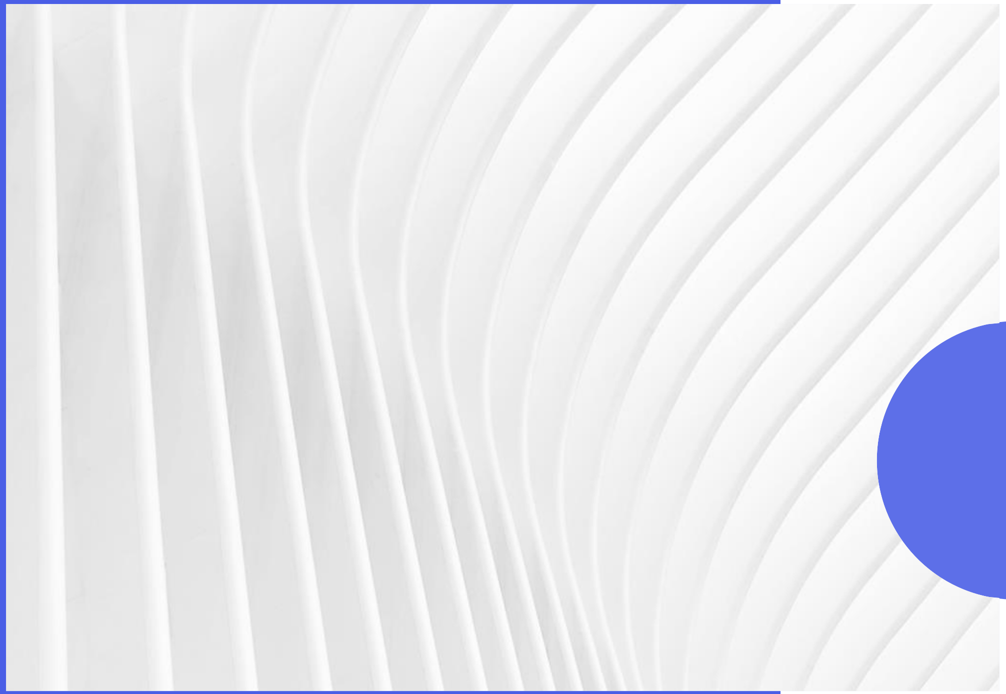
TEAM MEDBOTS

MED-NLP PROJECT

(PHASE-2)

UMBC

MOTIVATION



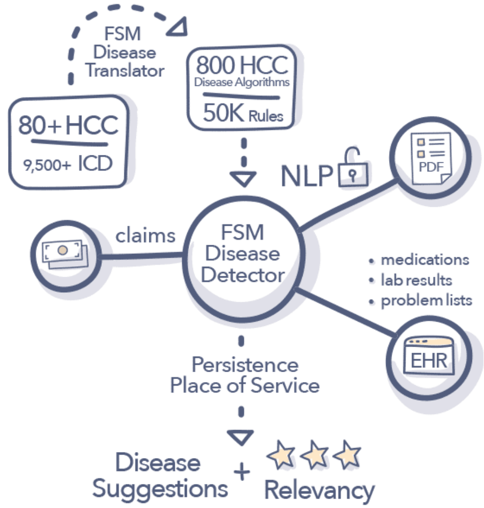
"Many online patient reports are not coded but are recorded in natural-language text that cannot be reliably accessed. Natural language processing (NLP) can solve this problem by extracting and structuring text-based clinical information, making clinical data available for use."[1]

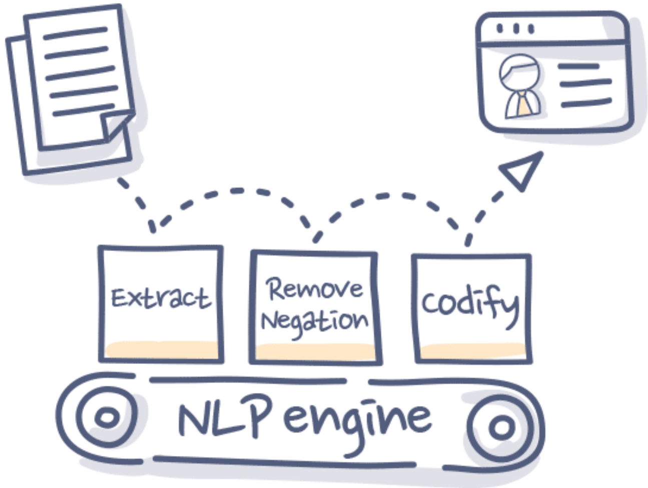
-Friedman C. & Hripcsak G.

PLAYERS IN THIS INDUSTRY

Foresee Medical

- ForeSee Medical's unique combination of machine learning technology and risk adjustment rules delivers industry leading NLP accuracy scores.
- How it works:

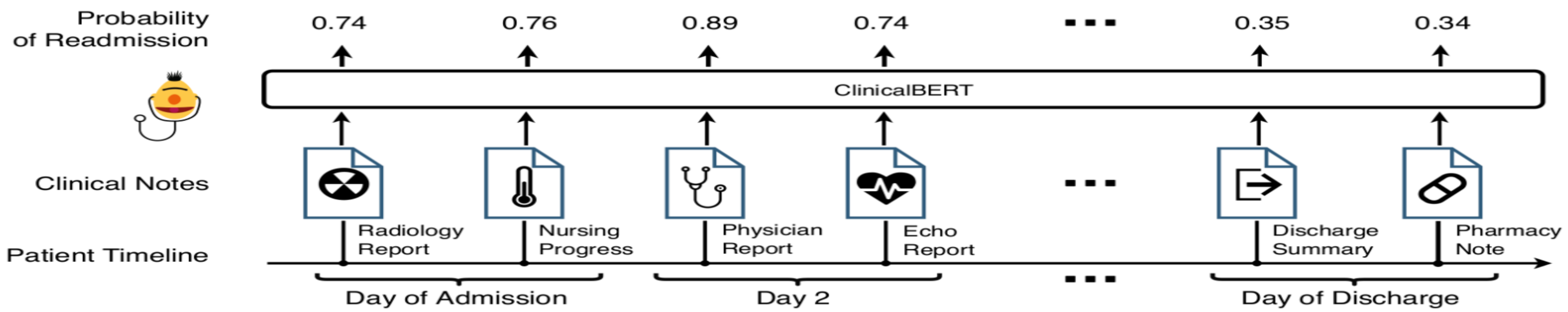


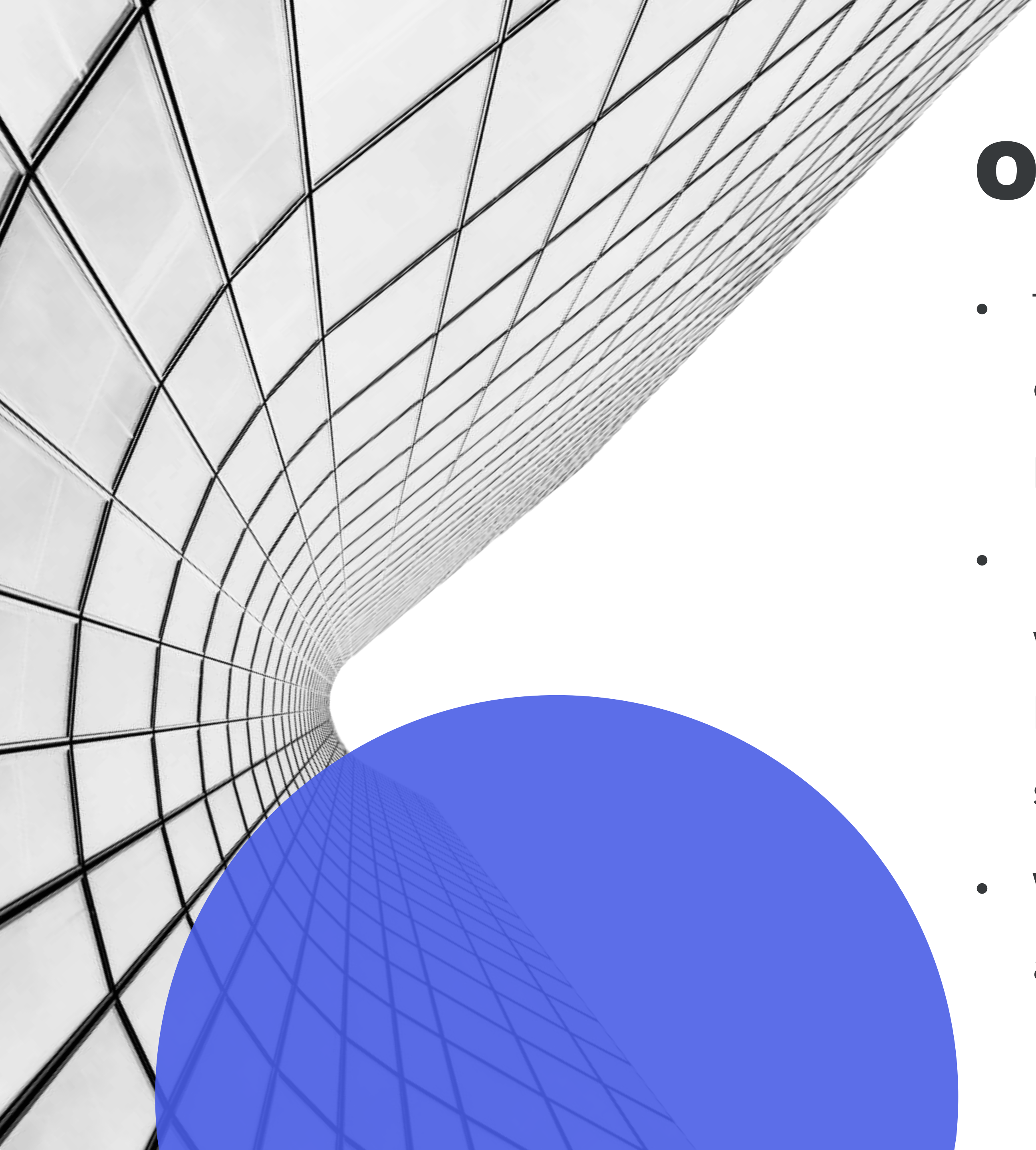


REFERENCE RESEARCH

ClinicalBERT

- ClinicalBERT is a machine learning model that uses clinical notes and electronic health records (EHRs) to predict hospital readmissions.
- The model is based on BERT pre-trained language model and fine-tuned on EHRs to capture specific language used in clinical notes.
- The evaluation of ClinicalBERT on a large dataset of patient records from two hospitals showed that it outperformed several other baseline models in predicting readmissions within 30 days of discharge.
- The research suggests that ClinicalBERT and other machine learning models have the potential to improve healthcare outcomes by helping clinicians identify high-risk patients and prevent hospital readmissions.





OUR PRODUCT

- The healthcare sector being very vast, we decided to focus on solving a specific problem at first.
- Hence, we plan on building a product which would be capable of predicting patient hospital readmission with their discharge summary.
- We believe this would help the patient as well as the hospital to plan its resources.

PRODUCT DEVELOPMENT PLAN

RESEARCH

Conduct brief **literature/industry research** to determine similar projects.

Learn from those studies' outcomes and differentiate their project than ours.

PLANNING

Plan and document the details of the planned implementation.

Get familiar with the **datasets** and carry out transformations and cleansing.

Carry out some basic **exploratory data analysis**.

DESIGN

Getting the dataset completely ready after appropriate **cleansing and transformations**.

Getting familiar with all the major **patterns and trends** in dataset.

DEVELOPMENT

Construct the model.

Produce concrete **outcomes**.

Focus on **performance** of the model.

LAUNCH

Deploy the tool on **Streamlit cloud.**

TIMELINE

FEB 6TH — **Team** creation and **project** selection

FEB 27TH — Discuss initial **EDA** findings

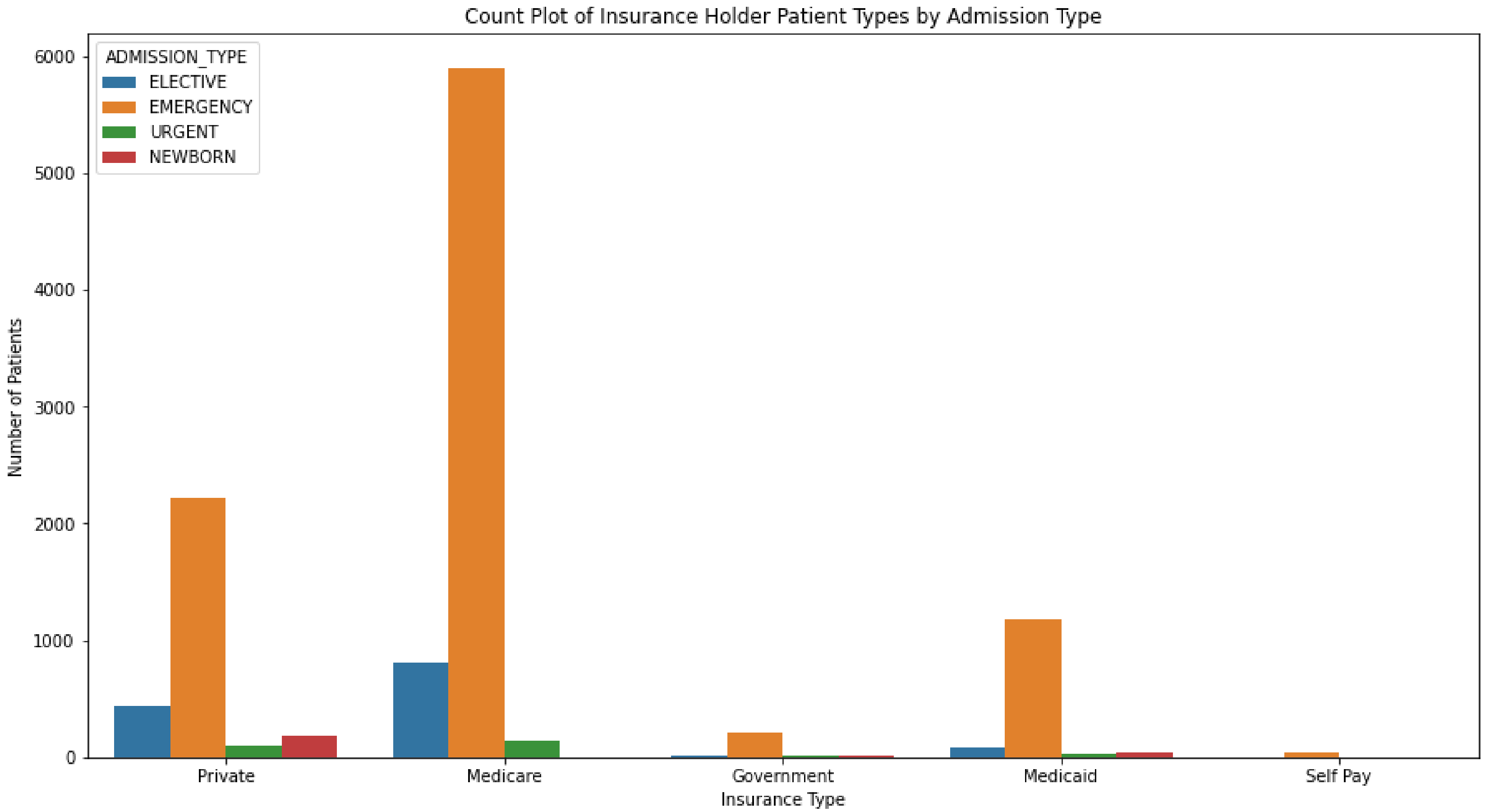
MAR 13TH — Present **Phase-1** presentation, start **design and development** stage

MAR 27TH — Present **Phase-2** presentation, complete **model evaluation**

MAY 1ST — Present **Phase-3** presentation, execution and interpretation, **deploy product**

ADDITIONAL EDA

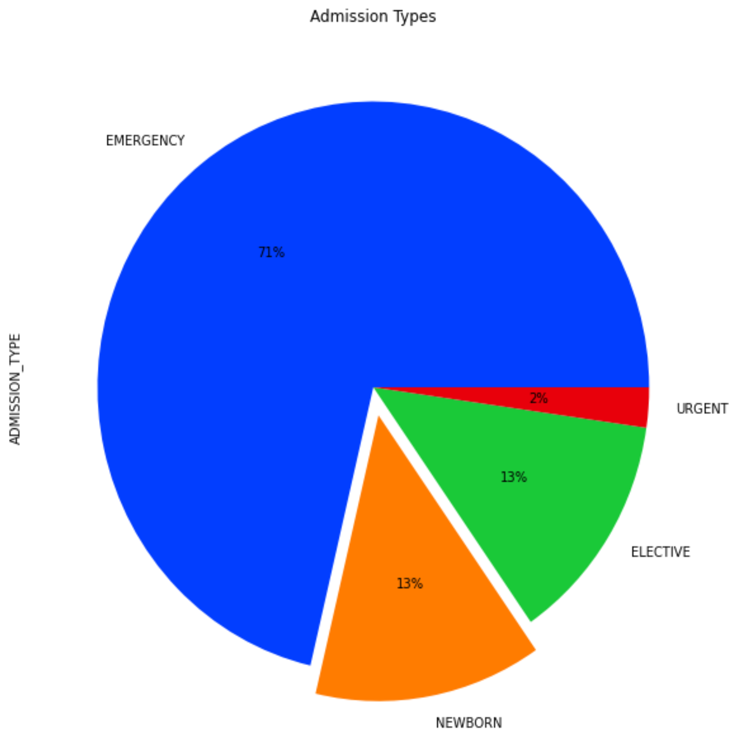
Insurance Holder Patient Types by Admission Type

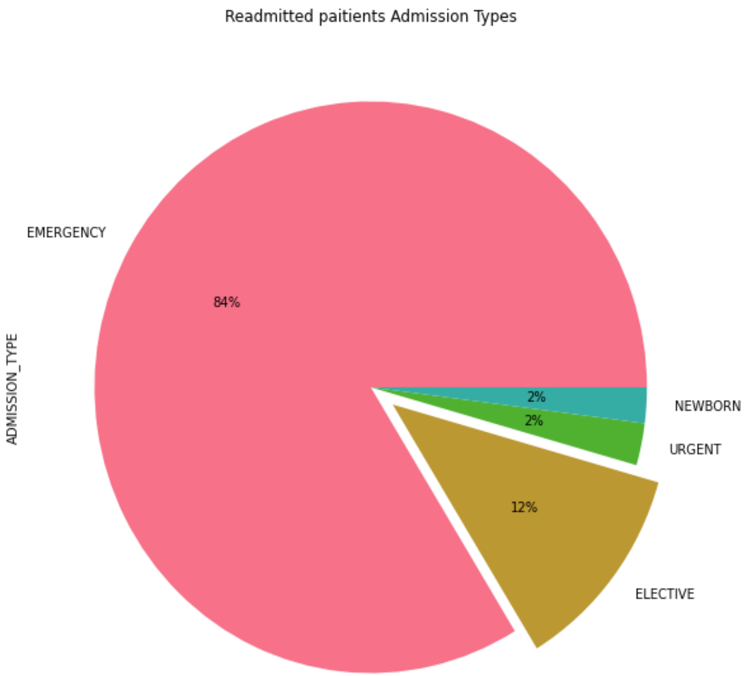


Interpretation:

- Medicare insurance is most common among Emergency, Elective, and Urgent admissions, while Private insurance is most common among Newborn admissions.
- Government insurance holders are few, and Self-paid patients are negligible across all admission types.

Admission Types of New and Readmitted Patients





Interpretation:

- Among the first-time admitted patients, around 71% were admitted as Emergency cases, followed by Elective and Newborn with 13% each. Only 2% of the first-time admissions were Urgent cases.
- On the other hand, in Readmitted patients, around 84% were Emergency readmissions (Unplanned readmissions), indicating that the majority of the readmitted patients require urgent medical attention. Elective readmissions remained unchanged with 12%, while Newborn and Urgent readmissions constituted only 2% each.

MODELLING

```
# logistic regression
from sklearn.linear_model import LogisticRegression
clf=LogisticRegression(C = 0.0001, penalty = 'l2', random_state = 42)
clf.fit(X_train_tf, y_train)
```

Interpretation:

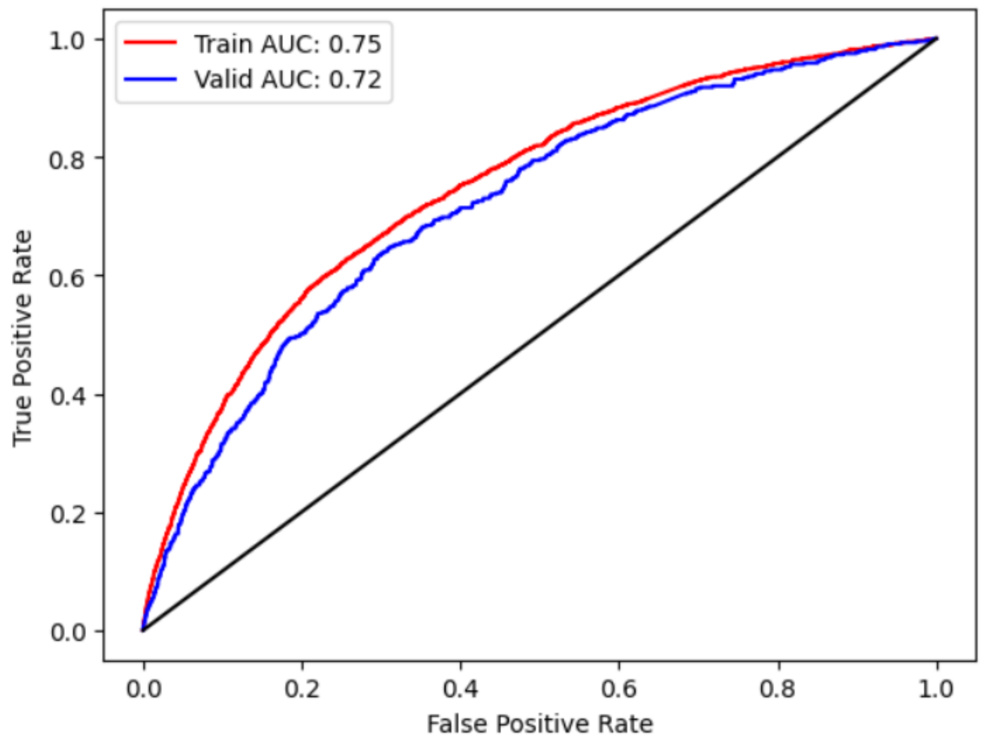
- Since our problem is a classification problem, we started off with preliminary data preprocessing.
- Followed by building a Logistic Regression Model. (Log Reg Model serves as a good baseline model)

EVALUATION METRICS

Following are the scores:

- Accuracy: 94.3%
- AUC: 71.9%

	precision	recall	f1-score	support
0	0.94	1.00	0.97	33687
1	0.50	0.00	0.01	2092
accuracy			0.94	35779
macro avg	0.72	0.50	0.49	35779
weighted avg	0.92	0.94	0.91	35779



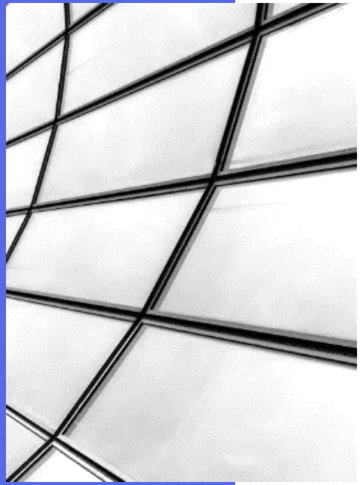
REFERENCES

• Friedman, C., & Hripcsak, G. (1999). Natural language processing and its future in medicine. Academic medicine : journal of the Association of American Medical Colleges, 74(8), 890–895. https://doi.org/10.1097/00001888-199908000-00012

• Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

DATA INFO

- MIMIC-III is a relational database consisting of 26 tables (https://physionet.org/content/mimiciii/1.4/).
- Database size is approximately 20GB, where as ADMISSIONS.csv is 12MB and NOTEEVENTS.csv is 3.9GB.
- 2.08 Million records present in NOTEEVENTS.csv



MEET OUR TEAM

Suhetu Ring

Student, UMBC

Harsha Vanga

Student, UMBC

Harshit Shrimali

Student, UMBC

THANK YOU

Team MedBots