Predicting the Likelihood of Acute Myocardial Infarction (MI) Diagnosis in ICU Patients Using Machine Learning Approaches

Student: Trang Tran

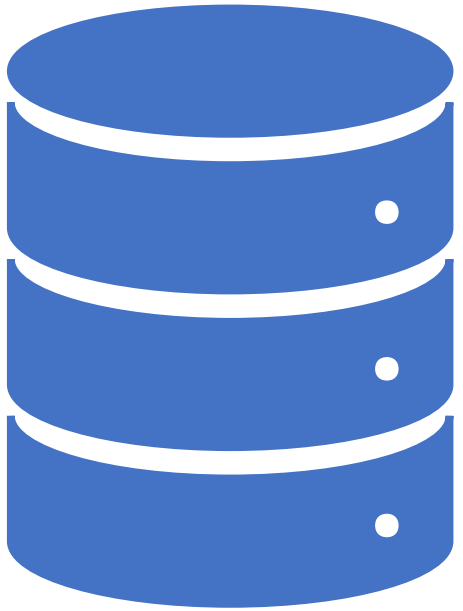
Teacher: Saeed Amal

Executive Summary

- Primary Objective: Develop and evaluate machine learning algorithms capable of predicting the likelihood of patients in the Intensive Care Unit (ICU) developing the diagnosis of Acute Myocardial Infarction (MI).
- Binary Classification:
  - Output Variable: Whether the patient develops MI in the ICU (Yes/No).
  - Input Variables: Diagnosis codes, procedure codes, lab results, and demographic information.
- Analysis Structure:
  - Data Processing and Cohort Construction
  - Model Development and Feature Importance
  - Model Evaluation

Data Summary

- Data Source: The MIMIC-III database contains multiple components relevant to ICU admissions.
- The datasets included in this project are:
    1. **DIAGNOSES_ICD**: Contains ICD codes for diagnoses.
    2. **PROCEDURES_ICD**: Contains ICD codes for medical procedures.
    3. **PATIENTS**: Contains patient demographic information like age, gender, and ethnicity.
    4. **ADMISSIONS**: Includes admission information such as admission type, admission time, and discharge time.
    5. **LABEVENTS**: Contains lab test results linked to patient admissions.
    6. **Lab_Item_Codes.txt**: Provides descriptions or categories of lab items.
    7. **Error-prone codes**: Provides the diagnosis codes of the ICU outcomes.



Data Preprocessing – Part 1

- Get the first admission date for each Patient from the Admissions file.
- Calculate their age at the first admission.
- Keep only patients under 120 years old (eliminate ~2000 errors).

	ROW_ID	SUBJECT_ID	GENDER	DOB	DOD	DOD_HOSP	DOD_SSN	EXPIRE_FLAG	First_Admission_Date	Age_at_First_Admission
0	234	249	F	2075-03-13	NaN	NaN	NaN	0	2149-12-17 20:41:00	75.0
1	235	250	F	2164-12-27	2188-11-22 00:00:00	2188-11-22 00:00:00	NaN	1	2188-11-12 09:22:00	24.0
2	236	251	M	2090-03-15	NaN	NaN	NaN	0	2110-07-27 06:46:00	20.0
3	237	252	M	2078-03-06	NaN	NaN	NaN	0	2133-03-31 04:24:00	55.0
4	238	253	F	2089-11-26	NaN	NaN	NaN	0	2174-01-21 20:58:00	84.0
...	...	...	...	...	...	...	...	...	...	...
46515	31840	44089	M	2026-05-25	NaN	NaN	NaN	0	2111-09-30 12:04:00	85.0
46516	31841	44115	F	2124-07-27	NaN	NaN	NaN	0	2161-07-15 12:00:00	37.0
46517	31842	44123	F	2049-11-26	2135-01-12 00:00:00	2135-01-12 00:00:00	NaN	1	2135-01-06 07:15:00	85.0
46518	31843	44126	F	2076-07-25	NaN	NaN	NaN	0	2129-01-03 07:15:00	52.0
46519	31844	44128	M	2098-07-25	NaN	NaN	NaN	0	2149-06-08 15:21:00	51.0

44529 rows × 10 columns

Patient DF after merging with some Admission info

Data Preprocessing – Part 1

Filter for MI cases in the Diagnoses dataset using ICD-9 codes starting with "410".

Merge the MI case- cohort with the Admissions file and choose the earliest admission date only.

Merge with the Patient file for full demographic data (4671 matching cases).

Create a control cohort by selecting patients not in the case cohort.

Create a random sample of 4671 rows from the control cohort for the balanced training data.

Combine both cohorts, assigning a label 1 for MI cases and 0 for controls.

Compute the STAYTIME (day) feature from two columns, “DISCHTIME” and “ADMITTIME”.

Keep only 11 relevant columns from 33 columns for the training process.

	SUBJECT_ID	HADM_ID	STAYTIME (day)	ADMISSION_TYPE	ADMISSION_LOCATION	DISCHARGE_LOCATION	MARITAL_STATUS	ETHNICITY	LABEL	GENDER	Age_at_First_Admission
0	3	145834	10	EMERGENCY	EMERGENCY ROOM ADMIT	SNF	MARRIED	WHITE	1	M	77.0
1	21	109451	13	EMERGENCY	EMERGENCY ROOM ADMIT	REHAB/DISTINCT PART HOSP	MARRIED	WHITE	1	M	87.0
2	24	161859	2	EMERGENCY	TRANSFER FROM HOSP/EXTRAM	HOME	SINGLE	WHITE	1	M	39.0
3	25	129635	3	EMERGENCY	EMERGENCY ROOM ADMIT	HOME	MARRIED	WHITE	1	M	59.0
4	37	188670	5	EMERGENCY	EMERGENCY ROOM ADMIT	HOME HEALTH CARE	MARRIED	WHITE	1	M	69.0
...	...	...	...	...	...	...	...	...	...	...	...
9337	19988	141007	3	NEWBORN	CLINIC REFERRAL/PREMATURE	HOME	NaN	WHITE	0	M	0.0
9338	99229	150893	27	EMERGENCY	CLINIC REFERRAL/PREMATURE	LONG TERM CARE HOSPITAL	WIDOWED	PATIENT DECLINED TO ANSWER	0	F	78.0
9339	24541	157578	4	EMERGENCY	EMERGENCY ROOM ADMIT	HOME	SINGLE	WHITE	0	M	67.0
9340	10992	145405	6	EMERGENCY	EMERGENCY ROOM ADMIT	SNF	SINGLE	WHITE	0	F	83.0
9341	18836	168022	8	ELECTIVE	PHYS REFERRAL/NORMAL DELI	SNF	MARRIED	UNKNOWN/NOT SPECIFIED	0	M	75.0

9342 rows × 11 columns

Data Preprocessing – Part 2

- Procedures DF: Perform one-hot encoding for 2009 procedure codes.
- Labevents DF: Convert lab item codes to descriptive names using the Lab Item Codes text file.
  => Pivot the data so that each lab test item becomes a column with the corresponding 'max' VALUENUM.
- Merge new ‘Procedures’ and ‘Labevents’ DFs with the case-control cohort DF from Part 1.
- One-hot encoding for remaining categorical columns in the merged DF: ['ADMISSION_TYPE', 'ADMISSION_LOCATION', 'DISCHARGE_LOCATION', 'MARITAL_STATUS', 'ETHNICITY', 'GENDER’].
- Exclude “Troponin I” and “Troponin T” features to avoid data leakage.

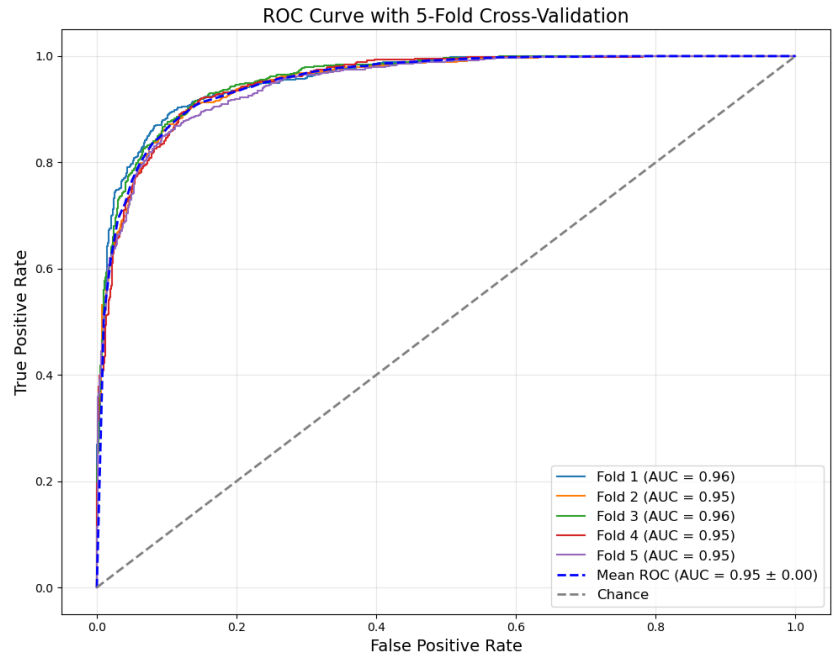
	SUBJECT_ID	HADM_ID	STAYTIME (day)	LABEL	Age_at_First_Admission	ICD9_10	ICD9_11	ICD9_12	ICD9_13	ICD9_14	...	ETHNICITY_PATIENT DECLINED TO ANSWER	ETHNICITY_PORTUGUESE	ETHNICITY_UNABLE TO OBTAIN	ETHNICITY_UNKNOWN/NOT SPECIFIED	ETHNICITY_WHITE
0	3	145834	10	1	77.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	True
1	21	109451	13	1	87.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	True
2	24	161859	2	1	39.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	True
3	25	129635	3	1	59.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	True
4	37	188670	5	1	69.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9337	19988	141007	3	0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	True
9338	99229	150893	27	0	78.0	0.0	0.0	0.0	0.0	0.0	...	True	False	False	False	False
9339	24541	157578	4	0	67.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	True
9340	10992	145405	6	0	83.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	True
9341	18836	168022	8	0	75.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	True	False

9342 rows × 2467 columns

Machine Learning Models

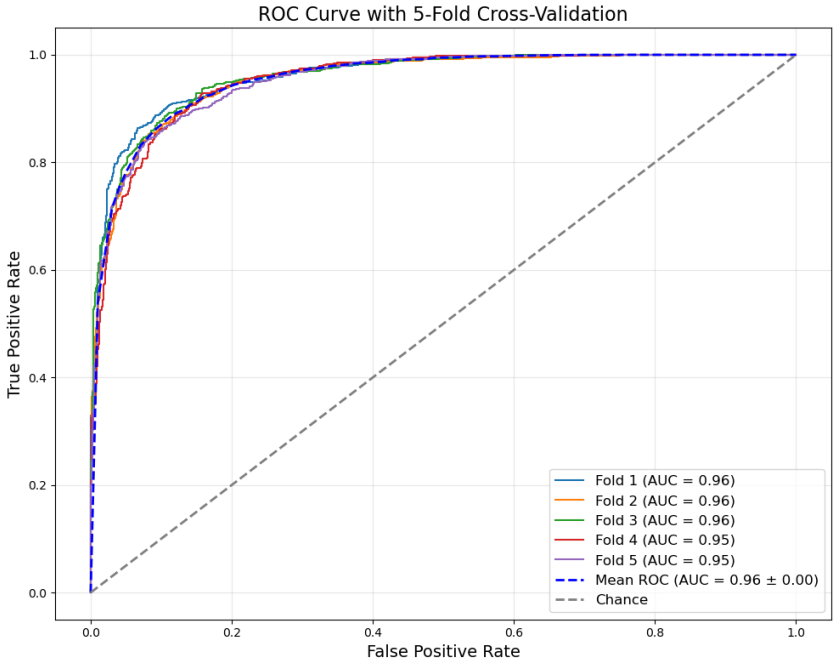
XGBoost

Average Metrics:	
Accuracy	0.884178
AUC	0.953687
Sensitivity	0.884819
Specificity	0.883538
PPV	0.883693
NPV	0.884738



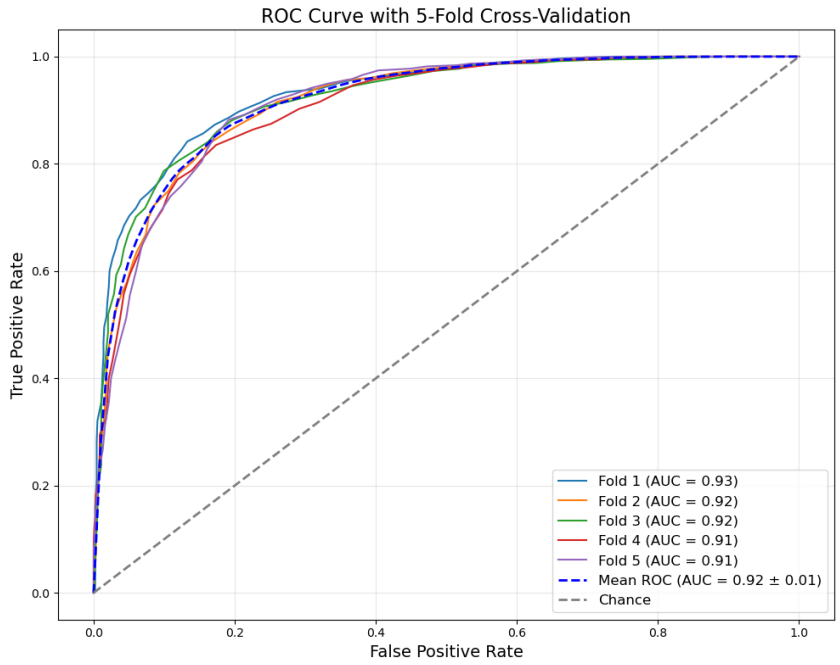
Histogram Gradient Boosting

Average Metrics:	
Accuracy	0.885248
AUC	0.956793
Sensitivity	0.883962
Specificity	0.886535
PPV	0.886256
NPV	0.884407

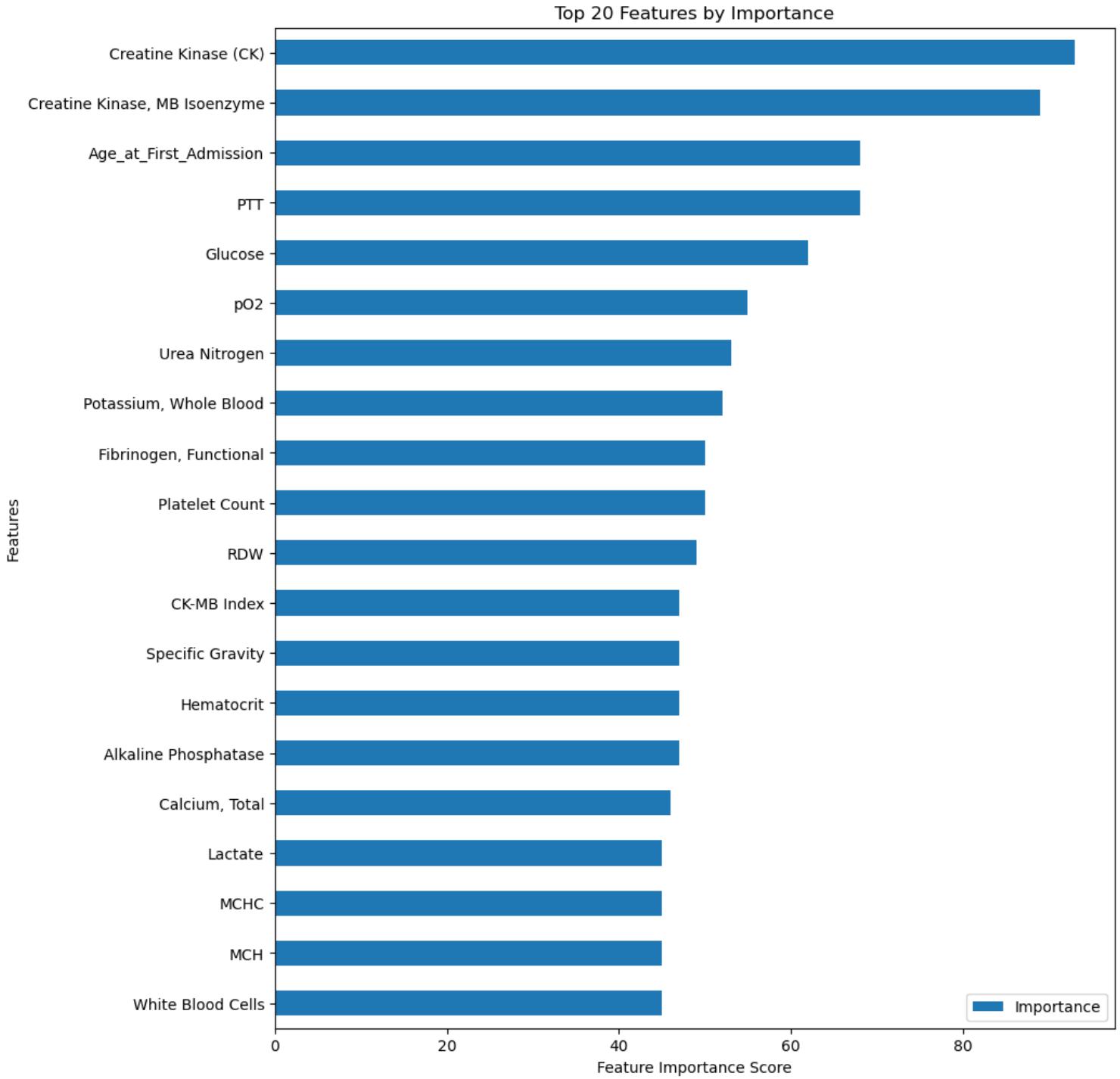


Random Forest

Average Metrics:	
Accuracy	0.740731
AUC	0.918201
Sensitivity	0.514638
Specificity	0.966819
PPV	0.941057
NPV	0.669103



Feature Importance in XGBoost Model



Conclusion

Models such as XGBoost, Histogram Gradient Boosting, and Random Forest effectively handle null values, large datasets, and high dimensionality. They all achieve high AUC and accuracy scores when performing classification tasks on the Mimic dataset.

We have gained better insights into the appropriate variables and model development by testing multiple aggregation functions for the lab test values.

There is still room to apply the LSTM deep learning model to the time series data. However, tree-based and ensemble models also perform well in predicting the diagnosis, with less complexity in layers and computation. Additionally, XGBoost offers greater interpretability through feature importance extraction.

Thank you!

Questions?