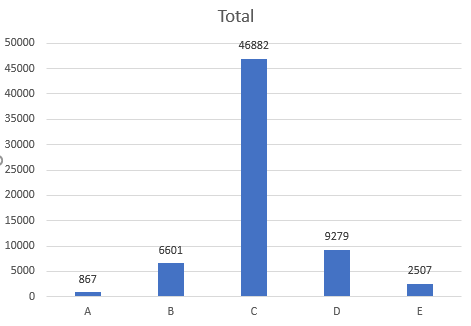
DI-Challenge: Neural Network

Initial Data Analysis

In order to have a nice overview of the whole data in rows and columns, I imported the data to Excel. With whopping 296 columns, feature selection is necessary to reduce the dimension. In addition, oversampling the data points is also necessary because it has imbalanced classes (see chart below).



Import dataset and relevant Python libraries

For the NN model I used the neural\_network library of Scikit Learn with estimator Multi-Layer Perceptron Classifier. After importing the csv file as a dataframe, I separated the target variable from the data.

Feature selection

Reducing the amount of features in the dataset x to save training time and avoid overfitting. The techniques performed are as follows:

Missing value ratio. Result: There are no value missing in the data points.

Low variance data. Filter out features with VarianceThreshold from Scikit Learn and set the threshold to be 0.05, i.e. Columns that have variance less than 0.05 will be removed. Result: 48 features are selected.

Create a new dataset

The target variable has first to be encoded, i.e. categorical values (A, B, C, D, E) are transformed to a numerical encoding (0, 1, 2, 3, 4), respectively. The new dataset is established from the selected columns in dataset x and encoded labels y.

So, the final dataset I will use for my model has a shape of (66, 49), where the 49th column is the labels.

Splitting the dataset

Now that a new cleaner dataset is done, I am ready to split the data I got to training and test dataset. As the name suggest, I’ll use training dataset to build my model from and keep the test dataset untouched and use it at the end to test the model I built. This is quickly done using train\_test\_split from Scikit Learn (test dataset set to be 20% of the whole data).

Last steps of data preprocessing

Back to the second problem I have: imbalanced classes. I opt for the oversampling technique SMOTE to oversample all the classes except the most represented one (class C), so that I have a dataset with equally-distributed classes (37484 data points per class in training dataset and 9398 data points per class in test dataset).

Then, scale the training data to feed to the Neural Network model.

Build the Neural Network model for predictions

Now I can start with building the model. As mentioned before, I use MLPClassifier from Scikit Learn. I choose 3 hidden layers with the same number of neurons as there are features with max. 500 iterations. Then, use the model to predict the test data. Despite of balanced classes, the prediction model always predicts all the data as class C.

>>> print(confusion\_matrix(sm\_y\_test, y\_pred))

[[2269 3556 479 1609 1485]

[ 831 3899 1243 2000 1425]

[ 103 669 7382 905 339]

[ 528 1890 1487 3590 1903]

[ 678 2062 627 3054 2977]]

>>> print(classification\_report(sm\_y\_test,y\_pred))

precision recall f1-score support

0 0.51 0.24 0.33 9398

1 0.32 0.41 0.36 9398

2 0.66 0.79 0.72 9398

3 0.32 0.38 0.35 9398

4 0.37 0.32 0.34 9398

micro avg 0.43 0.43 0.43 46990

macro avg 0.44 0.43 0.42 46990

weighted avg 0.44 0.43 0.42 46990

Model Improvement

To enhance prediction performance, I use Scikit Learn’s hyper-parameter optimization tool (GridSearchCV). Here I adjust different set of parameters, run them all, and take the best ones. I took various parameters for hidden layer sizes as well as the number of neurons, learning rate, solver and alpha.

Undersampling. With NearMiss method from the imblearn library, I obtained 708 data points per class in training dataset and 159 in test data.

>>> print(confusion\_matrix(nm\_y\_test, y\_pred))

[[60 27 19 23 30]

[23 31 30 41 34]

[11 29 58 39 22]

[14 30 27 45 43]

[22 22 11 47 57]]

>>> print(classification\_report(nm\_y\_test,y\_pred))

precision recall f1-score support

A 0.46 0.38 0.42 159

B 0.22 0.19 0.21 159

C 0.40 0.36 0.38 159

D 0.23 0.28 0.25 159

E 0.31 0.36 0.33 159

micro avg 0.32 0.32 0.32 795

macro avg 0.32 0.32 0.32 795

weighted avg 0.32 0.32 0.32 795