**Section 1:**

Introduction:

Diabetes mellitus (DM) constitutes a burgeoning public health challenge within the United States. Statistical estimates reveal that 9.3% of the U.S. population, equivalent to 29.1 million individuals, is afflicted by DM, with 28% of cases remaining undiagnosed [1]. DM manifests chronic complications, including retinopathy, neuropathy, nephropathy, and an elevated susceptibility to cardiovascular diseases, encompassing major cardiac events such as myocardial infarction and stroke [2]. The pervasive prevalence of DM and its associated complications position it as a prevalent comorbidity among hospitalized patients, necessitating frequent admissions for procedural interventions. Patients with DM reportedly experience prolonged lengths of stay (LOS), heightened incidences of hospital complications, and increased mortality rates during such hospitalizations [3].

Recently, governmental bodies and healthcare systems have increasingly directed their attention toward 30-day readmission rates as a metric for enhancing healthcare quality and discerning the intricacies of patient populations. The Centers for Medicare and Medicaid Services (CMS) have designated 30-day readmission rates as a pivotal gauge of healthcare quality, advocating for its reduction as a strategy to curtail healthcare costs while upholding quality standards [4]. In comparison to non-diabetic patients, those with DM exhibit a greater likelihood of readmission accompanied by additional comorbid conditions such as heart failure, myocardial infarction, and cardiac surgery [5]. Recent studies indicate that the 30-day readmission rate for hospitalized DM patients ranges from 14.4% to 22.7%, significantly surpassing the corresponding rates for all hospitalized patients (8.5–13.5%) [6].

Effectively addressing this incongruity in readmission rates necessitates a comprehensive understanding of the underlying causes of readmission in DM patients. Noteworthy factors include possession of health insurance [7], the nature of insurance coverage (governmental versus private or absence of insurance), male gender, duration of hospitalization, and the extent of comorbidities [7]. However, the existing literature lacks substantial evidence concerning clinically contributory factors and specific interventions aimed at mitigating readmission rates. Consequently, strategies to address these factors and diminish 30-day readmissions are not widely formulated or implemented.

**References:**

1. 2014 National DM Statistics Report. Centers for Disease Control and Prevention.
2. Fowler M. Microvascular and macrovascular complications of DM. Clin DM. 2008;26(2):77–82.
3. Umpierrez GE, et al. Hyperglycemia: an independent marker of in-hospital mortality in patients with undiagnosed DM. J Clin Endocrinol Metab. 2002;87(3):978–82.
4. Kocher RP, Adashi EY. Hospital readmissions and the affordable care act: paying for coordinated quality care. JAMA. 2011;306(16):1794–5.
5. Rubin DJ. Hospital readmission of patients with DM. Curr Diab Rep. 2015;15(4):17.
6. Kim H, et al. Scheduled and unscheduled hospital readmissions among patients with DM. Am J Manag Care. 2010;16(10):760–7.
7. Robbins JM, Webb DA. Diagnosing DM and preventing rehospitalizations: the urban DM study. Med Care. 2006;44(3):292–6.

**Literature survey:**

The paper “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records”, first introduced the “Diabetes 130-US hospitals for years 1999-2008” dataset. Therefore, in order to perform a literature survey we selected the following we searched the title of the paper on Scopus, then sorted all the papers that was citing the original paper by the number of citations. From these results, we selected the top 3 cited papers including the original paper and reviewed them herein. The literature review is structured in paragraphs, each discuss a different step in the analyses made in the corresponding papers.

**Data cleaning:**

The abovementioned dataset, were analyzed with two main aims: (1) to find determinants of early re-admission of hospital patients. (2) To find determinants of early re-admission of diabetic patients. Although the aims may vary, the sub-setting of the data did not differ between the three papers and the data were selected based on 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. In addition, in the original paper (REF), duplicated rows were removed based on multiple inpatient visits. Instead, the first encounter per patient was selected and the others removed. In this paper, patients that was discharged to hospice or the encounter was resulted in patient death were removed. Overall, the data contain 69,984 encounters for the final analysis.

**Feature Engineering:**

In the original paper, the features ‘Weight’, ‘Medical Specialty’ and ‘Payer Code’ were removed based on the high percentage of missing values, and their unlikeliness to be important for the prediction of the label. In all the papers, the label was modified to be binary with the values ‘readmitted < 30 days’ and ‘otherwise’. A new feature was created in the original paper and it included 4 values for HbA1c that was either not performed, performed and in normal range, and above 8% with or without change in medications. Further modifications of features were mostly removal of features based on their insignificant importance based on corresponding tests. On some occasions new features were introduced including: ‘gender\_male’, for male or not. ‘race\_caucasian’ for Caucasian or not, ‘change’ for change in medications with values of yes or no. Features were also selected based on their importance that was determined by Gradient Boosting technique. Notably, the feature engineering is not well detailed in the methods section of the above mentioned papers.

**Model selection and training:**

The split of the data to training and test set was performed in ratios of 70%:30% or 60%:40% for training and test set respectively. Cross validation was performed in 10 k fold cross validation. The analyses were defined as classification tasks and likewise the selected models were Multivariate Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Trees, Random Forest, k-Nearest Neighbors, J48 and Multi-Layer Perceptron. Hyper parameter tuning was mostly done using Grid Search. Performance estimation was assessed by Accuracy, AUC, Recall, Precision and F1\_score.

**Best fit models and parameters:**

The models that showed the best results in predicting the probability of <30 days readmission are Naïve Bayes and a Decision Tree that were used in two independent papers. Unfortunately, the hyper-parameters for the Naïve Bayes were not reported. The Decision Tree is Classification and Regression Tree (CART). The model was generated using the ‘gini’ function to evaluate the split quality of the tree. The final hyper-parameters of the decision tree were: min\_samples\_split = 30 and max\_depth = 15. Noteworthy is that the paper (REF) compared the results of Naïve Bayes and the Decision Tree CART on the same dataset and the Decision Tree outperformed Naïve Bayes with an accuracy of 0.87 compared to 0.63, respectively.

**Conclusions:**

With different models, the conclusions made by the different papers were somewhat similar. Diabetic patients who do not undergo vigorous lab assessments, diagnosis, medications are more likely to be readmitted when discharged without improvements and without receiving insulin administration, especially if they are women, Caucasians, or both. In general, the findings suggests that the probability of re-admission of hospital patients depends on whether their HbA1C was measured and on their primary diagnosis. The readmission rates for patients that were primarily diagnosed with circulatory diseases are the highest. Interestingly, readmission rates for patients with diabetes appeared to be associated with the decision to test for HbA1c, rather than the values of the HbA1c result. Across all studies the features that were deemed as important determinants of <30 days readmission are: Number of lab procedures, Number of medications, Number of inpatient, emergency, and outpatient visits, time in hospital, number of procedures, Caucasian race and more.

**References:**

(A) Strack et al., 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.

(B) Alajmani et al., 2019. Hospital Readmission Prediction using Machine Learning Techniques. (C) Alloghani et al., 2018. Implementation of machine learning algorithms to create diabetic patient re-admission profiles

**Section 2:**

Description of the entire dataset, the features and their meaning can be found in the following reference:

Strack et al., 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.

**Section 3:**

* Our objective is to predict readmission that occurs within 30 days. We define the objective as a supervised classification task.
* Based on the abovementioned literature, the current solutions are using supervised ML models like random forest or logistic regression for example.
* Our assumptions are that the important features for the prediction of readmission will be the diagnoses and number of medication because of their clinical relevance. However, the most important features we observed are number of emergency and number of outpatient visits.

**Section 4:**

The data were split using StratifiedGroupKFold (scikit-learn), in order to make sure that duplicate measurements from the same patient are either in the train or test set and therefore to avoid dependency of samples. Using this function we also stratified the label.

**Section 7:**

For training the models our performance measure was log loss. To describe the results our performance measures are precision and recall.

**Section 8:**

The selected models are LGBM, XGBoost and Random Forest Classifier, based on their performance i.e. log loss on the validation set. All models performed better on data that was balanced with CopulaGAN and therefore we proceed by using GAN synthesized data.

**Section 10:**

**Results:**

**Feature Engineering:**

**Add here information about the categorization of the features.**

|  |  |
| --- | --- |
| **Classifier ± balanced / imbalanced data** | **neg\_log\_loss** |
| CatBoostClassifier\_cop\_4\_num | -0.20011 |
| LGBMClassifier\_cop\_4\_num | -0.20079 |
| CatBoostClassifier\_cop | -0.20395 |
| XGBClassifier\_cop\_4\_num | -0.20471 |
| LGBMClassifier\_cop | -0.20763 |
| XGBClassifier\_cop | -0.20924 |
| RandomForestClassifier\_cop\_4\_num | -0.2284 |
| RandomForestClassifier\_cop | -0.2346 |
| CatBoostClassifier\_ct | -0.26172 |
| LogisticRegression\_cop | -0.2626 |
| XGBClassifier\_ct | -0.2743 |
| LGBMClassifier\_ct | -0.2962 |
| CatBoostClassifier | -0.32664 |
| LGBMClassifier | -0.32815 |
| LogisticRegression | -0.32877 |
| XGBClassifier | -0.34372 |
| LogisticRegression\_cop\_4\_num | -0.35529 |
| RandomForestClassifier\_ct | -0.38204 |
| RandomForestClassifier\_sm | -0.43401 |
| LogisticRegression\_ct | -0.49342 |
| CatBoostClassifier\_sm | -0.5033 |
| XGBClassifier\_sm | -0.52227 |
| LGBMClassifier\_sm | -0.52656 |
| LogisticRegression\_sm | -0.59681 |
| RandomForestClassifier | -0.64316 |

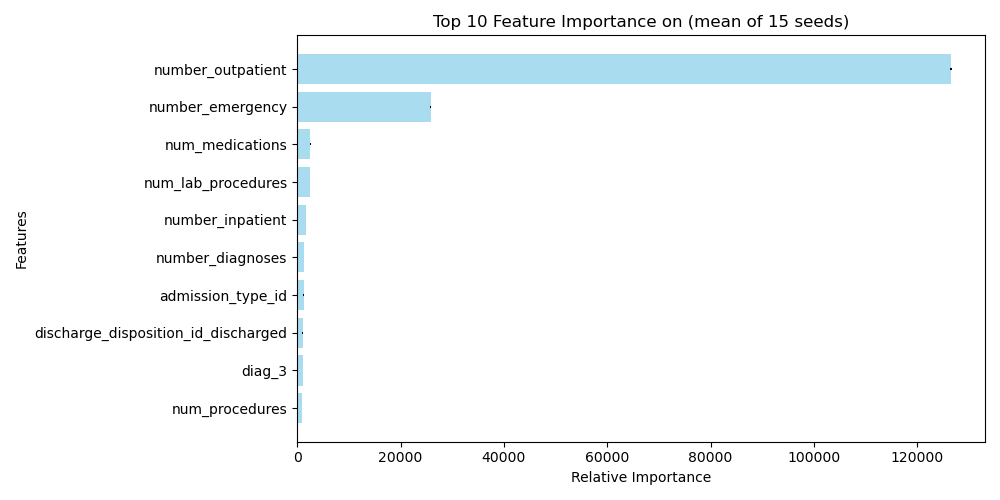
**Table 1. Cross validation scores of multiple classifier with default parameters, sorted by our performance measure of choice, neg\_log\_loss.** For the preliminary cross validation and model selection, classifiers were trained on unbalanced data unless otherwise specified. Data were balanced using CopulaGAN(cop), CTGAN(ct), SMOTE (sm). The suffix '4\_num', indicates that numeric features were analyzed as numeric, otherwise these were categorized.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mean ± STD, n=15 seeds | | | | |
| Model | Neg log loss | Precision | Recall | ROC AUC | Accuracy |
| LGBM | 0.36 ± 0.001 | 0.89 ± 0.0003 | 0.97 ± 0.0009 | 0.66 ± 0.002 | 0.8745 ± 1.1E-16 |
| XGBoost | 0.36±0.00047 | 0.893±0.00025 | 0.974±0.000323 | 0.675±0.00126 | 0.873484 |
| Random Forest |  |  |  |  |  |
| Dummy Classifier – “Most frequent” | 32.06 | 0 | 0 | 0.5 | 0.11 |
| Dummy Classifier – “Uniform” | 0.69 | 0.5 | 0.89 | 0.49 | 0.5 |

**Table 2. Prediction scores.** Data are presented as mean ± STD of 15 seeds. The presented models were applied with the 'best parameters'. The dummy classifier was applied with 'most\_frequent' strategy.

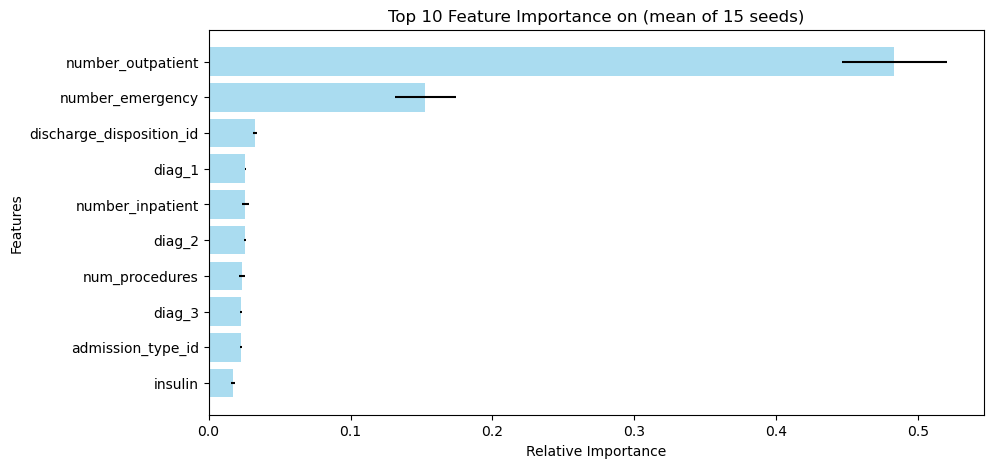
|  |  |  |
| --- | --- | --- |
| **Best Parameters** | | |
| **LGBM** | **XGBoost** | **Random Forest** |
| {'objective': 'binary',  'learning\_rate': 0.146,  'n\_estimators': 95,  'boosting\_type': 'gbdt',  'max\_depth': 12,  'num\_leaves': 31,  'colsample\_bytree': 1.0,  'min\_child\_samples': 80,  'subsample': 0.8,  'subsample\_freq': 1,  'reg\_alpha': 1e-07,  'reg\_lambda': 1e-08,  'is\_unbalance': 'True',  'scale\_pos\_weight': 1,  'min\_sum\_hessian\_in\_leaf': 0,  'min\_split\_gain': 0,  'min\_gain\_to\_split': 0,  'min\_child\_weight': 18,  'max\_bin': 742} | {'objective' : 'binary:logistic',  'device' : "cuda",  'max\_depth': 5,  'min\_child\_weight': 3,  'gamma': 1.26,  'colsample\_bylevel': 1.0,  'colsample\_bytree':0.9, 'subsample': 0.8,  'reg\_alpha': 1e-05,  'reg\_lambda': 1,  'booster': 'gbtree',  'grow\_policy': 'depthwise',  'max\_delta\_step': 0,  'scale\_pos\_weight': 1.0,  'learning\_rate': 0.07,  'num\_estimators' : 386} |  |

**Table 3. Best parameters obtained of each classifier.**



**Figure 1. Feature importance obtained by LGBM with 'importance\_type' = 'gain'.** Data are presented as mean ± STD of 15 seeds.

**Figure 2. Feature importance obtained by Random Forest.** Data are presented as mean ± STD of 15 seeds.

****

**Figure 3. Feature importance obtained by XGBoost.** Data are presented as mean ± STD of 15 seeds.

**Discussion and conclusions:**

The objective of this project was to predict readmission of diabetic patients. Specifically, our aim is to build a model that will classify patients that will or will not readmit within 30 days from their last admission (here termed, 30-days readmission). To do this we trained three different classifiers i.e. XGBoost, Random Forest and LightGBM, that are all ensemble learning methods, albeit they differ in their approach. The aforementioned models were selected from a list of classifiers that were trained using cross-validation with default parameters on imbalanced and balanced data, based on their negative log loss as our performance measure of choice (Table 1).

All tested models have a degree of overfitting (see Table 1 and 2). To improve it we could possibly improve the tuning of the models, try more models i.e. CatBoost, ensemble of ensembles etc., feature engineering and data processing and of course to collect more data. However, these were out of the scope of this course.

When performing on the test set, we selected Accuracy, Precision, Recall, Negative Log Loss and ROC AUC as our performance measures (Table 2). The mean values of theses metrics across all three classifiers are.

Based on the performance of the three models that were tested we selected X as the best model.

As above-mentioned, our model can predict 30-days readmission with an accuracy of XXX. The values of Precision XXX and Recall XXX, means that we can predict readmission of a patient with a probability of XX, and non-readmission with a probability of XXX.

Since the recall value of the model is higher than the precision, we conclude that our model is good at capturing most positive instances (recall value XXX). However, the lower precision value, means that chance of a false readmission prediction i.e. false positive is (precision value XXX). Therefore, we recommend this model as a screening tool for detecting patients with high risk of readmission, however in fields where the prediction probabilities of this model are insufficient, the positive results of the model may need to be consolidated by a complimentary tool.