SEPSIS PREDICTION USING MIMIC-III DATABASE

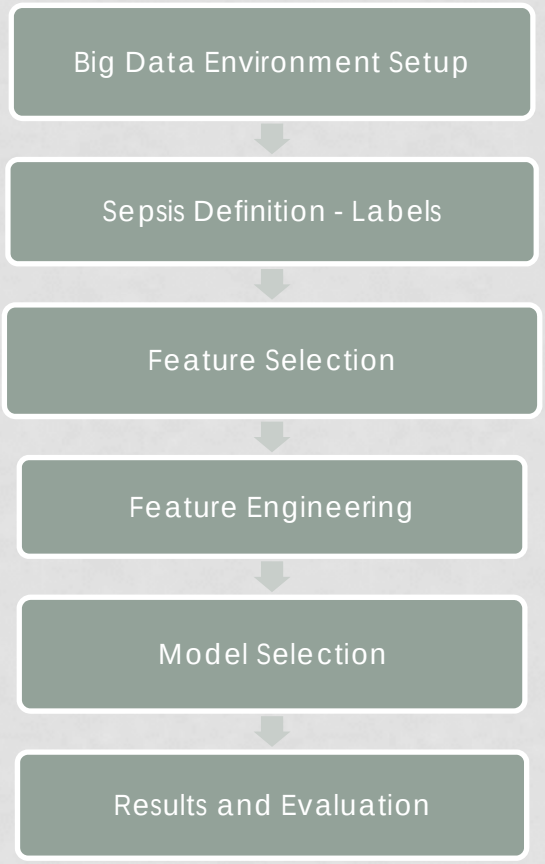
CSE 6250 – BIG DATA FOR HEALTH

GROUP 8 – VIRAL DOSHI

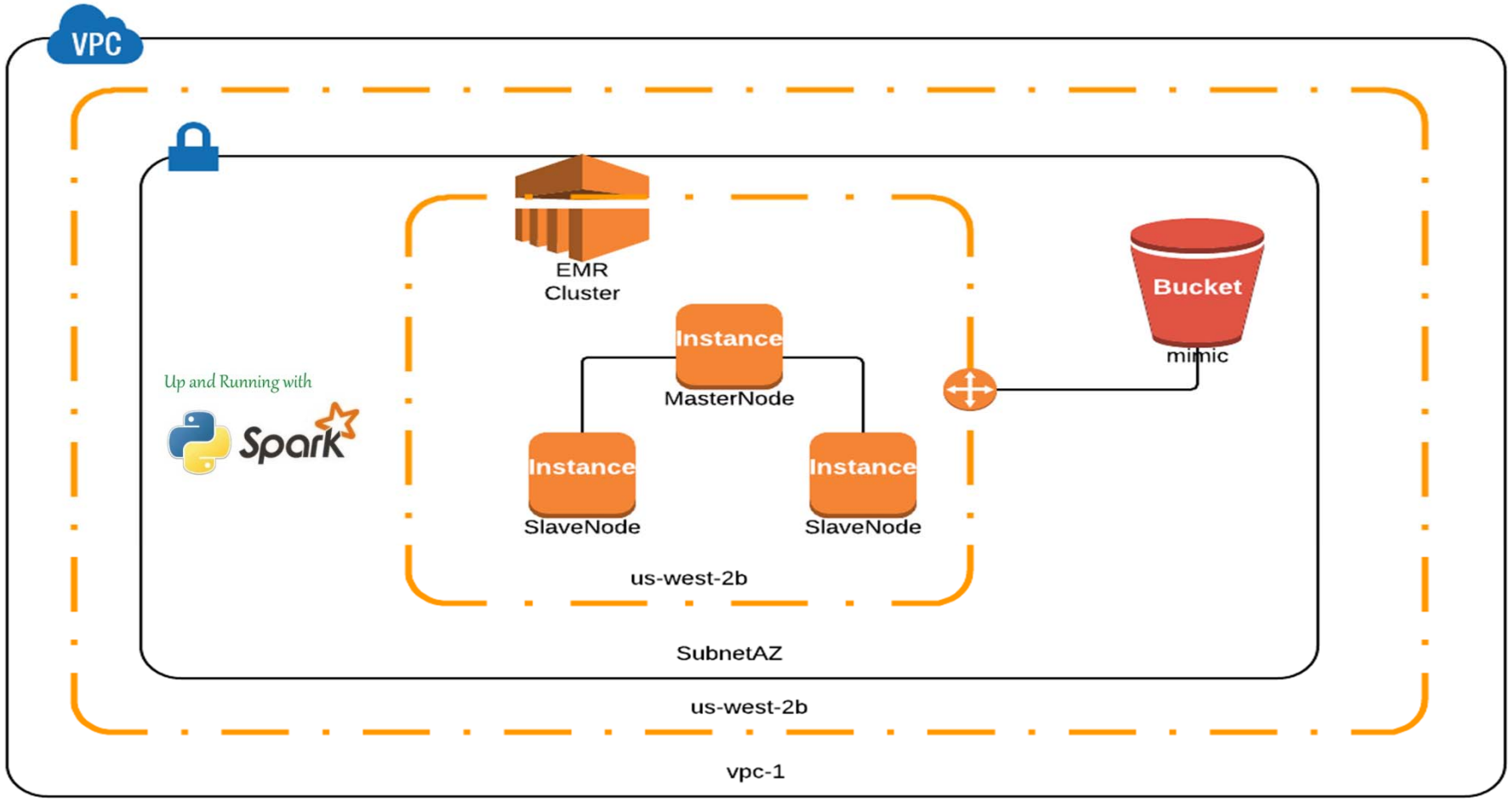
WHY SEPSIS PREDICTION

- Sepsis is one of the leading causes of mortality in hospitalized patients and responsible for placing an enormous cost burden on the health care system.
- Each day 2000 to 3000 new patients are identified with Sepsis in US hospitals amounting to approximately $24 billion in cost.
- Early identification of individuals at risk of developing life-threatening severe sepsis could enable early treatment and improve outcomes.

PROBLEM FORMULATION



BIG DATA MACHINE SETUP



DATA

- Medical Data (40 GB) from MIMIC-III database was used. This is the largest public database.
- Labels
  - 46535 admissions without Sepsis
  - 4085 admissions with Sepsis
- Feature Engineering

  About 320 features were generated using the following distinct tables from the database.
  - Admissions tables
  - Patients tables
  - Labevents table
  - CPTevents table
  - Prescriptions table
- The timeline was restricted to the first 24 hours after admissions.

FEATURE TABLE

Feature	Patient Characteristics	Laboratory Tests	Procedures	Medications
MIMIC Table from which it was extracted	ADMISSION & PATIENTS	LABEVENTS	CPTEVENTS	PRESCRIPTIONS
Feature description	Vector of demographics and patient characteristics	Count of different lab tests ordered for the patient. Limited to 100 most common labs	Count of different CPT codes. Limited to 100 most common labs	Count of medications that were administered. Limited to 100 most common medications.

FEATURES

ADMISSION_FEATURES

HADM_ID	ELECTIVE	EMERGEN	NEWBORN	URGENT	** INFO N	CLINIC RE	EMERGEN	HMO REFE	PHYS REFE	TRANSFER	TRANSFER
127847		1					1				
191743		1							1		
144378	1								1		
121893		1					1				
137154			1						1		
195486		1					1				

LAB_Features

Hadm_ID	% Hemogl	Alanine A	Albumin	Alkaline P	Alveolar-a	Amylase	Anion Gap	Anisocyto	Asparate A	Atypical L	Bacteria	Band
107083		1	2	1			2	1	1	1	1	
117650		1	1	1			3	1	1			
116430							1	2		2		
127929					1		1					
103828		1	1	1		1	4	1	1			
165185		1	2	1		1	2	2	1	2		
159268							1					
107893		1	1	1			1		1			
174680		1	1	1	3	1	1		1			
110914								1		1		

CPT_Features

HADM_ID	110	315	334	335	365	366	368	438	440	441	443	471	611	612
119984														
183978			2	2										
195438		1			1	1							1	1
155605	2							1	2	2				
174680					1									
179705												1		

Med_features

HADM_ID	0.9% Sodi	Acetamin	Albuterol	Aspirin	Aspirin EC	Calcium G	Captopril	D5 1/2NS	D5W	Docusate	Docusate	Furosemi	Glycopyrr	HYDROmc	Ha
131252							2	1		1				1	
122580															
128117				2					1			2			
126212			1									1			
193791						1									
129833		2		1	1	1			4	1			1		

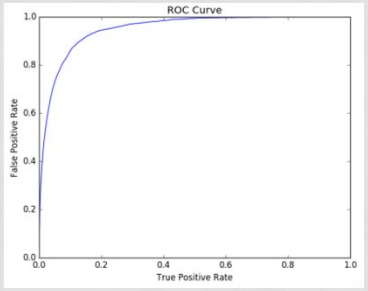
MODELS AND RESULTS

- We used two models to train and test the data
  - Logistic Regression
  - Random Forest

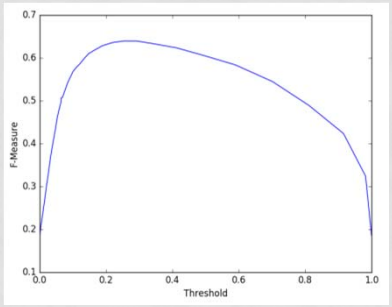
The data was split 80 /20.

RESULTS

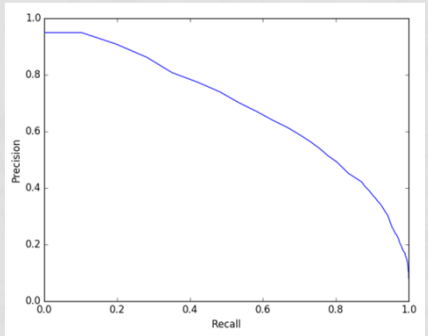
Model	F-1	Precision	Recall
Logestic Regression	0.937319695	0.9361777	0.942743932
Random Forest	0.89749176	0.91613343	0.925705794



ROC curve



Fmeasure



Precision Curve

ACKNOWLEDGEMENT

THANK YOU!