greatlearning

PGPDSE FT Capstone Project – Final Report

Group 1

Prediction of Patient Readmission due to Hyperglycaemia
Supervised Learning: Classification

Members:

Aishwarya S

Akshaya Ganesh

Nagarjun Gunapooti

Rahul Konwar

Rishi Tiwari

Udayan Jadhav

Mentor:

Mr. Animesh Tiwari

greatlearning

Table of Contents

Problem Statement:

Patient readmission is a major concern for hospitals in USA. The Centers for Medicare and Medicaid Services (CMS) is the U.S. federal agency that works with state governments to manage the Medicare program, and administer Medicaid and the Children's Health Insurance program. Unlike India, healthcare sector in USA is primarily run by private sector. Therefore, to keep a check on healthcare services quality and safeguard patient's rights CMS plays a vital role.

CMS runs a program called **Hospital Readmissions Reduction Program (HRRP)** whose prime focus is to statistically measure readmission rates of hospitals and provide this report to both hospital and Federal Government [1]. Hospitals have some time to challenge the report for corrections. If a hospital is found to have higher readmission rates than national average and it is unable to prove otherwise, it is penalized and it gets less incentives from the Federal Government. Along with this, patients who have time to select their healthcare provider check for CMS reports before choosing their healthcare provider. So, it becomes crucial for the hospitals to get their job done in **'First Time Right'** manner. Last year 2020 itself, CMS penalized half of US hospitals for too many readmissions [2].

The HRRP 30-day risk standardized unplanned readmission measures include:

- Unplanned readmissions that happen within 30 days of discharge from the index (i.e., initial) admission.
- Patients who are readmitted to the same hospital, or another applicable acute care hospital for any reason.

Readmissions to any applicable acute care hospital are counted, no matter what the principal diagnosis was. The measures exclude some planned readmissions [1].

If we are able to predict the probability of a patient to be readmitted within 30 days, it can save thousands, if not millions of US dollars for hospitals. Also, it is beneficial for the patients too in terms of time and savings. So, our project has both Business and Social value.

Objective:

To analyse 130 hospitals dataset and create a model which can predict whether a patient will be readmitted before discharging from the hospital. Our focus will be on recall for readmission, as if a patient who was at risk of readmission was skipped by the model, it will incur loss for the hospital.

Methodology:

We followed the CRISP DM methodology in which we had go back and forth many times to achieve a satisfactory performance.

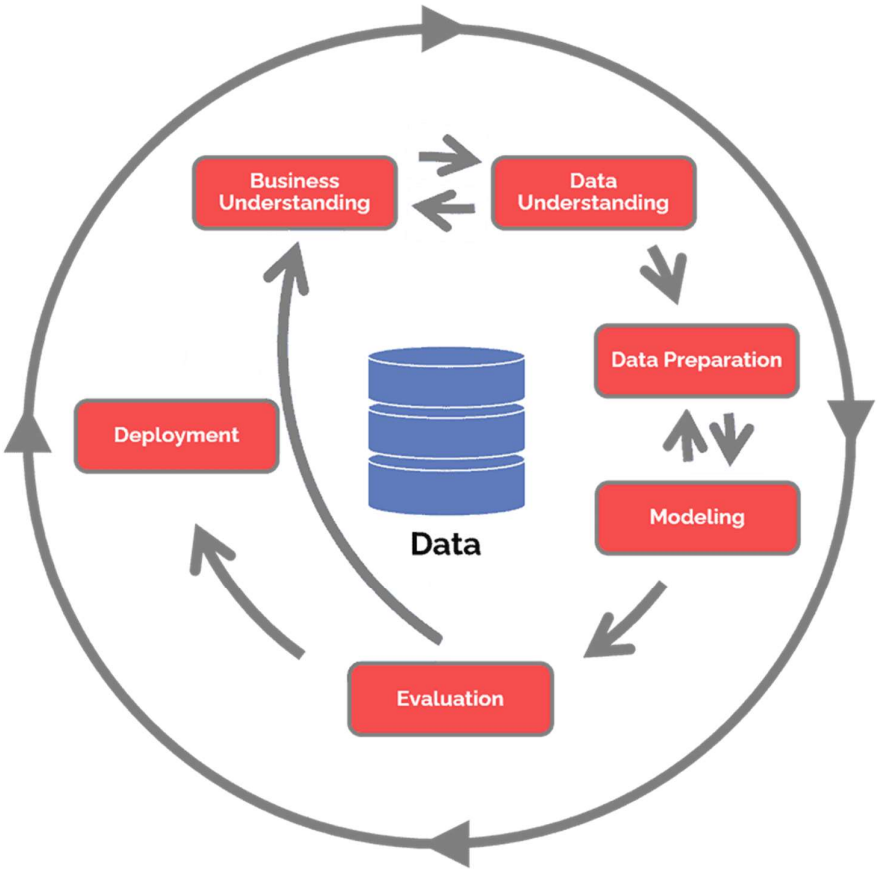

Figure 1: CRISP DM process Flow

In our journey towards completion of this project, we got the opportunity to learn a lot of new things which we will share in further reading. Initially, performance metrics were very poor for all algorithms we tried and despite numerous attempts we were unable to improve the performance. But by iterating through Business understanding, Data understanding, Data preparation, Modelling and Evaluation we were able to get better results which we will discuss as we read further.

Business Understanding:

Initially as we started working on the dataset our objective and goal was to analyse data and understand the features which affect patient readmission. At that time, we had the objective to reduce patient readmission through prediction based on data to reduce resource wastage and reduce financial burden on the patient. So, we were trying to model based on two different ways:

1. Multiclass classification where '<30','>30' and 'No readmission' was to be classified

2. Binary classification where '<30' and '>30' was one class and No readmission was another class.

But while reading and going through numerous articles and blogs about US healthcare system to gain domain knowledge, we came across Centers for Medicare and Medicaid Services website. CMS is a US Federal Government agency responsible for overseeing healthcare sector ensuring best practices are being practised. CMS runs a program called **Hospital Readmissions Reduction Program (HRRP)** which aims to reduce readmissions by penalizing hospitals with higher readmission rates. CMS includes the following six condition or procedure-specific 30-day risk-standardized unplanned readmission measures in the program:

- Acute myocardial infarction (AMI)
- Chronic obstructive pulmonary disease (COPD)
- Heart failure (HF)
- Pneumonia
- Coronary artery bypass graft (CABG) surgery
- Elective primary total hip arthroplasty and/or total knee arthroplasty (THA/TKA)

CMS calculates the payment reduction and component results for each hospital based on its performance during a rolling performance period. The payment adjustment factor is the form of the payment reduction CMS uses to reduce hospital payments. Payment reductions are applied to all Medicare fee-for-service base operating diagnosis-related group payments during the FY (October 1 to September 30). The payment reduction is capped at 3 percent (that is, a payment adjustment factor of 0.97).

CMS sends confidential Hospital-Specific Reports (HSRs) to hospitals annually. CMS gives hospitals 30 days to review their HRRP data as reflected in their HSRs, submit questions about the calculation of their results, and request calculation corrections. The Review and Correction period for HRRP is only for discrepancies related to the calculation of the payment reduction and component results.

After the Review and Correction period, CMS reports HRRP data in the Inpatient Prospective Payment System/Long-Term Care Hospital Prospective Payment System Final Rule Supplemental Data File on CMS.gov. In addition, CMS reports hospitals' HRRP data on Hospital Compare or the successor website [1].

So, after coming to know about CMS and HRRP, we decided to change our perspective and performed Binary classification in which '<30' was one class and the rest were another class. This step was very significant as before using this step our model was not performing well, no matter what method we tried. But as we realigned our model to the real-world logic of the problem and changed our classification based on CMS regulations, model started performing better than before.

Data Understanding

Dataset and Domain

The dataset we chose to work on is **Diabetes 130-US hospitals for years 1999-2008 Data Set** available on University of California, Irvine Machine Learning Data Repository. This dataset was submitted on behalf of the Centre for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grants UL1 TR00058 and a recipient of the CERNER data. John Clore, Krzysztof J. Cios, Jon DeShazo, and Beata Strack. This data is a de-identified abstract of the Health Facts database (Cerner Corporation, Kansas City, MO). It categorizes as healthcare domain dataset.

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria:

1.It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatients, inpatient, and emergency visits in the year before the hospitalization, etc [5].

We are grateful to University of California, Irvine, USA to allow students all across the world to let access their data repository.

Data Dictionary:

- **Encounter ID:** Unique identifier of an encounter
- **Patient number:** Unique identifier of a patient
- **Race:** Caucasian, Asian, African American, Hispanic, and other
- **Gender:** male, female, and unknown/invalid
- **Age:** Grouped in 10-year intervals like 0, 10), 10, 20), …, 90, 100)

- **Weight:** Weight in pounds
- **Admission type:** Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, new-born, and not available
- **Discharge disposition:** Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
- **Admission source:** Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
- **Time in hospital:** Integer number of days between admission and discharge
- **Payer code:** Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical
- **Medical specialty:** Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
- **Number of lab procedures:** Number of lab tests performed during the encounter
- **Number of procedures:** Numeric Number of procedures (other than lab tests) performed during the encounter
- **Number of medications:** Number of distinct generic names administered during the encounter
- **Number of outpatient visits:** Number of outpatient visits of the patient in the year preceding the encounter
- **Number of emergency visits:** Number of emergency visits of the patient in the year preceding the encounter
- **Number of inpatient visits:** Number of inpatient visits of the patient in the year preceding the encounter
- **Diagnosis 1:** The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
- **Diagnosis 2:** Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
- **Diagnosis 3:** Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
- **Number of diagnoses:** Number of diagnoses entered to the system 0%
- **Glucose serum test result:** Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
- **A1c test result:** Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.

- **Change of medications:** Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
- **Diabetes medications:** Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”
- 24 different kinds of medical drugs.
- **Readmitted:** Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission

The 130 hospitals dataset is a real-world data, which means it had all sorts of incorrect entries and missing values one would expect in real scenario. Most of the columns were categorical (some ordinal and some nominal). Some of the major data complexities are mentioned below:

1. Columns like payer code, weight, medical specialty had high number of missing values. Initially we had decided to drop these 3 columns but during interim report presentation we were told to make use of weight column which had 97% missing values. Although it seemed impossible, we accepted the challenge and did manage to reduce missing values to just about 300 rows by manual imputation.
2. Target variable readmitted was highly imbalanced. We applied both over sampling and under sampling methods out of which over sampling worked better, SMOTE to be specific.
3. Diagnosis 1 to 3 columns had ICD-9M codes which were revoked back in early 2000’s. We had to map them to their respective disease type and reduce number of classes.
4. Most columns were categorical with high number of sub classes, we were restricted by this to use dummy encoding which is mostly preferred. In this situation we used label encoding on ordinal features and maintained their order.
5. Due to the high number of rows and columns, hardware resource was a big issue. Even after using Colab Pro from Google which is known to have good performance for large data, we had to wait for hours to get tuned parameters and run ensemble algorithms.

At the beginning of our project, we decided to go for multiclass classification where we had to predict 3 different classes, namely readmission in less than 30 days, readmission in more than 30 days and no readmission. While iterating through steps of CRISP DM we changed our perspective a few times and finally managed to get the correct logic. The exact changes which we performed are discussed in detail in Business Understanding section of this report.

Data Preparation

Data Cleaning:

Table 1: Feature data type

Feature	Data Type
encounter_id	Categorical (Nominal)
patient_nbr	Categorical (Nominal)
race	Categorical (Nominal)
gender	Categorical (Nominal)
age	Categorical (Ordinal)
weight	Numerical (Continuous)
admission_type_id	Categorical (Nominal)
discharge_disposition_id	Categorical (Nominal)
admission_source_id	Categorical (Nominal)
time_in_hospital	Categorical (Ordinal)
payer_code	Categorical (Nominal)
medical_specialty	Categorical (Nominal)
num_lab_procedures	Numerical (Discrete)
num_procedures	Numerical (Discrete)
num_medications	Numerical (Discrete)
number_outpatient	Numerical (Discrete)
number_emergency	Numerical (Discrete)
number_inpatient	Numerical (Discrete)
diag_1	Categorical (Nominal)
diag_2	Categorical (Nominal)
diag_3	Categorical (Nominal)
number_diagnoses	Numerical (Discrete)
max_glu_serum	Categorical (Ordinal)
A1Cresult	Categorical (Ordinal)
24 Diabetes Medications	Categorical (Nominal)
change	Categorical (Nominal)
diabetesMed	Categorical (Nominal)
readmitted	Categorical (Nominal)

Summary

The data set comprises of 13 Numerical features and 37 categorical features. But some column data type may be changed in the EDA phase or if required even during the modelling phase.

Table 2: Null/Missing Values

Feature	Missing Values	Missing Percentage
encounter_id	0	0
patient_nbr	0	0
race	1918	2.741057
gender	0	0
age	0	0
weight	67185	96.015606
admission_type_id	0	0
discharge_disposition_id	0	0
admission_source_id	0	0
time_in_hospital	0	0
payer_code	30415	43.466766
medical_specialty	33639	48.074257
num_lab_procedures	0	0
num_procedures	0	0
num_medications	0	0
number_outpatient	0	0
number_emergency	0	0
number_inpatient	0	0
diag_1	10	0.014291
diag_2	293	0.418733
diag_3	1224	1.749246
number_diagnoses	0	0
max_glu_serum	0	0
A1Cresult	0	0
metformin	0	0
repaglinide	0	0
nateglinide	0	0
chlorpropamide	0	0
glimepiride	0	0
acetohexamide	0	0
glipizide	0	0
glyburide	0	0
tolbutamide	0	0
pioglitazone	0	0
rosiglitazone	0	0
acarbose	0	0
miglitol	0	0
troglitazone	0	0
tolazamide	0	0
examide	0	0
citoglipton	0	0
insulin	0	0
glyburide-metformin	0	0
glipizide-metformin	0	0
glimepiride-pioglitazone	0	0
metformin-rosiglitazone	0	0
metformin-pioglitazone	0	0
change	0	0
diabetesMed	0	0
readmitted	0	0

RED BARS INDICATE MISSING DATA

The data set had a lot of missing values and dependent observations due to repeated patient numbers. Therefore, the first step was to clean the data and make it suitable for doing any analysis. The steps taken were:

1. We dropped the duplicate patient number instances as it would have made it impossible to apply any machine learning algorithm due to dependent observations. We kept the first instance of every duplicate patient number and dropped the rest. We came down to 71518 observations from 101766 observations. Data lost: 29.72%.
2. There were some records where the patient was either dead or sent to hospice, such cases cannot be considered for readmission and were removed. Number of records came down to 69973. Data lost: 2%.
3. In the race features Caucasian and African Americans are the two major races dominating among patients. The other categories namely Hispanic, Asian and others are contributing only 4 % in total, hence we combine them into one single group called others.
4. In some of the records the gender is unknown/invalid. That is very low in percentage among all records, so we are dropping that record. After dropping we are losing 0.004287% of data in this stage.
5. Admission_type_id, discharge_disposition_id, admission_source_id columns were present as numerical in the data set, but are IDs. They should be considered categorical, so we have changed their type.
6. We have grouped Not Available, NULL and Not Mapped as one category 'Not Available'. Thereby, we have only six categories under Admission_type_id i.e., Emergency, Elective, New Born, Trauma Centre and Not Available. We further club together 4,5,6,7 and 8 admission_type to once category ‘Not Available’.
7. All the categories in discharge_disposition_id feature which contains the word 'home' are grouped into a category called 'Discharged to Home'. NULL, Not Mapped and Unknown/Invalid are grouped into a category called 'Unknown'. Rest all are grouped into a category called 'Other'.
8. Physician Referral, Emergency Room are the two major admission sources of all patients (80%), hence we merge all other sources as 'others'.
9. diag_1, diag_2, diag3 seems to be columns of continuous nature, but at a closer look they contain alpha-numeric and numeric values and each one of them specifies the ICD9-CM codes of procedure that was conducted for patients. Hence making these as categorical columns. As per the procedural norms, they are sequentially dependent in nature. So, there shouldn’t be values in diag_2 unless diag_1 is present, similarly there shouldn’t be values present in diag_3 column unless diag_2 is done for every patient.
10. There are literally hundreds of categories in the primary diagnosis columns, so we can rule off one hot encoding which only further leads to the curse of dimensionality. Hence, we set up limits and specify each value into the following category if that value is present in that limit, groups that covered less than 3.5% of encounters were grouped into “other” category.

    Circulatory --> 390-459,785 --> Diseases of circulatory system

Respiratory --> 460-519,786 --> Diseases of respiratory system
Digestive --> 520-579,787 --> Diseases of digestive system
Diabetes --> 250.xx --> Diabetes mellitus
Injury --> 800-999 --> Injury and poisoning
Musculoskeletal --> 710-7399 --> Disease of musculoskeletal
Genitourinary --> 580–629, 788 --> Diseases of the genitourinary system
Neoplasms --> 140–239 --> Neoplasms
Pregnancy --> 630–679 --> Complications of pregnancy, childbirth, and the puerperium
Other

11. Since three medications (examide, citoglipton, glimepiride-pioglitazone) are having only one category, indicates that medication is not prescribed to any patients. So, it not useful for our analysis, we are removing 3 features in this stage.
12. We have converted age feature to continuous from categorical by taking mean of maximum and minimum in each category.
13. From correlation matrix it was evident that there is no multi collinearity in the cleaned data.
14. There are scale imbalances and outliers in ‘age’, 'time_in_hospital', 'num_lab_procedures', 'num_procedures', 'num_medications', 'number_diagnoses', 'total_visits', we will do feature scaling if required at the time of model building.
15. None of the numeric columns are normally distributed.
16. In the readmitted feature, replaced ‘>30’ and ‘No’ with 0, and ‘<30’ with 1.

Exploratory Data Analysis:

First, we take a look at our target variable distribution through count plot.

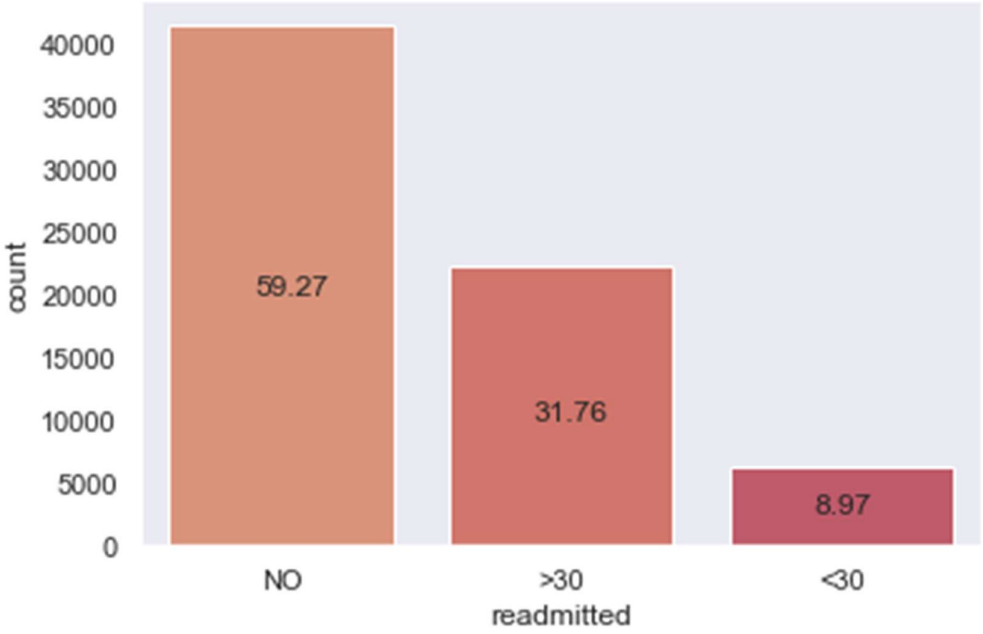

Figure 2: Readmitted Count plot

As is clear from the count plot, the target variable is highly imbalanced, we expect our models to not be able to correctly predict minority class without reducing this imbalance.

Gender:

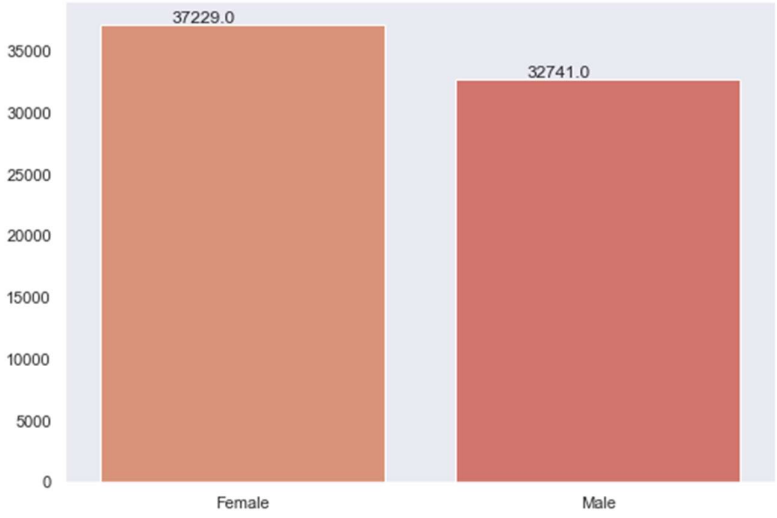

Figure 3: Gender Count plot

There are more female patients than male patients but the difference is not very significant.

Figure 4: Gender VS Readmitted Stacked Bar Chart

Both males and females are equally likely to be readmitted.

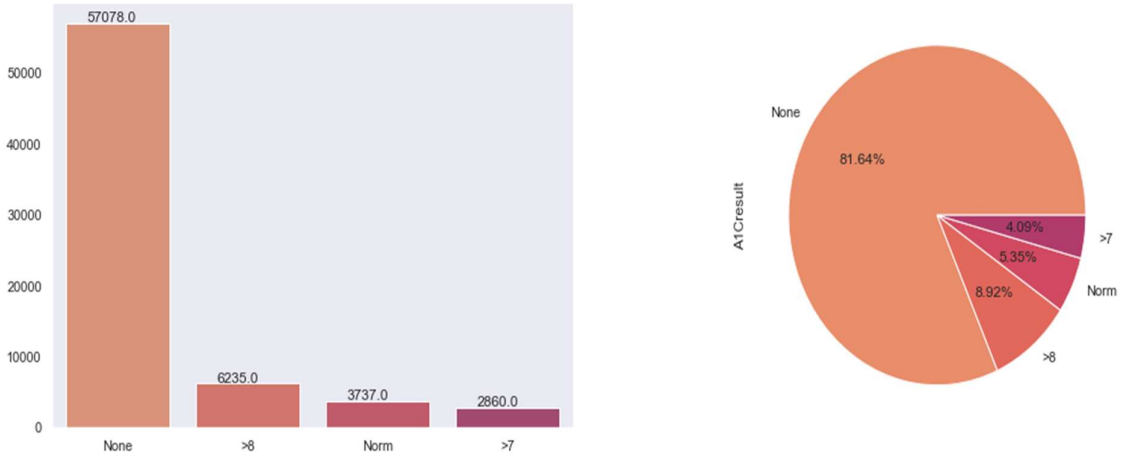

Figure 5: A1C(diabetes) test distribution

Most patients were not tested for diabetes, of those tested most belong to greater than 7 and greater than 8 categories.

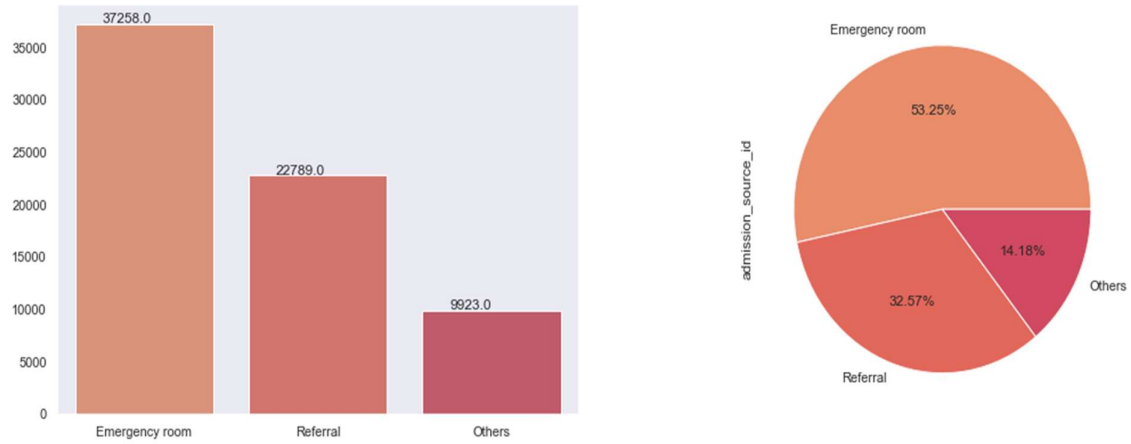

Figure 6: Admission source distribution

Most admissions are of emergency category followed by Referral and Others.

Table 3: Kruskal-Wallis test results for continuous features:

Column Name	P Values	Interpretation
age	3.6807831361951664e-97	Significant
time_in_hospital	3.4263709253832047e-96	Significant
num_lab_procedures	2.4522046342974965e-51	Significant
num_procedures	2.714690873037613e-18	Significant
num_medications	1.0581823172506986e-64	Significant
number_outpatient	4.163415382232466e-119	Significant
number_emergency	4.007131981024854e-125	Significant
number_inpatient	0.0	Significant
number_diagnoses	1.8822381933120828e-177	Significant

Table 4: Chi square test for categorical features:

Column Name	P Values	Interpretation
race	2.210285358836962e-37	Significant
gender	4.382243953761633e-05	Significant
admission_type_id	4.755094500032859e-66	Significant
discharge_disposition_id	5.99387316313625e-180	Significant
admission_source_id	2.0592947274029787e-54	Significant
medical_specialty	1.192720161475861e-24	Significant
diag_1	2.9054265334564585e-54	Significant
diag_2	8.53111747859959e-33	Significant
diag_3	8.970647658197248e-31	Significant
max_glu_serum	4.712898262609592e-07	Significant
A1Cresult	7.236372803358732e-09	Significant
metformin	7.630166138689623e-08	Significant
glimepiride	0.29586629127242703	Not Significant
glipizide	5.797853465758341e-11	Significant
glyburide	0.24042965399765617	Not Significant
tolbutamide	0.8628173149523063	Not Significant
pioglitazone	6.144171207702202e-05	Significant
rosiglitazone	2.5659100815467838e-08	Significant
insulin	2.471071802625426e-30	Significant
glyburide-metformin	0.3597380083023351	Not Significant
glipizide-metformin	0.3062417216016386	Not Significant
change	1.035354360424855e-20	Significant
diabetesMed	1.436778408726011e-57	Significant

Modeling and Evaluation

Modeling iteration 1:
Initially we treated the problem as multiclass classification with 3 categories to predict, namely readmission in less than 30 days, readmission in greater than 30 days and no readmission. We used random forest classifier because it can handle both categorical and numerical features well and we had both in our dataset. The results for this model were very poor as shown poor.

	precision	recall	f1-score	support
0	0.624011	0.906343	0.739133	12439.000000
1	0.476125	0.207135	0.288681	6643.000000
2	0.250000	0.002115	0.004195	1891.000000
accuracy	0.603347	0.603347	0.603347	0.603347
macro avg	0.450045	0.371865	0.344003	20973.000000
weighted avg	0.543447	0.603347	0.530192	20973.000000

After random forest we tried Ada boost, Gradient boost and XG boost, but the results were very similar to what we got in random forest. We had anticipated this due to high data imbalance as shown in figure 2. So now we had to go back to the data preparation stage and look for some solution. Also, during this time, we had Interim Project presentation where the faculty advised us to impute weight column which had 97% Null values. Although it seemed impossible to do so, we started working on how to impute 97% missing values.

We now had two things to focus on: 1. To increase model performance to acceptable level and 2. Impute weight column with 97% null values.

After a lot of iterations between modeling and data preparation we realised that maybe we have got something missing in the business understanding itself. So, we started learning about the domain, i.e., US healthcare system. During this time, we realised both our approaches, namely:

1. To consider readmission within 30 days, readmission in more than 30 days and no readmission as 3 different classes
2. To consider readmission within 30 days plus readmission in more than 30 days as one class and No readmission as another class

were both not as per the real-world logic. In US healthcare system, CMS penalizes hospitals when readmissions less than 30 days are high, and other two classes are not included in HRRP. So, this was one significant step in moving towards a better performing model as we realigned our assumptions and made them relevant to actual case.

At the same time, we sliced the data which had weight given and studied its relationship with other variables. We could not implement imputation methods like K Nearest Neighbours, Multivariate Imputation by Chained Equations because these cannot work with null values as

high as 97%. So, we studied relations between weight and other features and we did find relationship between age-weight, age-diabetesMed, age-gender. Based on these relationships we started imputing weight one by one class manually and were finally able to reduce Null value rows to around 300 rows. Whether this imputation helps in increasing model performance was yet to be seen.

Modeling iteration 2:

This time we considered two classes, one with readmissions within 30 days, and another with readmissions in more than 30 days plus No readmissions. Weight is still not considered as up to this point, we were still working on imputing weight. The results are as below:

Model_Name	Recall(0)	Recall(1)	F1_Score(0)	F1_Score(1)
RandomForest	100%	0%	95%	0%
Gradient_Boosting	100%	0%	95%	0%
XG_Boosting	100%	0%	96%	16%

Results were again not at all satisfactory and we had to consider over sampling techniques. So further down the line we used SMOTE to decrease data imbalance. We applied SMOTE step wise and did not directly go for completely balanced data.

Model Iteration 3:

SMOTE Iteration 1
We applied SMOTE to dataset with sampling strategy as 0:63638 and 1:30000. The results we obtained were:

Model_Name	Recall(0)	Recall(1)	F1_Score(0)	F1_Score(1)
RandomForest	99%	78%	95%	87%
Gradient_Boosting	99%	72%	93%	82%
XG_Boosting	98%	77%	94%	85%

This was a considerable improvement from previous models. In this SMOTE iteration, random forest gave the best recall for class 1. But this is not satisfactory as the project is from healthcare domain.

SMOTE Iteration 2
We applied SMOTE to dataset with sampling strategy 0:63638 and 1:35000. Results:

Model_Name	Recall(0)	Recall(1)	F1_Score(0)	F1_Score(1)
RandomForest	99%	81%	95%	89%
Gradient_Boosting	98%	76%	93%	85%
XG_Boosting	98%	80%	94%	87%

The performance improved further and best recall for class 1 was again given by random forest. But still, this is not satisfactory.

SMOTE Iteration 3
We applied SMOTE to dataset with sampling strategy 0:63638 and 1:40000. Results:

Model_Name	Recall(0)	Recall(1)	F1_Score(0)	F1_Score(1)
RandomForest	99%	83%	94%	90%
Gradient_Boosting	97%	78%	92%	86%
XG_Boosting	97%	82%	93%	88%

Recall for class 1 increased by few percent but still not satisfactory.

SMOTE Iteration 4
We applied SMOTE to dataset with sampling strategy 0:63638 and 1:63638. Results:

Model_Name	Recall(0)	Recall(1)	F1_Score(0)	F1_Score(1)
RandomForest	98%	90%	94%	94%
Gradient_Boosting	94%	88%	91%	91%
XG_Boosting	96%	89%	93%	92%

Recall for class one is now in satisfactory range, we can further tune the model now.

Final Tuned Model
We used grid search to tune our model with optimum hyperparameters and got the following results:

Model_Name	Recall(0)	Recall(1)	F1_Score(0)	F1_Score(1)
RandomForest	89%	78%	85%	83%
XG_Boosting	96%	89%	93%	92%

By this time, we were done with imputing null values in weight column, so we made a XG boost model to see if it helps:

Model_Name	Recall(0)	Recall(1)	F1_Score(0)	F1_Score(1)
XG_Boosting	96%	88%	91%	92%

Even though the recall for class 1 in XG boost with imputed weight is good, but it is 1% less than recall for class 1 without weight imputed. So, imputing the weight did not help in improving the results but it did give us a good challenge to learn more.

Final Model

The final model which we selected was XG boost. The details of hyperparameters used is given in the image below:

```
xg= XGBClassifier(n_estimators=200,learning_rate=0.4,max_depth=8,gamma=9)
xg.fit(xtrain_smote,ytrain_smote)
trainpred=xg.predict(xtrain_smote)
testpred=xg.predict(xtest)
print("-----------------TRAIN REPORT--------------")
print(classification_report(ytrain_smote,trainpred))
print("-----------------TEST REPORT--------------")
print(classification_report(ytest,testpred))
```

```
-----------------TRAIN REPORT--------------
              precision    recall  f1-score   support

           0       0.90      0.96      0.93     51004
           1       0.96      0.89      0.93     50816

    accuracy                           0.93    101820
   macro avg       0.93      0.93      0.93    101820
weighted avg       0.93      0.93      0.93    101820

-----------------TEST REPORT--------------
              precision    recall  f1-score   support

           0       0.89      0.96      0.93     12634
           1       0.96      0.89      0.92     12822

    accuracy                           0.92     25456
   macro avg       0.93      0.92      0.92     25456
weighted avg       0.93      0.92      0.92     25456
```

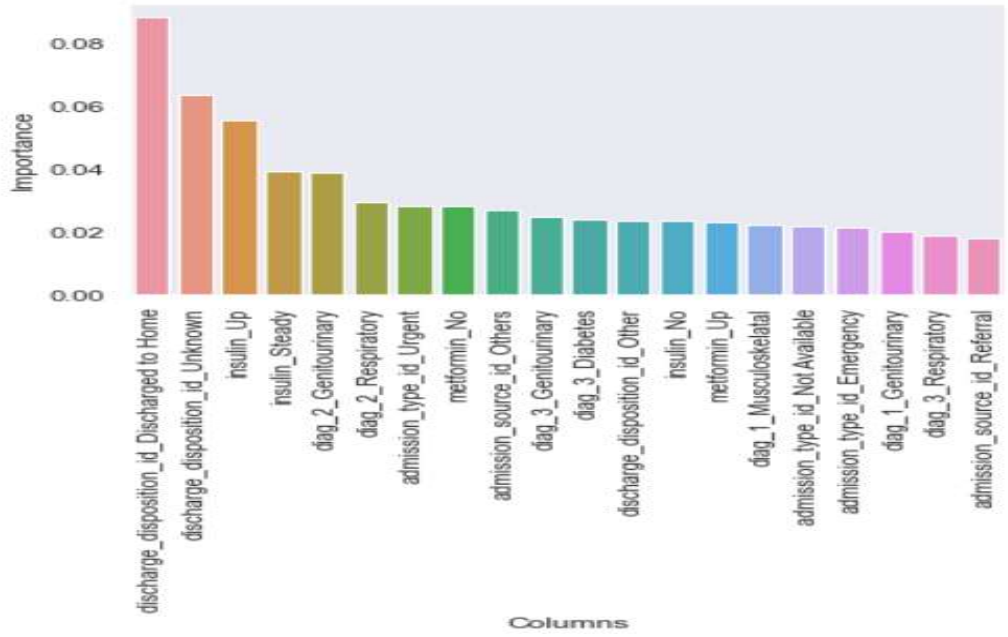

Figure 7: Important Features Bar chart

Implications

Since our focus was on reducing readmission rates so that hospitals can reduce their penalties, we had recall (1) as our prime statistic to consider for model performance. The reason being that if a patient who has high chance of getting readmitted within 30 days is discharged by the doctors and the patient is readmitted within 30 days, hospital would have to pay penalty for that. Some key notes to takeaway are:

1. The final model has acceptable recall score for both class 1 and 0.
2. Harmonic score or F1 score of the model is also in acceptable range.
3. The model can predict readmission of a patient reliably.

Limitations:

The model accuracy and performance are good but it can be better. This model can still not be as good as an experienced doctor who is good at his practice. Since recall for class 1 is .89, the doctor can do a recheck on patient if the model predicts that the patient is at risk of readmission within 30 days. But this model is in no way a substitute for domain experts(doctors). It can only suggest doctors while discharging patients if they need to recheck their patients, but the final call still depends on the doctor in charge.

Model is based on data from 1999–2008 which means it is 13 years old data, more than a decade. During this time period a lot of changes have been implemented by the CMS and other US Federal Health agencies like Obama care. Also, we must take into account the disease type prevalence in the population changes with time. Like during the 19th Century plague was more prevalent than other diseases, but towards the start of 20th Century plague was restricted to only some parts of the world.

Closing Reflections:

Prediction of readmission of patient within 30 days project gave us the opportunity to learn new things at every step. It was a very challenging yet rewarding project for us as we are beginners in the field of Data Science.

We managed to get recall from 0 to 0.89 which was something which we can confidently talk about as our first project. But there is always room for improvements. If we had a controlled data, or had the resources to validate data from its source, we might have got even better results.

The most important thing that we have learnt from here is that the code or the medium is not important, understanding business requirement and data is.

We are thankful to our mentor Mr. Animesh Tiwari for providing expert advice and recommending best practices to us during the project.

References

[1] CMS, “CMS.gov,” CMS,USA, [Online]. Available: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program.

[2] R. Leventhal, HealthcareInnovation, [Online]. Available: https://www.hcinnovationgroup.com/policy-value-based-care/readmissions-bundled-payments/news/21160954/medicare-again-penalizes-half-of-us-hospitals-for-too-many-readmissions.

[3] J. Clore, K. J. Cios, J. DeShazo and B. Strack, “UCI Machine Learning Depository,” University of California, Irvine, 03 05 2014. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008. [Accessed 08 08 2021].