Cloud Computing

•••

Group 1
Trinh Vu, Jane Zeng, Qing Ruan

Project Definition

Scope of the project

Number of diabetic patients is growing rapidly in the United States.

Problem: Hospital readmission for diabetic patients has become a priority and the cost associated with treatment when patients have to be readmitted is a concern.

Goals: Implement a Machine Learning problem to classify patients who are more likely to be readmitted using AWS resources.

Source of Data

The source of dataset is UCI

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

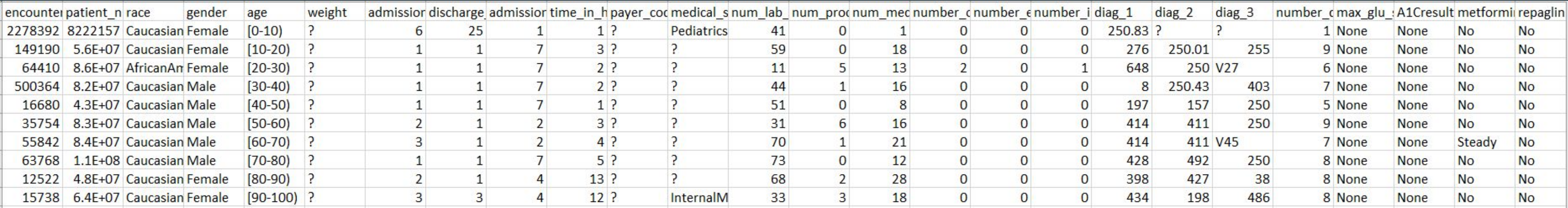
Project Definition

Features Implemented

AWS: S3, Amazon SageMaker, Python, Boto3, Jupyter Notebook, Tensorflow, Keras

Dataset has 101,766 rows and 54 columns, these are a few of the important features:

- Age
- Gender
- Race
- Admission type
- Time in hospital
- Treatments
- Readmitted



encounte	patient_n	race	gender	age	weight	admissior	discharge_	admissior	time_in_h	payer_coc	medical_s	num_lab_	num_proc	num_mec	number_c	number_e	number_i	diag_1	diag_2	diag_3	number_c	max_glu_	A1Cresult	metformi	repaglin
2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	?	Pediatrics	41	0	1	0	0	0	250.83	?	?	1	None	None	No	No
149190	5.6E+07	Caucasian	Female	[10-20)	?	1	1	7	3	?	?	59	0	18	0	0	0	276	250.01	255	9	None	None	No	No
64410	8.6E+07	AfricanAn	Female	[20-30)	?	1	1	7	2	?	?	11	5	13	2	0	1	648	250	V27	6	None	None	No	No
500364	8.2E+07	Caucasian	Male	[30-40)	?	1	1	7	2	?	?	44	1	16	0	0	0	8	250.43	403	7	None	None	No	No
16680	4.3E+07	Caucasian	Male	[40-50)	?	1	1	7	1	?	?	51	0	8	0	0	0	197	157	250	5	None	None	No	No
35754	8.3E+07	Caucasian	Male	[50-60)	?	2	1	2	3	?	?	31	6	16	0	0	0	414	411	250	9	None	None	No	No
55842	8.4E+07	Caucasian	Male	[60-70)	?	3	1	2	4	?	?	70	1	21	0	0	0	414	411	V45	7	None	None	Steady	No
63768	1.1E+08	Caucasian	Male	[70-80)	?	1	1	7	5	?	?	73	0	12	0	0	0	428	492	250	8	None	None	No	No
12522	4.8E+07	Caucasian	Female	[80-90)	?	2	1	4	13	?	?	68	2	28	0	0	0	398	427	38	8	None	None	No	No
15738	6.4E+07	Caucasian	Female	[90-100)	?	3	3	4	12	?	InternalM	33	3	18	0	0	0	434	198	486	8	None	None	No	No

Project Definition

Expected Outcome

Data is stored in S3

Machine Learning models are built and trained using Python script that will be run in Notebook instance in Amazon SageMaker

Results of Machine Learning models with their scores

Select best model to classify readmission of diabetic patients

Project Architecture

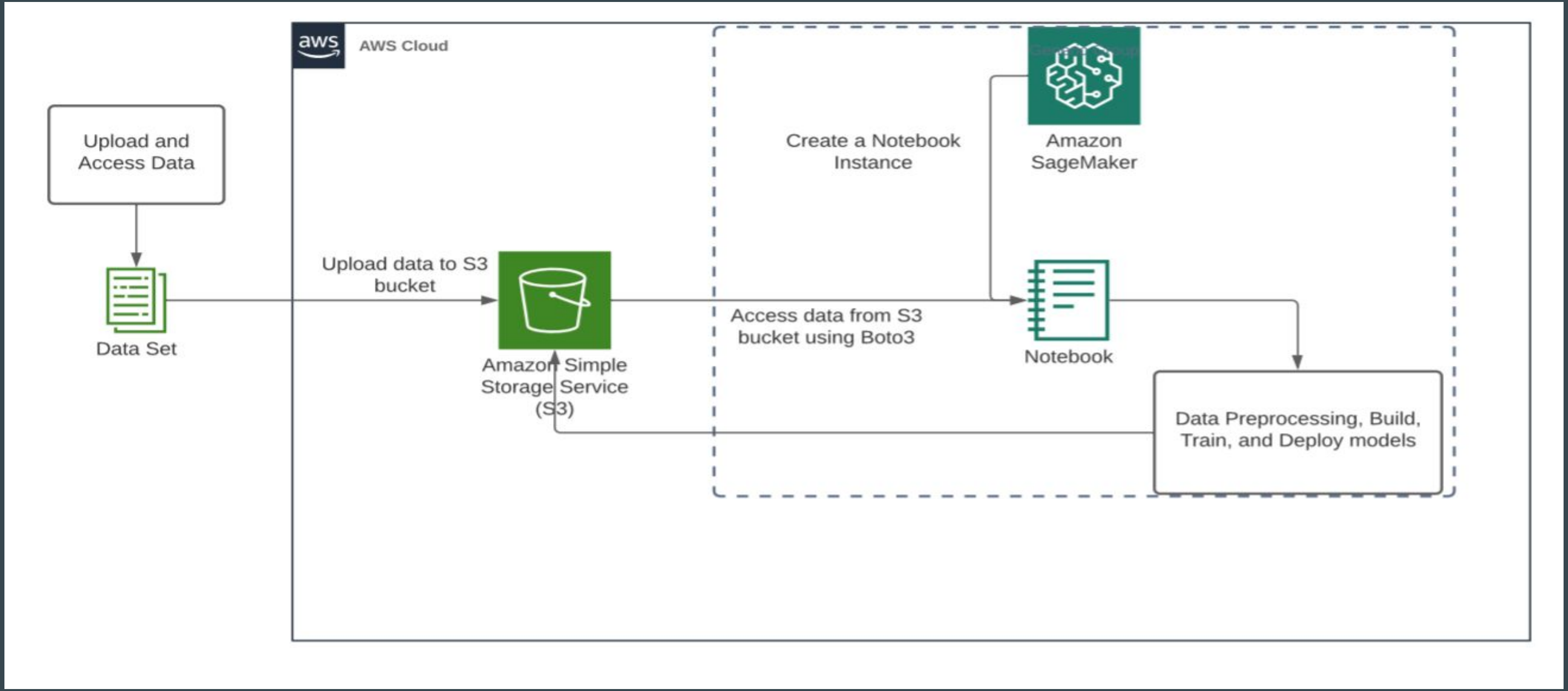
Logical Architecture

S3 is used to store data and results.

Amazon SageMaker Notebook instance is used to run Data Preprocessing, Exploratory Data Analysis, Machine Learning models

Project Architecture

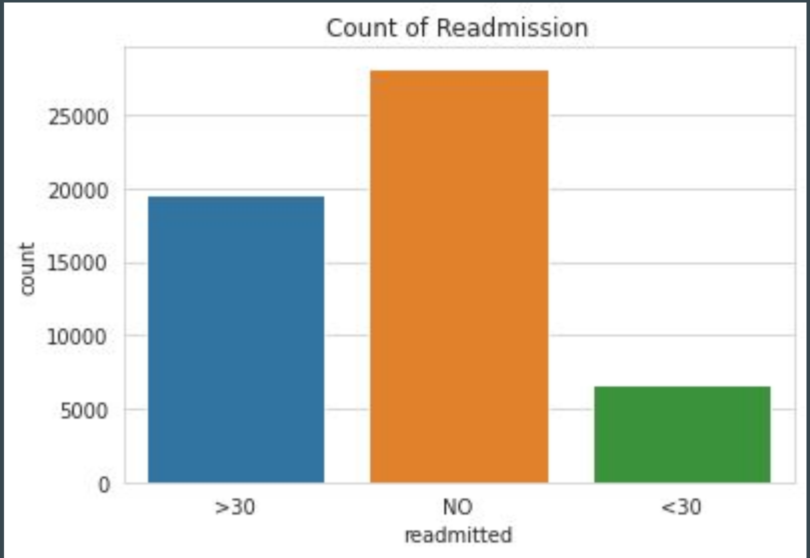
Data FLow

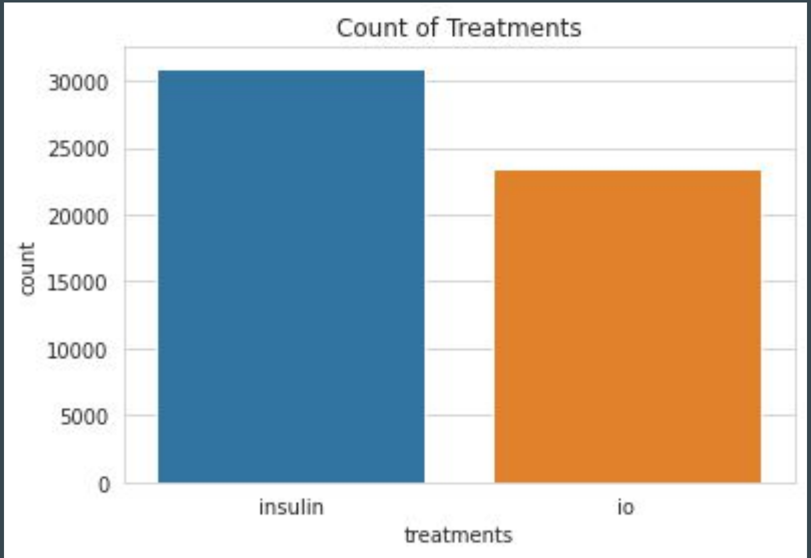

AWS Cloud
Upload and Access Data
Data Set
Upload data to S3 bucket
Amazon Simple Storage Service (S3)
Access data from S3 bucket using Boto3
Create a Notebook Instance
Amazon SageMaker
Notebook
Data Preprocessing, Build, Train, and Deploy models

DATA PREPROCESSING

1. Remove observations with ‘diabetesMed’==’No’
2. Remove columns with high proportion of missing values (“weight”, “medical_specialty”, and “payer_code”)
3. Remove categorical features that have large number of categories (diag_1, diag_2, diag_3)
4. Group 24 treatments into 2 categories “insulin” and “io” (stands for insulin+others) and then create 1 column for treatment

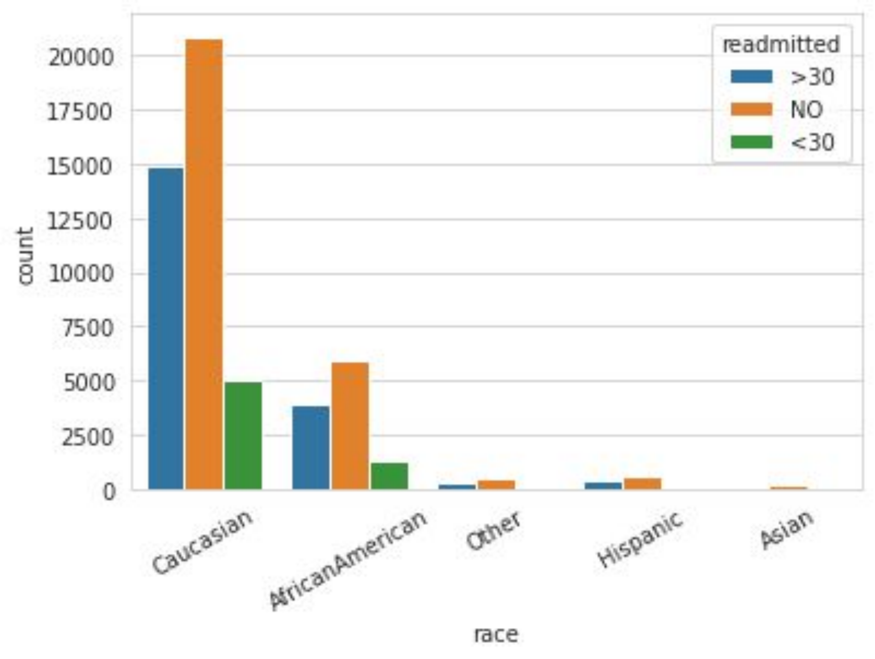
EXPLORATORY DATA ANALYSIS





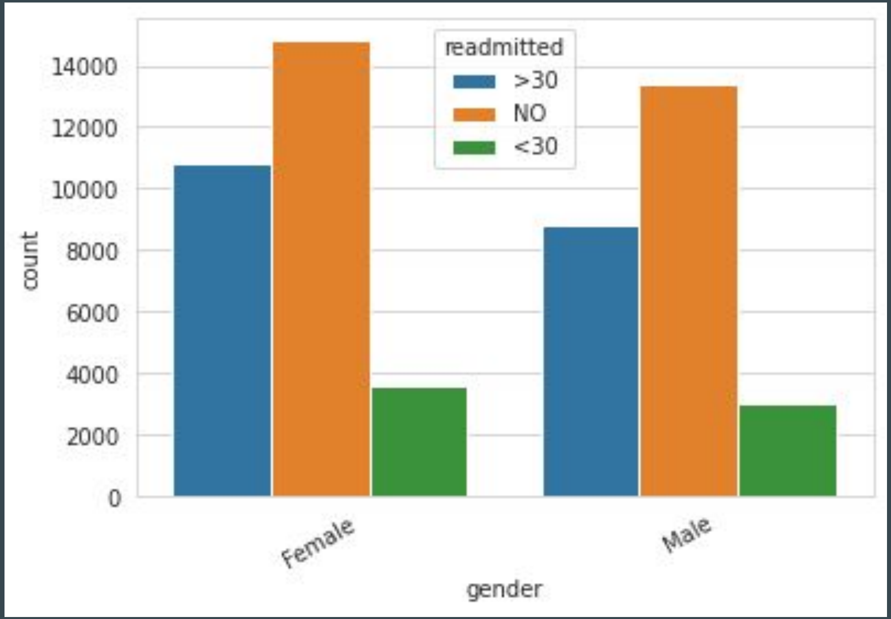
EXPLORATORY DATA ANALYSIS

Race and Readmitted



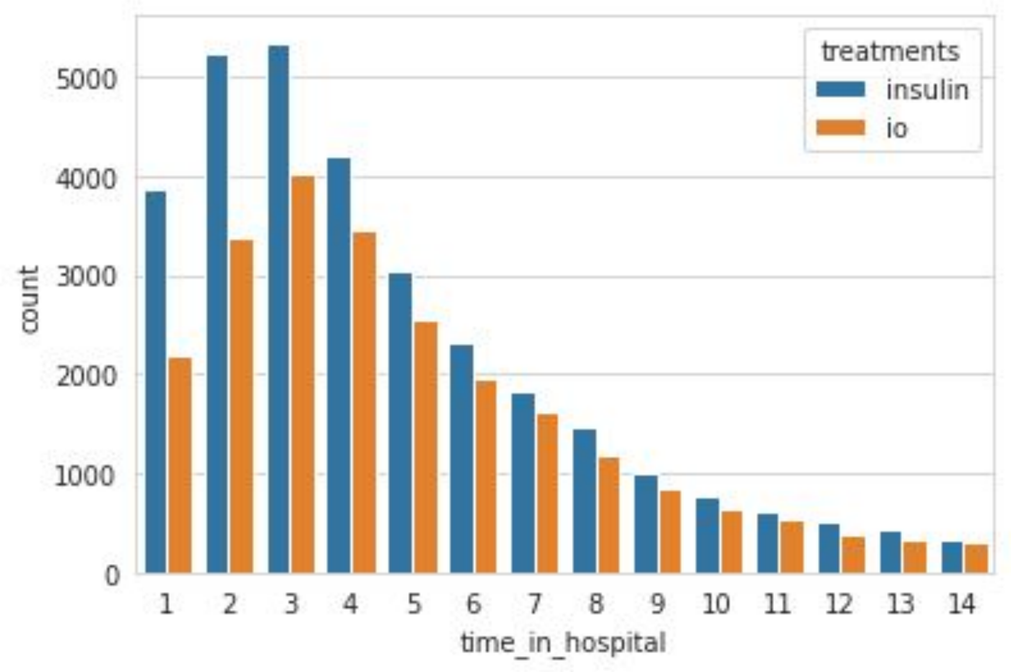
EXPLORATORY DATA ANALYSIS

Gender and Readmitted



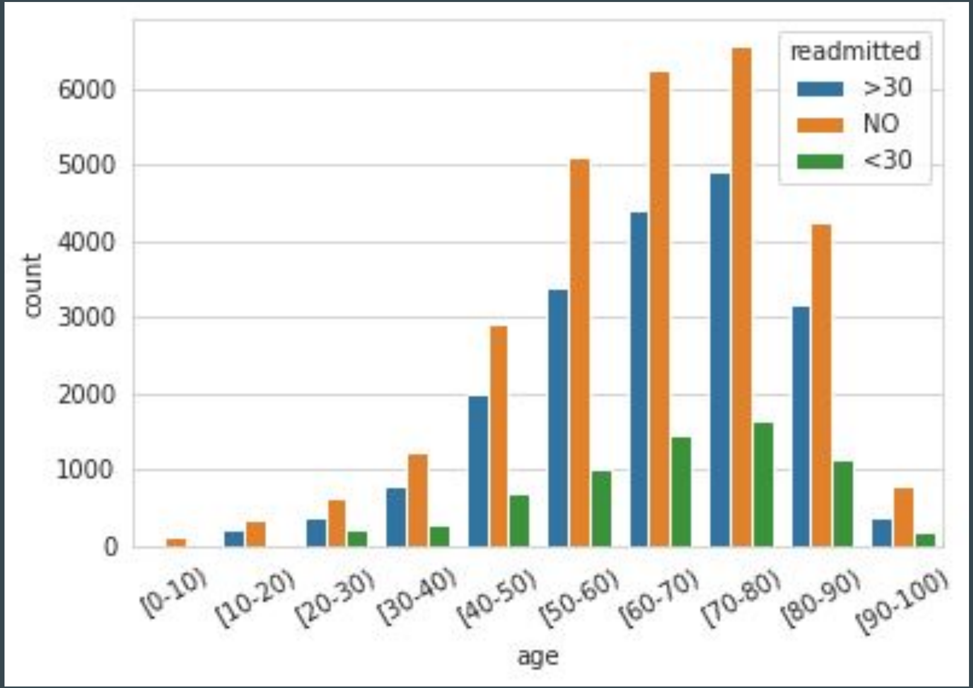
EXPLORATORY DATA ANALYSIS

Time in hospital



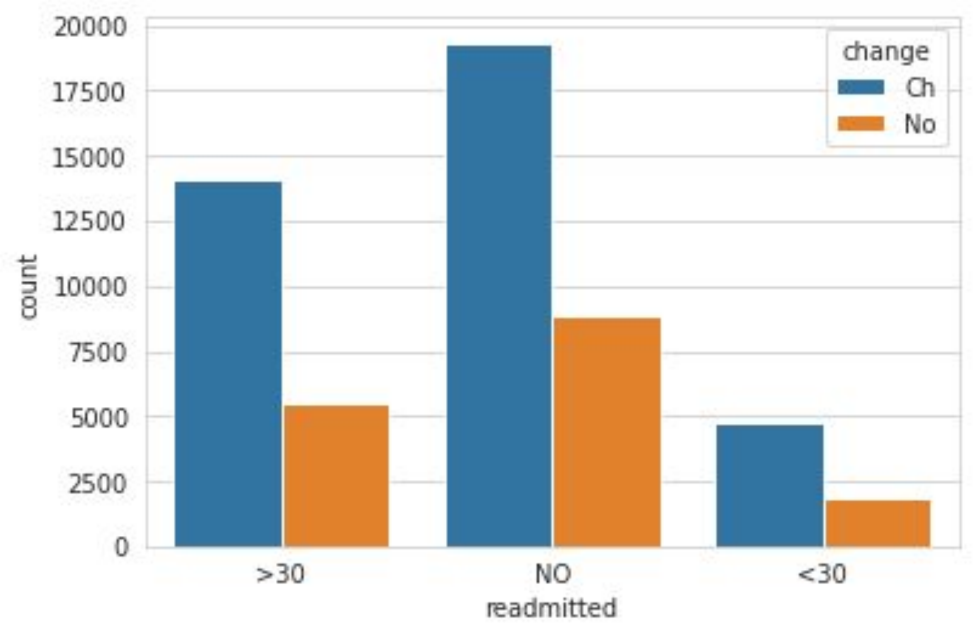
EXPLORATORY DATA ANALYSIS

Age and Readmitted



EXPLORATORY DATA ANALYSIS

Change in medication



MODELS AND PREDICTION

Target is Readmitted.

Encoding:

- "Readmitted" target: Group patients with "No" and ">30" into 1 group and label this group as 0. The other group ("<30") is labeled as 1.
- Use label encoder for target "Readmitted" and one-hot encoding for the remaining categorical features

Split the data into train, validation, and test sets. Then standardize the data.

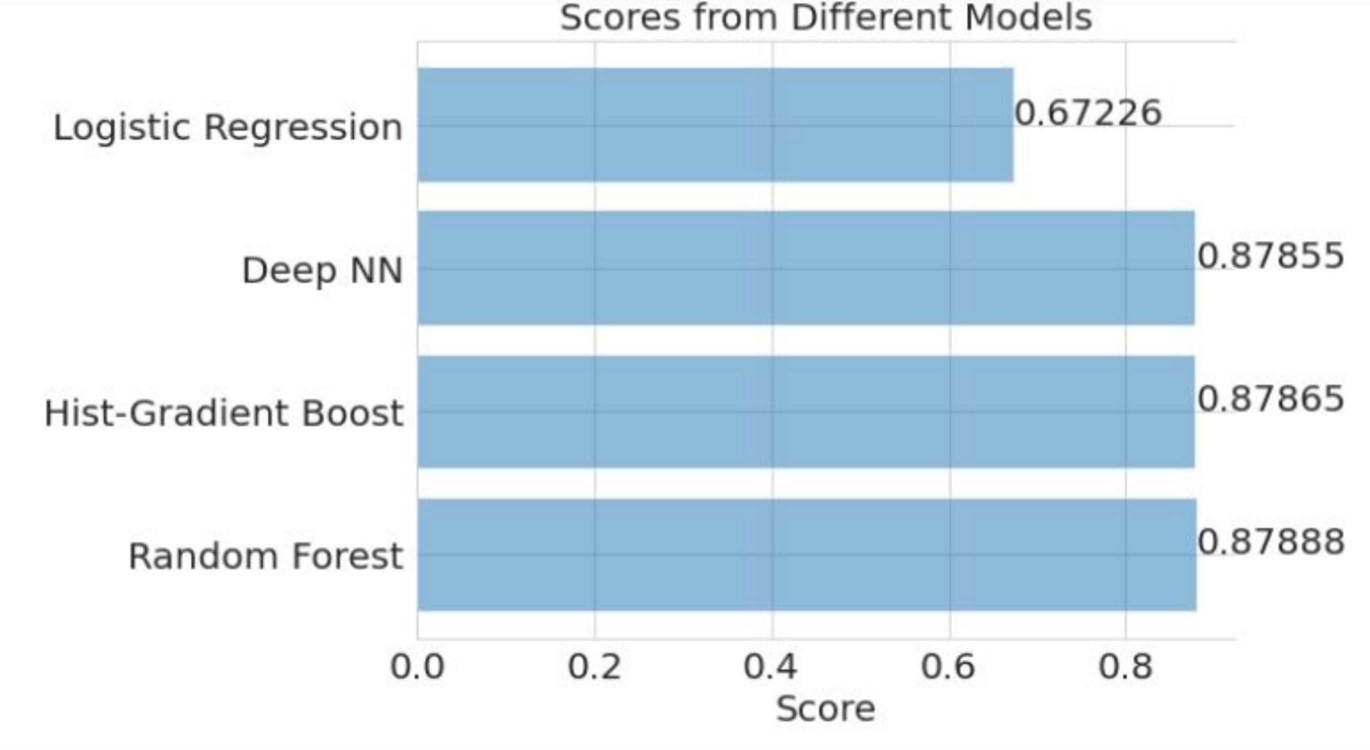
Models for Readmission prediction includes: Logistic Regression, Random Forest, Gradient Boosting, and Deep Neural Network.

Scores for Different Models for Readmission Prediction

The best model:

Random Forest

Score: 87.9%



CONCLUSION

Conclusion:

- Random Forest gives the best score for predicting readmission (87.9%).
- Amazon SageMaker is a good resource to do Machine Learning without complicated installations of Python and other software.

Future work:

- Perform the models on a complete dataset with more data (no missing values, no nulls)
- Explore other resources from AWS for Data Science