**Capstone Project-Final Report**

**Group-4**

**Abstract**

Hospital readmission is an indicator of the quality of care and is a driver for the increasing cost of healthcare. Like other chronic diseases, Diabetes is associated with a higher risk of hospital readmission. In this research, we evaluate several machine learning approaches to predict the probability of hospital re-admissions for diabetic patients. The data set used for this study contains more than 100,000 diabetic patient data and 55 variables including length of stay, insulin, and in-patient visits from hospitals in the United States. We leverage several pre-processing techniques and investigate the performance of the various models. The significant variables contributing to the analysis are the number of in-patients, length of stay, number of medications, number of diagnoses, and age. The results demonstrate the viability of the techniques in providing a better understanding of factors influencing hospital re-admission.

Although, identifying patients who are expected to be readmitted in 30 days of discharge is a complex task for hospitals. Techniques that can help to predict the likelihood for readmission and to identify the factors contributing to readmission can be of significant value to healthcare providers. Specifically, such techniques, allows provider to optimize their interventions for high-risk patients and to ultimately reduce readmission rates via improved processes.

Recent developments in machine learning have been successful at predicting readmissions from the medical history of the diabetic patient. However, these approaches rely on a large number of clinical variables thereby requiring deep learning techniques. This article presents the application of simpler machine learning models achieving superior prediction performance while making computations more tractable.

**Business Problem: Prediction on Hospital Readmission**

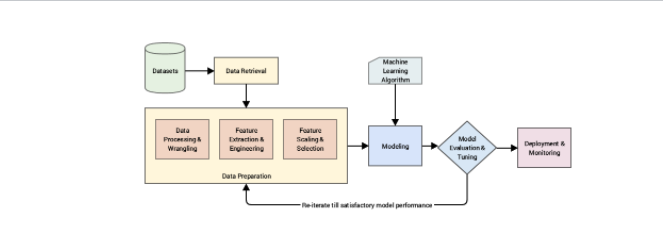
A hospital readmission is when a patient who is discharged from the hospital, gets re-admitted again within a certain period of time. Hospital readmission rates for certain conditions are now considered an indicator of hospital quality, and also affect the cost of care adversely.

For this reason, Centres for Medicare & Medicaid Services established the Hospital Readmissions Reduction Program which aims to improve quality of care for patients and reduce health care spending by applying payment penalties to hospitals that have more than expected readmission rates for certain conditions.

Although diabetes is not yet included in the penalty measures, the program is regularly adding new disease conditions to the list, now totalling 6 for FY2018. In 2011, American hospitals spent over $41 billion on diabetic patients who got readmitted within 30 days of discharge. Being able to determine factors that lead to higher readmission in such patients, and correspondingly being able to predict which patients will get readmitted can help hospitals save millions of dollars while improving quality of care. So, with that background in mind, we used a medical claims dataset.

**Data Description:**

* **Encounter ID** Unique identifier of an encounter
* **Patient number** Unique identifier of a patient
* **Race** Values: Caucasian, Asian, African American, Hispanic, and other
* **Gender** Values: male, female, and unknown/invalid
* **Age** Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)
* **Weight** Weight in pounds
* **Admission type** Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, 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 features for medications For the generic names: **metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride- pioglitazone, metformin-rosiglitazone, and metformin- pioglitazone**, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed
* **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
* Initially the data set had 50 features and 101766 records
* We have dropped the insignificant columns which contained ids in the initial stage itself – Encounter ID and Patient NBR Number
* We checked for missing values and removed ‘weight’ feature from the dataset as it has 96% values missing. And imputed missing values in other features by using median for numerical and mode for categorical feature
* Data Processing – data contained 22 columns on medication and had categorical values, for the sake of simplicity we calculated the sum of the medication category for each record and converted the 22 columns into 4 columns
* In the base model we have used 23 features - 8 numerical , 15 categorical
* The target variable “READMITTED” is to be found out from the other observations and need to be classified as the chances of readmission in >30 days , <30 days or No.
* The target variable is the categorical variable so we can go for classification model.



METHODS-This section will highlight and rationalize the methods used in this study in order to achieve the objectives. While the selected dataset will be presented in section A particular attention will be given to the pre-processing stage in section B. In this section, details over cleaning, data reduction, transformation techniques but also pre-processing performance evaluation will be outlined . Namely, key preprocessing steps include feature imputation with Approximate Bayesian Bootstraps, ICD-9-DM clustering and feature selection using the Random Forest algorithm. Furthermore, the need for class balancing will be presented in section D and emphasize on Synthetic Minority Over-sampling Technique (SMOTE) algorithm. Particular attention will also be given to the modeling part in section E where Multilayer Perceptron (MLP) will be described.

1. **Dataset**

This study uses the Health Facts National Database (Cerner Corporation, Kansas City, MO), gathering extensive clinical records across hundreds of hospitals throughout the US [18]. The data subset used for analysis covers 10 years of diabetes patient encounter data (1999 – 2008) among 130 US hospitals with over 100,000 diabetes patient. Moreover, all the encounters used for analysis satisfy five key criteria:

• It is a hospital admission.

• The inpatient was classified as diabetic (at least one of three initial diagnoses included diabetes).

• The length of stay was comprised from 1 to 14 days.

• The inpatient underwent laboratory testing.

• The inpatient received medication during its stay.

**B .Data Pre-processing:**

In order to empower the dataset and improve the model performance, the importance of each input feature against the output variable will be assessed while non-important features will be excluded.

1. Missing values: The first step in cleaning the data consist of handling missing values. Missing values refers to the absence, voluntary or not, of data in a record. While the initial step is to identify and encode missing values, the second step consists in addressing the missing values.

Each variable comprising missing values were independently analyzed, as the methods to be applied differs based on statistics but also best practices and industry knowledge. In this particular case, the missing values are encoded as “?”, which is not a standard missing value format. Therefore, the first step in addressing missing values will be to encode them properly.

weight 96.858479

medical\_specialty 49.082208

payer\_code 39.557416

race 2.233555

diag\_3 1.398306

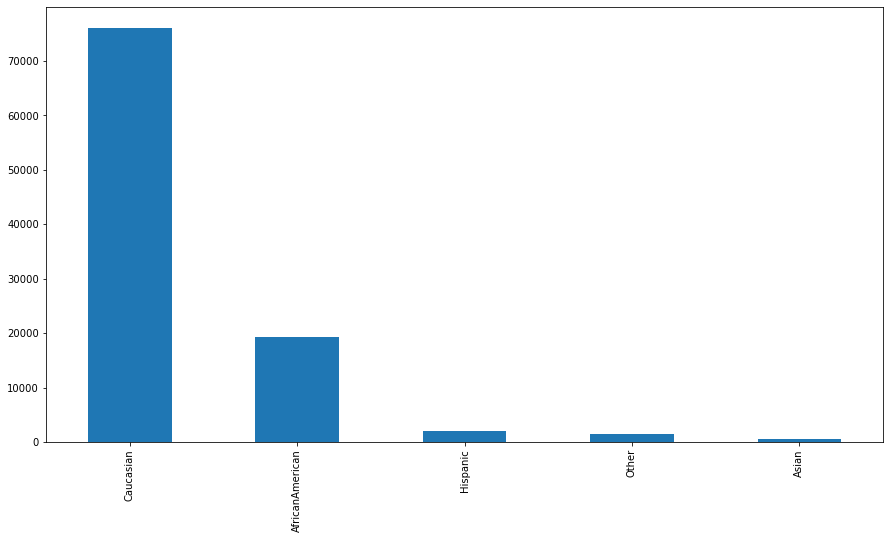
diag\_2 0.351787

diag\_1 0.020636

As a general rule, variables with 50% or more missing values should be dropped from the analysis. The variable medical specialty comprises 49% of missing observations. In term of proportion, the whole column should be dropped. However, based on background understanding and recommendation from previous researches such variable is of prime importance when predicting readmission .

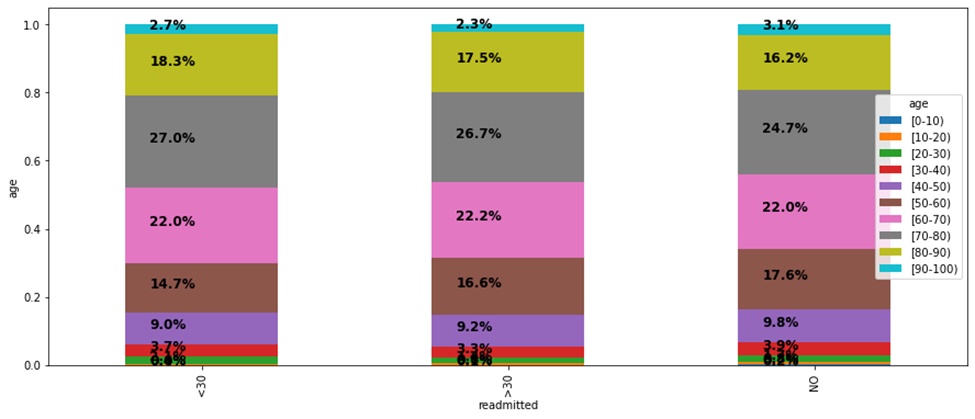
2)Uni-Variate and Multi-variate analysis-

Perfomed univariate analysis to derive the data, define and summarize it, and analyse the pattern present in it. In a dataset, it explores each variable separately. It is done for two kinds of variables- Categorical and Numerical.



Uni-variate analysis on ‘race’ column

When the data involves **three or more variables**, it is categorized under multivariate



**C. FEATURE ENGINEERING:**

Several feature engineering steps were taken in order to perform both feature creation, encoding, feature selection, and data scaling. Indeed, while some feature engineering steps are based on the data and business understanding others such as variable encoding taken into account the requirements of future algorithms to be applied. In this study, the feature engineering step will be subdivided into 4 categories

1. Feature Creation:

The initial set of data includes 23 medication related features, each associated with 4 classes, namely “No”, “Steady”, “Up” and “Down”. Such categories aim at assessing whether a change of medication occurred during the patient’s admission. Several studies highlighted medication change as an influential factors towards readmission. Hence, a new feature label “Med\_No”,”Med\_Steady”,”Med\_Up”,”Med\_Down” will be engineered by counting all changes in medication for all records.

b)Feature Encoding:

As neural networks will be applied in the later model stage, features will need to be encoded as numeric. The encoding process is discussed below. Reduce Output Class to Binary: The objective of this study is to predict whether a patient will be readmitted or not within the next 30-days after discharge. Therefore, the scope of the study is limited to discrimination between

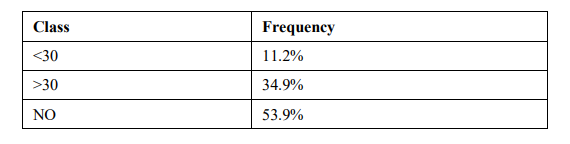
Encode other variables: The studied dataset encoded most of the variable in string format including race, gender, medication change, and all the medication used features. Hence medication change will be re-encoded into 0 and 1 values respectively for “No” (no change) and “Ch” (change). Moreover, all medication used features will be simplified as “Change” and “No Change” and will be encoded as 0 and 1.

1. Feature Selection:

Several techniques were tested to perform feature selection. The initial test using logistic regression and regularization technique coefficient’s p-values appeared to drop highly significant features both from a medical perspective but also as per highlighted by the body of literature. Hence, the random forest feature selection was performed. The variable importance was then computed during 60 iterations. A total of 20 predictors were hence selected after 60 iterations

**D. CLASS IMBALANCE:**

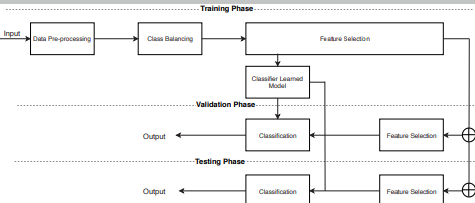
The output variable “readmitted” appeared to be relatively imbalanced with the below distribution, refer to Table 2. Such distribution should be addressed as it may alter the generalizability of the model. The SMOTE algorithm was found particularly efficient, generating a target class distribution close to 50%. The distribution after and before is summarized in the Tables .



|  |  |
| --- | --- |
| **Output Class** | **Observations** |
| **0** | **46.15%** |
| **1** | **53.85%** |

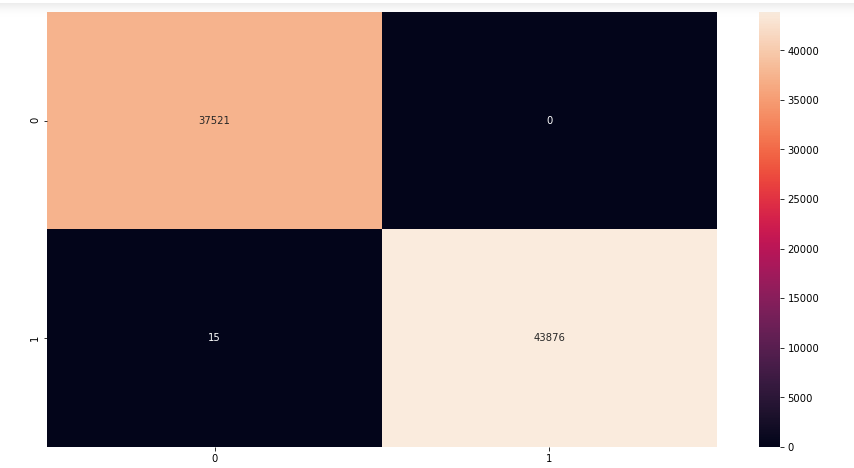
**E. MACHINE LEARNING METHODS**:

Identification of high-risk diabetic patients was posed as the problem of classifying whether a patient would be readmitted within 30-days of being discharged. Best practice is to make use of several machine learning algorithms, which is part of this study. Prior to training the classification algorithms, we randomly split our dataset into two distinct sets – the training and the test set. The training and test set consisted of 80% and 20% of the data. The parameters of each algorithm were chosen based on the classification performance evaluated by 10-fold cross-validation on the training set. The performance of all algorithms was evaluated on the test set.



Tree-based models calculate feature importance by keeping the best performing features as close to the root of the tree.

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **F1\_Train** | **F1\_Test** | **Roc\_Auc** |
| Logistic Regression | 0.69 | 0.69 | 0.61 |
| Decision Tree | 0.65 | 0.65 | 0.55 |
| Random Forest | 0.69 | 0.69 | 0.60 |
| XGBoost | 0.69 | 0.68 | 0.62 |
| LightGBM | 0.70 | 0.68 | 0.62 |



**Conclusion-**

In this study, we evaluated various machine learning models to predict readmissions of high-risk patients. Extending prior research, we performed class balancing considering the skewness of data. Our results show LightGBM slightly out-performing logistic regression and decision trees which were widely adopted in the literature. Some of the key features that drove readmissions are number of in-patients, length of stay, number of medications and number of diagnosis. Extending this research, we plan to further investigate the performance of classifiers with the goal to improve the accuracy of prediction of readmission risk. Further research is also warranted to explore the relevant feature space particularly with the variability of findings across research studies. The latter is particularly important as it can translate to proactive processes and policies aimed at addressing the factors that significantly influence hospital readmissions. Given the increasing cost of hospital readmissions and the increased emphasis on quality of care, the accuracy and validity of prediction models remain an important, yet elusive goal.