How to implement Logistic Regression model for binary classification in python

Import libraries

We start by creating a jupyter notebook in python. We want to import pandas, numpy, sklearn and seaborn. These libraries are used for loading the dataset into our development environment, preprocess the data to prepare it for a training and training the data.

Text

Description automatically generated

Next we will import and explore the data set which is found here (<https://raw.githubusercontent.com/madmashup/targeted-marketing-predictive-engine/master/banking.csv>)

EDA stands for exploratory data analysis. This is typically the first stage in the data mining process. Here we will explore all the variables.

1. The first column is age which is numeric and has minimum value 17, maximum value 98, range 81, average 40.04, standard deviation 10.4, we can also see that the data type is stored as int64

Text

Description automatically generated

1. The second column is job which is categorical. The unique values of this column are blue-collar, technician, management services, retired, admin., housemaid, unemployed, entrepreneur, self-employed, unknown, student. Here is a way to print out the different categories and their counts as a percentage of the whole data set. This is important because if we don’t have enough records associated with a certain value we may not want to train a machine learning model on that attribute value.

Text

Description automatically generated

1. In the same style here we see that most people are either married or singleText

   Description automatically generated
2. Now for education level there are several different categories basic.4y, unknown, university.degree, high.school, basic.9y, professional.course, basic.6y, illiterate

Text

Description automatically generated

1. No adhering to the DRY(don’t repeat yourself) principle I will create a function that I can reuse to print the counts and percent of the categories. We can see that very few customers have defaulted. This data set does not have enough support to predict default

A screenshot of a computer

Description automatically generated with medium confidence

1. The last two variables I will work with are housing and loan. These indicate whether the individual has a house loan or personal loan.

Text

Description automatically generated

1. Next we will aggregate the education levels. Basic.4yr, basic.6y and basic.9y should be a single category basic