CASE STUDY 2: DIABETES

By: Andreea Carolina Craus

1 | INTRODUCTION

The goal of this case study is to predict the readmission of a diabetes patient within 30 days of hospitalization. The data provided 10,1766 unique entries based on the encounter ID and associated with a unique patient number. the target variable “readmission” can be classified into three different categories:

1. “No” if the patient has no readmission record
2. “>30” if the patient was readmitted in more than 30 days
3. “<30” if the patient was readmitted in less than 30 days

2 | METHODS

DATA PREPROCESSING

From an initial look at the features, the columns “counter_id” and “patient_nbr” were removed as they are irrelevant for our analysis. The only columns containing missing values were “max_glu_serum” and “A1Cresult”. However, it was noticed that columns “race”, “weight”, “payer_code”, “medical_specialty”, “diag_1”, “diag_2”, and “diag_3” contained “?” responses which can be visualized in Figure 2. These responses were treated as missing values. Seeing as the majority of the “weight” responses were unknown, this column was removed, as well as the “medical_specialty” column which contained nearly 50% unknown values, and “payer_code”, which was irrelevant to this analysis.

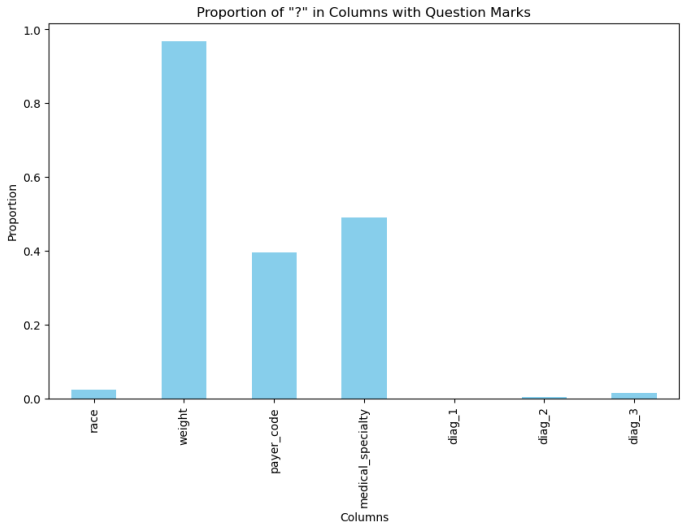

Figure 1

As can be seen in Figure 2, the proportion of missing or unknown values in "max_glu_serum" and "A1Cresult" were very large with a proportion of 0.947 and 0.833, respectively. These features were removed as they are not informative when containing such a large amount of missing data. The remaining features had a very small proportion of missing data but as this data was more difficult to impute, the rows containing missing values were removed. This brought the dataset down to 43 features and 98,053 responses to work with, which is a sufficient amount of data for this analysis.

	diag_1	diag_2	diag_3	race	max_glu_serum	A1Cresult
proportion	0.000206	0.003518	0.013983	0.022336	0.947468	0.832773

Figure 2

EXPLORATORY DATA ANALYSIS (EDA)

As can be seen from the histogram in Figure 2, the target, "readmission", is not equally distributed, with about half of the readmission responses being "No". Since we are only interested in determining readmission within 30 days, the "No" responses were combined with the ">30" day responses and encoded as either 0 for "No" or 1 for "Yes", which can be seen in Figure 3.

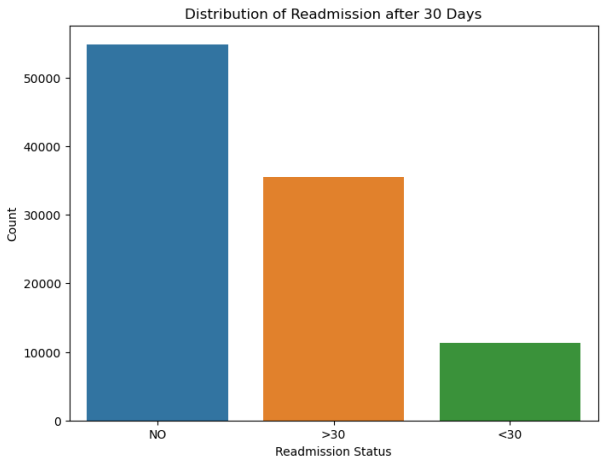

Figure 3

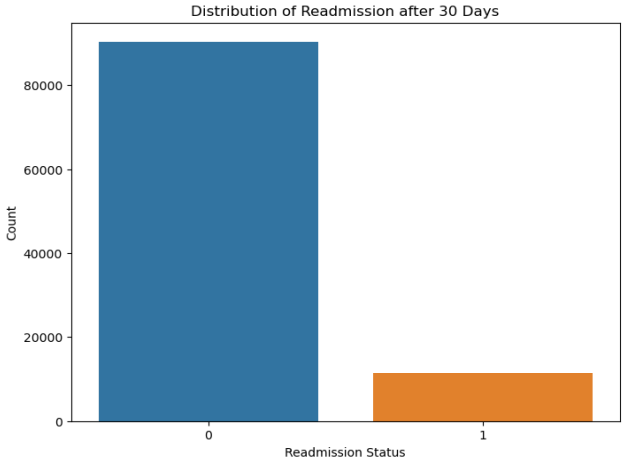

Figure 4

FEATURE ENCODING

In order to be able to use all the features in the machine learning models, which requires numerical input, the categorical variables needed to be encoded into numerical values. This was done using the sklearn "LabelEncoder" which assigns a unique integer to each category, effectively transforming the categorical data into numerical data.

3 | RESULTS

Three different machine learning algorithms were assessed to determine the model that best fits this analysis: Naïve Bayes, Random Forest, and Logistic Regression. To assess the performance and generalization ability of the model, cross-validation was used, which divides the dataset into multiple subsets, training the model on different combinations of these subsets, and then evaluating its performance. The performance metric used was accuracy, which represents the ratio of correctly predicted instances to the total number of instances in the dataset. In addition, to assess performance, the ROC (Receiver Operating Characteristic) Curve was used, which shows a representation of the trade-off between true positive rates and false positive rates for different threshold, which shows the model's ability to discriminate between classes across a range of thresholds. The area under the ROC curve (AUC) is the metric used to assess performance, with a higher AUC indicating better performance.

NAÏVE BAYES

Naïve Bayes is a family of probabilistic classification algorithms based on Bayes' theorem with the "naïve" assumption of independence between algorithms, which calculates the probability of an event based on prior knowledge of conditions related to the event. This algorithm is known for its simplicity, efficiency, and effectiveness, however, the assumption of feature independence does not hold in all scenarios, and not enough EDA was performed to determine the independence of all features. The ROC curve for the Naïve Bayes model can be seen in Figure 5.

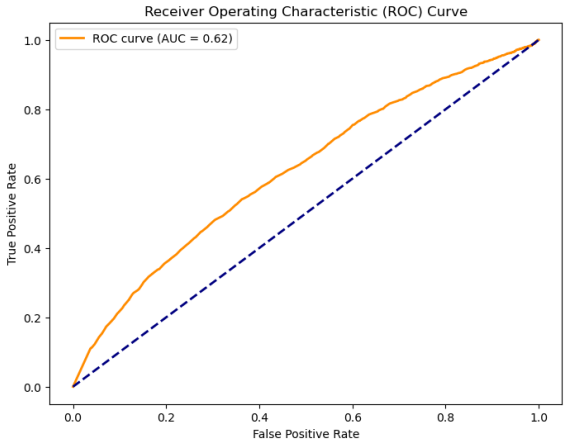

Figure 5

RANDOM FOREST

Random Forest is an ensemble learning algorithm that belongs to the family of decision tree-based methods. The main idea behind this algorithm is to build a multitude of decision trees at training time and output the classification of the individual trees as the final prediction.

Decision trees are simple tree-like structures where each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents the final prediction. Random Forest tends to be robust to overfitting due to its ensemble approach and use of bootstrap sampling and feature randomization. In addition, it can handle a variety of data types, including both categorical and numerical variables. The ROC curve for the Random Forest model can be seen in Figure 6.

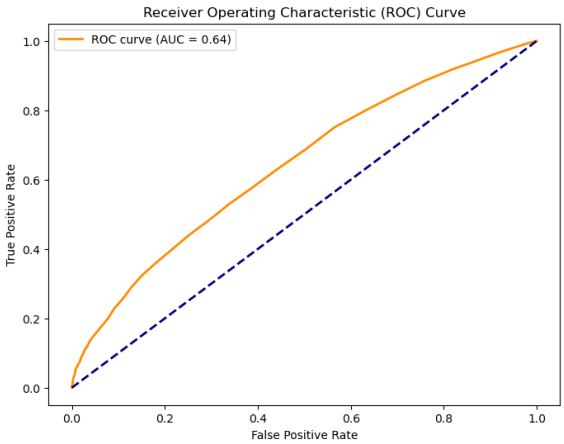

Figure 6

LOGISTIC REGRESSION

Logistic regression is a statistical method used for binary classification problems, where the outcome, or target variable, is categorical and has two classes, as in our case. Logistic regression uses the logistic function, also known as the sigmoid function, to model the relationship between the independent variables and the probability of a particular outcome. This sigmoid function maps any real-valued number to the range[0,1], which is suitable for representing probabilities. The ROC curve for the Logistic Regression Model model can be seen in Figure 7.

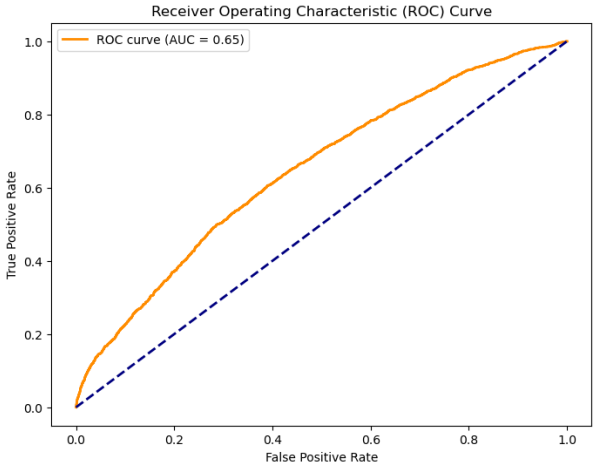

Figure 7

4 | CONCLUSION

The results of the performance of the three models can be seen in Figure 8. The accuracy scores were calculated based on the average of 10 cross-validation folds. As can be seen, the AUC scores were very close to each other, with Logistic Regression performing slightly better with an AUC of 0.65. A visual inspection of the ROC curves do not show any significant differences either. The accuracy scores of the Random Forest and Logistic Regression model were exactly the same and fairly high at 0.887, compared to the Naïve Bayes model. Based on just these results, it is hard to say if Random Forest or Logistic Regression are better fit models for this scenario and this needs to be further looked into. However, for the purpose of this case study, the Logistic Regression model was chosen to move forward with.

Model	Accuracy Score	AUC
Naïve Bayes	0.724	0.62
Random Forest	0.887	0.64
Logistic Regression	0.887	0.65

Figure 8

Using the Logistic Regression model, the top 10 most important features for predicting the readmission of a patient within 30 days are seen below in Figure 9 in descending order with the most important being number_impatient. Without very much additional information about what the features represent, it is difficult to interpret these results more in depth.

	Feature	Coefficient	Absolute_Coefficient
12	number_inpatient	0.324798	0.324798
4	discharge_disposition_id	0.126563	0.126563
41	diabetesMed	0.086261	0.086261
16	number_diagnoses	0.080967	0.080967
36	glipizide-metformin	-0.076265	0.076265
17	metformin	-0.070271	0.070271
6	time_in_hospital	0.055204	0.055204
8	num_procedures	-0.049196	0.049196
30	troglitazone	-0.045705	0.045705
9	num_medications	0.043965	0.043965

Figure 9

In addition, a confusion matrix was calculated for the results of the Logistic Regression classification, which can be seen in Figure 10, and interpreted as follows:

- True Positive, TP (Top Right): The instances correctly predicted as positive
- False Positive, FP (Top Left): The instances correctly predicted as negative
- False Negative, FN (Bottom Right): The instances incorrectly predicted as positive (also known as Type I error or false alarm)
- True Negative, TN (Bottom Left): The instances incorrectly predicted as negative (also known as Type II error)

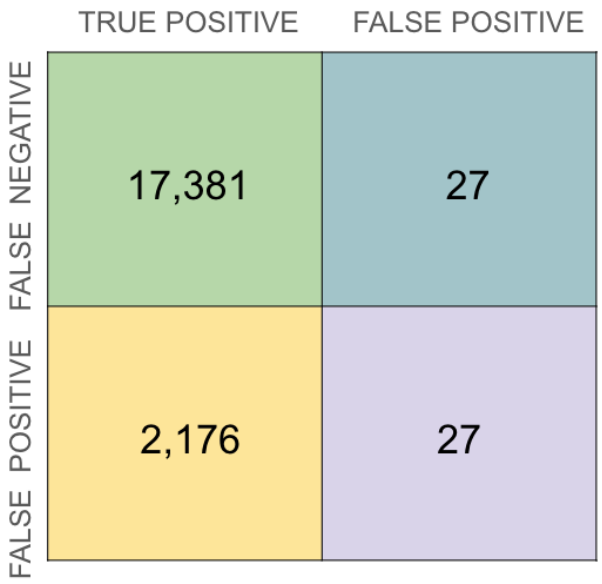

Figure 10

5 | CODE

The associated code can be found in the attached file CarolinaCraus_CaseStudy2.ipynb.