PROJECT REPORT

ITCS 6190 –Cloud computing for Data Analysis | Guided by Professor Srinivas Akella

PREDICTION ON HOSPITAL READMISSION OF DIABETIC PATIENT

Submitted by:

Akash Savaliya | Riya Jain

TABLE OF CONTENTS

Contents

LIST OF FIGURES 2

LIST OF TABLES 2

ABSTRACT 3

1. INTRODUCTION 4

2. DATASET 4

3. METHODOLOGY 5

3.1 Data Preprocessing 5

3.2 Exploratory Data Analysis 8

4. EXPERIMENT 11

4.1 Logistic Regression 11

5. RESULT 12

6. CONCLUSION 12

7. WORK DISTRIBUTION 12

8. REFERENCES 13

LIST OF FIGURES

Figure 1 Box-plot 8 numerical variables 7

Figure 2 Correlation Matrix for numerical variable 8

Figure 3 Histogram for time_in_hospital & num_lab_procedures 8

Figure 4 Histogram for num_procedures,num_medications,number_outpatient & number_emergency .. 9

Figure 5 Box-plot for time_in_hospital,num_lab_procedures,num_procedures & num_medications..... 10

Figure 6 Box-plot for number_outpatient,number_emergency,number_inpatient & number_diagnoses 10

LIST OF TABLES

Table 1 Descriptive statistics of numeric variables 5

Table 2 Glipizide-metformin count 6

ABSTRACT

Diabetes is a widespread chronic disease that is accompanied with irregularities of blood glucose levels due to problems related to insulin. Some of the diabetes patients face a situation of getting readmitted to the hospital due to deterioration of their health condition even after getting discharged, which causes severe inconvenience and distress to the patients. Increase in Hospital readmission rates are an indication of poor hospital quality, poor treatment and they result in the hospital getting penalized. Hence by identifying the factors that lead to higher readmission and being able to predict if a patient is going to be readmitted, the treatment provided by the hospital can be changed, to avoid a readmission and thereby the quality of healthcare provided to the patient can be largely improved, as well as billions of dollars can be saved. For diabetes; the cost analysis estimates that $250 million can be saved across 98,000 diabetic patients by incorporating predictive modeling and prompting greater attention to those who were predicted to get readmitted.

The goal of this study is to assist hospitals in properly predicting whether a patient will be readmitted following discharge. The projections will aid the hospital in making educated judgments about treatment process changes that are required to lower patient readmission rates. The hospital will be able to make required treatment changes to avoid the patient being readmitted if it can reliably anticipate whether or not a patient would be readmitted. This decrease in patient readmissions will save the hospital billions of dollars while also improving customer satisfaction.

Prediction on Hospital Readmission Classification using Diabetes dataset is obtained from the Kaggle. The data set represents 10 years (1999-2008) of clinical care at 130 US hospitals and

integrated delivery networks. The data has around 100k+ observations and 53 attributes with a mix of both categorical and numerical variables.

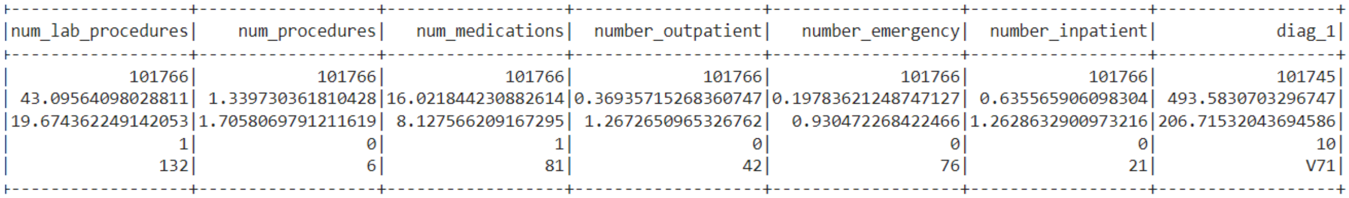
1. INTRODUCTION

As mentioned above, The purpose of this project is to assist hospitals in precisely predicting whether a patient will be readmitted post - discharge. The predictions will help the hospital in making sensible decisions about treatment process changes that are required to lower patient readmission rates. The hospital will be able to make necessary treatment changes to avoid the patient being readmitted if it can accurately predict whether or not a patient would be readmitted. This decrease in patient readmission rates will save the hospital billions of dollars while also enhancing patient satisfaction and the health-care system to make it more dependable. This prediction will help the hospital in identifying all patients who are likely to be readmitted after discharge as 'red,' and focusing on giving special/additional treatment to such patients. Furthermore, by recognizing the circumstances that lead to a patient's readmission, the hospital can take the necessary corrective actions to modify the patient's treatment, preventing the patient from being readmitted.

2. DATASET

The dataset was obtained from the Kaggle, as mentioned above. It's a patient-level dataset with 101,766 rows corresponding to 101,766 patients and 53 columns corresponding to 53 different patient attributes for each patient. Primary and foreign key attributes such as 'patient number,' 'admission id,' and so on are included in the dataset. The information includes the dates of the patient's admission, release, and readmission. We also have features that record essential data such as the number of surgeries conducted on the patient, the total time spent in the hospital, changes in medications, and changes in insulin levels, among other things. Other categorical attributes such as race, admission type, admitting medical specialty, and binary variables such as gender, Diabetes Medication used, and other numeric variables such as age, time in hospital, number of lab tests performed, number of medications, number of outpatient, inpatient, and emergency visits in the year prior to the hospitalization, etc. are also included in the data. All of these characteristics have the potential to be quite important in deciding the output variable. "No readmission," "A readmission in less than 30 days," and "A readmission in more than 30 days" are the three classes of the goal variable. However, because we are only interested in predicting whether the patient will be readmitted or not, we will combine the 30 and > 30 day classes into one and name it 'readmitted.'

The statistics of the numeric variables in the dataset are summarized in the table below. The'mean' (to indicate the central tendency),'standard deviation' (to indicate the spread), and min-max values (to indicate the range/spread) for the numeric variables in the dataset are all included in the summary table. We can see that the 'number outpatient' and 'number emergency' variables have a right skew since the maximum values are far from the mean value.



```
+------------------+------------------+------------------+-------------------+-------------------+------------------+------------------+
|num_lab_procedures|    num_procedures|   num_medications|  number_outpatient|   number_emergency|  number_inpatient|            diag_1|
+------------------+------------------+------------------+-------------------+-------------------+------------------+------------------+
|            101766|            101766|            101766|             101766|             101766|            101766|            101745|
| 43.09564098028811| 1.339730361810428|16.021844230882614|0.36935715268360747|0.19783621248747127| 0.635565906098304| 493.5830703296747|
|19.674362249142053|1.7058069791211619| 8.127566209167295| 1.2672650965326762|  0.930472268422466|1.2628632900973216|206.71532043694586|
|                 1|                 0|                 1|                  0|                  0|                 0|                10|
|               132|                 6|                81|                 42|                 76|                21|               V71|
+------------------+------------------+------------------+-------------------+-------------------+------------------+------------------+
```

Table 1 Descriptive statistics of numeric variables

Next, we will discuss the methodology we followed to complete this project. We have employed the CRISP-DM iterative process, which is the Cross Industry Standard Process of Data Mining. The process starts off with Data Cleaning to handle the data quality issues, followed by Data Preprocessing, Exploratory Data Analysis and feature selection, which is followed by model building, model performance evaluation, and producing the results, inferences and conclusion. To achieve our project goal of predicting whether a patient will be readmitted or not, we created a binary classification machine learning model. These models were created with the help of a supervised machine learning algorithm. The existence of a target variable distinguishes supervised learning from unsupervised learning. We will discuss each stage of the process and all of the supervised learning algorithms that we have used in detail one by one in the upcoming sections.

3. METHODOLOGY

In this section we will cover every stage of the Data Science methodology that we have used to accomplish our project objective.

3.1 Data Preprocessing

The first step in the Data Preprocessing stage is data wrangling, sometimes referred to as Data Cleaning. Data wrangling is the process of cleaning, organizing, and enriching raw data into a desired format in order to make better decisions in less time and to address all data quality issues. Duplicate values, missing values, noise/outliers, and attributes with minimal to no variance are some examples of data quality difficulties that people frequently find when working on data science projects. Let's talk about the data quality challenges we ran across and how we dealt with them.

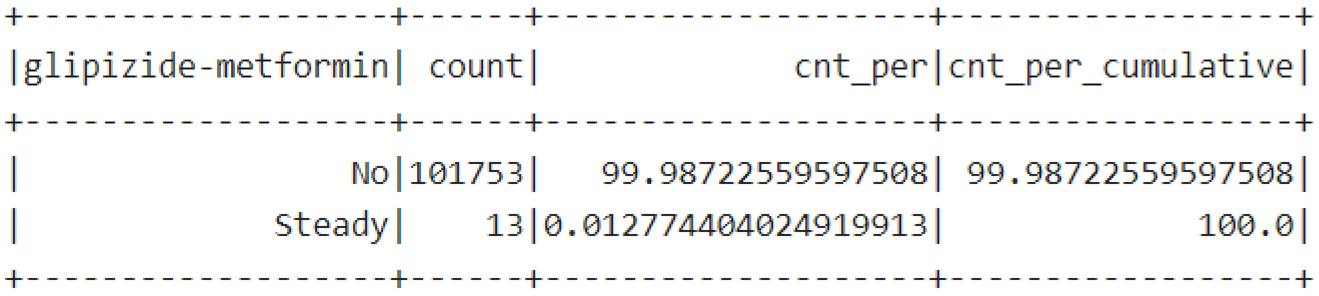
Firstly, we witnessed the presence of special characters and invalid values in some of the attributes in the dataset. We declared all such values as missing values and set it as Null for future treatment. Then we dropped all the attributes which had more than 10 % of missing values.

Secondly, we identified 4 columns ('race', 'diag_1' ,'diag_2', 'diag_3') which had only a few missing values. Since all these 4 attributes are categorical, we calculated the percentage of missing values and replaced all missing values with new category name OTHERS if the column had total missing values greater than

1%; otherwise, we replaced missing values with the Mode (most frequently occurring value) of that attribute, which helped us reduce the number of missing values in the dataset to zero.

Next, we identified which primary key and foreign key attributes, such as 'patient number,' 'encounter id,' and others, have no bearing on our target variable and are therefore insignificant. As a result, these attributes were removed from the dataset.

We next identified all the attributes with No Variance or close to No Variance, i.e. all the attributes with the same value in the majority of rows. When we look at the distribution of the '`glipizide-metformin`' variable, for example, we can see that the bulk of the values are in the "No" category, with only three in the "Steady" category. Such qualities with No variance are of little assistance in describing our target variable, thus they can be eliminated.



```
+-------------------+------+--------------------+------------------+
|glipizide-metformin| count|             cnt_per|cnt_per_cumulative|
+-------------------+------+--------------------+------------------+
|                 No|101753|   99.98722559597508| 99.98722559597508|
|             Steady|    13|0.012774404024919913|             100.0|
+-------------------+------+--------------------+------------------+
```

Table 2 Glipizide-metformin count

We identified 15 such attributes (‘'acarbose', 'acetohexamide', 'chlorpropamide', 'citoglipton', 'examide', 'glimepiride-pioglitazone', 'glipizide-metformin', 'glyburide-metformin', 'metformin-pioglitazone', 'metformin-rosiglitazone', 'miglitol', 'nateglinide', 'tolazamide', 'tolbutamide', 'troglitazone') and dropped them.

Lastly, we plotted histograms of all the numeric variables to detect the presence of outliers in the data by looking at their distributions. We did not find a sufficiently strong reason to declare any value as outlier.Thus We shifted to box plot analysis of these numeric variables.

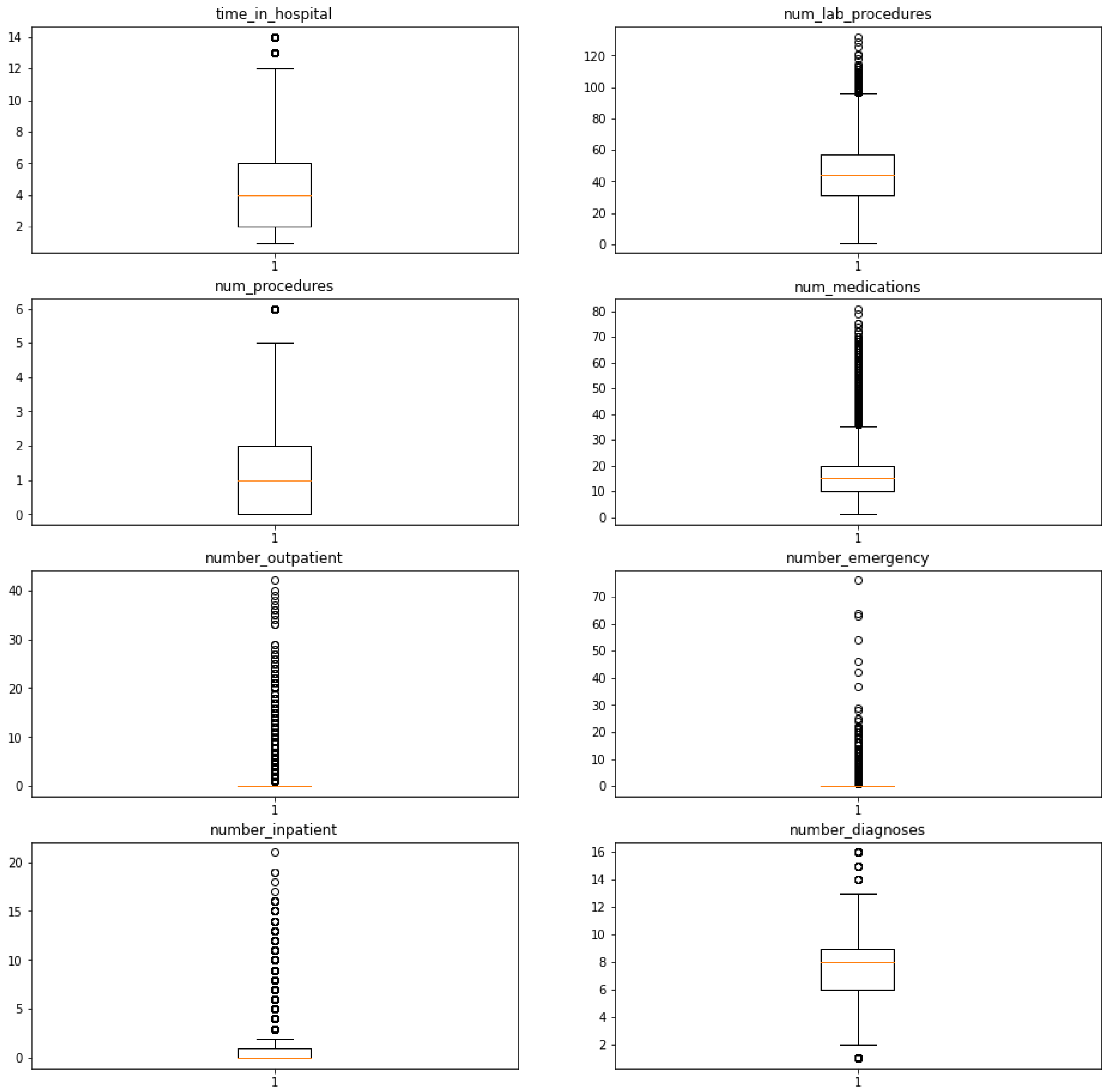


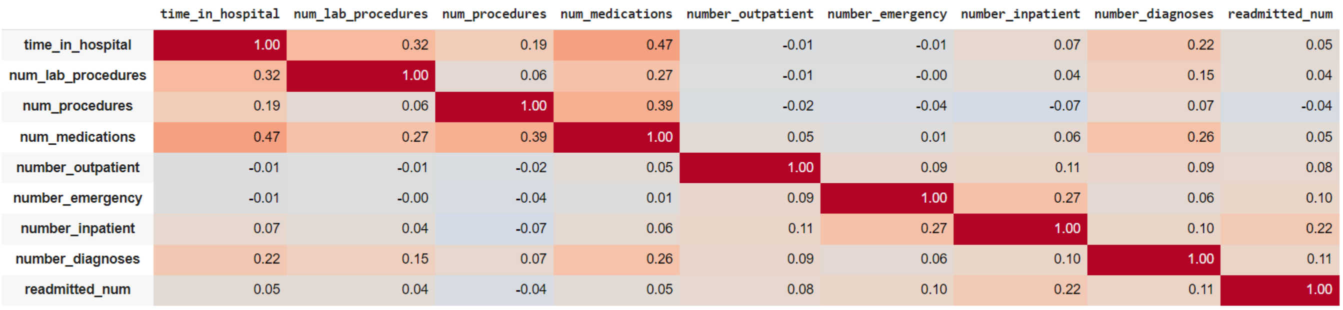
Figure 1 Box-plot 8 numerical variables

As part of Data Preprocessing, we used data transformation techniques including One Hot Encoding and Standardization after we finished Data Wrangling. The technique of transforming all category variables into numeric dummy variables is known as one hot encoding. Because all of our data must be numeric in order to develop machine learning algorithms, just one hot encoding is required. Because the variables have distinct units and values in completely separate ranges, standardization is done. Every variable undergoes standardization, which results in new values known as Z score values. The mean of these new Z score values will be 0 and the standard deviation will be 1. It is optional to set the standard deviation to 1, and it is dependent on whether the values have the same units. We decided to set the standard deviation to 1 in our situation because all of the variables have different units. Hence the Z score values obtained after standardization are calculated as z = (Xi - Mean)/standard deviation.

3.2 Exploratory Data Analysis

Exploratory Data Analysis is the process of looking for hidden connections or insight in data. These connections will aid us in the development of models. We construct several univariate, bivariate, and multivariate plots during the exploratory data analysis process in order to uncover the intriguing pattern in the dataset. Following the principles of explanatory data analysis, all of the patterns are presented as a tale.

We implemented three different plot to complete the detailed eda process. First plot which we used for the data description was correlation matrix. From the correlation time_in_hospital and num_medications have comparatively very high correlation also time_in_hospital has a good correlation with num_lab_procedures. Some of the correlation like time_in_hospital and num_lab_procedures have to be correlated with each-other since lab-procedures does increase as the time in hospitals is more. Also, num of medications have to highly correlated with time in hospital.



	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	number_diagnoses	readmitted_num
time_in_hospital	1.00	0.32	0.19	0.47	-0.01	-0.01	0.07	0.22	0.05
num_lab_procedures	0.32	1.00	0.06	0.27	-0.01	-0.00	0.04	0.15	0.04
num_procedures	0.19	0.06	1.00	0.39	-0.02	-0.04	-0.07	0.07	-0.04
num_medications	0.47	0.27	0.39	1.00	0.05	0.01	0.06	0.26	0.05
number_outpatient	-0.01	-0.01	-0.02	0.05	1.00	0.09	0.11	0.09	0.08
number_emergency	-0.01	-0.00	-0.04	0.01	0.09	1.00	0.27	0.06	0.10
number_inpatient	0.07	0.04	-0.07	0.06	0.11	0.27	1.00	0.10	0.22
number_diagnoses	0.22	0.15	0.07	0.26	0.09	0.06	0.10	1.00	0.11
readmitted_num	0.05	0.04	-0.04	0.05	0.08	0.10	0.22	0.11	1.00

Figure 2 Correlation Matrix for numerical variable

One of the other plot which is highly used in the exploratory data analysis process is histogram. Histogram gives the data range and data distribution. From the histogram plotted below, time_in_hospital histogram is right-skewed since most of the patients stay lesser than 6 days. Num_lab_procedures histogram shows us that the data is uniformly distributed with mean around 50 lab procedures.

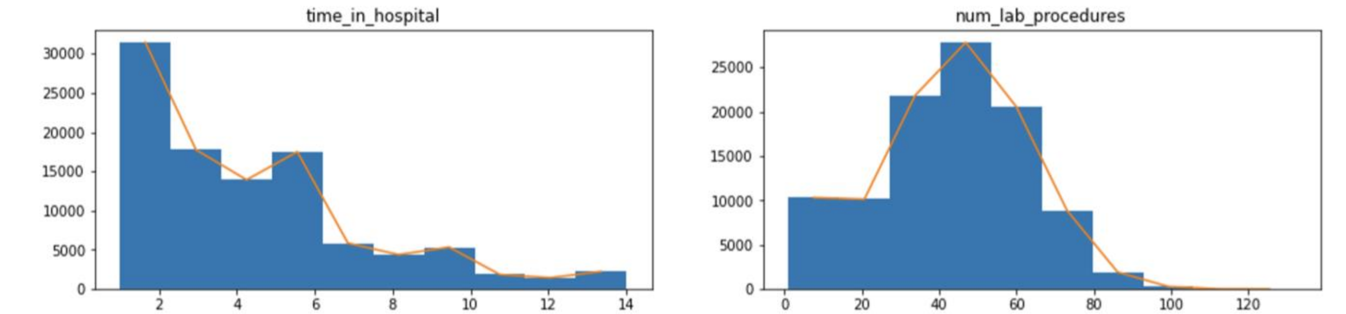


Figure 3 Histogram for time_in_hospital & num_lab_procedures

num_medications seems to right skewed due to outliners but is uniformly distributed with mean around 20. Num_emergency and number_outpatient have very high number of outlinears and the rest of the data is have constant value and hence excluded from the further analysis.

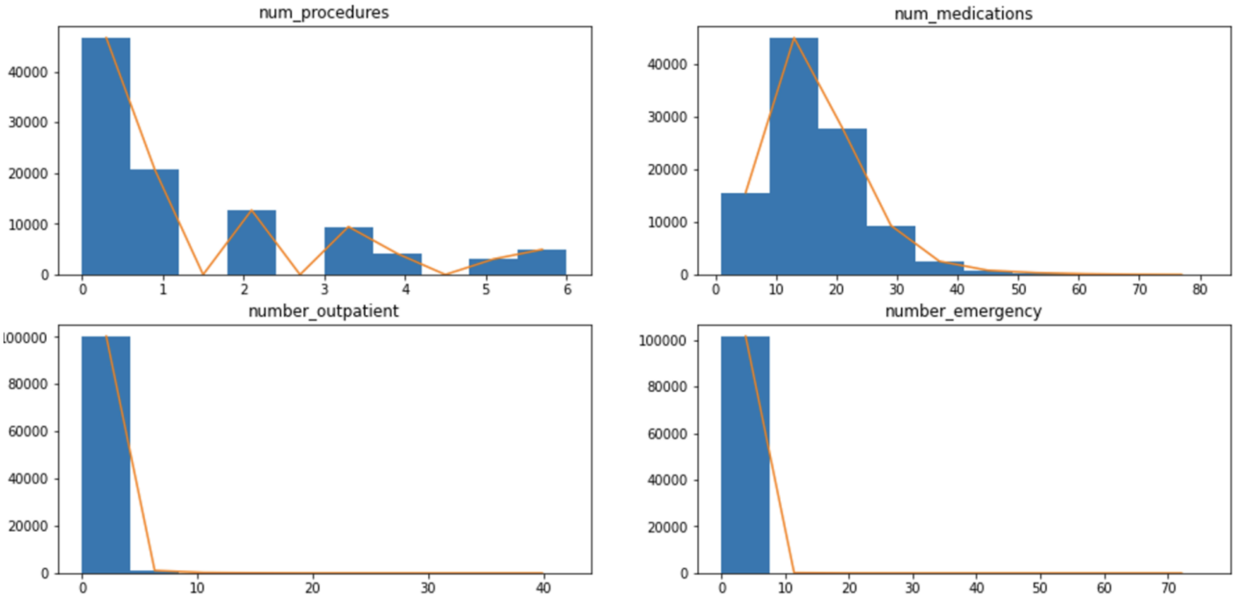


Figure 4 Histogram for num_procedures,num_medications,number_outpatient & number_emergency

The final plot type we have used is boxplot. Boxplot is highly effective to analyze the data for the outliers and data distribution in terms of quartiles. The following four boxplots are time_in_hospital, num_lab_procedures, num_procedures and num_medications. Time_in_hospital has a mean of 4 and few outliers which extends till 14 days. This feature has first quartile as 2 and third quartile as 6. Most of data lies between 2 and 6. In num_lab_procedures boxplot, we have mean value of 40 lab procedures and q1 as 30 and q3 as 50. It has high number of outliners which extends the data range to more than 130 lab procedures.

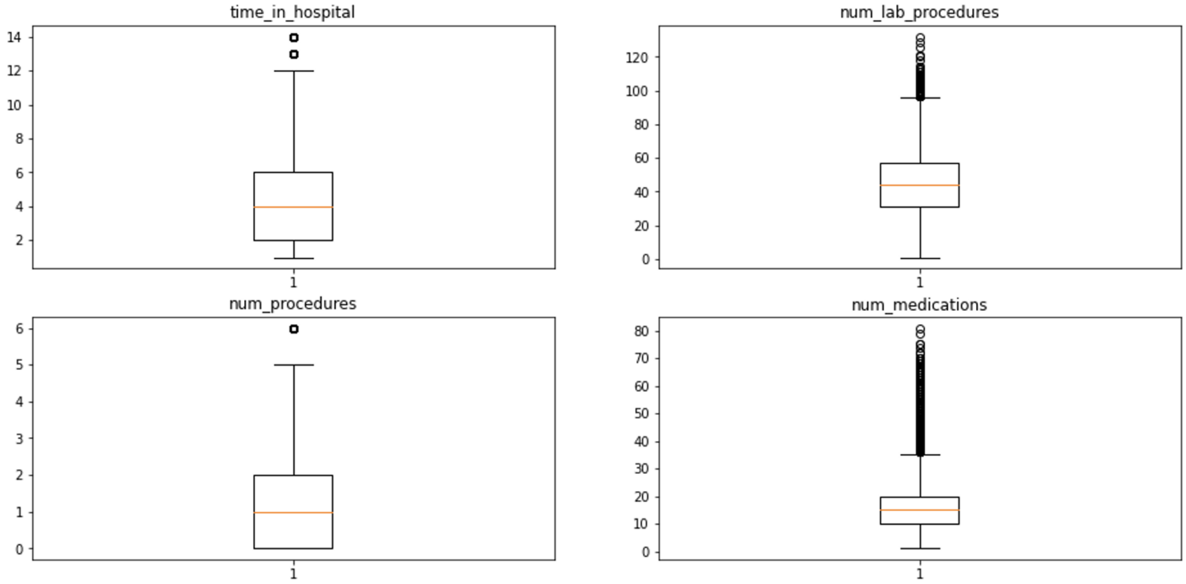


Figure 5 Box-plot for time_in_hospital,num_lab_procedures,num_procedures & num_medications

The following 4 boxplot shows us that the number_outpatient, number_emergency and numer_inpatient have very high outliners and data does not show any kind of correlation with out target variables.

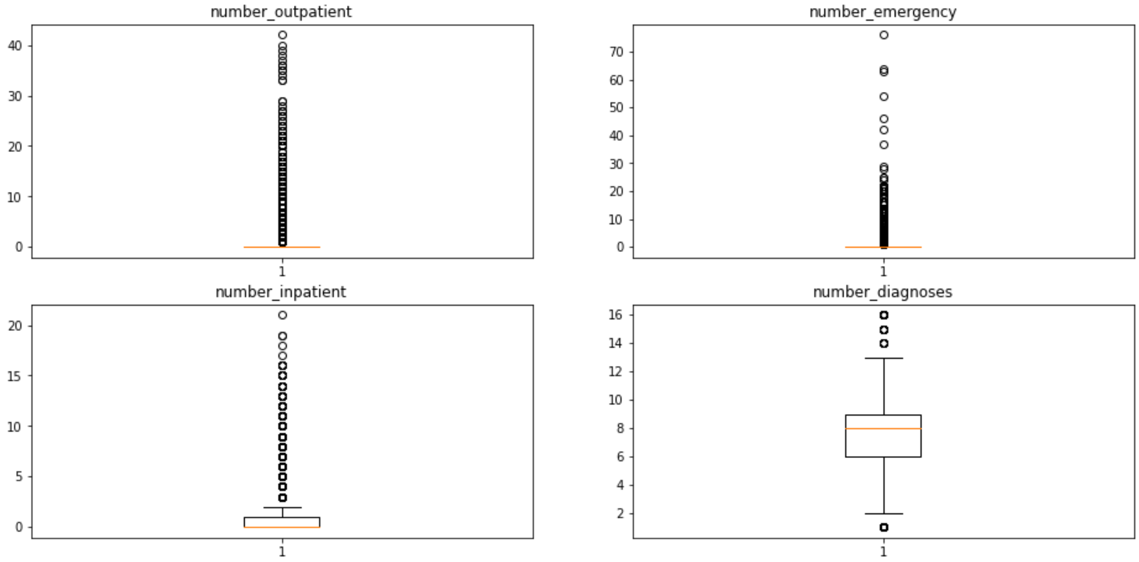


Figure 6 Box-plot for number_outpatient,number_emergency,number_inpatient & number_diagnoses

4. EXPERIMENT

Now coming to the model building and evaluation process. Since ours is a binary classification problem, we have built a classification model to predict our binary target variable (patient readmission). The classification algorithm which we have used for building our model is:

- Logistic Regression

Our major model performance evaluation measures have been **Accuracy**, **Recall**, **Precision**, and **F1-Score**. Because our dataset is well-balanced, Accuracy is a viable option for assessing our model's performance. In our instance, recall is particularly vital because we want to avoid False Negatives as much as possible. False Negatives are situations in which the patient is expected to be readmitted but our model predicts that the patient will not be readmitted. Such situations are quite undesirable, and they must be avoided by concentrating on raising our Recall scores. Precision is used to reduce the number of False Positives. The F1-Score is used to calculate the Precision vs. Recall tradeoff.

4.1 Logistic Regression

A machine learning technique called logistic regression fits a linear decision boundary between positive and negative samples. To determine the probability of the classes, this linear function is run via a sigmoid function. When the characteristics are linearly separable, logistic regression is a good model to apply. One advantage of logistic regression is that the model is interpretable, which means we can see which characteristics are crucial for predicting good or negative outcomes.

Logistic regression helped us understand the relative impact and statistical significance of each factor on the probability of readmission by starting with the assumption that the impact of factors and their interactions may be described as a log likelihood of outcome. We dealt with missing values by deleting columns or going through the preparation processes. We used L2 regularization with Logistic Regression.

We have Processed numerical variables using quartile values such that q1 denotes 25 percentile q2 denotes 50 percentile and q3 denotes 75 percentile. The formula for the inter quartile range is q3-q1 and after getting the iqr value we found the lower and upper value using 1.5-iqr and 1.5+iqr respectively. After treating the outliears we created data preprocessing pipeline for vector assembler, standard scaling and one hot encoding.

Precision reflects the percentage of positive predictions that are correct. If the model classifies 100 observations as positive, and 45 of those were in fact positive, then the model's precision is .45 (or 45%).Recall captures the percent of true positives that are classified as positive. Returning to our model's context, if there were 1,000 patients readmitted within thirty days and the model detected 850 of them, then the model's recall would be .45 (or 45%).Finally, the F1-score (the harmonic mean) shows the weighted average of precision and recall. It is a useful metric because it factors in both false positives and false negatives. The model here has an F1-score of .45, which is not good enough considering that 1.00 would be a perfect score and .5 would be as good as random guessing. The final model performed with an accuracy of 49 %, a recall of 45 %.

5. RESULT

For our hospital readmission dataset, Logistic Regression gave Accuracy of accuracy of 49 %, a recall of 45 % and F-1 Score is 0.45.

6. CONCLUSION

We gained hands-on experience developing machine learning models using pipelines, estimators, and transformers provided in the PySpark libraries as part of this project. We've worked with datasets that are really huge in size and include a significant number of rows and columns. We learned the value of using github for version control and collaboration as a result of this project.

The purpose of this research was to create predictive machine learning/deep learning models capable of binary classification to reliably categorize every patient into one of two categories: 'Will be Readmitted' or 'Will not be Readmitted.' We were able to achieve this target with a prediction accuracy of 45 percent, which means that out of 100 patients, 45 of our patient forecasts will most likely be right.

We can utilize pre-built Machine Learning Libraries for such data and Predictive Analysis for such a problem statement. We can also look at more complex feature selection techniques and learning algorithms that can outperform our current model and produce higher Accuracy and Recall.

7. WORK DISTRIBUTION

Riya:- Data Preprocessing ,Exploratory Data Analysis, Documentation

Akash : - Model Building(Logistic Regression) , Outlier Treatment

8. REFERENCES

- https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.plot
- https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.boxplot
- https://spark.apache.org/docs/3.1.1/api/python/reference/api/pyspark.ml.feature.OneHotEncoder.html
- https://spark.apache.org/docs/3.1.1/api/python/reference/api/pyspark.ml.feature.StringIndexer.html#pyspark.ml.feature.StringIndexer
- https://spark.apache.org/docs/3.1.1/api/python/reference/api/pyspark.sql.functions.percentile_approx.html