Diabetic Patients Data Of 130-US hospitals for years 1999-2008

Group 15:

Anurag Kosana

Sri Manasa Valluru

Mohit Kota

Vignan Vennampally

Abstract

The main intention to choose Healthcare data is to understand the characteristics and importance of each feature in the dataset and how better models can be built using the right set of data. Around 7% of the population worldwide are suffering with Diabetes. It is a chronic disease characterized by elevated levels of blood glucose which increases the risk of stroke by 1.8 times, increases the mortality rate by 1.8 times when compared to undiagnosed diabetic patients. We will be determining the readmission of patients to the hospitals after their medication during the first visit. We have analyzed diabetic data from 130 US hospitals. Thereby considering different parameters of the patients. The class of interest is People readmitting before 30 days. We have considered the categories into people readmitting <30 days and >30 days. Different types of Machine learning models were trained to understand the patterns. Results showed some interlink between number of inpatients admits and patients admitting before 30 days. Neural Networks models showed better accuracy than any others.

INTRODUCTION

Through observing and controlling the readmission of patients suffering from Diabetic data investment in money and other resources can be saved. Our aim is to observe and reduce the readmission of diabetic patients into the hospitals. Few criteria were imposed to make sure the collected data was accurate and related to diabetes.

The present analysis of a large clinical database was undertaken to examine historical patterns of diabetes care in patients with diabetes admitted to a US hospital and to inform future directions which might lead to improvements in patient safety.

https://www.hindawi.com/journals/bmri/2014/781670/#B8

Databases of clinical data contain valuable but heterogeneous and difficult data in terms of missing values, incomplete or inconsistent records, and high dimensionality understood not only by number of features but also their complexity.

https://www.hindawi.com/journals/bmri/2014/781670/#B8

The data contains Numerical, Categorical columns from which few columns are inconsistent and missing few entries. Since the data comes from 130 US hospitals it is expected to have some anomalies in the data. Few columns in the dataset have the records of medications suggested for the patients. They have been included to see any change in those variables might affect the readmission rates.

Data Description:

The dataset represents 10 years (1999-2008) data of clinical care at 130 US hospitals. It includes 55 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 into 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 tests performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

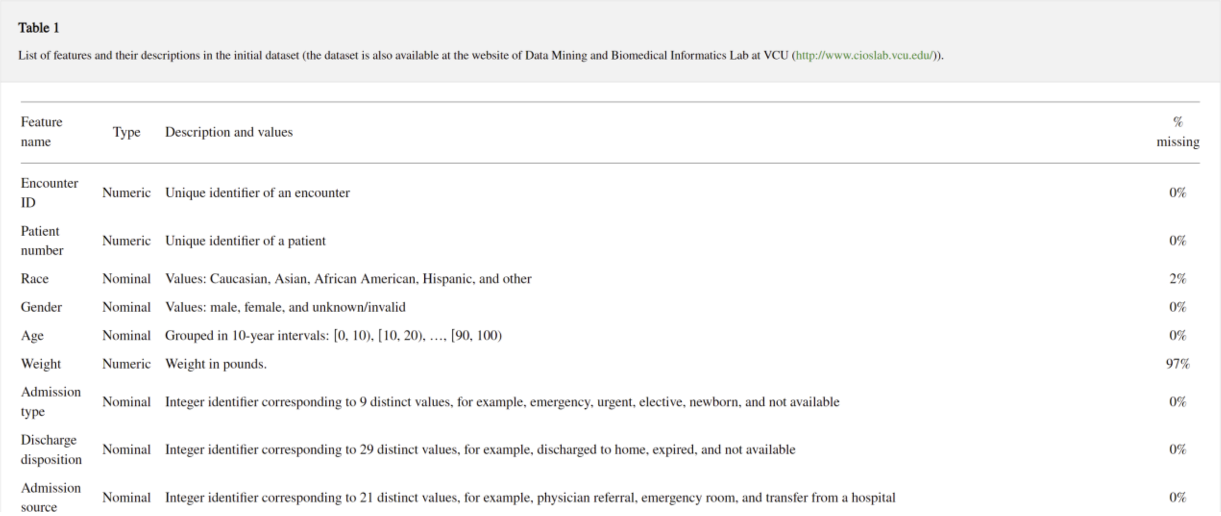


Table 1

List of features and their descriptions in the initial dataset (the dataset is also available at the website of Data Mining and Biomedical Informatics Lab at VCU (http://www.cioslab.vcu.edu/)).

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), …, [90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%

Link to Complete information on Features: Features Description, Feature Mapping

Exploratory Data Analysis

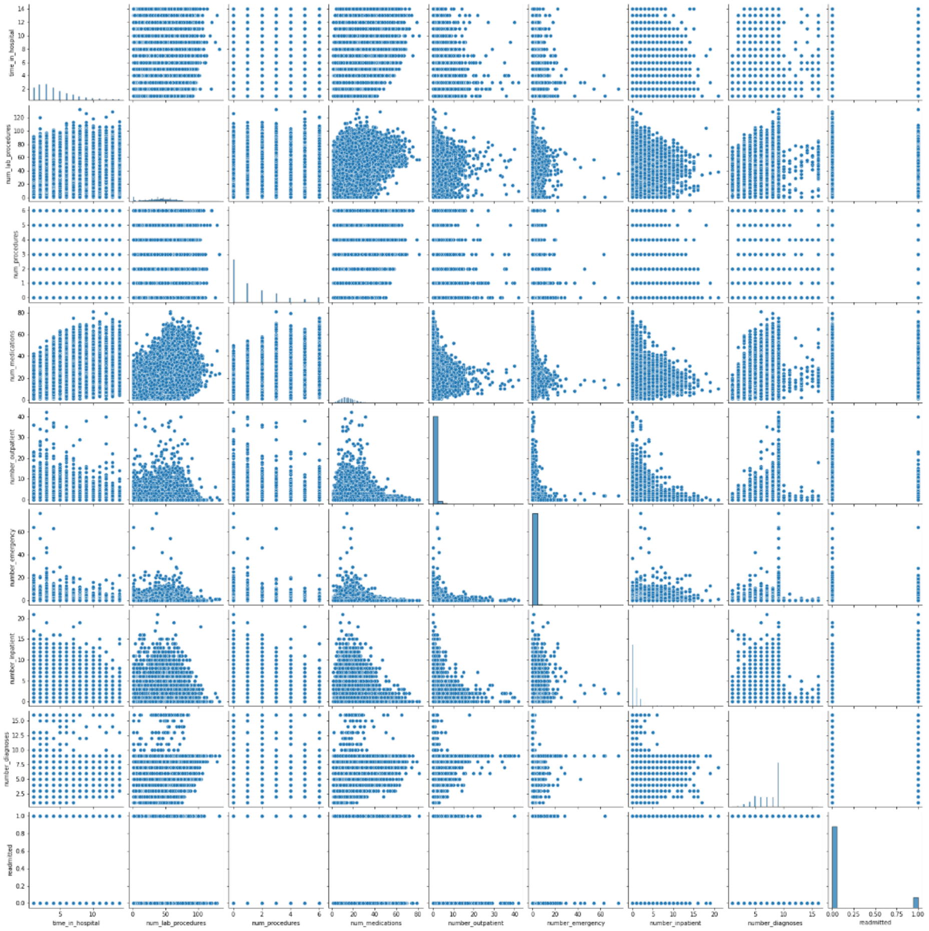
Understanding of data:

1. The data contains 13 Numerical features and 37 categorical features.
2. Few columns are unique to each row like the encounter_id, patient_nbr
3. Weight column in the Dataset has nearly 97% missing values which is dropped eventually.
4. Columns admission_type_id, discharge_disposition_id, admission_source_id are the categorical columns each entity mapping to a special category.
5. There are majority of the columns that represent medications administered by the diabetic patients. This is important to observe the change in medications that might affect the readmission rates.
6. As our focus is on hospital readmissions, we must disregard the entries of patients who might not return to the hospital. The discharge reason of the patient has been described in the discharge_disposition_id. We are removing the entries of patients whose discharge_disposition_id is Expired, Hospice / home, Hospice / medical facility, Expired at home. Medicaid only, hospice, Expired in a medical facility. Medicaid only, hospice, Expired, place unknown. Medicaid only, hospice.
7. Columns named 'examide','citoglipton' have only single value for all the entries In the data which does not add any value to the model training and hence removed
8. Since our focus is on People who readmitted before 30 days. The problem is assumed to have two classes '<30' as class label '1' and '>30' as class label '0'.
9. Columns named 'payer_code', 'medical_speciality' have almost 50% of null values. These missing values are assumed to be 'unknown' category instead of removing them

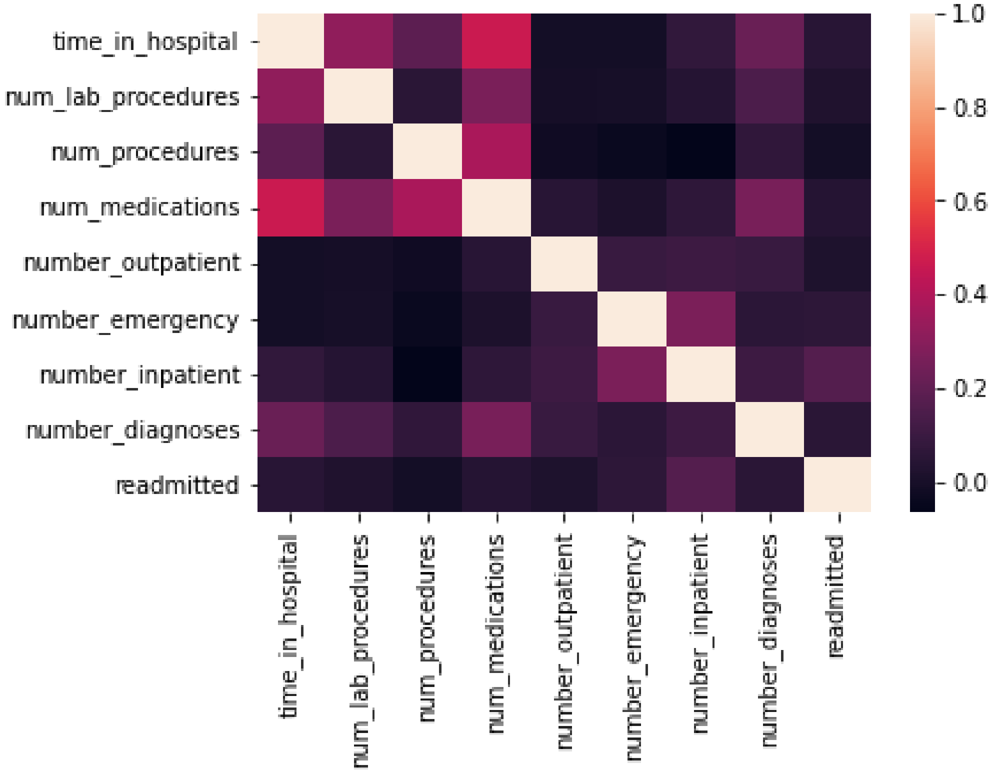
Statistical Analysis:

Multivariate Analysis Using pairplot()

1. **Sns.pairplot(df_num_cols)**


time_in_hospital
num_lab_procedures
num_procedures
num_medications
number_outpatient
number_emergency
number_inpatient
number_diagnoses
readmitted
0.00
0.25
0.50
0.75
1.00

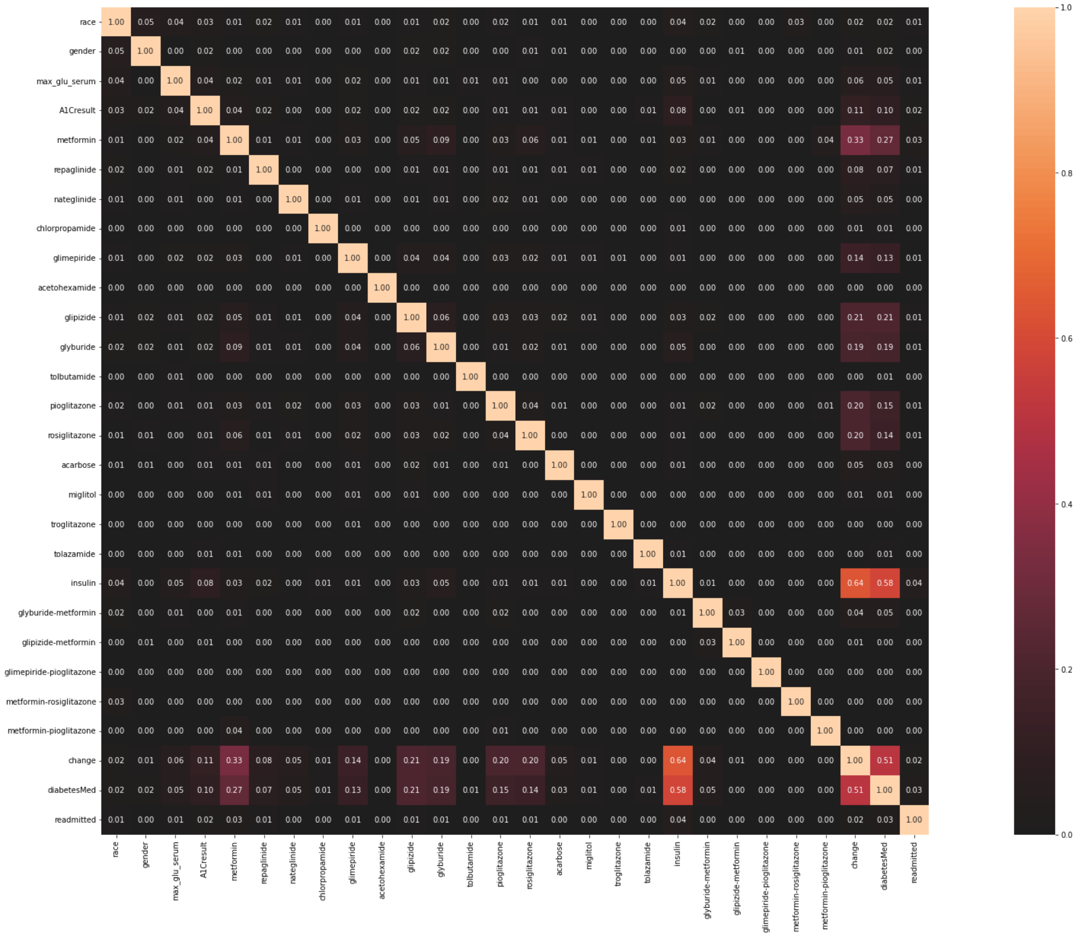
2. **Sns.histogram(df_num_cols)**



Observations:

1. It shows that the column 'number_in_patient' is mostly correlated with the outpur readmitted. Might be patients who are admitted in the hospital have higher chance of readmission within 30 days.

3. **nominal.associations(df_categorical_cols,figsize=(40,20))**



Observations:

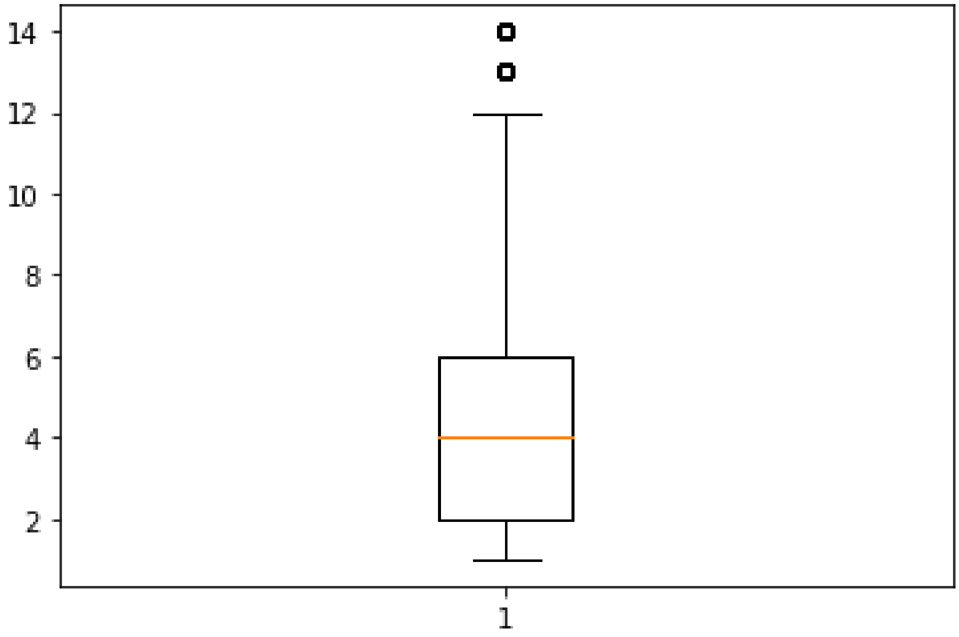
1. There is a little dependency among the independent features of the data which is a good sign to train the model.
2. The highest correlation is among the Features ‘insulin’ and ‘Change’ which also highlights change in insulin dosage had significant effect on the readmission.
3. Also Features ‘diabetesMed’ and ‘insulin’ are closely correlated as Insulin is the most prescribed and prevalent medicine for diabetes

Outlier Handling:

1. Boxplots are plotted for each of the columns to observe the outlier’s presence. It has been observed that major portion of the numerical data lies within a range and there are no outliers that differ largely from the rest of the data

```
plt.boxplot(df_num_cols['time_in_hospital'])
```

We can observe that values 12, 14 outside the 3 S.D from mean cannot be the outliers as there is possibility of people staying in hospitals for 12 or 14 days.



Feature Engineering:

1. Column ‘medical_speciality’ has 73 unique categories of values. Using one-hot encoding would create columns equal to number of different categories. To avoid this top-10 categories of the ‘medical_speciality’ has been considered which have almost 93% of data.

```
 medical_specialty
UKN                             48616
InternalMedicine                14237
Emergency/Trauma                 7419
Family/GeneralPractice           7252
Cardiology                       5279
Surgery-General                  3059
Nephrology                       1539
Orthopedics                      1392
Orthopedics-Reconstructive       1230
Radiologist                      1121
dtype: int64
```

2. ‘Age’ column in the dataset consists of bins. To make it easy for the model to train ‘Age’ values have been converted to numerical values. Example: [70-80) is converted to 75

```
age_id = {'[0-10)':5,
          '[10-20)':15,
          '[20-30)':25,
          '[30-40)':35,
          '[40-50)':45,
          '[50-60)':55,
          '[60-70)':65,
          '[70-80)':75,
          '[80-90)':85,
          '[90-100)':95}
dff['age_group'] = dff.age.replace(age_id)
```

3. Label column ‘readmitted’ has ‘<30’, ‘>30’, ‘NO’ values initially. Since our class of interest is ‘<30’ which is ‘1’. We are assuming the category ‘NO’ to be ‘>30’.

   df['readmitted']=df['readmitted'].replace(['NO','>30','<30'],['0','0','1'])

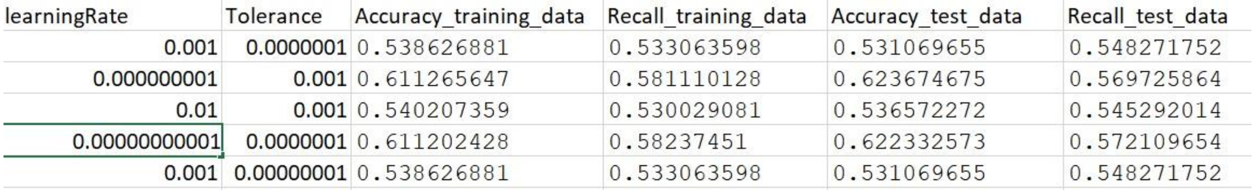
4. Features ‘ encounter_id’, ‘patient_nbr’ are removed as they represent unique row and do not contribute to model building

Model Performance

1. Logistic Regression

Our base model is chosen as Logistic Regression. Logistic Regression is an algorithm that can be used to model the probability of a certain class.
Reason for selection of Logistic Regression as model -
It is used when the data is linearly separable, independent, no outliers and the outcome is binary. It provides the binary output using the sigmoid function.

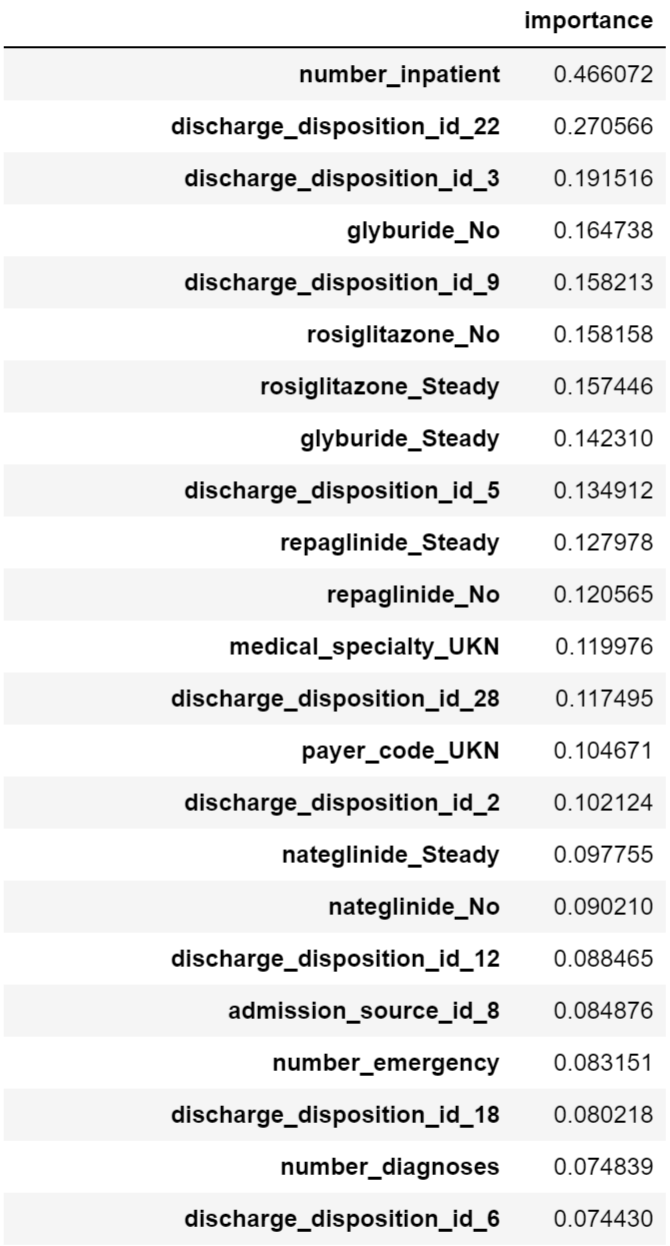
Model Performance -



learningRate	Tolerance	Accuracy_training_data	Recall_training_data	Accuracy_test_data	Recall_test_data
0.001	0.0000001	0.538626881	0.533063598	0.531069655	0.548271752
0.000000001	0.001	0.611265647	0.581110128	0.623674675	0.569725864
0.01	0.001	0.540207359	0.530029081	0.536572272	0.545292014
0.0000000001	0.0000001	0.611202428	0.58237451	0.622332573	0.572109654
0.001	0.00000001	0.538626881	0.533063598	0.531069655	0.548271752

We tried multiple combinations of hyperparameters (i.e. learning rate and tolerance) and displayed some combinations to highlight the best and worst hyperparameters. Our main metric of focus is Recall, as we must reduce the False Negatives in the data. The model has tolerance to predict class label ‘0’ as ‘1’ but not ‘1’ as ‘0’. Hence, reducing the occurrences of False Negatives results in an increase in the Recall value. Since we have considered the balanced class label data, we can consider Accuracy as a performance metric. Keeping in mind the importance of both the metrics, we can clearly see that for learning rate = ‘0.00000000001’ and tolerance = ‘0.0000001’are fetching better accuracy and recall. So, we are using these hyperparameters in our final model which results in accuracy of 62.2 % and recall of 57.2 %

Feature Importance:



	importance
number_inpatient	0.466072
discharge_disposition_id_22	0.270566
discharge_disposition_id_3	0.191516
glyburide_No	0.164738
discharge_disposition_id_9	0.158213
rosiglitazone_No	0.158158
rosiglitazone_Steady	0.157446
glyburide_Steady	0.142310
discharge_disposition_id_5	0.134912
repaglinide_Steady	0.127978
repaglinide_No	0.120565
medical_specialty_UKN	0.119976
discharge_disposition_id_28	0.117495
payer_code_UKN	0.104671
discharge_disposition_id_2	0.102124
nateglinide_Steady	0.097755
nateglinide_No	0.090210
discharge_disposition_id_12	0.088465
admission_source_id_8	0.084876
number_emergency	0.083151
discharge_disposition_id_18	0.080218
number_diagnoses	0.074839
discharge_disposition_id_6	0.074430

2. **Neural Networks:**
   A neural network is a set of algorithms that attempts to recognize underlying relationships in a batch of data using a method that mimics how the human brain works. Neural networks, in this context, refer to systems of neurons that can be organic or artificial in nature.

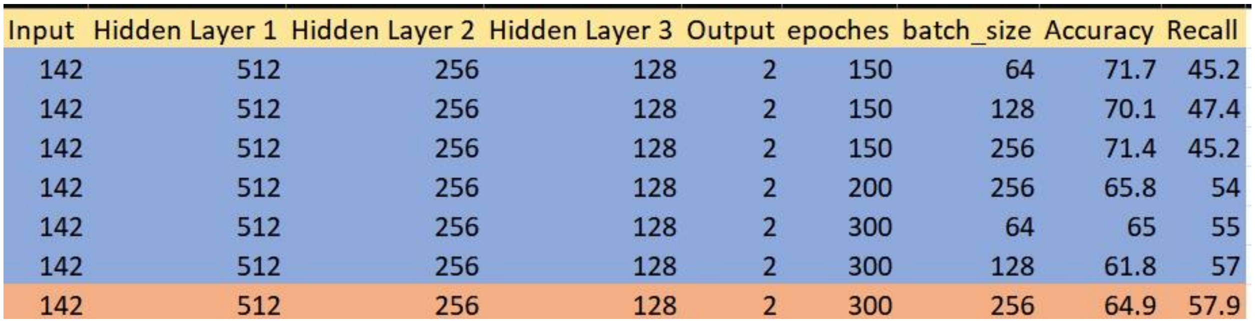
   Reference - (https://www.investopedia.com/terms/n/neuralnetwork.asp)

We have used RELU activation function in our hidden layers, because the main reason why ReLu is used is because **it is simple, fast, and empirically it seems to work well**. Empirically, early papers observed that training a deep network with ReLu tended to converge much more quickly and reliably than training a deep network with sigmoid activation.

Reference - (https://stats.stackexchange.com/questions/126238/what-are-the-advantages-of-relu-over-sigmoid-function-in-deep-neural-networks)

We have used Adam optimizer; it combines the best properties of the AdaGrad and RMSProp algorithms to **provide an optimization algorithm that can handle sparse gradients on noisy problems**. Adam is relatively easy to configure where the default configuration parameters do well on most problems.

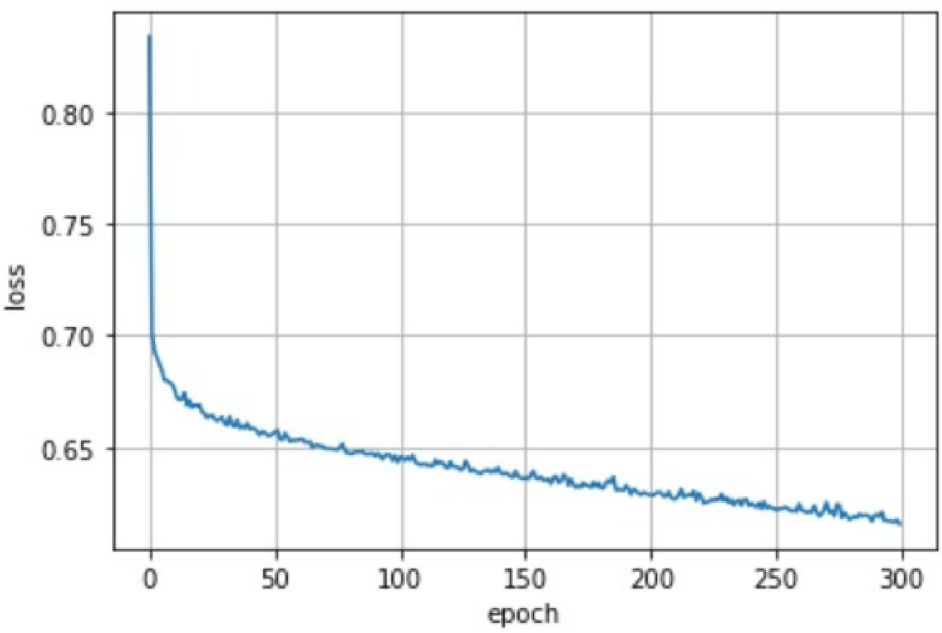
Reference - (https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/)



Input	Hidden Layer 1	Hidden Layer 2	Hidden Layer 3	Output	epoches	batch_size	Accuracy	Recall
142	512	256	128	2	150	64	71.7	45.2
142	512	256	128	2	150	128	70.1	47.4
142	512	256	128	2	150	256	71.4	45.2
142	512	256	128	2	200	256	65.8	54
142	512	256	128	2	300	64	65	55
142	512	256	128	2	300	128	61.8	57
142	512	256	128	2	300	256	64.9	57.9

We tried different combinations of the number of neurons in hidden layers, number of epochs and batch size. Here, we can see that we achieved accuracy of around 72% in our first combination which is very high, but at the same time recall is 45% which is very low. As recall is important in our problem, we tuned some hyperparameters. In all combinations, we found that the best results are given by the last record in the above-mentioned combinations. We achieved accuracy of 65% and recall of 58% which are better than the results obtained by our baseline model (Logistic Regression).

Loss function on training data vs epoch



Bias and Variance Tradeoff

1. **Using Validation Set Approach**

   We have used validation dataset in our model without revealing our test data to the model. This way we can try different models with our validation data and predict the values of the test data with the best fit model. We have used 70% data as train and remaining 15% of data as test and other 15% of validation data

2. **Random shuffling of data**
   Before train, test split we have random sampled the data which will avoid the case of training the data on only one set of class labels. Also, while training the data balanced data is considered in order not to train the model on biased data to overcome the generalization error

3. **Hyper Parameter Tuning**

   We have evaluated the data on different combinations of Hyperparameters like learning rate and tolerance for logistic regression and Number of neurons in different hidden layers, batch size in Neural Network models, epochs. Tuning these hyperparameters helped in maximizing the model performance and minimizing the error

Discussion

Since the data is health care data we cannot neglect or eliminate any of the features that may seem unwanted. Also, we have observed some interesting correlations that seemed logical.

We observed a chance to do feature engineering through different combinations of the features which will reduce the number of features and increase the model performance

It is possible to identify readmissions based on few features like number_inpatients and few others which has highest feature importance