­­­A white paper with black text

Description automatically generated

MLWP: Question 3

1. The independent variables seem suitable for the problem however, the data collection does have a few biases. Firstly, the “Ethnicity” column seems odd as it only looks at whether a patient is Hispanic or not which. For those labelled “not Hispanic” , what are they then? This kind of data collection does not give us much information and is quite biased as the majority (94%) of patients are “not Hispanic”.

Furthermore, if you look at the “Age Bucket” column you will notice that the age category only goes as low as “<55” with no further granularity of age ranges specified. To me this indicates a potential bias towards selecting older patients

1. The Logistic Regression Model had an accuracy of 81.22568093385215 % when used on the test dataset while the Support Vector Machine Model only had an accuracy of 74.5136186770428 %.

The regression model has an ROC Area Under Curve value of 0.882. When a value is closer to 1 it indicates that the model performs well. In this case, it indicates that it can effectively differentiate between patients who are persistent in drug usage and those who are not.

Using the confusion matrix (from Kaggle) shown below we can further investigate the logistic regression performance.

A yellow blue and purple squares with red text

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | **Actual persistent** | **Actual non-persistent** |
| **Predicted persistent** | 920 (TP) | 255(FP) |
| **Predicted non-persistent** | 369(FN) | 1880(TN) |

* The accuracy is a measure of how often the model correctly predicts the outcome. This can be calculating by dividing the (true positives + true negative) by the total number of predictions

Accuracy = ((920 + 1880) / 3424)\*100

Accuracy = 81.8%

* Precision measures the accuracy of the positive predictions made by the model.

Precision = (TP)/ (TP + FP)

Precision = 920 / (920 + 255)

Precision = 0.783

* + This means that when the model predicts a patient will be persistent, we expect It to be correct 78.3% of the time.
* Recall is the ratio of true positives to all the actual positive observations (TP + FN)

Recall = (TP)/(TP + FN)

Recall = 920/(920 + 369)

Recall = 0.714

* + The model was only able to identify 71.4% of the patients who were persistent.
* F1- score metric is a weighted average of the precision and recall, where an F1 score closer to 1 is best and closer to 0 is worse.

F1 = 2\*(Precision \* Recall) / (Precision + Recall)

F1 = 2\*(0.783 \* 0.714) / (0.783 + 0.714)

F1 = 0.75

1. Multicollinearity has been addressed with this line below:



As you can see from the comment the logistic Regression with a “newton-cg” solver will have a penalty parameter set to “elasticNet”. This will regularise the data using a combination of the popular regularization techniques of L1 (Lasso) and L2 (Ridge). It

helps to stabilize the model by distributing the weights more evenly among correlated

features.

(NOTE TO MARKER: Regularisation was implicitly stated to not be included in the coursework)

Yes, there are highly correlated independent variables. “Race” and “Ethnicity” are highly correlated.

1. The data could be limited. This can be because the size of the dataset is not big enough. Furthermore, as stated before, the age ranges may not be fully represented.
2. Yes, there are external factors which could influence medication adherence. Life in general leads to people making different decisions. Maybe they move away to country that does not have access to the drug. Maybe they just decided they do not want to take it anymore for no other reason than they just don’t like it. Maybe they are easily influenced by social media and their feed tells them the drug is bad for them (regardless of whether it is or not) so they stop. The model will never be able to consider everything.