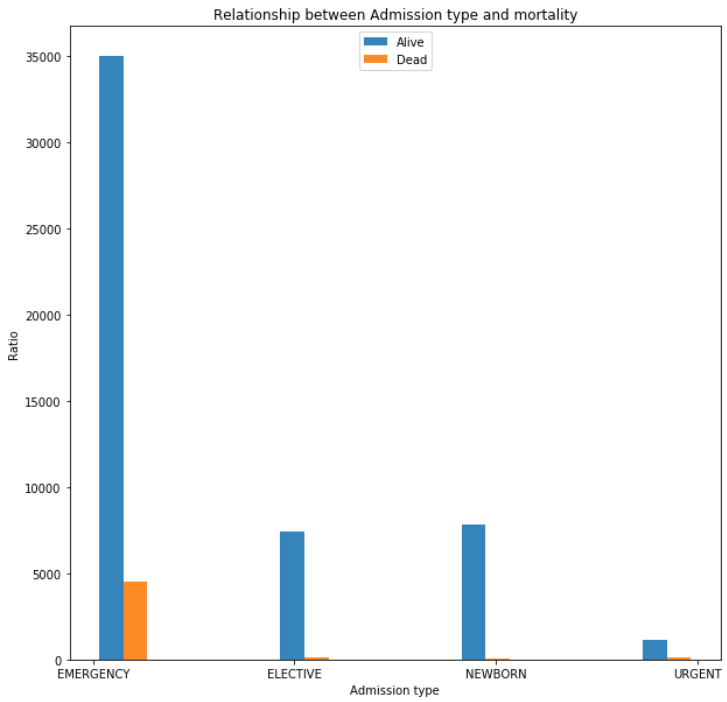
Age was not explicitly mentioned as a feature in any of the tables. So, we constructed it ourselves by subtracting ADMITTIME from DOB. The age or death age of anyone over 89 was corrected as per HIPAA rules. It was done by subtracting 211 years from the values presented. Then we did some basic EDA to understand the structure of our data and get useful insights about the features.



One observation that I made from this graph is that in the given cohort of patients, the incidence of CAD is high among neonates, though overall mortality is low. The highest mortality rate is among the age group 75-80. The gradual increase in the mortality period with age also indicates that this is an ageing disease.
Insurance type was next plotted against mortality and the results are shown below.



This shows that Medicare insurance and Private insurance holders dominate the other insurance categories. Older people are more inclined to have Medicare insurance policies than the new generation who mostly prefer Private insurance policies. Type of admission was the next variable to be explored. As expected, Emergency admission type had the highest fraction of people.



The noteevents table

The noteevents table contains free text clinical notes from discharge summaries. Basic preprocessing (like replacing na values with empty string ‘ ’ etc) was applied on each note. There were almost 0.02% of notes missing that were filled with blank spaces. We can’t directly input the strings as features and they need to be converted to a form that can be used for model building. All the sentences were tokenized into individual words and punctuations and numbers were replaced with spaces. The tokens were all converted to lower case letters for ease of analysis.

CountVectorizer: The CountVectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary. It works as follows:

1. Create an instance of the CountVectorizer class.
2. Call the fit() function in order to learn a vocabulary from one or more documents.
3. Call the transform() function on one or more documents as needed to encode each as a vector.

As an example consider you have three sentences in your corpus:
“I love dogs”, “I hate dogs and knitting”, “Knitting is my hobby and my passion”
Then the vectorized way to represent them would be:

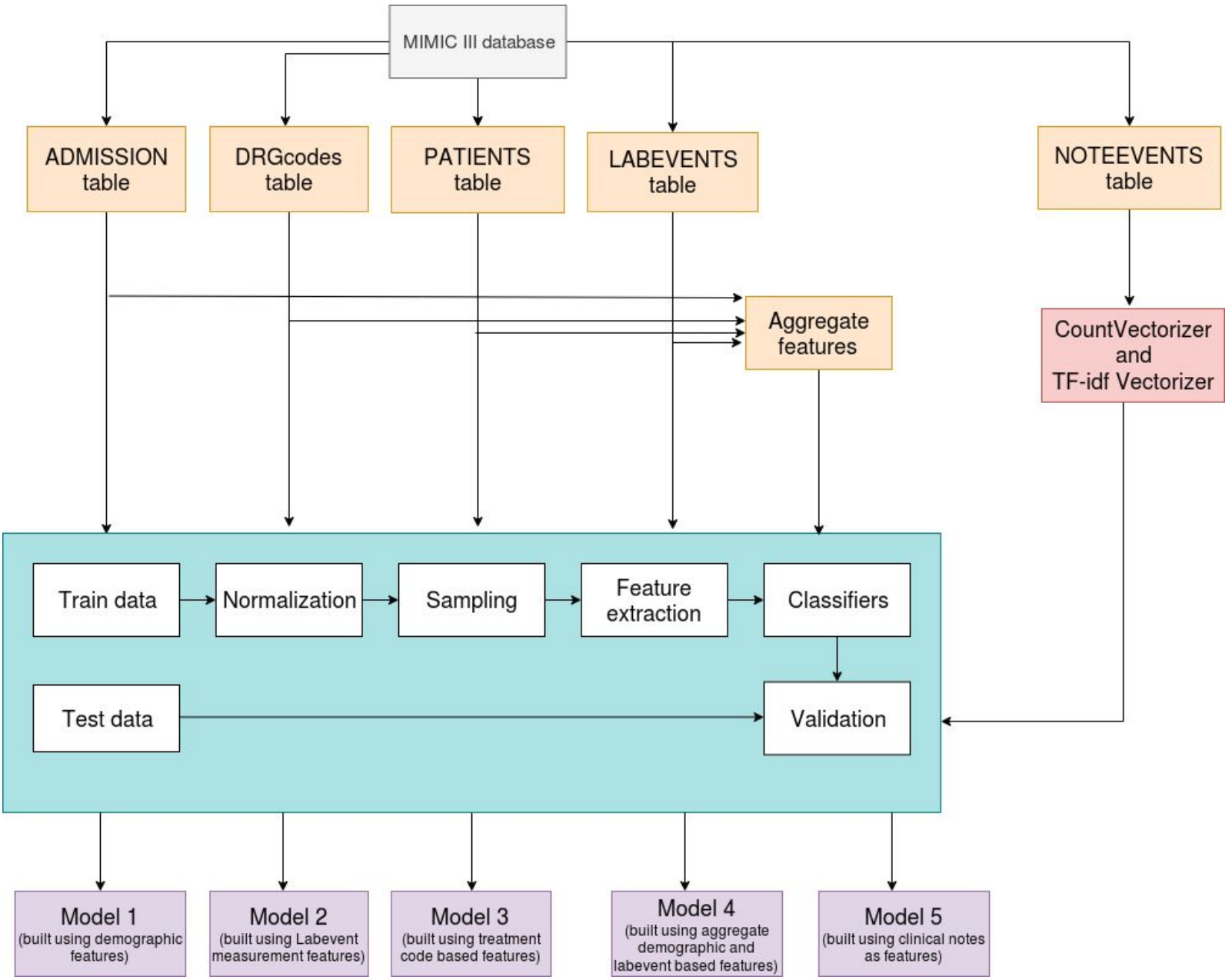
	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

(Image source: Link)

Feature description

Table	Variables
Admission information and demographics (One-hot encoding was done)	Religion, Language, Admission type, Insurance, Ethnicity and Age
Treatment information (One-hot encoding was done)	Cardiac catheterization, cardiac defibrillator and heart assist anomaly, cardiac defibrillator implant with cardiac catheterization, cardiac defibrillator implant without cardiac catheterization, cardiac valve and other major cardiothoracic procedures with cardiac catheterization, cardiac valve and other major cardiothoracic procedures without cardiac catheterization, cardiac valve procedures with cardiac catheterization, cardiac valve procedures without cardiac catheterization, coronary bypass with cardiac catheterization, coronary bypass with cardiac catheterization or percutaneous cardiac procedure, coronary bypass with PTCA, coronary bypass without cardiac catheterization, coronary bypass without cardiac catheterization or percutaneous cardiac procedure, other cardiac pacemaker implantation, other major cardiovascular procedures, other permanent cardiac pacemaker implant or PTCA with coronary artery stent implant, percutaneous cardiac procedure with drug-eluting stent, percutaneous cardiac procedure with non-drug-eluting stent, percutaneous cardiac procedure without coronary artery stent, percutaneous cardiovascular procedure, and permanent cardiac pacemaker implant
Lab measurements (Continuous values were standardized)	Average lab values included a lipid profile (cholesterol ratio, LDL cholesterol, HDL cholesterol, total cholesterol, and triglycerides), liver and kidney function tests (alanine transaminase, aspartate transaminase, alkaline phosphatase, albumin, bilirubin, blood urea nitrogen, creatinine, gamma-glutamyltransferase, L-lactate dehydrogenase, and total protein), cardiac function tests (N-terminal prohormone of B-type natriuretic peptide, C-reactive protein, creatine kinase, creatine phosphokinase-MB, cortisol, homocysteine, troponin I and troponin T), and electrolytes (bicarbonate, calcium, chloride, potassium, and sodium), glucose, hematocrit, hemoglobin, and white blood count
Aggregate	All of the above
Notes (CountVectorizer was used)	Free text discharge notes for patients suffering from CAD

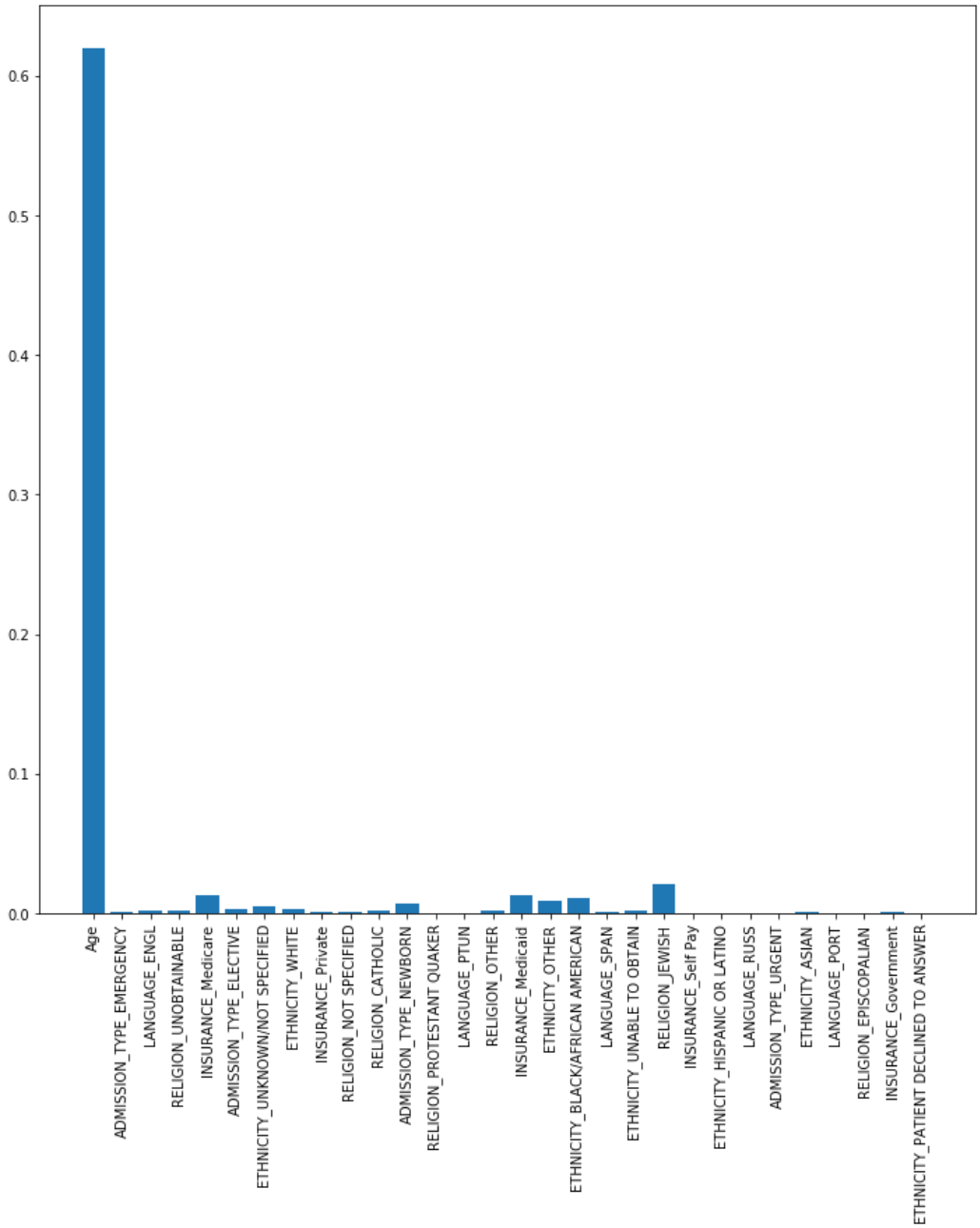
Workflow


MIMIC III database
ADMISSION table
DRGcodes table
PATIENTS table
LABEVENTS table
NOTEEVENTS table
Aggregate features
CountVectorizer and TF-idf Vectorizer
Train data
Normalization
Sampling
Feature extraction
Classifiers
Test data
Validation
Model 1 (built using demographic features)
Model 2 (built using Labevent measurement features)
Model 3 (built using treatment code based features)
Model 4 (built using aggregate demographic and labevent based features)
Model 5 (built using clinical notes as features)

Results

The following are the best results for each of the sampling technique-classifier pair.

1. Admission information and demographics



Feature importance

Undersampling
(Technique: Random Undersampler, Classifier: Linear Discriminant Analysis)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.96	0.59	0.77	0.73	0.68
Deceased	0.15	0.77	0.59	0.25	

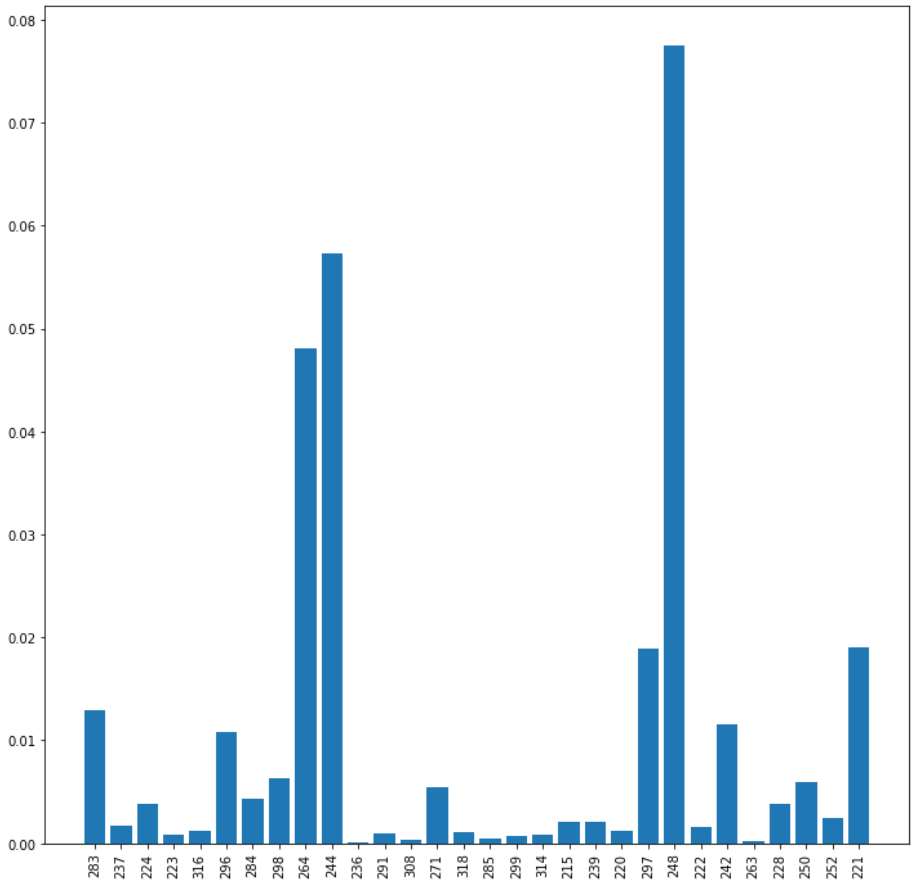
Oversampling
(Technique: Random Oversampler, Classifier: AdaBoost)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.96	0.60	0.76	0.74	0.68
Deceased	0.15	0.76	0.60	0.25	

Ensemble
(Technique: Balanced Random forests)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.97	0.57	0.79	0.72	0.679
Deceased	0.15	0.79	0.57	0.25	

2. Treatment information



Feature importance

Some of the top chosen features are (MSDRG codes):

a. **248: Perc cardiovasc proc w non-drug-eluting stent w MCC or 4+ ves/stents**
b. **244: Permanent cardiac pacemaker implant w/o CC/MCC**
c. **264: Other circulatory system O.R. procedures**
d. **297: Cardiac arrest, unexplained w CC**
e. **221: Cardiac valve & oth maj cardiothoracic proc w/o card cath w/o CC/MCC**

Undersampling
(Technique: Random Undersampler, Classifier: Random Forests)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.98	0.67	0.75	0.80	0.71

Deceased	0.12	0.75	0.67	0.21	

Oversampling

(Technique: Random Oversampler, Classifier: Random Forests)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.98	0.71	0.72	0.82	0.7138
Deceased	0.13	0.72	0.71	0.22	

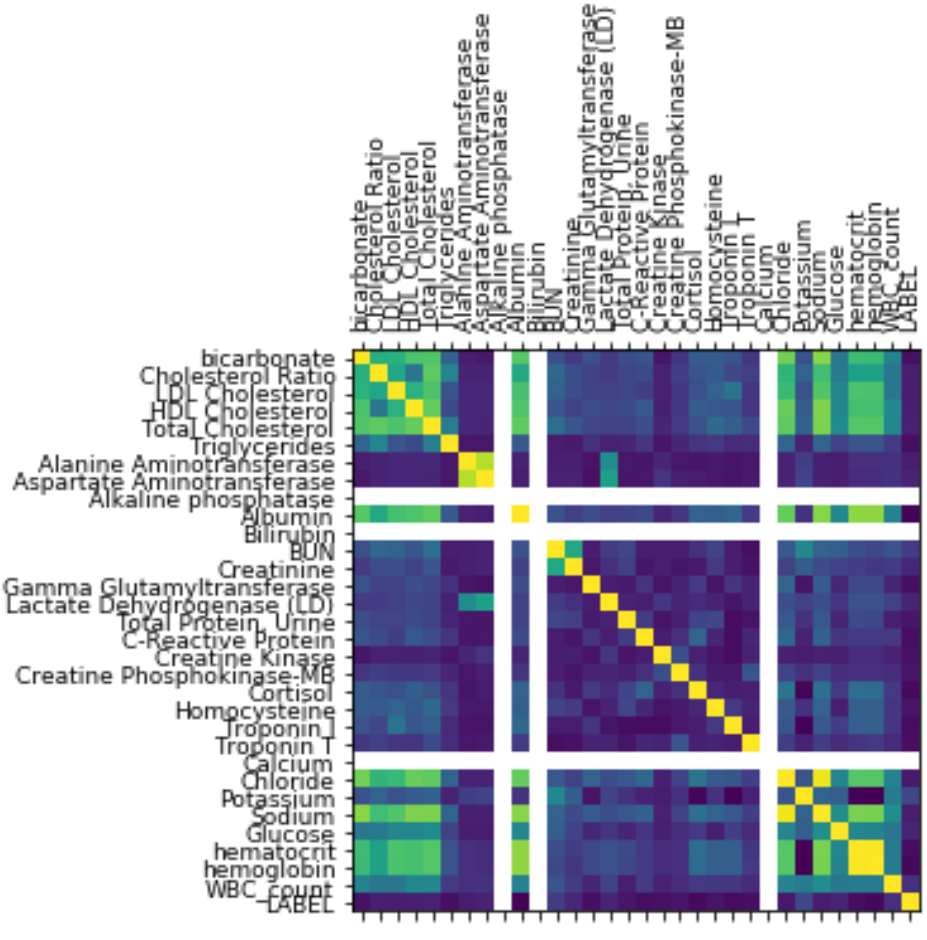
Ensemble

(Technique: Easy Ensemble classifier)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.98	0.69	0.76	0.81	0.723
Deceased	0.13	0.76	0.69	0.22	

Lab measurements

Correlation plot (LABEL as target variable)



Undersampling
(Technique: Random Undersampler, Classifier: Random Forests)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.96	0.65	0.709	0.77	0.68
Deceased	0.16	0.70	0.65	0.26	

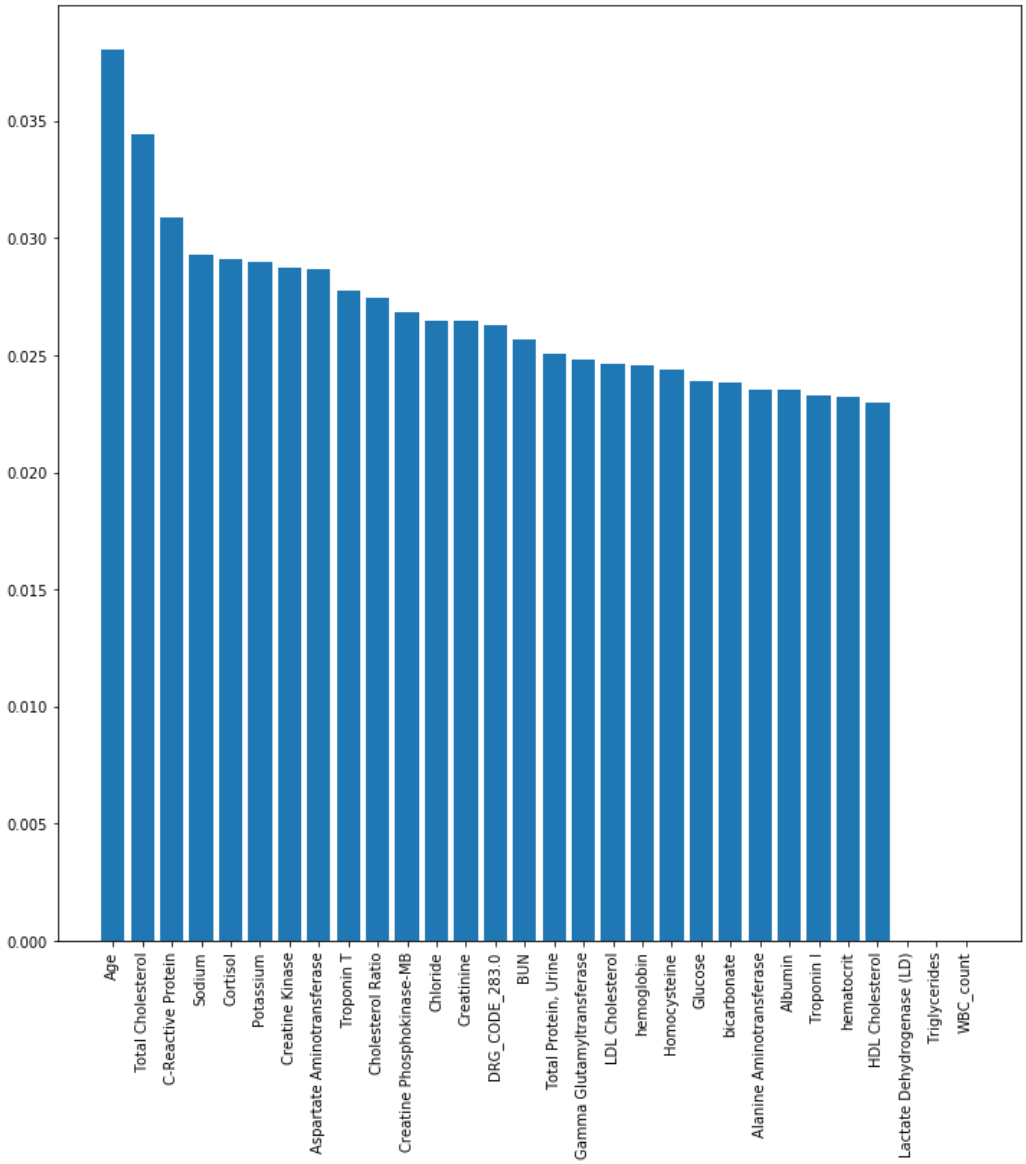
Oversampling
(Technique: Random Oversampler, Classifier: Random Forests)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.95	0.74	0.63	0.83	0.68
Deceased	0.18	0.63	0.74	0.28	

Ensemble
(Technique: Easy Ensemble classifier)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.96	0.65	0.72	0.78	0.68
Deceased	0.16	0.72	0.65	0.27	

Combined feature set



Feature importances (Top 30)

Undersampling
(Technique: Random Undersampler, Classifier: Linear Discriminant Analysis)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.99	0.73	0.83	0.84	0.78
Deceased	0.15	0.83	0.73	0.26	

Oversampling
(Technique: SMOTE, Classifier: Linear Discriminant Analysis)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.98	0.80	0.74	0.88	0.77
Deceased	0.18	0.74	0.80	0.29	

Ensemble
(Technique: Balanced RF)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.98	0.71	0.81	0.83	0.76
Deceased	0.15	0.81	0.71	0.25	

Text features

Undersampling
(Technique: Random Undersampler, Classifier: GradientBoosting)

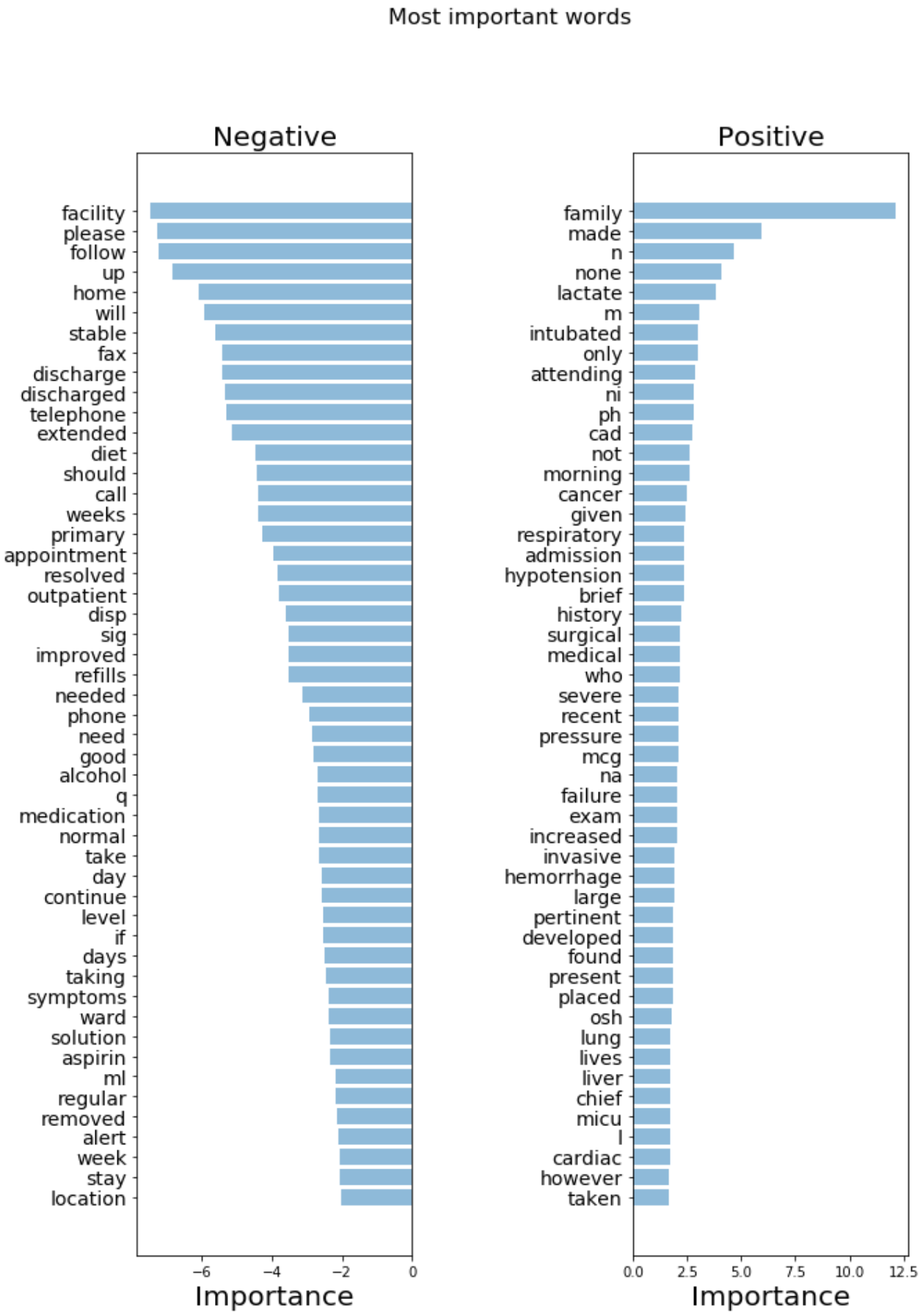
Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.99	0.97	0.88	0.93	**0.93**
Deceased	**0.76**	**0.88**	0.97	0.82	

Oversampling
(Technique: SMOTE, Classifier: Linear Discriminant Analysis)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.99	0.98	0.87	0.98	**0.93**
Deceased	**0.82**	**0.87**	0.98	0.84	

Ensemble
(Technique: Balanced RF)

Class	Precision	Recall	Specificity	F1 score	AUROC
Survived	0.99	0.97	0.88	0.98	**0.92**
Deceased	**0.71**	**0.88**	0.97	0.79	



Most important words from the text-based model

Label	Some important Words
Deceased	Family, lactate, intubated, attending, ni (no improvement), PH, CAD (Coronary artery disease), cancer, hypotension, respiratory, surgical, Na (sodium), failure, invasive, hemorrhage, pertinent, OSH (Occupational safety and health).
Survivor	Facility, follow, up, home, discharge, telephone, extended, weeks, resolved, appointment, outpatient, improved, medication, normal, take, continue, days, ward, solution, regular, location, stay

Discussion

There are many factors that affect mortality rates following a diagnosis of coronary atherosclerosis. Figuring out the factors that mostly influence this outcome, will assist clinicians and doctors to stratify high-risk patients well in advance. In this study, we saw that demographic, lab measurements and treatment based codes were able to classify high-risk patients with an AUC of 0.78. Note based features gave an AUC of 0.92 but these are discharge notes and won't be available to the doctors well in advance. Text-based features were chosen to see whether the model was able to capture important words that were indicative of complications for the deceased and recovery for the survivors.

There are some limitations of this model. First, some important features from the CHARTEVENTS table (like blood pressure (systolic and diastolic), heart rate etc were not included here and is part of another deep learning based model that I am currently working on. Second, although note based features gave the highest AUC, these are discharge notes and can't be used to predict disease onset or stratify complications well in advance. Thirdly, there were several missing values from the lab measurements table that were imputed using a K-nearest neighbors based imputation technique. Other imputation techniques or removing rows where the number of missing values exceed a certain threshold, need to be tried. Finally, temporal studies can be performed to use minute by minute measurements as features and predict the onset of complications well in advance.

This is the Github repository containing the codes that can be used to reproduce the results.