Part 1: Data Processing

How did you load and clean the data, and why?

Firstly, I load data by a function of 'pandas.read_csv' and give each column a specific name.

Secondly, making sure that the data are in the right format, so I replace the 'question mark' data with 'nan'.

Thirdly, removing any nonsensical data, so I drop the row that contains the Nan data.

It is important to keep data clean because is important because clean data can help the Regression model get a better result, wrong data will mislead the model. In Logistic Regression, the wrong data will lead us to get the wrong θ.

How did you split the data into test/train sets and why?

I split the data into 25 percentage of testing data and 75 percentage of training data.

I set the 'stratify' parameter to 'data_output' which is data for classification. 'stratify' parameter is to split data into testing data and training data based on the percentage of distribution of data classification.

It is important to set the testing and training data like what I did because the training data is representative of the main data set.

How did you process the data including encoding, conversion, and scaling, and why?

encoding: I encoding data by Enumeration, for example, I change the sign of '+' into '1' and change the sign of '-' into '0'. Also, dummy variables are used because dummy variables enable to use of a single regression equation to represent multiple groups of data.

conversion: I convert the dataset by selecting the most correlative features. Compared with correlated features, some of the other features might be rather random and unrelated to what we're trying to predict. So, we might be a way to eliminate such unrelated data to prove the accuracy of the model.

scaling: I scale the original data into a range of zero to one because Data Scaling is useful in Logistic regression, support vector machines, deep neural networks, nearest neighbor. However, the performance of logistic regression did not improve significantly by data scaling.

How did you ensure that there is no data leakage?

I avoid using the test set until everything is done.

I split the test set first before finding mean data and selecting the standard deviation of features because systems would essentially minimize our prediction error in process of the standard deviation of features and standard deviation should not have a test set.

Part 2: Training

Train using penalty = 'none' and class weight=None. What is the best and worst classification accuracy you can expect and why?

Best: the best classification accuracy maybe is high than 95 percentage. Because the classification of original data is equally distributed, so it does not matter any value of class weight.

Worst: the worst classification accuracy maybe is lower than 40 percent. Because parameter penalty = 'none', so the trained model maybe is overfitted. Also, if the classification of original data is not equally distributed, so the trained model still lacks training of one of the classifications. As a result, this trained model is not outstanding.

Explain each of the following parameters used by LogisticRegression in your own words: penalty, tol, max iter.

penalty: The choice of penalty parameter will affect the choice of loss function optimization algorithm.

Tol: is the standard to stop solving, float type, defaults to 1e-4. So, when you get to a value, the model stops and thinks an optimal solution is found.

max iter: The algorithm converges to the maximum number of iterations.

Train using balanced class weights (setting class weight='balanced'). What does this do and why is it useful?

'**weights '** is used to indicate the weights of various types in a classification model.

If class_weight selects 'balanced', then the class library calculates the weight based on the training sample size. 'balanced' class_weight can help the model to be still trained evenly by unbalanced data. The larger the sample size of a certain type, then the lower the weight. And the smaller the sample size, the higher the weight. So, it will be helpful to get a higher accuracy regression model (Yadav, 2021).

LogisticRegression provides three functions to obtain classification results. Explain the relationship between the vectors returned by predict(), decision function(), and predict proba().

predict () will give either 0 or 1 as output and returns a class decision using the rule f(x)>0.5

predict proba () will give the only probability of 1 and it equals $1 / (1+ \exp (-1*(\beta_0+\beta_1x_1+\cdots+\beta_kx_k)))$

decision function () returns term inside the exponential which equals $\beta_0+\beta_1x_1+\cdots+\beta_kx_k$

Part 3: Evaluation

What is the classification accuracy of your model and how is it calculated? Give the formula.

Correct classification: True Positives (TP) + True Negatives (TN)

Incorrect classification: False Positives (FP) + False Negatives (FN)

FP = Number of False Positives

FN = Number of False Negatives

P = Total number of instances of Class 1 = TP + FN

N = Total number of instances of Class 2 = FP + TN

classification accuracy = Classification Rate = (TP+TN) / (P+N)

What is the balanced accuracy of your model and how is it calculated? Give the formula.

False Positive Rate (FPR)= FP / total negatives = FP / (TN+FP)

False Negative Rate (FNR)= FN / (TP+FN)

Specificity: 1 - FPR

Sensitivity: 1 - FNR

balanced accuracy = (Specificity + sensitivity)/2

Show the confusion matrix for your classifier for both unbalanced (2a) and balanced (2b) cases. Discuss any differences.

```
LogIsticG = LogisticRegression(penalty='none', class_weight='none', )
```

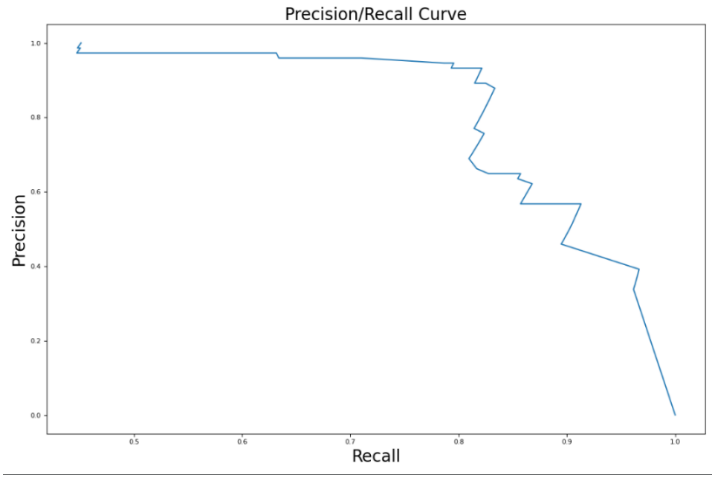
```
[[71 19]
 [ 4 70]]
```

```
LogIsticG = LogisticRegression(penalty='none', class_weight='balanced', )
```

```
[[71 19]
 [ 4 70]]
```

The output is the same. So, I guess the percentage of data classification is the same.

Plot the precision-recall curve and report the Average Precision (AP) for your algorithm. What is the relationship between AP and the PR curve?



the Average Precision:

```
0.8795612587745983
```

Relationship: Area under the Precision-Recall curve is Average Precision (AP)

References list:

Yadav, D., 2021. *Weighted Logistic Regression for Imbalanced Dataset*. [online] Medium. Available at: <https://towardsdatascience.com/weighted-logistic-regression-for-imbalanced-dataset-9a5cd88e68b> [Accessed 2 March 2021].