The Thesis Committee for Sanjana Tripathi
Certifies that this is the approved version of the following Thesis:

Exploring the effect of Discharge Summaries for the Prediction of 30-day unplanned patient readmission to the ICU

APPROVED BY
SUPERVISING COMMITTEE:

Ying Ding, Supervisor

James Howison

Exploring the effect of Discharge Summaries for the Prediction of 30-day unplanned patient readmission to the ICU

by

Sanjana Tripathi

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Information Studies

The University of Texas at Austin

May 2021

Acknowledgements

I wish to express my gratitude to my supervisor and mentor, Dr. Ying Ding, who guided my throughout this project. I would also like to thank Dr. James Howison who agreed to read the thesis and provide his valuable feedback.

I also wish to acknowledge the help provided by everyone at the School of Information at the University of Texas at Austin. I would also like to show my deep appreciation to my family and friends who supported me throughout the research study in these tough times.

Abstract

Exploring the effect of Discharge Summaries for the Prediction of 30-day unplanned patient readmission to the ICU

Sanjana Tripathi, MSInfSt

The University of Texas at Austin, 2021

Supervisor: Ying Ding

Healthcare is transforming into a data-intensive industry with the expectation to double its own data every 73 days by 2020. Electronic Health Records hold a vast amount of information that has the potential of improving care delivery ranging from management tasks in hospitals to inferring diagnoses from X-ray images. The massive volume of data, such as demographic data, diagnoses, tests, prescribed medications, and procedures, can be used to predict health risk or diagnose diseases. But few pay attention to the medical notes which contain abundant and critical information written by healthcare service providers during a patient's stay or visit to the hospital. Because of the unstructured feature in these notes, they are usually underutilized to build prediction models. This project incorporates medical notes (e.g., discharge notes) along with demographic data available in the MIMIC-III dataset, to visualize patterns and finally train a prediction model for readmission of patients in the ICU.

Table of Contents

Chapter 1. Introduction……01

1.1 Clinical Relevance of Hospital Readmission Predictions……03

1.2 Discharge Summary……04

Chapter 2. Literature Review……..08

2.1 Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach ……08

2.2 Effect of Discharge Summary Availability During Post-Discharge Visits On Hospital Readmission……..09

2.3 Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory……10

2.4 Racial/Ethnic Disparities in Readmissions in US Hospitals: The Role of Insurance Coverage……10

2.5 Patient Readmission Rates for all Insurance Types after Implementation of theHospital Readmissions Reduction Program……12

Chapter 3. Data Exploration and Visualization……..14

3.1 MIMIC-III Dataset……..14

3.2 ADMISSIONS table……16

3.2.1 Readmission Distribution……17

3.3 NOTEVENTS Table and Clinical Notes……19

3.4 Demographic Data and Readmission……..22

Chapter 4. Data Preparation……30

Chapter 5. Natural Language Processing……32

Chapter 6. Data Pre-processing…….. 34

6.1 scispaCy Model……34

6.2 Word Vectors……36

6.3 Zipf's Law……38

Chapter 7. Prediction Models……..41

7.1 Logistic Regression……..41

7.2 Gaussian Naive Bayes Model……..42

7.3 Support Vector Machine……..43

7.4 AdaBoost Classifier…….45

7.5 BERT…….. 46

Chapter 8. Method……..48

Chapter 9. Results……..51

Chater 10. Conclusion……61

Appendix……63

A. Discharge Summary…….63

B. MIMIC-III Schema……. 64

References……..……65

Chapter 1: Introduction

Medical notes have held enormous amounts of important clinical information but went underutilized by the data scientists in healthcare. Fortunately, these notes are now available in an electronic form as a part of EMR (Electronic Medical Record) ready for analysis. With the current advancement in deep learning and the availability of state-of-the-art Natural Language Processing (NLP) tools, medical notes can be utilized for improving patient care as well as hospital management (Long, 2018). Researchers, over the years, have published models and articles around predicting length-of-stay and hospital readmission for patients utilizing EHR data about patient history (Desautels et al., 2017), medical procedures (Lin et al., 2019), and demographic information (Lin et al., 2019;Ferro et al., 2019). Few have included the diagnosis notes into their prediction models (Walraven et al., 2002). This project aims to demonstrate the value of adding diagnosis notes into the hospital readmission prediction for patients using EHR data, especially patients' discharge summaries. "A discharge summary is a clinical report prepared by a health professional at the conclusion of a hospital stay or series of treatments. It is the primary mode of communication between the hospital care team and aftercare providers, which can be considered as a legal document and has potential to jeopardize the patient's care if errors are made" (Kamalodeen, 2020, para 1).

Rajkomar et al. (2018) built state-of-the-art deep learning models to predict: 1) In-Hospital Mortality (AUC = 0.93–0.94), 2) 30-Day Unplanned Readmission (AUC =

0.75–76), 3) Prolonged Length of Stay (AUC = 0.85–0.86), and 4) Discharge Diagnoses (AUC = 0.90). AUC (Area under the ROC Curve) measures the area underneath the entire ROC curve with the range from 0 to 1 (Google Developers, 2020), with 1 as the best. Predicting readmissions is the hardest task out of the four since the AUC value is lowest (Rajkomar et al., 2018; Deschepper et al., 2019; Zhao, 2021). This project predicts readmissions for patients with features from discharge summaries and shows the improved performance.

While there is not a standard, pre-defined performance measure for the clinical interpretation of readmission prediction, AUC score for a ROC curve is a widely used metric for evaluating the prediction models in the domain (Rajkomar et al., 2018). AUC value of a ROC curve helps determine the model's ability to correctly classify instances in the different classes (here, readmission and no readmission) (Google Developer, 2020). The choice of AUC of ROC as the performance measure is useful and relevant since identifying the patients with a risk of short-term readmission is a first step to identifying and strategizing after-care practices. Although the exact clinical use of these models is as yet unknown, one possibility is to conduct greater follow contact with the group assessed as likely to be readmitted. For this reason, it may be more important to contact as many patients as possible who will likely be re-admitted, even at the cost of contacting more of those who are less likely to be admitted. This implies trading higher recall for lower precision in predicting readmissions.

1.1 Clinical Relevance of Hospital Readmission Predictions

The financial cost of hospital readmissions is estimated to be about $26 billion annually (Wilson, 2019). The emotional cost of these hospital readmissions is generally ignored. It is noteworthy that patients felt that most of their readmissions were caused by issues in "discharge timing, follow-up, home-health, and skilled services" which could be prevented (Smeraglio et al., 2019; Healthstream, 2020, para. 5). Smeraglio et al. (2019) found that review by a Registered Nurse (RN) case manager found that 49% of readmissions the hospital system had some amount of opportunity to improve the discharge process. The RN case managers more often agreed with the patient's perspective of readmission than the provider's (Healthstream, 2020, para. 5). Furthermore, the burnout of a care provider with high patient volumes and inadequate support, could cause problems in discharge planning, care transitions, and patient education; which leads to the increased probability of hospital readmissions (Healthstream Blog, 2020).

Additionally, the rate of mortality associated with ICU readmissions ranges from 26% to 58%. There has been a growth trend in the ICU readmissions seen from 1989 to 2003 with the readmission rate rising from 4.6% to 6.4% (Lin et al., 2019). The rate of readmission to the ICU reflects poorly on the performance of the ICU facility and service. In order to reduce ICU readmissions, patients at high risk of readmission should be identi-

fied beforehand and taken care of. This will also save manpower and other medical resources incurred by the hospital during readmission (Lin et al. 2019). A model that can accurately predict the chance of a patient readmission can be beneficial for both healthcare service providers and patients.

A patient discharged from the hospital remains in highly vulnerable and stressful state marked by physiological distress, health impairments and psychological impact of the illness and hospitalization (Lehn et al., 2019). Short term or 30-day readmission is usually categorized by worsening of the existing conditions, new impairments as a result of improper after-care, longer stays in the ICU, increased risk of mortality and higher financial costs (Li et al., 2019). A 14-day window has been identified where most of the unplanned readmissions occur due to improper or no follow-up with the Primary Care Physician (Meyers, & Brady, 2020). Various regulations around discharge procedures and costs around readmissions and related insurance have been passed to regulate the rate of readmission in the United States. In the MIMIC-III dataset used in the project, the days between readmissions peaks around the window of 30 days (see Figure. 3).

1.2 Discharge Summary

Clinical notes are written by healthcare professionals in the form of free text which are unstructured but contain a richer and denser profile of a patient than other kinds of EHR data. There are numerous clinical notes associated with a patient's stay or treatment. This project choses clinical notes in the form of *Discharge Summaries* to predict the possibili-

ty of hospital readmission. Discharge summaries are believed to improve the efficiency of hospital readmission prediction models, especially in shorter time frames.

Discharge summary is a kind of clinical note prepared by a healthcare professional for a patient at the end of a patient's stay at the hospital or series of treatment. Discharge summaries are particularly important as they are the primary source of information for the aftercare service providers about the patient's treatment at the hospital and the primary mode of communication between the patient's healthcare service providers and aftercare service providers. Appendix A shows the image of a discharge summary for a patient with a ruptured appendix. A discharge summary typically includes: patient information, healthcare provider's information, patient history, allergies, diagnoses, medication, investigations and procedures, management of the patient stay or treatment, and any complications that arose. The joint commission has mandated that all hospitals in the United States follow a structure while creating a discharge summary (Kind & Smith, 2008). Kind & Smith (2008) in their work mention that the mandated summary structure must have the following six components:

A. Reason for hospitalization
B. Significant findings
C. Procedures and treatments provided
D. Patient and family instructions
E. Attending physician's signature. (p. 1)

An example of discharge summary in the MIMIC-III dataset is shown below:

> *"Admission Date: [**2151-7-16**] Discharge Date: [**2151-8-4**] Service: ADDENDUM: RADIOLOGIC STUDIES: Radiologic studies also included a chest CT, which confirmed cavitary lesions in the left lung apex consistent with infectious process/tuberculosis. This also moderate-sized left pleural effusion. HEAD CT: Head CT showed no intracranial hemorrhage or mass effect, but old infarction consistent with past medical history. ABDOMINAL CT: Abdominal CT showed lesions of T10 and sacrum most likely secondary to osteoporosis. These can be followed by repeat imaging as an outpatient. [**First Name8 (NamePattern2) **] [**First Name4 (NamePattern1) 1775**] [**Last Name (NamePattern1) **], M.D. [**MD Number(1) 1776**] Dictated By:[**Hospital 1807**] MEDQUIST36 D: [**2151-8-5**] 12:11 T: [**2151-8-5**] 12:21 JOB#: [**Job Number 1808**]"* (extracted from MIMIC-III dataset)

Other key information in the discharge summaries that makes them an indispensable part of the EHR and of a biomedical prediction model for patients are: "identification of unresolved medical issues at the time of discharge, test results requiring follow-up, and the presence of an accurate discharge medication list" (Legault et al., 2012, para. 1). It is also anticipated that post discharge adverse drug events are a factor in morbidity and mortality

and while the event is predictable it requires better medication documentation, reconciliation, and management (Legault et al., 2012). A study conducted by van Walraven and colleagues (2002) found that patients whose discharge summaries arrive at the PCP office before the first outpatient visit were at 0.74 times relative lower risk of hospital readmission.

Chapter 2: Literature Review

2.1 Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach

Desautels et al. (2017) trained and tested a model for the prediction of unplanned readmission of patients to ICU and deaths inside the hospital within 48 hours of the first time ICU discharge from the hospital. The study uses a dataset comprising 3,326 ICU episodes from the Cambridge University Hospitals NHS Foundation Trust (CUH) collected between 2014 and 2016 for patients above the age of 16 years old and with at least one episode of admission to the ICU. The feature set for training the model is made up of age of the patient, vital signs (blood pressure, heart rate, temperature, pulse pressure, respiration rate, SpO2 level), lab tests records (bilirubin, creatinine, international normalized ratio (INR), lactate, WBC count, platelet count, pH level), FiO2 and total Glasgow Coma Score (GCS). Each patient in the dataset had at least one of the vital sign measurements and GCS. An ensemble of classification models using AdaBoost was trained on the dataset by dividing the data into 10 cross-validation folds and the results across all folds were combined for the evaluation of model performance using the AUROC curve value. In the work while the choice of model is a classic yet smart, the data used for the prediction is insufficient in terms of patient demographic information and is totally categorical ignoring the large amounts of data hidden in clinical notes written and prepared by healthcare service providers during the course of treatment and after the stay.

2.2 Effect of Discharge Summary Availability During Post-Discharge Visits on Hospital Readmission

In this study, Walraven et al. (2002) studied the data collected from patients who participated in a clinical trial between 1996 and 1997 at the Ottawa Civic Hospital. Walvaren et al. collected the discharge summaries of each patient and as a part of the experiment determined if the discharge summary successfully reached the patient's Physician before the first outpatient visit. The study observed and recorded the first non-elective readmission of each patient to the hospital within 90 days of discharge. The outcome was determined when the patient died, was urgently readmitted to the hospital, or at least three months after discharge of the patient (Walraven 2002). The nature of this study is exploratory and observatory, and does not predict any outcome. The authors analyzed the correlation between the post-discharge communication between the hospital and the physician through discharge summaries of the patients and their non-elective readmission to the ICU within 90 days. The factors observed to determine the association are: admission and discharge dates, patient age, patient gender, if the patient lived in a nursing home or not, active medical problems, admission diagnoses, procedure complications, and socioeconomic status of the patient. While the feature set is balanced with categorical and non-categorical data instances, the focus is heavily shifted in the favor of patient problems at the time of admission. It is interesting to see that socioeconomic background of the patient is taken into consideration but it is too fuzzy a variable to base trained predictions upon.

2.3 Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory

The study uses (Lin et al., 2019) supervised machine learning models on "comprehensive, longitudinal clinical data" in the MIMIC-III dataset for the 30-day unplanned readmission of a patient to the ICU. The features extracted from the dataset include chart events, patient demographic information, and chronic diseases. Through chart events, the authors extracted patients' physiological conditions like blood pressure and categorical items (e.g. capillary refill rate). For chronic diseases, embeddings (vectors) of ICD-9 codes are used, and for demographic information, the patient's gender, age, race and ethnicity are considered. Three different types of models are used, namely Logistic Regression with L1 and L2 regularization, Convolutional Neural Network and a bidirectional Long Short-Term Memory model. The LSTM model reaches a sensitivity of 0.742 with the current feature set for the prediction of short-term unplanned readmission. With complex and hardware intensive models like LSTM and CNN, the pipeline still reaches an efficiency that is achievable by using baseline methods but with a more advanced feature set as shown in this study.

2.4 Racial/Ethnic Disparities in Readmissions in US Hospitals: The Role of Insurance Coverage

This study investigates the rate and risk of readmissions and its association with the race and ethnicity of the patients which could be due to "limited access to post-discharge care,

disparities in healthcare quality, and socioeconomic factors" (Basu et al., 2018). Another factor being considered in the study of readmission rate is insurance held by the patient and how the issues posed by insurance for readmission differ for minority patients based on their racial and ethnic identities. The dataset is obtained from Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID) of the agency for Healthcare Research and Quality (Basu et al., 2018). The abstracted data consists of patient's discharge data in five US states (e.g., California, Florida, Missouri, New York and Tennessee). This data is linked to the contextual and provider availability data from Area Resource File (Health Resources and Service Administration) and American Hospital Association's (AHA) Annual Survey of hospitals for data on hospital characteristics (Basu et al., 2018). The study analyzes the dependency of 30-day readmission excluding *elective* readmissions for patients 18 years and above on several independent variables focused on patient data and hospital characteristics. The independent variables related to patients used for the analysis are: age, sex, race/ethnicity, insurance type, income, number of chronic diseases on record, and health status indicators for disease severity and risk of mortality from All Patient Refined Diagnosis Related Group (APR-DRG). For the hospital characteristics, bed size, mortality rate and teaching status were considered in the dependence analysis. Some Primary Care Service Area (PCSA)-level factors like, primary care provider density, population density, PCSAs with urban/rural residence status were also accounted for in the analysis. The analysis was done and divided into two parts: 1)

direct association between probability of readmission and race/ethnicity of the patient; 2)association of readmission risk to a patient's race/ethnicity and how it varies by the insurance status considering the interactions between insurance and race/ethnicity, and insurance and readmission risk (Basu et al., 2018).

2.5 Patient Readmission Rates for all Insurance Types after Implementation of the Hospital Readmissions Reduction Program

"Hospital Readmissions Reduction Program (HRRP) was implemented in October 2012 as a part of the Affordable Care Act (ACA)" (Ferro et al., 2019, p. 2). In this research, Ferro and the co-authors (2019) studied the trend in readmission rates for patients with all types of insurance after the implementation of HRRP and compared it to the trend before the program to find a causal relationship. For the study, the dataset used is taken from the "Nationwide Readmissions Database which contains discharge data from twenty-two states accounting for 51.2% of the U.S. population and 49.3% of hospital admissions in 2014" (Ferro et al., 2019, p. 3). The authors take a statistical approach and conduct a difference-in-differences study to determine the trend and relationship between the HRRP and readmission rates for patients with different types of insurance (Medicare, Medicaid, and private). The different characteristics used in the study were: patient's age, patient's sex, twenty-nine comorbidities collected by the Nationwide Readmissions Database based on the Elixhauser Comorbidity Index, insurance status, length-of-stay, costs of index admission and readmission, hospital size and teaching and ownership status (Ferro et

al. 2019). While the study reveals some interesting insights of trends in the readmission rates, it does not shed much light on the role insurance plays in the plausible readmission of patients.

Chapter 3: Data Exploration and Visualization

3.1 MIMIC-III Dataset

MIMIC-III is the third version of MIMIC dataset which is a freely accessible critical care dataset created by MIT Lab containing anonymized health-related data from over 40,000 patients in the ICUs of the Beth Israel Medical Center between 2001 and 2012 (MIMIC-III Critical Care Database. (n.d.) MIMIC-III v1.4 documentation. https://mimic.physionet.org/about/mimic/). The birth and death dates of the patients have been timeshifted to the future to protect the identities of the patients, but the time between two consecutive events for a patient is kept the same as the original in the database. The patient information in the MIMIC-III dataset is made of demographics, vital signs, ICD code for diagnosed diseases, procedures, notes, medications, lab tests, and more. The dataset comprises of 26 different tables that can be largely grouped into four broad categories based on the kind of data they hold -

1. Patient Tracking

   (ADMISSIONS, ICUSTAYS, PATIENTS, CALLOUT, TRANSFERS)

2. ICU Data

   (CHARTEVENTS, INPUTEVENTS_CV, INPUTEVENTS_MV, DATETIMEEVENTS, OUTPUTEVENTS, PROCEDUREEVENTS_MV)

3. Hospital Data

(CAREGIVERS, CPTEVENTS, DIAGNOSES_ICD, DRGCODES, LABEVENTS, MICROBIOLOGYEVENTS, NOTEEVENTS, PRESCRIPTIONS, PROCEDURES_ICD, SERVICES)

4. Dimension Tables

(D_CPT, D_ICD_PROCEDURES, D_ITEMS, D_ICD_DIAGNOSES, D_LABITEMS)

Table	Children	Parents	Columns	Rows	Comments
admissions	18	1	19	58,976	Hospital admissions associated with an ICU stay.
callout		2	24	34,499	Record of when patients were ready for discharge (called out), and the actual time of their discharge (or more generally, their outcome).
caregivers	7		4	7,567	List of caregivers associated with an ICU stay.
chartevents		5	15	330,712,483	Events occuring on a patient chart.
chartevents_1			15	38,033,561	Partition of chartevents. Should not be directly queried.
chartevents_10			15	9,584,888	Partition of chartevents. Should not be directly queried.
chartevents_11			15	470,141	Partition of chartevents. Should not be directly queried.
chartevents_12			15	265,413	Partition of chartevents. Should not be directly queried.
chartevents_13			15	39,066,570	Partition of chartevents. Should not be directly queried.
chartevents_14			15	100,075,138	Partition of chartevents. Should not be directly queried.
chartevents_2			15	13,116,197	Partition of chartevents. Should not be directly queried.
chartevents_3			15	38,657,533	Partition of chartevents. Should not be directly queried.
chartevents_4			15	9,374,587	Partition of chartevents. Should not be directly queried.
chartevents_5			15	18,201,026	Partition of chartevents. Should not be directly queried.
chartevents_6			15	28,014,688	Partition of chartevents. Should not be directly queried.
chartevents_7			15	255,967	Partition of chartevents. Should not be directly queried.
chartevents_8			15	34,322,082	Partition of chartevents. Should not be directly queried.
chartevents_9			15	1,274,692	Partition of chartevents. Should not be directly queried.
cptevents		2	12	573,146	Events recorded in Current Procedural Terminology.
d_cpt			9	134	High-level dictionary of the Current Procedural Terminology.
d_icd_diagnoses	1		4	14,710	Dictionary of the International Classification of Diseases, 9th Revision (Diagnoses).
d_icd_procedures	1		4	3,898	Dictionary of the International Classification of Diseases, 9th Revision (Procedures).
d_items	8		10	12,487	Dictionary of non-laboratory-related charted items.
d_labitems	1		6	753	Dictionary of laboratory-related items.
datetimeevents		5	14	4,485,937	Events relating to a datetime.
diagnoses_icd		3	5	651,047	Diagnoses relating to a hospital admission coded using the ICD9 system.
drgcodes		2	8	125,557	Hospital stays classified using the Diagnosis-Related Group system.
icustays	8	2	12	61,532	List of ICU admissions.
inputevents_cv		4	22	17,527,935	Events relating to fluid input for patients whose data was originally stored in the CareVue database.
inputevents_mv		5	31	3,618,991	Events relating to fluid input for patients whose data was originally stored in the MetaVision database.
labevents		3	9	27,854,055	Events relating to laboratory tests.
microbiologyevents		5	16	631,726	Events relating to microbiology tests.
noteevents		3	11	2,083,180	Notes associated with hospital stays.
outputevents		5	13	4,349,218	Outputs recorded during the ICU stay.
patients	19		8	46,520	Patients associated with an admission to the ICU.
prescriptions		3	19	4,156,450	Medicines prescribed.
procedureevents_mv		5	25	258,066	Procedure start and stop times recorded for MetaVision patients.
procedures_icd		3	5	240,095	Procedures relating to a hospital admission coded using the ICD9 system.
services		2	6	73,343	Hospital services that patients were under during their hospital stay.
transfers		3	13	261,897	Location of patients during their hospital stay.
40 Tables			**534**	**728,556,685**	

Figure 1. Overview of the tables in MIMIC-III Dataset generated by SchemaSpy. Sourced from https://mit-lcp.github.io/mimic-schema-spy/. For the legend refer to Appendix C.

The complete schema of the MIMIC-III dataset sourced from MIT in collaboration with SchemaSpy is available in Appendix B. The model in this project uses two tables - ADMISSIONS which contains data about a patient admission to the hospital, demographic data, discharge/death timings etc., and NOTEEVENTS which holds detailed notes like reports, discharge summaries etc.

3.2 ADMISSIONS Table

Column	Type	Size	Nulls	Auto	Default	Children	Parents	Comments
row_id	int4	10						Unique row identifier.
subject_id	int4	10					patients	Foreign key. Identifies the patient.
hadm_id	int4	10				callout chartevents cptevents datetimeevents diagnoses_icd drgcodes icustays inputevents_cv inputevents_mv labevents microbiologyevents noteevents outputevents prescriptions procedureevents_mv procedures_icd services transfers		Primary key. Identifies the hospital stay.
admittime	timestamp	22						Time of admission to the hospital.
dischtime	timestamp	22						Time of discharge from the hospital.
deathtime	timestamp	22	√		null			Time of death.
admission_type	varchar	50						Type of admission, for example emergency or elective.
admission_location	varchar	50						Admission location.
discharge_location	varchar	50						Discharge location
insurance	varchar	255						Insurance type.
language	varchar	10	√		null			Language.
religion	varchar	50	√		null			Religon.
marital_status	varchar	50	√		null			Marital status.
ethnicity	varchar	200						Ethnicity.
edregtime	timestamp	22	√		null			
edouttime	timestamp	22	√		null			
diagnosis	varchar	255	√		null			Diagnosis.
hospital_expire_flag	int2	5	√		null			
has_chartevents_data	int2	5						Hospital admission has at least one observation in the CHARTEVENTS table.

Figure 2. Schema of the ADMISSIONS table in the MIMIC-III dataset. Schema generated by SchemaSpy. For the legend refer to Appendix C.

This table houses information about a patient's admission data, discharge date, death date (if applicable) and demographic data like *Ethnicity, Marital Status, Gender, Insurance* etc. ADMISSIONS table has 58,976 unique admissions for 46,520 patients. Most of the patients are only admitted once. Table 1 shows the frequency of the number of admissions per patient.

Number of Admissions	Patient Count
1	38983
2	5160
3	1342
4	508
>4	527

Table 1: Number of patients that were admitted once, twice, thrice, four times or more in the ICU

Here, aligning with our goal of predicting unplanned hospital readmission within the next 30 days, this research considers the first readmission within a month (30 days) after discharge as a positive prediction.

3.2.1 Readmission Distribution

To train a model for readmission prediction, a set of ground truth has been built. The ground truth for this data is the True (1) or False (0) label for readmission per each hospital admission which were generated using the data in the ADMISSIONS table. The labels generated focus on readmission within 30 days for this project. If a patient has 1 or True as his/her readmission label, then all notes associated with that patient, in the NOTEEVENTS table, will be assigned the readmission label as True or 1. Table 2 shows the total number of patients in the dataset and the number of readmissions.

Total number of admissions	58976
Number of Readmissions	**11399**

Table 2: Total number of admissions and readmissions.

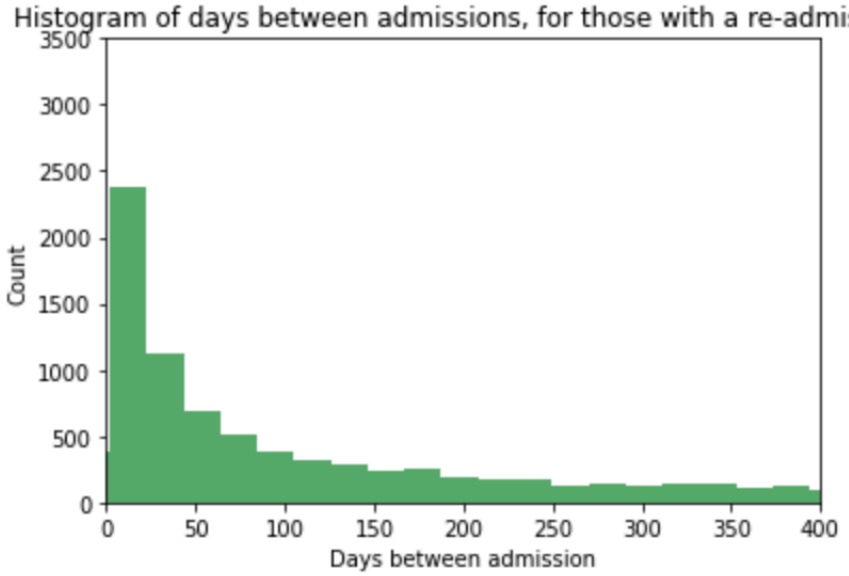


Figure 3. Readmission counts based on days between admissions

Figure 3 shows the counts of number of readmissions based on the number of days between two admissions for patients who were readmitted to the hospital due to emergency. This figure shows that most readmissions happened within 30 days of discharge. This inference, supported with the claim stating the use of discharge summaries in predicting readmission for a shorter time frame, I decided to generate and use labels for readmission within 30 days. No NEWBORN admissions are included in the dataset.

3.3 NOTEEVENTS Table and Clinical Notes

Column	Type	Size	Nulls	Auto	Default	Children	Parents	Comments
row_id	int4	10						Unique row identifier.
subject_id	int4	10					patients	Foreign key. Identifies the patient.
hadm_id	int4	10	√		null		admissions	Foreign key. Identifies the hospital stay.
chartdate	timestamp	22	√		null			Date when the note was charted.
charttime	timestamp	22	√		null			Date and time when the note was charted. Note that some notes (e.g. discharge summaries) do not have a time associated with them: these notes have NULL in this column.
storetime	timestamp	22	√		null			
category	varchar	50	√		null			Category of the note, e.g. Discharge summary.
description	varchar	255	√		null			A more detailed categorization for the note, sometimes entered by free-text.
cgid	int4	10	√		null		caregivers	Foreign key. Identifies the caregiver.
iserror	bpchar	1	√		null			Flag to highlight an error with the note.
text	text	2147483647	√		null			Content of the note.

Figure 4. Schema of the NOTEEVENTS table in the MIMIC-III dataset. Schema generated by SchemaSpy. For the legend refer to Appendix B.

The NOTEEVENTS table contains clinical notes associated with each admission. These notes are unstructured texts grouped under different categories. Table 3 shows the different categories of clinical notes in the table and their counts for each hospital stay.

Category of Associated Clinical Note	Number of Admissions
Case Management	954
Consult	98
Discharge summary	59652
ECG	138190
Echo	34037
General	8209
Nursing	220758
Nursing/other	821258
Nutrition	9378
Pharmacy	102
Physician	140100
Radiology	378920
Rehab Services	5409

Respiratory	31667
Social Work	2612

Table 3: Distribution of different types of clinical notes by admissions in the ICU.

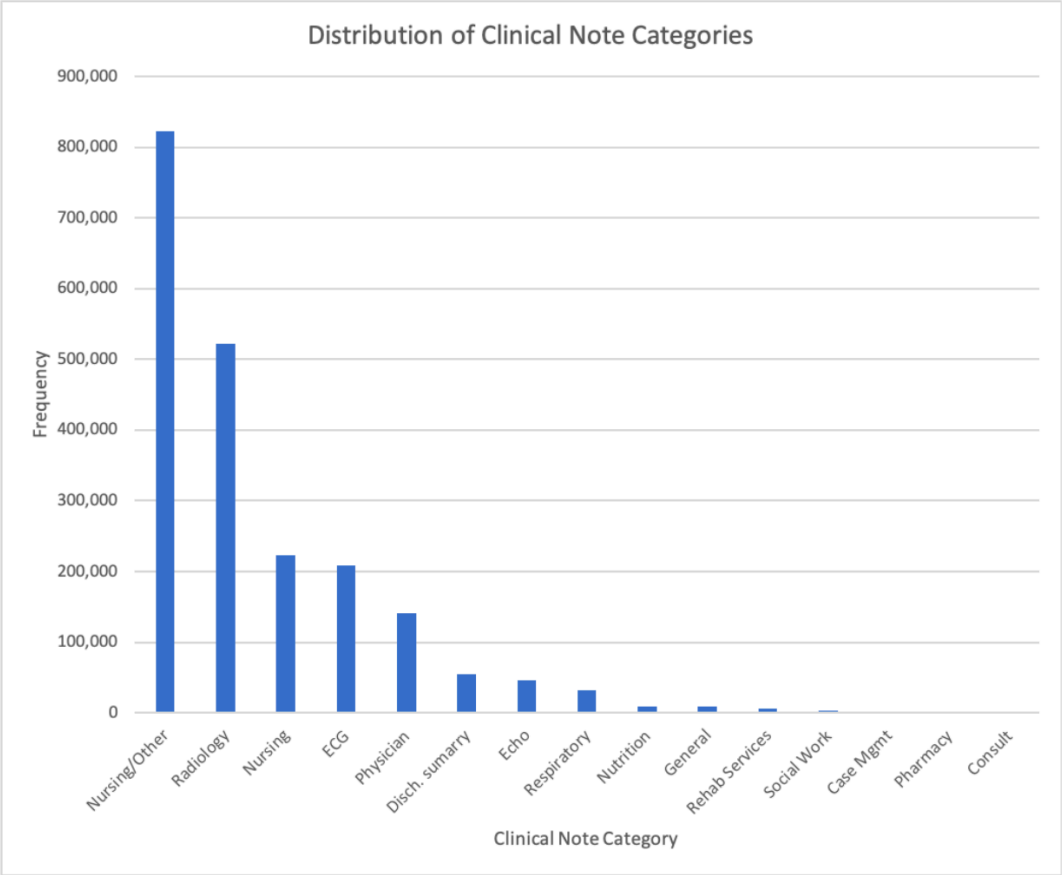


Figure 5. Graphical representation of the frequency distribution of different types of clinical notes by admissions in the ICU

Most categories in the clinical notes focus on the categorical variables and have already been used in previous research mentioned in the literature review. While each category is important in its own standing, and researchers can argue for and against each, one of the clinical notes often ignored are the discharge summaries. Many researchers have explored and successfully established that patient's physicians not having access to

patient's discharge summary before their first outpatient visit after discharge from the ICU is related to the unplanned readmission of patients to the ICU (Walraven et al., 2002). Taking into account the clinical relevance of discharge summaries for after patient care, in predicting readmission for shorter time frame and the presence of a sufficient number of discharge summaries in the dataset, I decided to move forward with using discharge summaries as my chosen unstructured text data for readmission prediction. All patients without a readmission had one or more discharge summaries associated with them (see Figure 6).

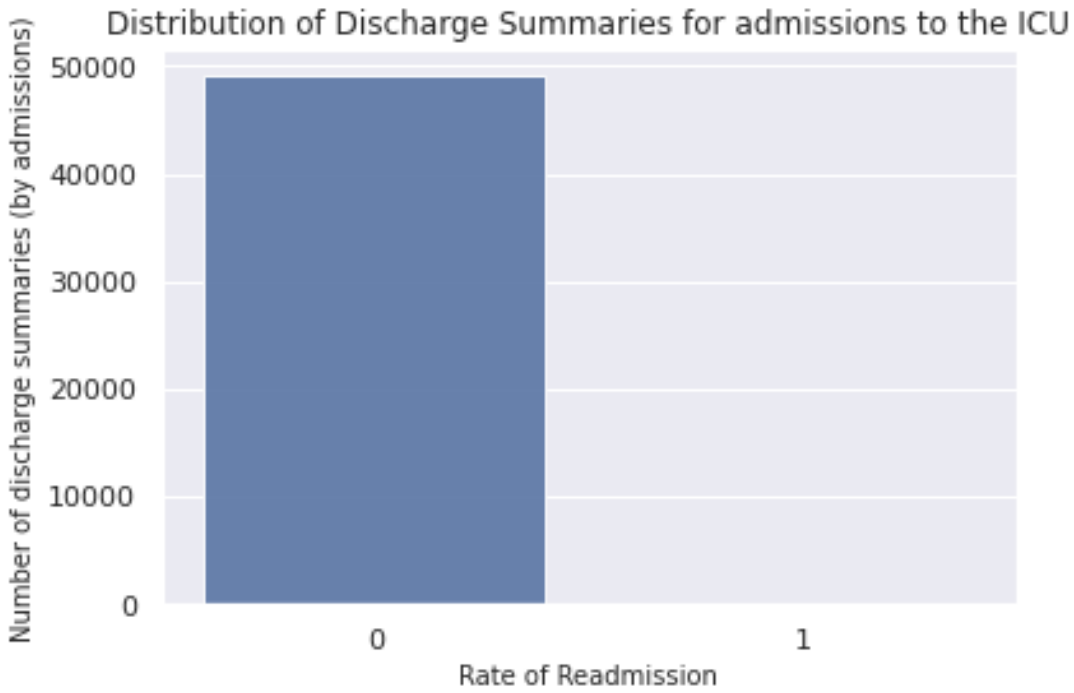


Figure 6. Distribution of discharge summaries for readmitted and not readmitted patients.

3.4 Demographic Data and Readmission

To get a better understanding of the data, specifically how the rate of readmissions relates to the different attributes of the patients' data, I started with creating various visualizations of the data contained in the MIMIC III dataset. Figure 7 shows the distribution of patients based on their insurance providers. We can see there is a clear majority of patients with Medicare, followed by privately insured patients. The number of self-paying or patients with no insurance is significantly less and nearly tending to zero.

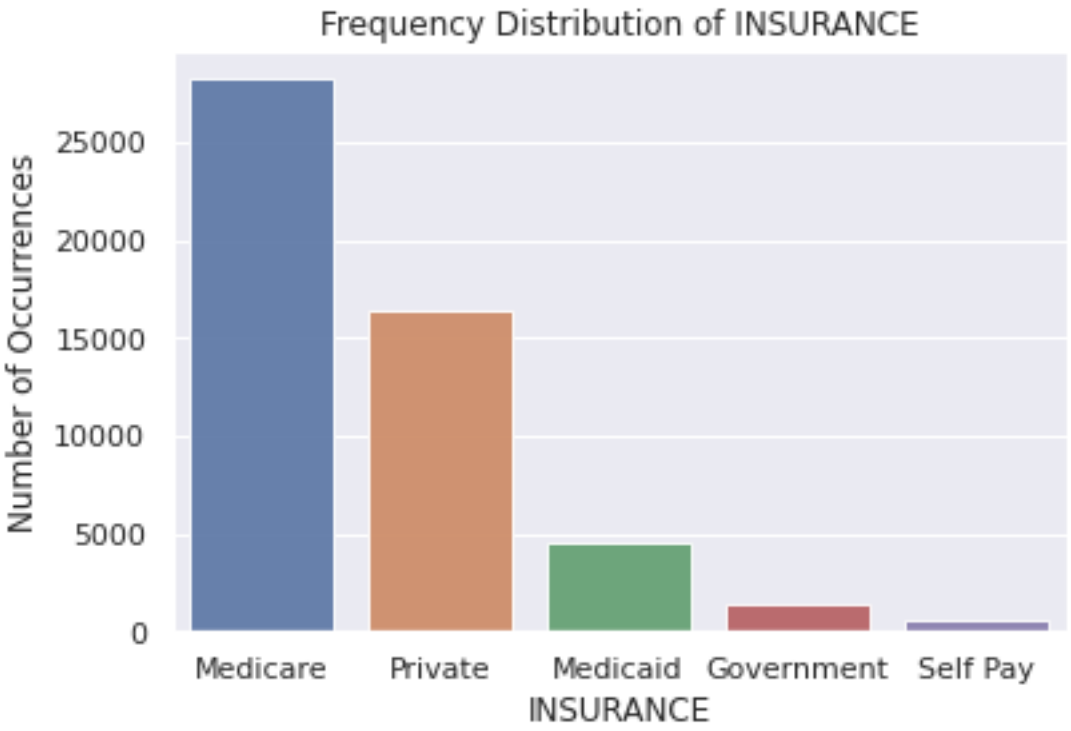


Figure 7. Graphical representation of patients by the type of insurance providers

Starting with simple visualization approaches, I chose to create the correlation heatmap of different features. Correlation heatmap is popularly used in statistical analysis and machine learning models. The features considered here were on the basis of existing

research on patient ethnography and readmission relationship. Figure 8 shows the correlation between the demographic attributes and the patient readmission prediction variable. There is very little correlation between the different variables considered here.

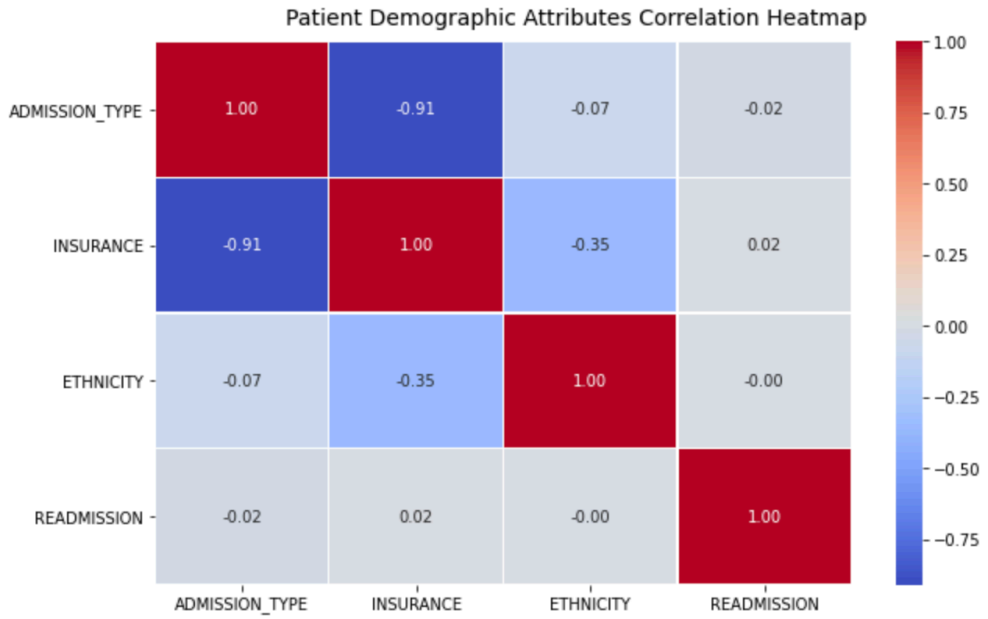


Figure 8. Correlation heatmap of patient demographic data and readmission prediction variable.

Whitney and Chuang (2016) found that patients in a hospice program eligible for both Medicare and Medicaid have lower rates of 30-day readmission because insurance that makes post-discharge custodial care accessible decreases the possibility of readmission in older patients. Ferro et al. (2019) found that the implementation of HRRP (Hospital Readmissions Reduction Program) was associated with a decrease in readmissions for both Medicare and Medicaid patients with target condition however patients with private insurance (with the lowest aggregate readmission rates during the course of the study) did

not see a decrease in the readmission rate after the HRRP implementation on a composite level.

In a study conducted by Jayasree Basu, Amresh Hanchate and Arlene Bierman (2018), the authors explored through a regression analysis the association between the likelihood of 30-day readmission for any disease or cause and insurance type and race of patients (above 18 years of age) in California, Florida, Missouri, New York and Tennessee. When comparing the insurance types, patients without insurance had a less likelihood of readmission in all states as compared to patients with private insurance. And patients with Medicaid and Medicare were much more likely to be readmitted in contrast with the privately insured patients in all five states.

	California (n = 2 443 046)	Florida (n = 1 930 337)	New York (n = 1 807 383)	Tennessee (n = 515 960)	Missouri (n = 609 560)
Mean probability of 30-day readmission (%)	0.05	0.06	0.04	0.06	0.06
Patient characteristics					
Demographics					
Female (%)	0.62	0.59	0.59	0.61	0.60
Age (year)	55.24	58.55	58.05	57.07	57.03
White (%)	0.53	0.67	0.62	0.79	0.83
African American (%)	0.08	0.16	0.17	0.18	0.14
Hispanic (%)	0.28	0.12	0.12	0.02	0.01
Other race (%)	0.11	0.05	0.09	0.01	0.02
Privately insured (%)	0.29	0.25	0.30	0.36	0.28
Medicare (%)	0.39	0.49	0.45	0.50	0.48
Medicaid (%)	0.23	0.14	0.19	0.15	0.16
Uninsured (%)	0.04	0.06	0.04	0.07	0.05
Other pay (%)	0.05	0.06	0.02	0.02	0.03
No. of chronic conditions (n)	3.90	4.60	3.95	4.71	4.72

Figure 9. Mean of independent attributes/variables categorized by state (Base et al,. 2018)

In the same study, the data for race or ethnic groups is shown in Figure 9. Comparing Hispanic and non-Hispanic white patients, Hispanic patients had significantly lower rates of readmission in four out of five states (e.g., California, Florida, Missouri and Tennessee, Basu et al. 2018). For black patients, the risk-adjusted likelihood of readmission as compared to whites was higher on a composite level, and higher specifically in California, New York and Tennessee (Basu et al. 2018). The ADMISSIONS table in the MIMIC-III dataset being used for the readmission prediction in this project contains patient ethnographic attributes like: race/ethnicity, language, marital status, religion and insurance, and the diagnosis of the patient. Since the correlation heatmap did not provide

any insight into the dependence of readmission on demographic and ethnographic attributes of a patient, visualization of the association of readmission labels to the aforementioned independent attributes using PCA (Principal Component Analysis), Parallel Coordinates and t-SNE provided some clarity.

For PCA dimensionality reduction and visualization, we encode the values using a Label Encoder and then use min-max scaling as the normalization technique. Figure 10 visualizes the output variables for a feature set with dimensionality reduced to two components.

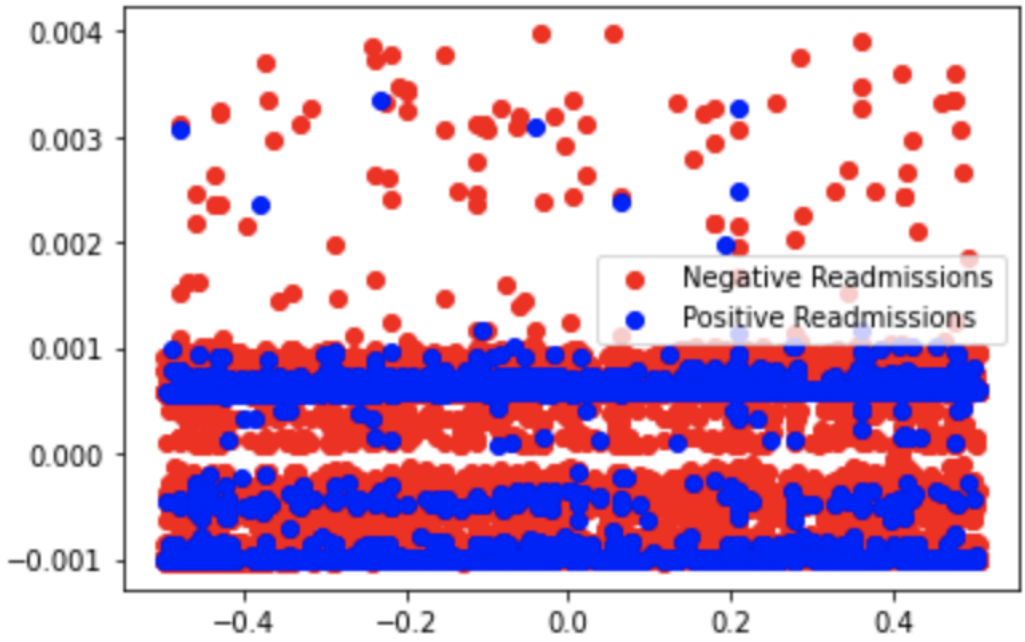


Figure 10. PCA graph showing the analysis between the demographic features and readmission variable reducing the demographic features set to a dimensionality=2

Figure 11 shows a parallel coordinate visualization of the readmission output variable (0: negative or no readmission, 1: positive readmission) in association with the ethnographic variables - insurance, ethnicity and language, and patient diagnosis. Each vertical line

represents one of the four attributes - Language, Insurance Type, Ethnicity and Diagnosis, and each green line across the 2-D graph represents a patient with the light green lines representing patients with a readmission and the one with darker green are the patients without readmission. Even after compressing the feature space to a 2-D representation there is no substantial trend visible in between the features and the rate of readmissions.

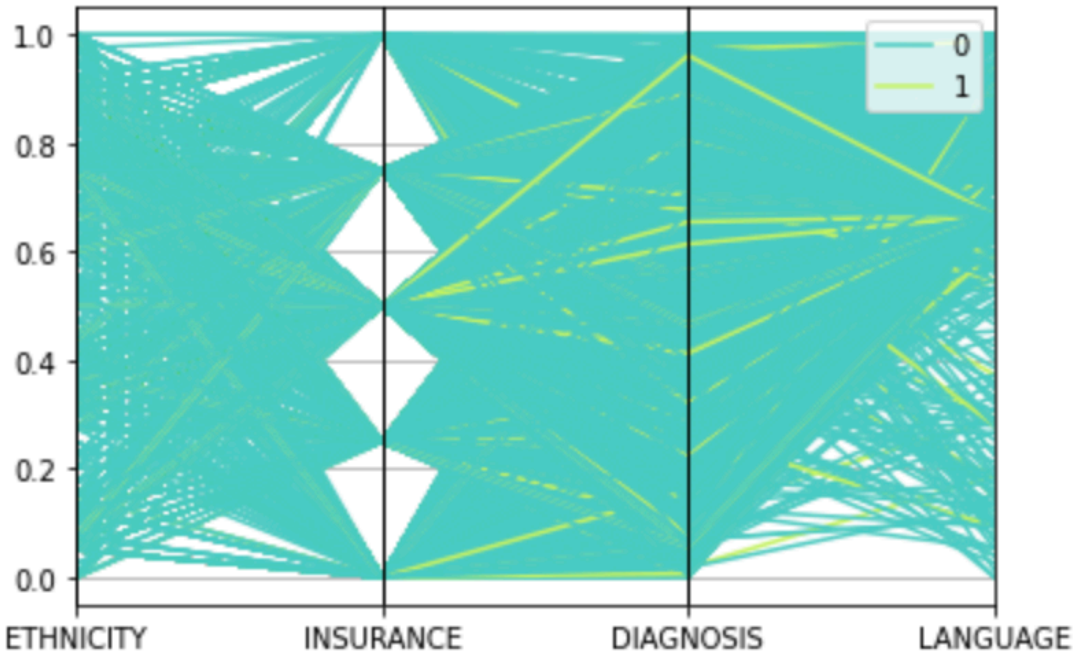


Figure 11. Parallel Coordinates graph between the different demographic attributes of patients and the readmission prediction variable. There is no clear clustering of readmitted patients when analysed for demographic attributes.

Figure 12 shows a t-SNE scatterplot visualization of the same feature set and unlikely it does not show much improvement over the PCA plot. The negative and positive readmission instances are dispersed all over the place and however there is some visual clustering of a single class data instance together, there is not a clear difference between two clus-

ters of negative and positive readmissions. This is perhaps because t-SNE works well for high-dimensional data with complex polynomial relationships in between attributes which does not seem like the case (see Figure 8, Figure 10) for the attributes being used for the t-SNE plot.

```
(<Figure size 576x576 with 1 Axes>,
 <matplotlib.axes._subplots.AxesSubplot at 0x7fe3a56f9190>,
 <matplotlib.collections.PathCollection at 0x7fe3a57223d0>,
 [Text(-0.21823052, -0.73487604, '0'), Text(1.7772219, 4.6015816, '1')])
```

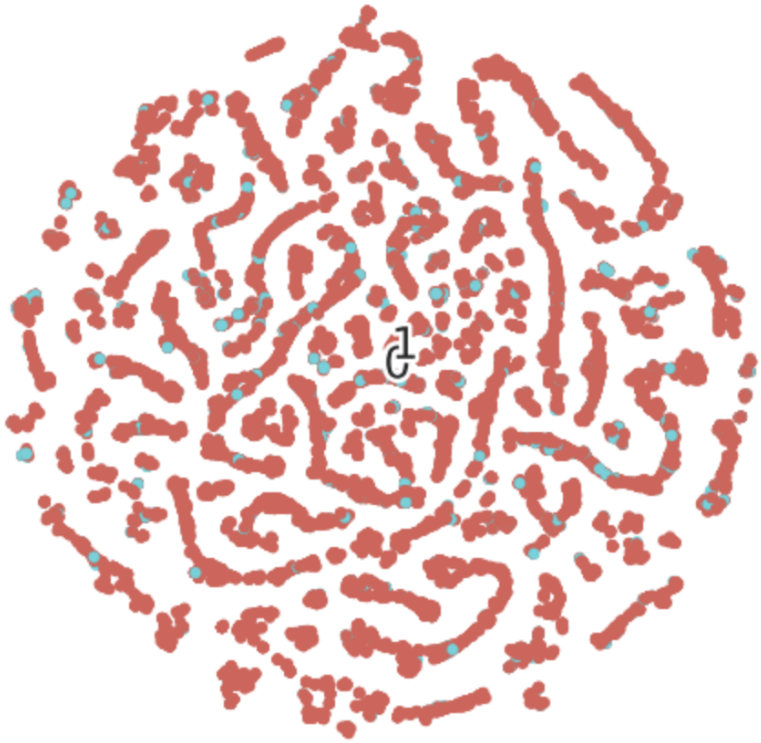


Figure 12. t-SNE scatter plot of 30-day unplanned readmission output labels plotted for a feature set comprising of patient ethnicity, language, marital status, religion and insurance type.

In all of the visualizations exploring the association or dependence of the output variable of 30-day unplanned readmission on the demographic and ethnographic features of the

patient available in the MIMIC-III dataset, there is not a direct correlation or strong association that exists between the readmission variable and the feature set. This is another evidence-based reason that analysing the effect of data hidden in clinical notes for prediction of long-term and short-term readmission to the ICU is crucial.

Chapter 4: Data Preparation

As already mentioned, data from ADMISSIONS and NOTEEVENTS tables was used to predict the probability of 30-day readmission for each patient. To prepare the data for training the model, data-type mismatch, missing data and existing biases in the dataset had to be accounted for.

The process started by converting all the dates (e.g., admit date, discharge date and death date) in the ADMISSIONS table to a manipulatable format and then used the dates to find the next 'unplanned' admission for each patient and the days between admissions.

Next, to merge this data with the clinical notes data in the NOTEEVENTS table we started by filtering and merging only the discharge summaries for an admission. For this project, only the last discharge summary per patient was used. The reason behind this being combining all the discharge summaries for a patient made the training dataset too large to train for the available hardware and memory. And given the structure of the discharge summaries in the dataset, which is very close to the structure mandated by the Joint Commission (Kind & Smith, 2008), the last discharge summary was sufficient in determining the history, patient problems, allergies, medication and any medical outliers which might or might not be present in the discharge summaries earlier. Some investigation into the data revealed that about 10.6% of the admissions were missing discharge summaries. Diving deeper into the missing notes, I discovered that about 53% of the

NEWBORN admissions were missing discharge summaries as compared to a very small percentage of the other categories. Thus, I decided to remove all NEWBORN admissions from the data.

Then to create the ground truth for our training model, we needed an output label which I created by using the time between admissions for a patient. If the days to next admission was less than 30, the patient was assigned a label of 1 (= TRUE); otherwise 0 (=FALSE). The merged dataset with the output labels was biased towards the negative samples with a difference of about 45,000 samples. We needed to address the imbalance in the dataset before using it to train our model to prevent the trained model from predicting negative heavily. I first split the data into training, validation and test sets because you always want the validation and test set to be as close to real data as possible. Then, I sampled the negative values to balance the training dataset. I also tried over-sampling the positive values and it led to the similar prevalence and model performance. I refrained from balancing the data by creating synthetic data (SMOTE) because of the hardware restrictions.

Chapter 5: Natural Language Processing

Natural Language Processing is the branch of computer science and artificial intelligence that helps computers and digital systems to interpret human language in textual form (Garbade, 2018; IBM Cloud Education, 2020; Yse, 2019). NLP has gained impetus in the past few years and the domain of healthcare has also started to realise its potential. There are enormous amounts of data buried, unused and untapped in the EHR of patients in the form of notes from doctors, care-takers and other healthcare service providers. Since this unstructured form of data and language in itself is very complex to understand and model, the true potential of NLP for healthcare data is yet to be explored. Structured data like Consolidated Clinical Document Architecture (CCDA) and Fast Healthcare Interoperability Resources (FHIR) gives a very limited insight into the actual patient record, which doctors spend a lot of time inputting and recording in the charts and other clinical notes (Foreseemed, n.d.).

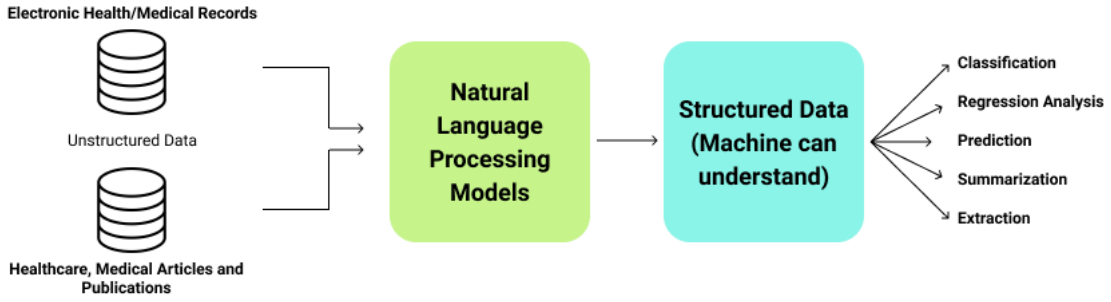


Figure 13. Natural Language Processing models help with improve the efficiency and accuracy of biomedical tasks by converting the information in the unstructured data into structured data understandable by machines. (Foreseemed, n.d.)

Some of the latest and more popular problems being explored in the NLP in healthcare space are focused on EHR usability, predictive analytics, phenotyping, and quality improvement (Health Catalyst Editors, 2019). This project uses NLP to understand and analyse the data contained in the discharge notes of patients for the prediction of short-term (30-days) unplanned readmission to the ICU.

Chapter 6: Data Pre-Processing

In addition to the demographic features, the project uses the associated discharge summaries to predict 30-day unplanned readmission. This text data is in an unstructured format and to process it, the project uses the BC5CDR variant of the scispaCy model over the more popular and widely used NLTK model. For the project, scispaCy model is chosen over NLTK, because scispaCy affords bio-entity extraction, while NLTK just uses a Bag-of-Words approach to process text data. In the medical domain, using a Bag-of-Words approach will prioritize the words representing regular real-world entities over the biomedical entities. The Named Entity Recognition (NER) function afforded by the model was used to identify chemicals/drugs and diseases in the discharge summaries. The recognized entities were converted to word vectors and a CountVectorizer was trained on these converted tokens to create a sparse matrix of the count of the different tokens in the summaries.

To continuously improve the model, the stop words list was iteratively updated based on token occurrence ranking in the discharge summaries using Zipf's Law.

6.1 scispaCy Model

Neumann et al. (2019) introduced scispaCy which is a Python library containing models developed and trained for real-time biomedical text processing. The model is built on the spaCy library and offers features like Part of Speech Tagging, Dependency Parsing, Named entity Recognition and Sentence Segmentation. The robust POS Tagger and De-

pendency Parser features of the model are trained and tested using the GENIA 1.0 corpus as well as the OntoNotes corpus, which perform just as well as the other state-of-the-art models/packages.

GENIA corpus contains abstracts and texts extracted from the articles in the MEDLINE database and the title and abstracts are annotated specifically for biomedical text processing and data mining (Kim et al., 2003). Along with the annotations in the corpus, the semantic associations using the GENIA ontology (e.g., 47 relevant nominal categories in the biomedical domain) in the extracted biomedical terms are also a part of the annotated corpus (Kim et al., 2003). OntoNotes corpus is a large annotated corpus of text ranging from (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows) in three different languages (e.g., English, Chinese and Arabic) (Weischedel et al., 2017). In addition to the annotation the corpus also contains some base-level semantic associations like ontology in the text and structural information on subject and predicate (Weischedel et al., 2017).

For more accurate and fine-grained Named Entity Recognition (NER) models, scispaCy released additional packages (*en_ner_bc5cdr_md, en_ner_craft_md, en_ner_jnlpba_md, en_ner_bionlp13cg_md*) which are trained on four different datasets - *BC5CDR* (for chemicals and diseases; Li et al., 2016), *CRAFT* (for cell types, chemicals, proteins, genes; Bada et al., 2011), *JNLPBA* (for cell lines, cell types, DNAs, RNAs, proteins; Collier and Kim, 2004) and *BioNLP13CG* (for cancer genetics; Pyysalo et al.,

2015) respectively. This project selected the scispaCy BC5CDR model to detect names of drugs/chemicals and diseases in the discharge summaries and use them as a part of the feature set for readmission prediction.

Admission Date: [**2149-1-10**] Discharge Date: [**2149-1-14**] Service: NEUROSURGERY Allergies: Patient recorded as having No Known Allergies to Drugs Attending:[**First Name3 (LF) 1271**] Chief Complaint: R Subdural hemorrhage Major Surgical or Invasive Procedure: None History of Present Illness: 83yM tripped and fell while taking out trash landed on face. T'ferred from [**Hospital 27217**] Hosp. where CT head showed R temporal-parietal SDH measuring 5.1cm AP diameter and 6.2mm in thickness. There is no mass effect seen on CT. Patient was given ASA for CP at OSH and later given 2U FFP to reverse. Past Medical History: Atrial fibrillation and MI Social History: Lives alone Family History: Noncontributory Physical Exam: O: T: 98.3 BP:160 /82 HR: 83 R 14 O2Sats 95% Gen: WD/WN, comfortable, NAD. HEENT: Pupils: EOMs Neck: Supple. Lungs: CTA bilaterally. Cardiac: RRR. S1/S2. Abd: Soft, NT, BS+ Extrem: Warm and well-perfused. Neuro: Mental status: Awake and alert, cooperative with exam, normal affect. Orientation: Oriented to person, place, and date. Language: Speech fluent with good comprehension and repetition. Naming intact. No dysarthria or paraphasic errors. Cranial Nerves: I: Not tested II: Pupils equally round and reactive to light, to mm bilaterally. Visual fields are full to confrontation. III, IV, VI: Extraocular movements intact bilaterally without nystagmus. V, VII: Facial strength and sensation intact and symmetric. VIII: Hearing intact to voice. IX, X: Palatal elevation symmetrical. [**Doctor First Name 81**]: Sternocleidomastoid and trapezius normal bilaterally.

Allergies, Allergies, Subdural hemorrhage, ASA, CP, Atrial fibrillation, NAD, dysarthria, Extraocular movements, nystagmus, fasciculations, abnormal movements, tremors, RBC-3.72, Hgb-12.4, Ct-156, Ct-156, Glucose-181, UreaN-20 Creat-1.0 Na-136, HCO3, CK-MB-NotDone, hematoma, subfalcine herniation, uncal herniation, acute infarction, fracture, Cspine, fractures, dislocations, hematoma, decrease mass effect and midline shift, Subfalcine, subdural hemorrhage, hemorrhage, Dilantin, seizure, hemorrhage, Dilantin, Digoxin, Diltiazem, Pravastatin, Atenolol, Lasix, Qinapril, Protonix, Prednisone, Pantoprazole, Folic Acid, Allopurinol, Digoxin, Diltiazem HCl, Pravastatin, Atenolol, Furosemide, Quinapril, Phenytoin Sodium, Protonix, Prednisone, Pantoprazole, FoLIC Acid, Multivitamins, Allopurinol, Subdural hemorrhage, neurologic deficits , Dilantin, headache, nausea, vomiting, weakness, numbness, tingling, by:[**2149-4-7**]

Figure 14. The image above shows a snippet of the discharge summary in the MIMIC-III database. Image below show the entities scispaCy extracted.

6.2 Word Vectors

Some alternate approaches to textual data representation in Natural Language Processing include **Bag-of-Words** and **One-hot encoding**. Both these models focus on multiplicity

of the words and ignore the word association, relationship, context and meanings of the words. Contrary to the traditional models, word vectors enable us to analyze the relationships across words, sentences, and documents. These vectors are modeled in the space based on the meaning and context of the word.

In simple terms, word vectors are numerical representations of words that take into account the meaning of a word while representing it numerically in the form of a vector. These vector representations of the words allow a model to train on textual data for prediction or classification. “A word vector is a row of real-valued numbers (as opposed to dummy numbers) where each point captures a dimension of the word's meaning and where semantically similar words have similar vectors. Words that are used in a similar context will be mapped to a proximate vector space” (Ahire, 2018, para 4). When words are represented as vectors in the aforementioned way, mathematical operators can be used on the words for manipulation, thus rendering them as an even more useful and meaningful feature for a prediction model.

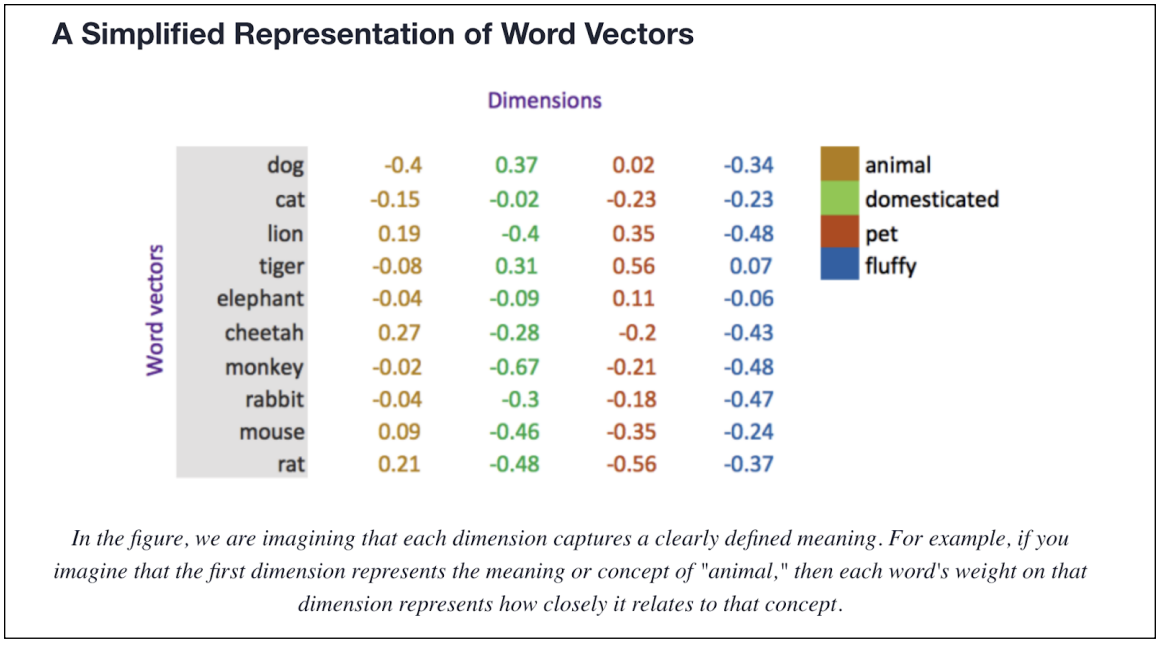


Figure 15. Word vector representation of different animals where the different colored number represent how closely the animal is associated with the attribute (legend on top right) (Ahire, 2018).

Figure 14 shows what the vectors for some words (animal names) look like. From this figure, we know that the dimensions in the vector represent a meaning and the value for the dimension is an indication of the word numerical weight on that dimension which in turns represents the word's association with and to the dimension meaning (Ahire, 2018).

6.3 Zipf's Law

Zipf's Law is an empirical law that states that "in a large sample of words the frequency of a word is inversely proportional to its ranking in the frequency table" ("Zipf's Law", n.d., para. 2). In other words, the *r*-th most frequent word will have frequency *f(r)* which can be determined using (Piantadosi, 2014):

$$f(r) \propto 1/r^{\alpha}$$

where α is approximately 1; r=frequency rank of the word; f(r)=frequency in the sample.

The law was proposed by and named after George Kingslay Zipf. In Natural Language Processing terms, the law provides a probability distribution that helps predict the probability of a word in a given sample text. The probability mass function in Zip's Law can be written as (Hasan, 2019):

$$f(k;\alpha,N) = (1/k^{\alpha})/(\sum_{n=1}^{N} 1/n^{\alpha})$$

where, k: rank of the word whose probability of appearance in the corpus i being calculated

N: size of the vocabulary of the corpus

α: probability mass function distribution parameter. Normally set to 1.

(paras. 4-5).

When working with NLP models, the training datasets are huge and contain an enormous collection of textual data where even the more frequent word is only a very small fraction of the entire corpus (Hasan, 2019). Most of the latest NLP models choose to represent the tokens extracted from the text in a multi-dimensional vector format. Now given the massiveness of the corpus and the high-dimensional word vector representations our models perform well for predicting more common words and perform worse for rare words since the rare words occur less or have lesser examples than common words (Zipf's Law) in the corpus but are modeled in the same dimension in the vector space as

the popular or common words. Therefore, Zipf's law can be used to address some of these biases by taking notice of word frequencies and accounting for over-fitting.

Chapter 7: Prediction Models

7.1 Logistic Regression

Logistic Regression is a regression analysis model used for classification of the dependent variable based on one or more independent variables (Swaminathan, 2018; Thanda, 2021).

Model Output = **0 or 1**; *Hypothesis:* **Z=WX+B** ; $h\Theta(x) = sigmoid(Z)$

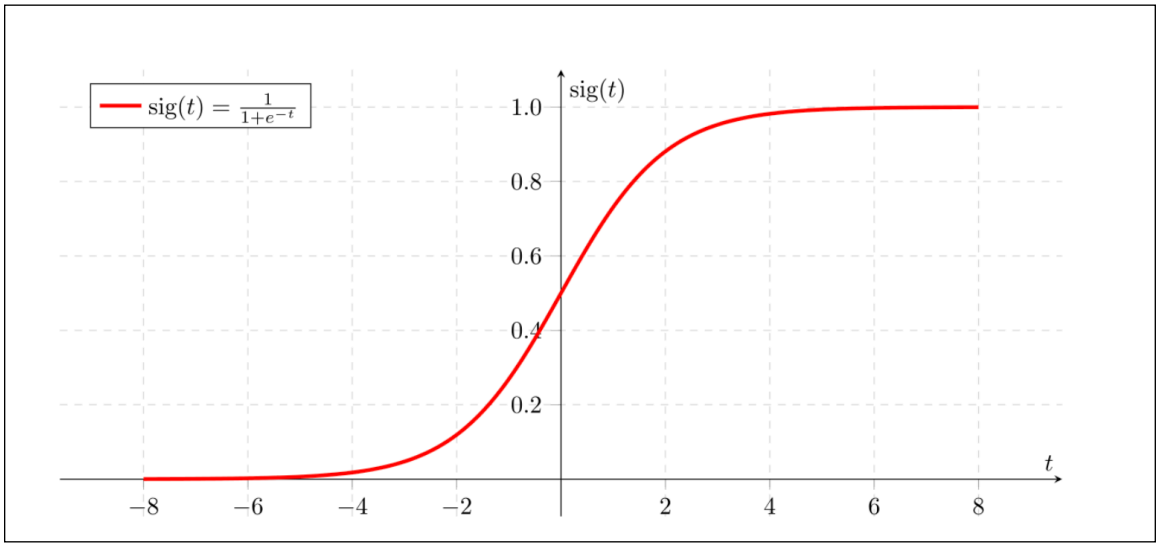


Figure 16. Sigmoid Activation function for Logistic Regression (Swaminathan, 2018)

When Z → ∞, Y (dependent/prediction variable) becomes 1; and when Z → −∞, Y becomes 0. This hypothesis gives us the estimated probability of a certain prediction, meaning how confident is the correctness of the predicted value as compared to the actual value of the independent variable. Mathematically,

$$h\Theta(x) = P(Y = 1 \mid X; theta)$$

$$P(Y = 1 \mid X; theta) + P(Y = 0 \mid X; theta) = 1$$

$P(Y = 0 \mid X; theta)$ = 1 - $P(Y = 1 \mid X; theta)$

$P(Y = 1 \mid X; theta)$: Probability of Y being 1 given X is parameterised by 'theta'

Cost function for a logistic regression variable is defined as:

$$Cost(h\Theta(x), Y(actual)) = -log(h\Theta(x)) if y = 1$$

$$-log(1 - h\Theta(x)) if y = 0$$

Given the large sample size of our dataset, the dichotomous nature of the predictor variable and the little correlation between the independent variables, binary logistic regression is a wise choice of model to be used for the prediction of 30-day unplanned readmission for the project.

7.2 Gaussian Naive Bayes Model

Bayes theorem provides a way of selecting the best hypothesis (*h*) for a given data (*d*). The probability of hypothesis is calculated based on the prior information (Brownlee, 2016). The model is based on the Bayes theorem defines as:

$$P(h \mid d) = (P(d \mid h) * P(h)) / P(d)$$

where, *P(h|d) is* probability of hypothesis given data d (conditional probability)

P(d|H) : probability of data d if the hypothesis were true (conditional probability)

P(h) : probability of hypothesis h

P(d) : probability of data

Once the conditional probability of different hypotheses given the dataset is known, the hypothesis with maximum posterior probability (probability of hypothesis given the data) is chosen to classify the data. The Naive Bayes model is used for a classification problem to classify data instances into different classes (or hypotheses) based on the maximum probability of a class given the dataset. This model can be extended to real-time data and attributes by assuming that the data follows a Gaussian (or normal) distribution. The probability of hypothesis being true for a given data instance is defined by the Probability Density Function that makes use of mean and standard deviation of the dataset and is given by (Brownlee, 2016):

```
pdf(x, mean, sd) = (1 / (sqrt(2 * PI) * sd)) * exp(-((x-mean^2)/(2*sd^2)))
```

Because of its straightforward approach, Gaussian Naive Bayes model is simple and fast and can be used on complex, large datasets for classifications. It is also widely used as a model of choice in sentiment analysis and text classification problems (Kelly & Johnson, 2021; Jurafsky & Martin, 2009).

7.3 Support Vector Machine

Support Vector Machines is a supervised-learning classification model that attempts to find a hyperplane in the data distribution that categorizes the data instance into different classes (“Support-vector machine”, n.d., para. 1; Gandhi, 2018).

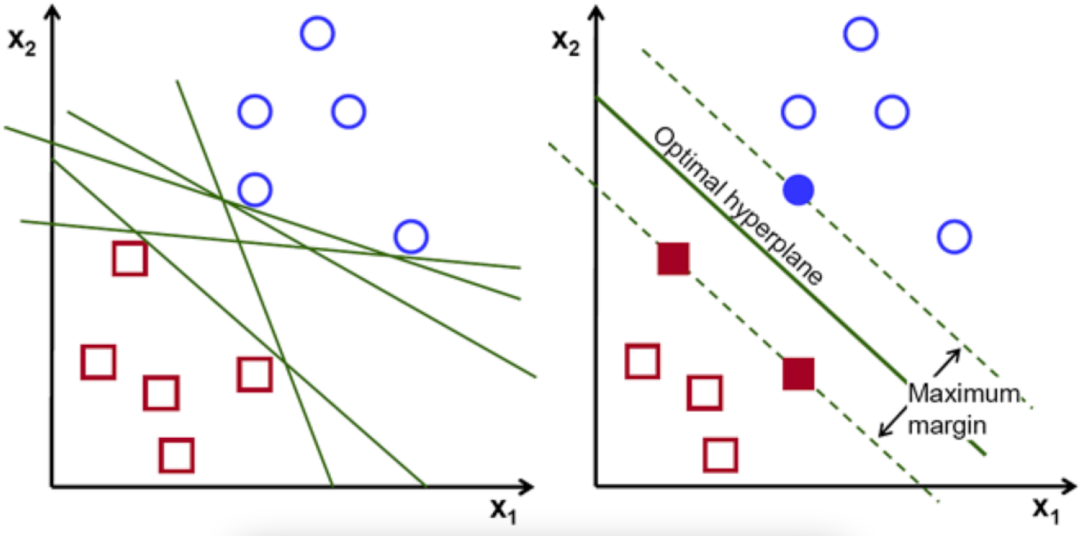


Figure 17. Possible hyperplanes for Support Vector Machine (Gandhi, 2018)

To classify the data points correctly, the possibility of hyperplanes can be endless but the model tries to find a hyperplane with maximum margin, i.e, maximizing the distance between the plane and nearest data-points or instances on each side. These data points that are near or on the hyperplane determine the position and width of the hyperplane and are called Support Vectors. SVMs help capture complex relationships between data instances without a lot of manual transformation required. Using the correct kernel and optimal parameters, it helps provide accurate predictions.

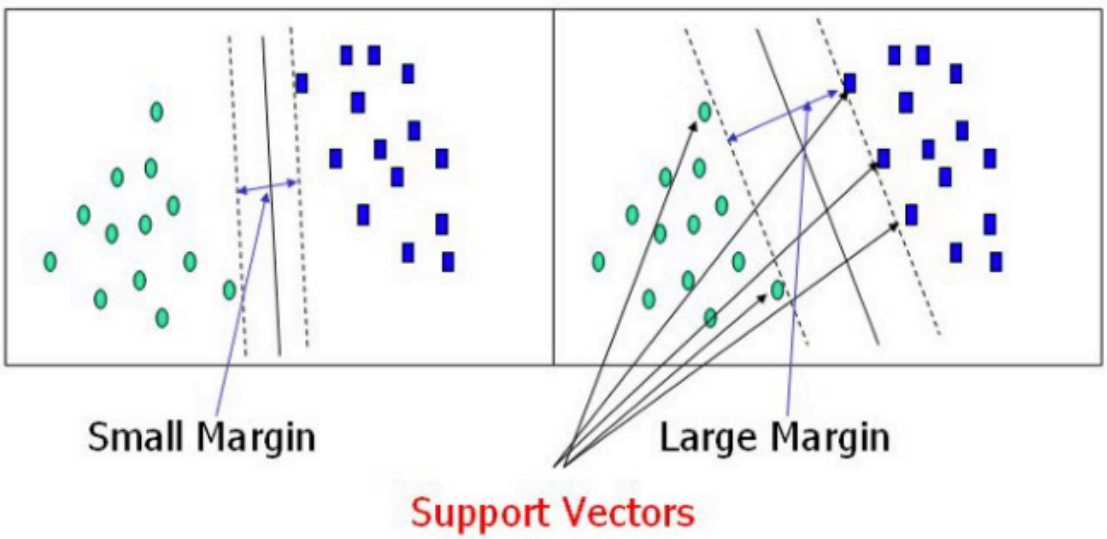


Figure 18. Support Vectors - data instances or points that are closest to the hyperplane or lie on the hyperplane (Gandhi, 2018).

7.4 AdaBoost Classifier

AdaBoost (Adaptive Boosting) is a boosting ensemble classifier proposed by Yoav Freund and Robert Schapire in 1996 (Navlani, 2018). An ensemble machine learning model follows one of the three approaches: Bagging, Boosting and Stacking to improve the prediction accuracy of the final model. Adaboost is an ensemble of multiple classification models, using a Boosting approach, whose performance is improved through iterative training and adjusting of weights in the model based on the training error to account for any unusual instances in the training dataset. Boosting helps address bias in the dataset and avoids overfitting which makes it a good fit for this project.

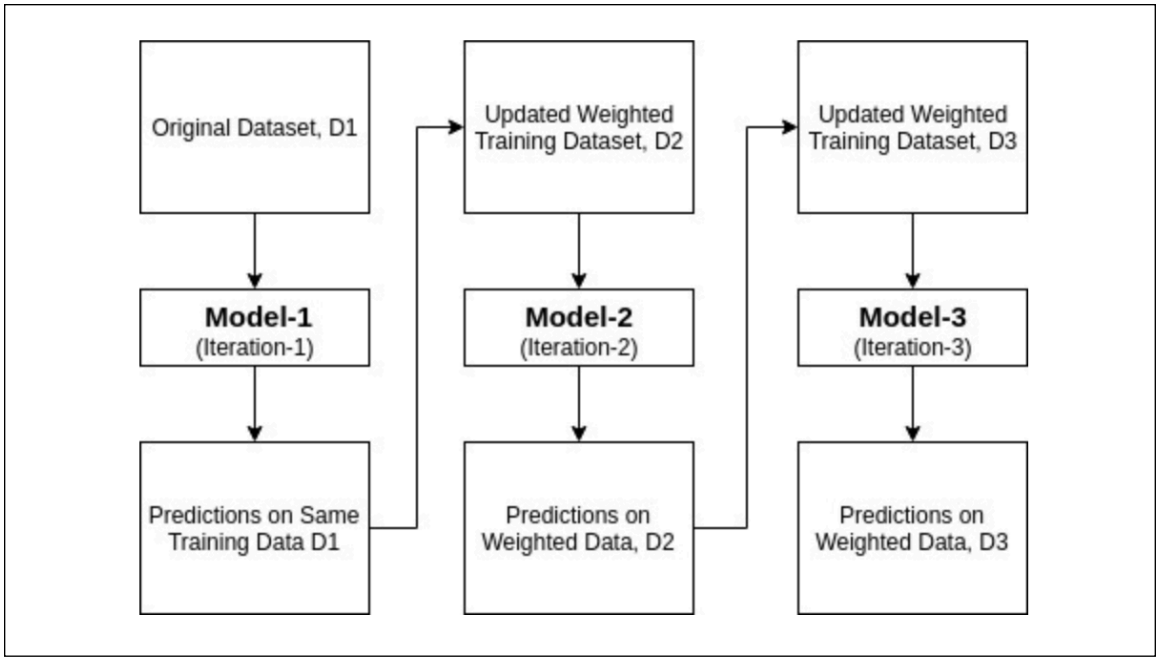


Figure 19. Working mechanism of a typical AdaBoost ensemble model (Navlani, 2018)

7.5 BERT

BERT is the latest state-of-the-art model in the NLP and NLU domain proposed by Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova in 2019. BERT stands for *Bidirectional Encoder Representations from Transformers* (Devlin et al., 2019). BERT is a bidirectional transformer model that pretrains on bidirectional representations from unlabeled text data by conditioning on right as well as left context in all layers (Devlin et al., 2019). The BERT model can be used and fine-tuned for a specific problem by just adding another training layer on top of the existing model. BERT is highly accurate and efficient for many NLP and NLU tasks without the need of any architectural modifica-

tions to the model (Devlin et al., 2019). There are many pre-trained versions of BERT available and this project fine-tunes blueBERT for the readmission prediction task.

BlueBERT or NCBI BERT was developed by the National Library of Medicine and National Institute of Health specifically for tasks in the clinical domain. The *BLUE* in BlueBERT stands for Biomedical Language Understanding Evolution (Peng et al., 2019). This variant of the BERT model is pretrained on MIMIC-III clinical notes and abstracts PubMed dataset and performs better than other variants of BERT for biomedical tasks (Peng et al., 2019). The project utilizes the built-in pipeline with the pre-trained weights in the model for prediction and evaluation.

Chapter 8: Method

In the project, all models mentioned in the previous section are trained and evaluated based on AUC score for 30-day unplanned readmission prediction. The final dataset used for training and evaluating the models is prepared by cleaning and processing the data in the table NOTEEVENTS and ADMISSIONS of the MIMIC-III dataset, as described in the *Chapter 4: Data Preparation* and *Chapter 5: Data Pre-Processing*. The processed feature-set is transformed into a sparse matrix before being used for training the five models: Logistic Regression, SVM with kernel, Adaboost ensemble and Guassian Naive Bayes. Each model is trained separately and evaluated based on accuracy and AUC score to identify scopes for improvement. Using the zipf's law as the guiding principles, the list of stop words is updated iteratively and the vectorizer is trained again to get an updated and more efficient feature set.

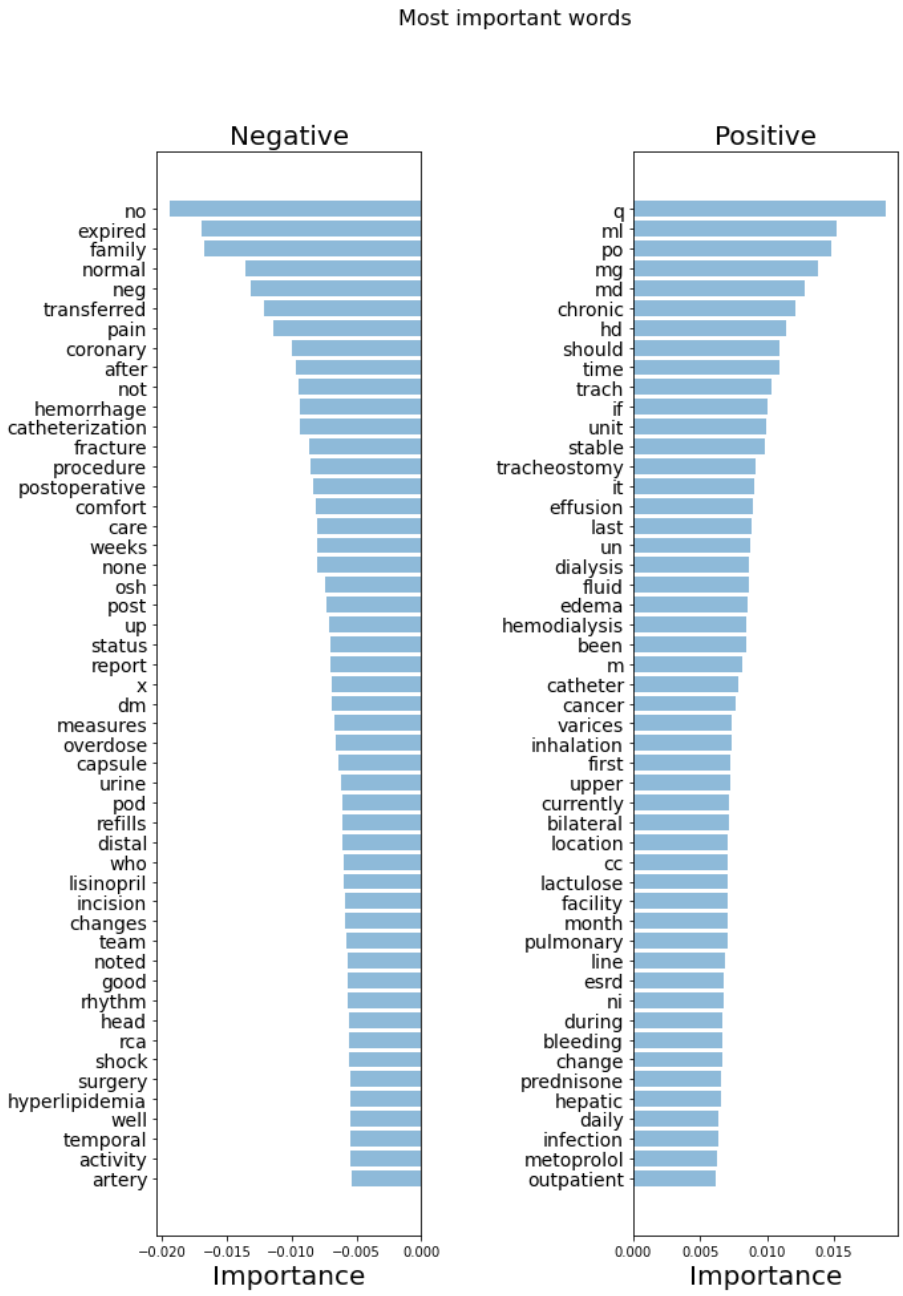


Figure 20. Importance of words in the discharge summaries calculated and plotted based on the Zipf's law

To improve the model efficiency and performance, each model cross-validated over five iterations to get a smoother AUC curve on the training and validation dataset. Finally, the vectorizer is re-trained on a range of max_features value and the models are trained and evaluated on the feature set of each dimension for comparison.

To try prediction using a transformer model, BERT is used with Huggingface transformers for training on the created feature set. BERT model is trained on created feature-set abstracted from the MIMIC-III dataset for 30-day unplanned readmission and the pre-trained checkpoints available in the blueBERT model are updated. This is done by downloading the pre-trained checkpoints for BERT and adding another layer on top of the core model and updating the pre-trained checkpoints. The pretrained checkpoints used are derived from the blueBERT model which is trained on MIMIC and PubMed datasets.

For the baseline results, a logistic regression and support vector machine models are trained and evaluated on the basic demographic and ethnographic attributes of patients data available in the ADMISSIONS table of the dataset. The output labels are generated in the same way as described in the *Chapter 5: Data Preparation* and the categorical data is encoded first using a label encoder and then encoded using the one-hot encoding technique before feeding into the model for training and evaluation.

Chapter 9: Results

The baseline model developed as a part of the project uses the demographic information of patients as it's feature set and even though has an accuracy high enough, does not perform distinguishing between the two classes - positive and negative readmissions. Table 4 shows the AUC scores attained by the baseline models on the MIMIC-III dataset.

Model	Train AUC	Valid AUC
Logistic Regression	0.500	0.500
Poly-kernel SVM	0.500	0.500

Table 4. AUC scores for the baseline models on the training and validation dataset.

Figure 21 shows the AUC graph for both the models (which look exactly the same) which outlines an AUC score of 0.5 with no ability to distinguish between the positive and negative classes.

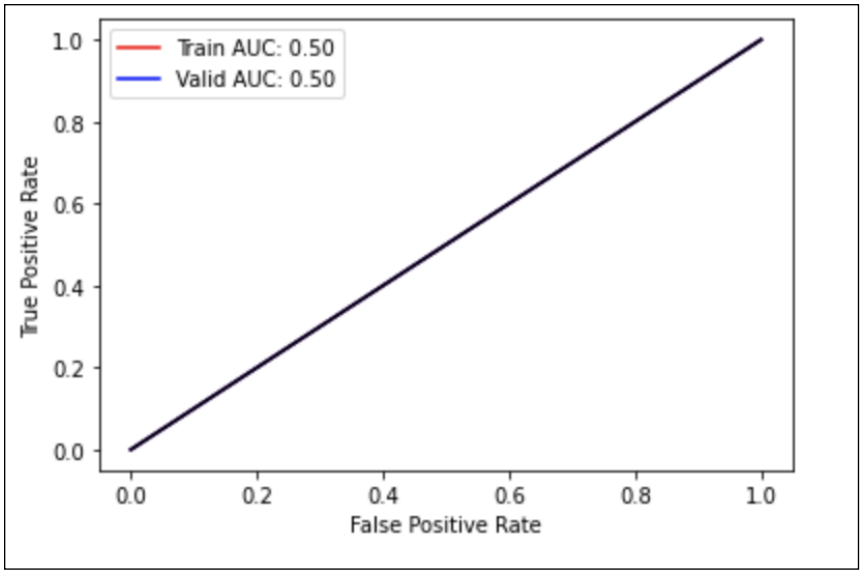


Figure 21. ROC curve for the baseline logistic regression and poly -kernel SVM models at a threshold value of 0.5

The project trains and compares a variety of models to see how they perform against each other. The hyperparameter tuning is primarily done manually and performance was evaluated based primarily on the AUC curve. Table 5 and Table 6 show the model performance without any improvement or modifications on the train dataset.

Model	AUC	Accuracy	Recall	Precision	Specificity	Prevalence
Logistic Regression	0.754	0.685	0.624	0.712	0.747	0.500
SVM	0.913	0.808	0.635	0.982	0.980	0.500
Adaboost	0.781	0.710	0.691	0.718	0.729	0.500
Guassian Naive Bayes	0.714	0.664	0.566	0.703	0.761	0.500

Table 5. Model evaluation on the training dataset

Model	AUC	Accuracy	Recall	Precision	Specificity	Prevalence
Logistic Regression	0.706	0.714	0.595	0.114	0.721	0.057
SVM	0.668	0.749	0.449	0.114	0.767	0.057
Adaboost	0.680	0.658	0.604	0.097	0.662	0.057

Guassian Naive Bayes	0.668	0.709	0.535	0.104	0.720	0.057

Table 6. Model evaluation on the validation dataset.

Figure 22, 23, 24 and 25 show the cross-validation training and the regular training AUC score of the model's while learning on the training dataset.

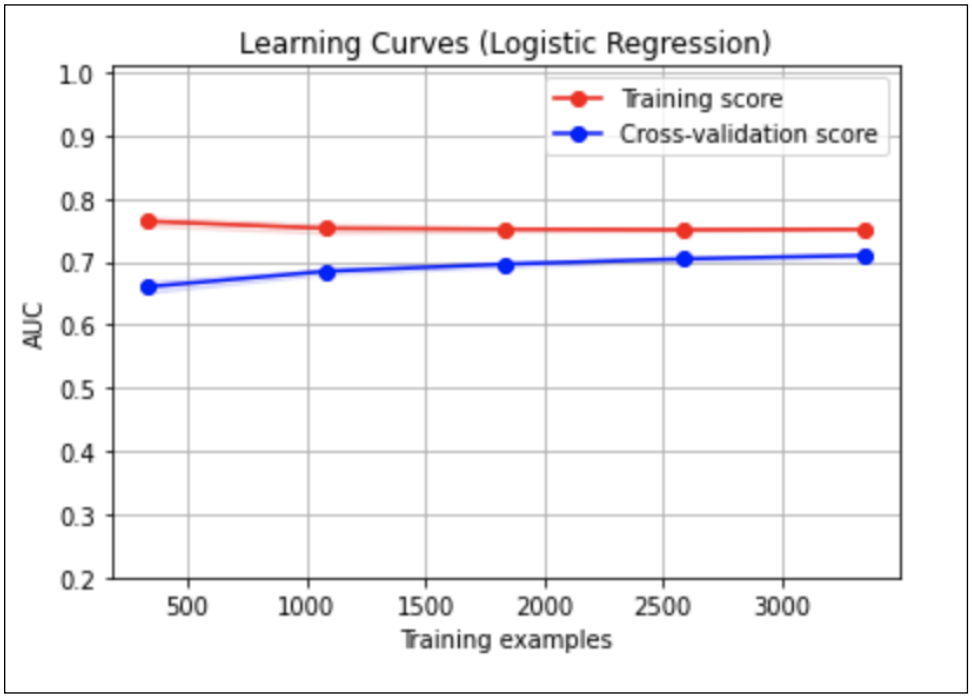


Figure 22. Learning curve showing AUC scores on the training dataset for the Logistic Regression model

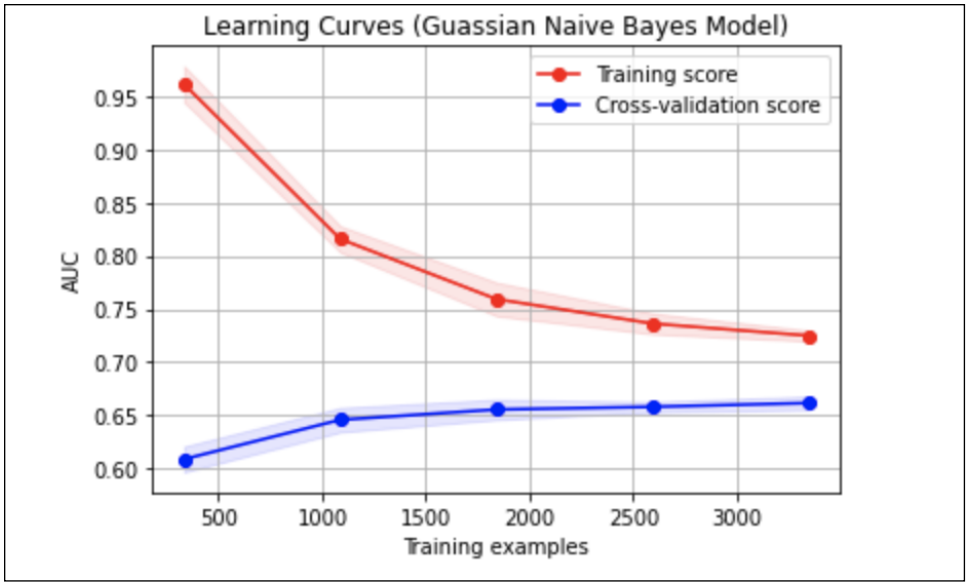


Figure 23. Learning curve showing AUC scores on the training dataset for the Gaussian Naive Bayes model

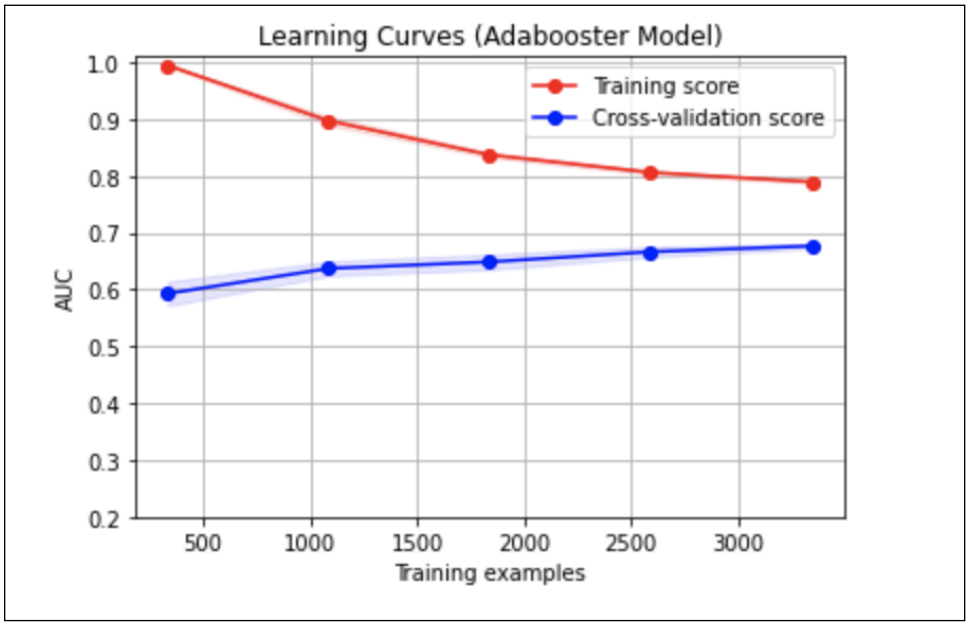


Figure 24. Learning curve showing AUC scores on the training dataset for the AdaBoost model

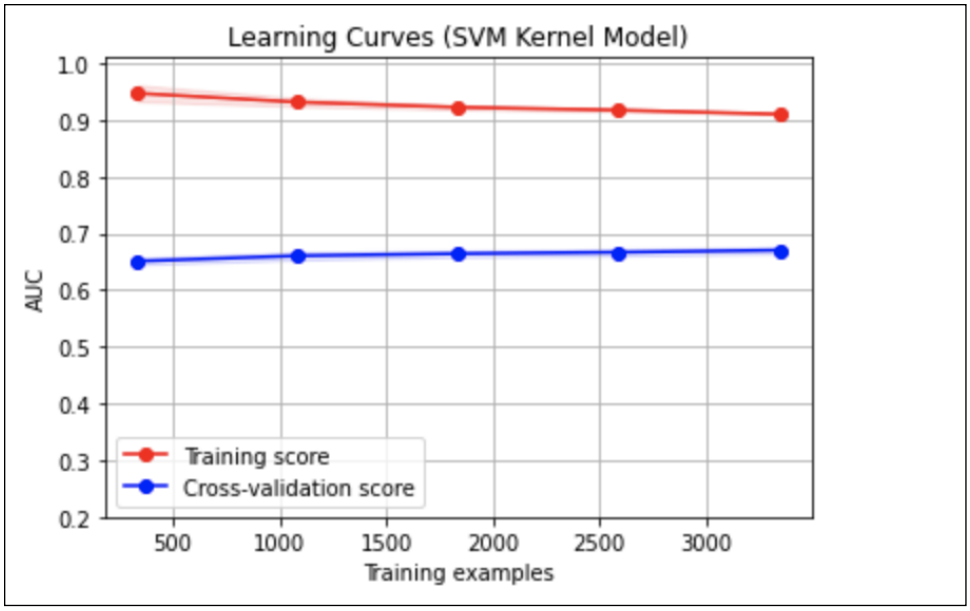


Figure 25. Learning curve showing AUC scores on the training dataset for the Poly-kernel SVM model

Next, the project evaluates the model by varying the number of features on which the word vectors are trained and evaluates each of the four models iteratively by varying the number of features used in the word vector and plotting a learning curve to see how each model learns and the learning trend they follow. The logistic regression model performs the best, reaching the highest AUC score of 0.708 on the validation dataset and does not show any signs of over-fitting on the training dataset.

Model	Number of features	Training AUC Score	Validation AUC Score
Logistic Regression	100	0.686	0.667
	300	0.712	0.684
	1000	0.736	0.700
	3000	0.753	0.707
	10,000	0.758	0.708
	30,000	0.759	0.708

Table 7. Evaluation (AUC scores) for the Logistic Regression model on the training and validation dataset by varying the features in the word vector.

In the Figure 26 (below), the learning of logistic regression models over a range of features is plotted against the fine-tuning of the cost-function parameter or decay variable (C) of the model.

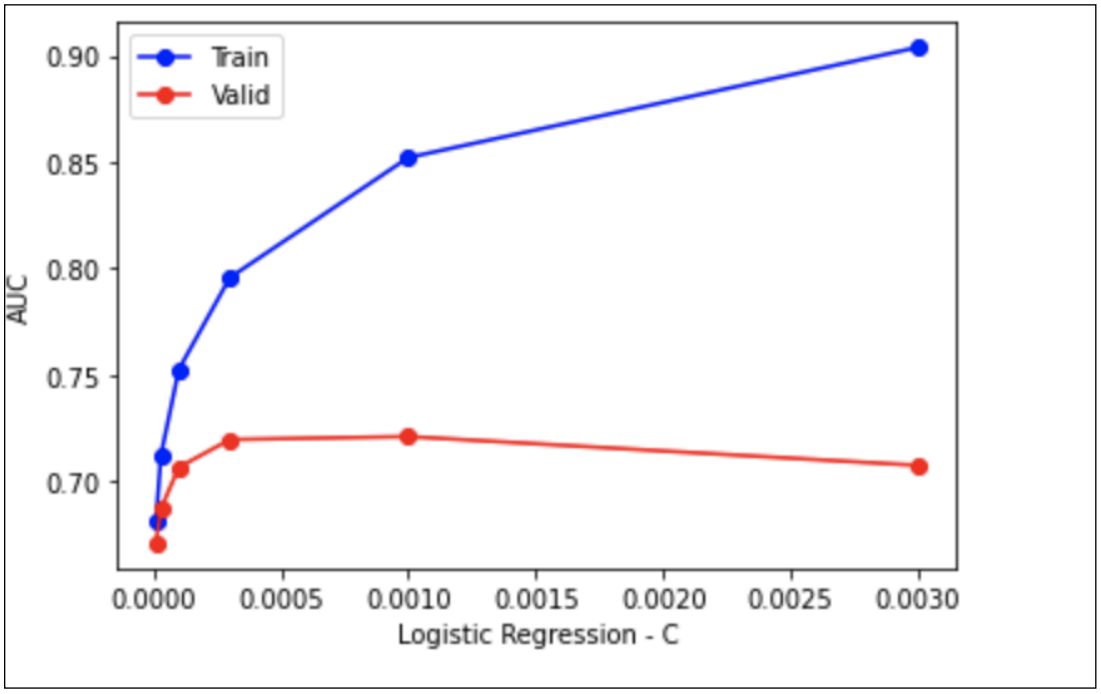


Figure 26. Graphical representation of the performance (AUC score) of the Logistic Regression model as the decay factor, C, is fine-tuned over a range of values.

The fine-tuning of the other models in the project is done through GridSearchCV to optimize fitting of the model and parameters on the train dataset. Table 8, 9, 10 show the AUC scores obtained by the different models on train and validation dataset for a range of feature values after parameter fine-tuning.

Model	Number of features	Training AUC Score	Validation AUC Score
Gaussian Naive Bayes	100	0.629	0.632
	300	0.628	0.635
	1000	0.652	0.644
	3000	0.714	0.668
	10,000	0.835	0.617
	30,000	0.925	0.559

Table 8. Evaluation (AUC scores) for the Gaussian Naive Bayes model on the training and validation dataset by varying the features in the word vector.

Model	Number of features	Training AUC Score	Validation AUC Score
Poly Kernel SVM	100	0.825	0.642
	300	0.867	0.652
	1000	0.900	0.661
	3000	0.912	0.668
	10,000	0.910	0.670
	30,000	0.908	0.670

Table 9. Evaluation (AUC scores) for the Poly-kernel SVM model on the training and validation dataset by varying the features in the word vector.

Both the Poly kernel SVM and Gaussian Naive Bayes model do not show a significant improvement in the AUC score as the feature values are varied. The models peak at an optimal hyper-parameter value and show only a slight increase in the AUC score as the number of features are increased. While the Gaussian model does not overfit and performs consistently, the poly-kernel SVM model tends to overfit as the features are increased reaching a peak in AUC value because of the 5-fold cross-validation training technique.

Model	Number of features	Training AUC Score	Validation AUC Score
Adaboost	100	0.908	0.670
	300	0.908	0.670
	1000	0.908	0.670
	3000	0.908	0.670
	10,000	0.908	0.670
	30,000	0.908	0.670

Table 10. Evaluation (AUC scores) for the AdaBoost model on the training and validation dataset by varying the features in the word vector.

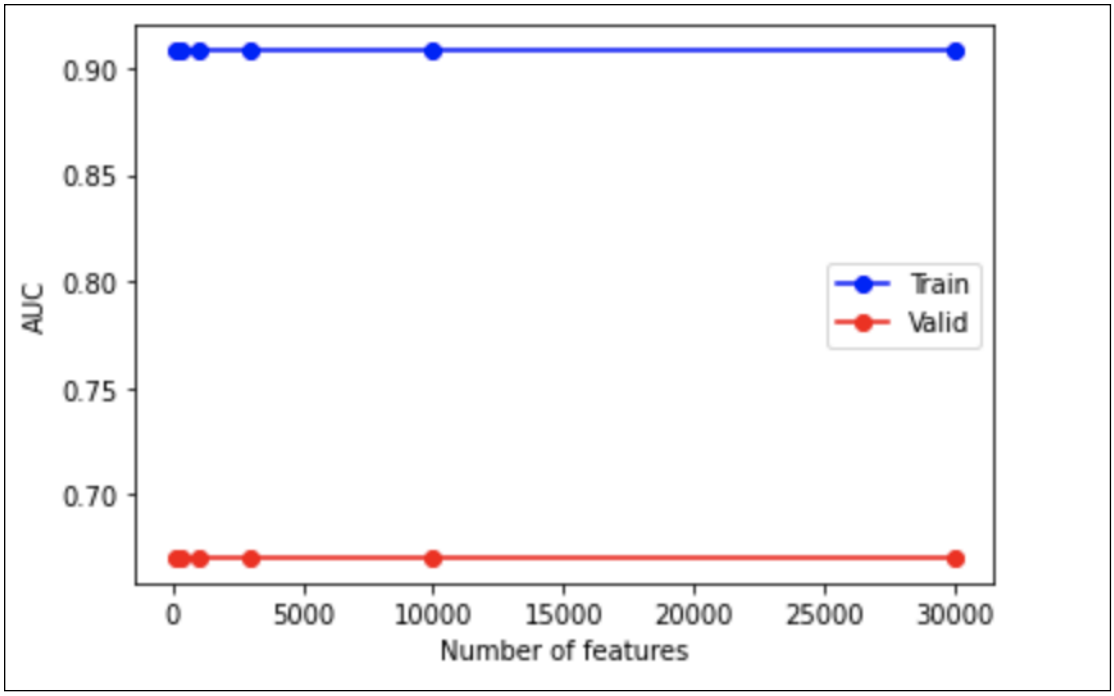


Figure 27. Graphical representation of the performance (AUC score) of the AdaBoost model the number of features in the word vector are varied.

We see that the learning and performance of Adaboost show no change and AUC score remains constant across the range of feature values (see Figure 27). A valid explanation behind this is the tendency of an Adaboost model to not overfit (Malfanti et al., 2017), but the difference between the AUC scores for train and validation sets shows that the model is biased which is a problem with shallow decision trees used in the Adaboost ensemble approach (Misra & Li, 2019).

Using the BlueBERT (base) pre-trained model with Huggingface transformers, the model reaches an accuracy level of 72% on the validation dataset. A disadvantage with using the pre-trained BERT models is that these models are trained for performing on predefined tasks and thus the weights in the outer layer of the model are biased to the task the model

was trained for during development. Adjusting and fine-tuning the weights and embedding in different layers of the model is computationally and time intensive, and was beyond the scope of this project.

10. Conclusion

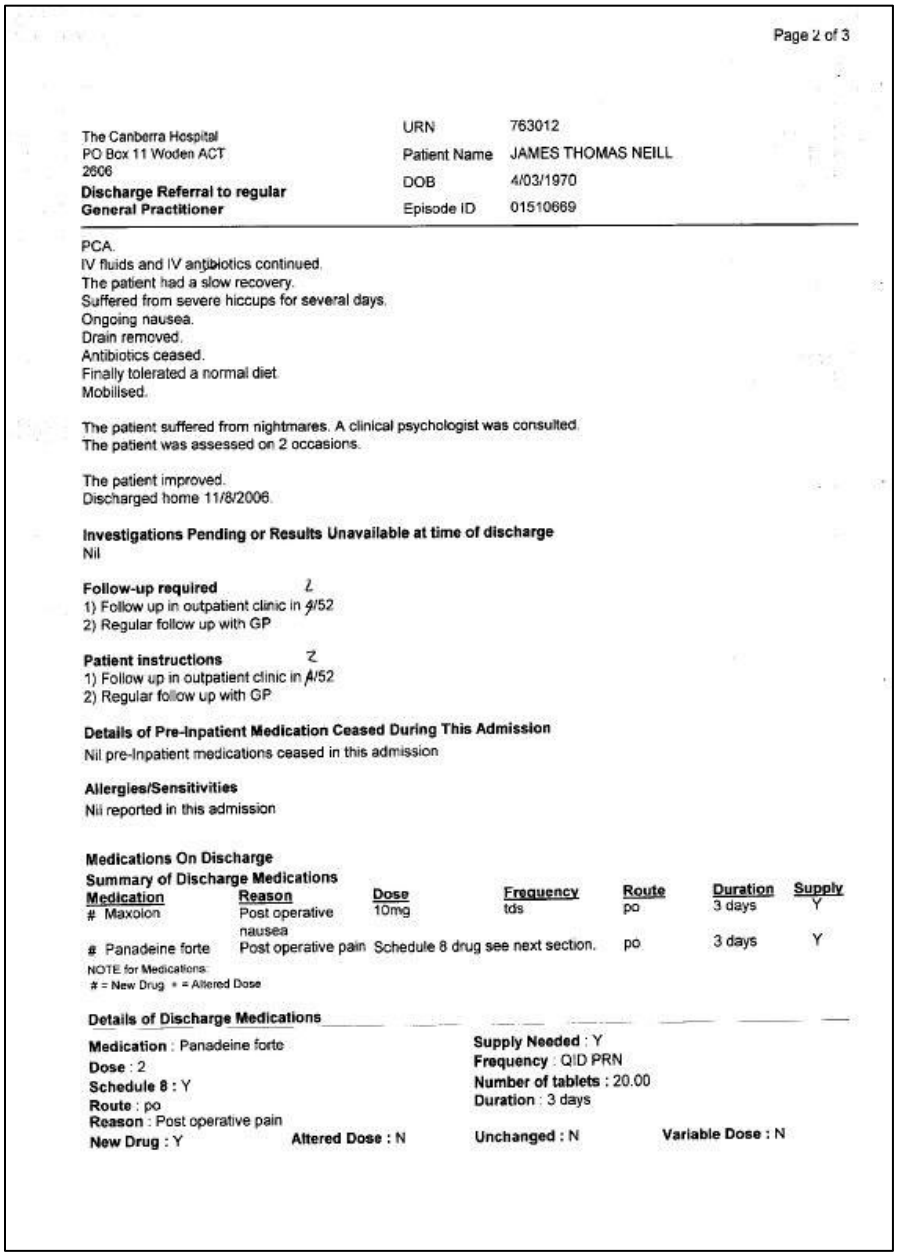
The existing research, data exploration and model evaluations done in the project bear evidence to the importance of unstructured data stored in the clinical notes for the prediction of 30-day unplanned readmission to the ICU. The prediction model performs better when trained on the data in the clinical notes, specifically discharge summaries used in this project. Ther performance in terms of AUC scores sees a growth from 0.5 using categorical demographic and ethnographic features to 0.71 on the validation data using discharge summaries. Future work in the space needs to focus on using other clinical notes data like nursing notes/charts, nutrition data, physician notes, daily chart data, case management, consultation prescriptions etc. for the prediction of unplanned readmission. Harnessing the information stored in the unstructured data could open possibilities for better and more accurate prediction of not just readmission but also plausible diagnosis and prognosis. It would also be interesting to analyse and correlate how often patients with different cultural and demographic backgrounds are readmitted to the ICU for similar disease diagnosis and prognosis and how is the readmission rate influenced by the length of stay of a patient in the ICU for a certain disease.

scispaCy (Neumann et al., 2019) and blueBERT (Peng et al., 2020) models used and trained in this project have been developed specifically for biomedical data, but they need to be trained and evaluated on larger amounts of data for a variety of prediction and clas-

sification problems to realise their full potential. The information hidden in EHR/EMR has the power of revolutionizing the healthcare industry through prediction of variables that can help patients, their family and friends and the hospital management and staff plan better for a more efficient case management which is financially effective and medically beneficial. For the future, it would be interesting to evaluate the model based on clinically accepted performance measures and use them for model fine-tuning. The project though calculates the precision and recall for the trained models, the values are not used for fine-tuning and improving the models. It might be beneficial to see how a tradeoff between precision and recall impacts the model performance for 30-day unplanned readmission prediction. Often for problems in the medical and clinical domain, given the loss and cost associated with a *False Negative* prediction, recall is valued over precision. Considering the influence of demographic and socioeconomic factors on patient after-care, thus impacting readmission possibility, it is important that for the given feature set a precision versus recall tradeoff is accounted for parameter fine-tuning and model evaluation in future works.

Appendix

A. An example image of the discharge summary for a ruptured appendix case.



Page 2 of 3

The Canberra Hospital
PO Box 11 Woden ACT
2606
Discharge Referral to regular General Practitioner

URN 763012
Patient Name JAMES THOMAS NEILL
DOB 4/03/1970
Episode ID 01510669

PCA.
IV fluids and IV antibiotics continued.
The patient had a slow recovery.
Suffered from severe hiccups for several days.
Ongoing nausea.
Drain removed.
Antibiotics ceased.
Finally tolerated a normal diet.
Mobilised.

The patient suffered from nightmares. A clinical psychologist was consulted.
The patient was assessed on 2 occasions.

The patient improved.
Discharged home 11/8/2006.

Investigations Pending or Results Unavailable at time of discharge
Nil

Follow-up required
1) Follow up in outpatient clinic in 4/52
2) Regular follow up with GP

Patient instructions
1) Follow up in outpatient clinic in 4/52
2) Regular follow up with GP

Details of Pre-Inpatient Medication Ceased During This Admission
Nil pre-Inpatient medications ceased in this admission

Allergies/Sensitivities
Nil reported in this admission

Medications On Discharge
Summary of Discharge Medications

Medication	Reason	Dose	Frequency	Route	Duration	Supply
# Maxolon	Post operative nausea	10mg	tds	po	3 days	Y
# Panadeine forte	Post operative pain	Schedule 8 drug see next section.		po	3 days	Y

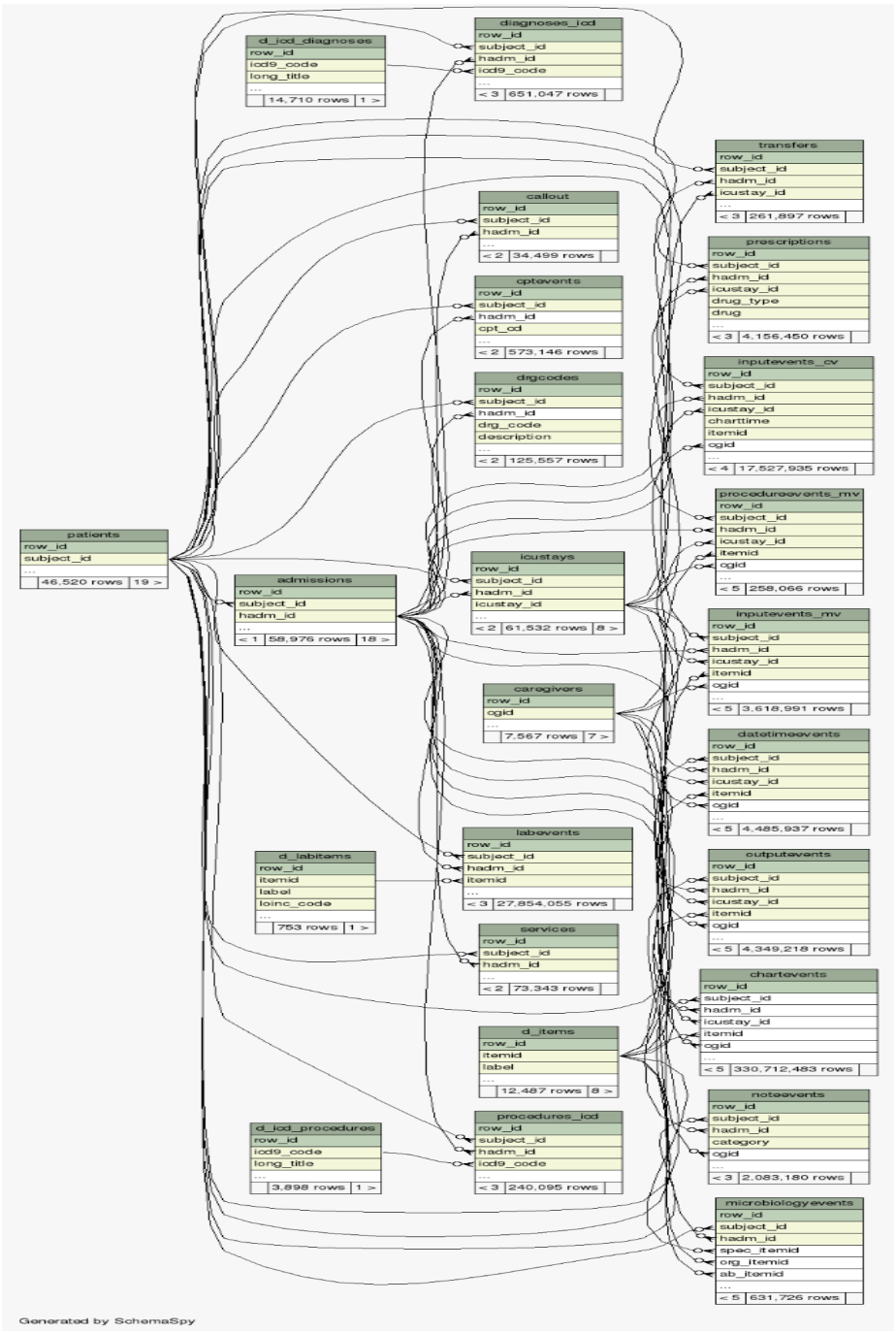
NOTE for Medications:
= New Drug * = Altered Dose

Details of Discharge Medications

Medication : Panadeine forte
Dose : 2
Schedule 8 : Y
Route : po
Reason : Post operative pain
New Drug : Y
Altered Dose : N
Supply Needed : Y
Frequency : QID PRN
Number of tablets : 20.00
Duration : 3 days
Unchanged : N
Variable Dose : N

Fig 1. An example of discharge summary retrieved from Flickr. Ruptured Appendix - Discharge Letter p2 by Jimee, Jackie, Tom & Asha is licensed under **CC BY-SA 2.0**

B. The complete schema of the MIMIC-III dataset. Generated by SchemaSpy



References

Ahire, J.B. (2018). Introduction to Word Vectors. DZone. https://dzone.com/articles/introduction-to-word-vectors#:~:text=Word%20vectors%20are%20simply%20vectors,the%20meaning%20of%20a%20word.&text=In%20essence%2C%20traditional%20approaches%20to,in%20a%20very%20naive%20way.

Baruah, P. (2020). Predicting Hospital Readmission using Unstructured Clinical Note Data.

Basu, J., Hanchate, A., & Bierman, A. (2018). Racial/ethnic disparities in readmissions in US hospitals: the role of insurance coverage. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 55, 0046958018774180.

Brownlee, J. (2016, April 11). Naive Bayes for Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/naive-bayes-for-machine-learning/

Deschepper M, Eeckloo K, Vogelaers D, Waegeman W. A hospital wide predictive model for unplanned readmission using hierarchical ICD data. *Comput Methods Programs Biomed*. 2019;173:177-183. doi:10.1016/j.cmpb.2019.02.007

Ferro, E. G., Secemsky, E. A., Wadhera, R. K., Choi, E., Strom, J. B., Wasfy, J. H., ... & Yeh, R. W. (2019). Patient readmission rates for all insurance types after implementation of the hospital readmissions reduction program. Health Affairs, 38(4), 585-593.

Foresee Medical. (n.d.). Natural Language Processing in Healthcare. https://www.foreseemed.com/natural-language-processing-in-healthcare#:~:text=Natural%20language%20processing%20(NLP)%20is,human%20speech%20terms%20and%20text.&text=The%20adoption%20of%20-natural%20language,mammoth%20amounts%20of%20patient%20datasets.

Gandhi, R. (2018). Support Vector Machine - Introduction to Machine Learning Algorithms. Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Garbade, M.J. (2018, October 15). A Simple Introduction to Natural Language Processing. Medium. https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32#:~:text=Natural%20Language%20Processing%2C%20usually%20shortened,a%20manner%20that%20is%20valuable.

Hasan, S. (2019). Using Zipf's Law to improve Neural Language Models. Medium. https://medium.com/@_init_/using-zipfs-law-to-improve-neural-language-models-4c3d66e6d2f6

Healthstream Blog (2020). The economic & emotional cost of hospital readmissions. Healthstream. https://www.healthstream.com/resources/blog/blog/2020/06/02/the-economic-emotional-cost-of-hospital-readmissions

Health Catalyst Editors. (2019). Healthcare NLP: The Secret to Unstructured Data's Full Potential. HealthCatalyst. https://www.healthcatalyst.com/insights/how-healthcare-nlp-taps-unstructured-datas-potential

Hugging Face. (n.d.). BERT. Hugging Face Transformers. https://huggingface.co/transformers/model_doc/bert.html

IBM Cloud Education. (2020, July 2). Natural Language Processing. IBM. https://www.ibm.com/cloud/learn/natural-language-processing

Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition

Kamalodeen S. (2020). How to write a Discharge Summary. GeekyMedics. https://geekymedics.com/how-to-write-a-discharge-summary/#:~:text=A%20discharge%20summary%20is%20a,care%20team%20and%20aftercare%20providers

Kelly, A., & Johnson, M. A. (2021). Investigating the Statistical Assumptions of Naïve Bayes Classifiers. In 2021 55th Annual Conference on Information Sciences and Systems (CISS) (pp. 1-6). IEEE.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. I. (2003). GENIA corpus—a semantically annotated corpus for bio-text mining. Bioinformatics, 19(suppl_1), i180-i182.

Kind, A. J., & Smith, M. A. (2008). Documentation of mandated discharge summary components in transitions from acute to subacute care. Advances in patient safety: new directions and alternative approaches (Vol. 2: culture and redesign).

Legault, K., Ostro, J., Khalid, Z., Wasi, P., & You, J. J. (2012). Quality of discharge summaries prepared by first year internal medicine residents. BMC medical education, 12(1), 1-6.

Lehn, S. F., Zwisler, A. D., Pedersen, S. G. H., Gjørup, T., & Thygesen, L. C. (2019). Development of a prediction model for 30-day acute readmissions among older medical patients: the influence of social factors along with other patient-specific and organisational factors. *BMJ open quality*, *8*(2), e000544.

Li, Z., Xing, X., Lu, B., & Li, Z. (2019). Early Prediction of 30-day ICU Re-admissions Using Natural Language Processing and Machine Learning. *arXiv preprint arXiv:1910.02545*.

Lin, Y. W., Zhou, Y., Faghri, F., Shaw, M. J., & Campbell, R. H. (2019). Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PloS one, 14(7), e0218942.

Louden, K. (2009, March). Creating a better Discharge Summary. ACP Hospitalist. https://acphospitalist.org/archives/2009/03/discharge.htm

Malfanti, F., Panaro, D., & Riccomagno, E. (2017). An Online Algorithm for Online Fraud Detection: Definition and Testing. In Adaptive Mobile Computing (pp. 83-107). Academic Press.

Marcus, R. W. E. H. M., Palmer, M., Ramshaw, R. B. S. P. L., & Xue, N. (2017). OntoNotes: A Large Training Corpus for Enhanced Processing.

Meyers, D., & Brady, J. (2020, March). Rethinking the Role of Primary Care in Reducing Hospital Readmissions. AHRQ. https://www.ahrq.gov/news/blog/ahrqviews/rethinking-role-of-primary-care.html

Misra, S., & Li, H. (2019). Noninvasive fracture characterization based on the classification of sonic wave travel times. Machine Learning for Subsurface Characterization, 243-287.

Navlani A. (2018). AdaBoost Classifier in Python. DataCamp. https://www.datacamp.com/community/tutorials/adaboost-classifier-python (visited on XXX)

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). Scispacy: Fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669.

Peng, Y., Chen, Q., & Lu, Z. (2020). An empirical study of multi-task learning on BERT for biomedical text mining. arXiv preprint arXiv:2005.02799.

Piantadosi S. T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. Psychonomic bulletin & review, 21(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

SchemaSpy. (2017, October 12). SchemaSpy Analysis of mimic.mimiciii - All Relationships. https://mit-lcp.github.io/mimic-schema-spy/relationships.html

SchemaSpy. (2017, October 12). Table mimic.mimiciii.admissions. https://mit-lcp.github.io/mimic-schema-spy/tables/admissions.html

SchemaSpy. (2017, October 12). Table mimic.mimiciii.noteevents. https://mit-lcp.github.io/mimic-schema-spy/tables/noteevents.html

Smeraglio, A., Heidenreich, P. A., Krishnan, G., Hopkins, J., Chen, J., & Shieh, L. (2019). Patient vs provider perspectives of 30-day hospital readmissions. BMJ open quality, 8(1), e000264.

Support-vector machine. (2021, April 30). In Wikipedia. https://en.wikipedia.org/wiki/Support-vector_machine

Statistic Solutions. (n.d.). What is Logistic Regression. ***https://www.statisticssolutions.com/what-is-logistic-regression/***

Swaminathan, S. (2018, March 15). Logistic regression - detailed overview. Towards Data Science. ***https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc***

Thanda, A. (2021, April 12). What is Logistic Regression? A Beginner's Guide. CareerFoundry. ***https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#:~:text=Logistic%20regression%20is%20used%20when,it%20into%20two%20seperate%20classes***.

Van Walraven, C., Seth, R., Austin, P. C., & Laupacis, A. (2002). Effect of discharge summary availability during post-discharge visits on hospital readmission. Journal of general internal medicine, 17(3), 186-192.

Whitney, P., & Chuang, E. J. (2016). Relationship between insurance and 30-day readmission rates in patients 65 years and older discharged from an acute care hospital with hospice services. Journal of hospital medicine, 11(10), 688–693. https://doi.org/10.1002/jhm.2613

Wilson, L. (2019, June 26). MA patients' readmission rates higher than traditional Medicare, study finds. HealthcareDive. ***https://www.healthcaredive.com/news/ma-patients-readmission-rates-higher-than-traditional-medicare-study-find/557694/***

Yse, D.L. (2019, January 15). Your Guide to Natural Language Processing. Towards Data Science. https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

Zhao, P., Yoo, I., & Naqvi, S. H. (2021). Early Prediction of Unplanned 30-Day Hospital Readmission: Model Development and Retrospective Data Analysis. *JMIR Medical Informatics*, *9*(3), e16306.

Zipf's law. (2021, April 6). In Wikipedia. https://en.wikipedia.org/wiki/Zipf's_law