Social Network Ads

Data Set

The Data Set proceeds from https://www.kaggle.com/. It contains 4 columns (User ID, Gender, Age, Estimated Salary, Purchased) and 400 rows.

Purchased takes 0 if the user did not buy the product/service advertised and 1 if the user did.

Purpose of the project

The purpose is to fit a model that classifies the instances, in order to predict whether a specific user will buy the advertised product/service.

We are working with labelled data, then the model fit in the category of *supervised learning.*

We are going to fit the following models and observe the results:

- Logistic regression model
- SVM

Work performed

Taking a look at the data set we find the feature "Gender". It is a column with two options, "Female" and "Male". The first step is to convert it into a binary column. We decide to put "0" if the user is female and "1" if it is male. We call the transform column as "gen_bin".

Fitting logistic regression

We fit the model:

$$\Pr(Y = 1) = \frac{e^{\beta_0 + \beta_1\, Age + \beta_2\, EstimatedSalary + \beta_3\, Gender}}{1 + e^{\beta_0 + \beta_1\, Age + \beta_2\, EstimatedSalary + \beta_3\, Gender}}$$

The estimation gives the following result:

$$\Pr(Y = 1) = \frac{e^{-1.63110274e^{-09} + \beta_1 - 4.20972046925316.e^{-09} - 4.516654372896413e^{-06}\, EstimatedSalary - 1.0098590781750962e^{-09}\, Gender}}{1 + e^{-1.63110274e^{-09} + \beta_1 - 4.20972046925316.e^{-09} - 4.516654372896413e^{-06}\, EstimatedSalary - 1.0098590781750962e^{-09}\, Gender}}$$

The first analysis to do is to check if the coefficients signs are the expected. As they are all negative, **the relation between each feature and the probability of a purchase to be done is inverse**. In the case of "Age" feature as the user is older the probability of purchasing the product/service is less.

The first problem arises when we analyse the Salary. Theoretically, **the amount of salary must be positively related with the probability of purchasing but our regression estimates a negative relation**. Finally, if the user is "Male" the probability of purchasing it is reduced.

The **accuracy of the model is 58%.** Not good at all.

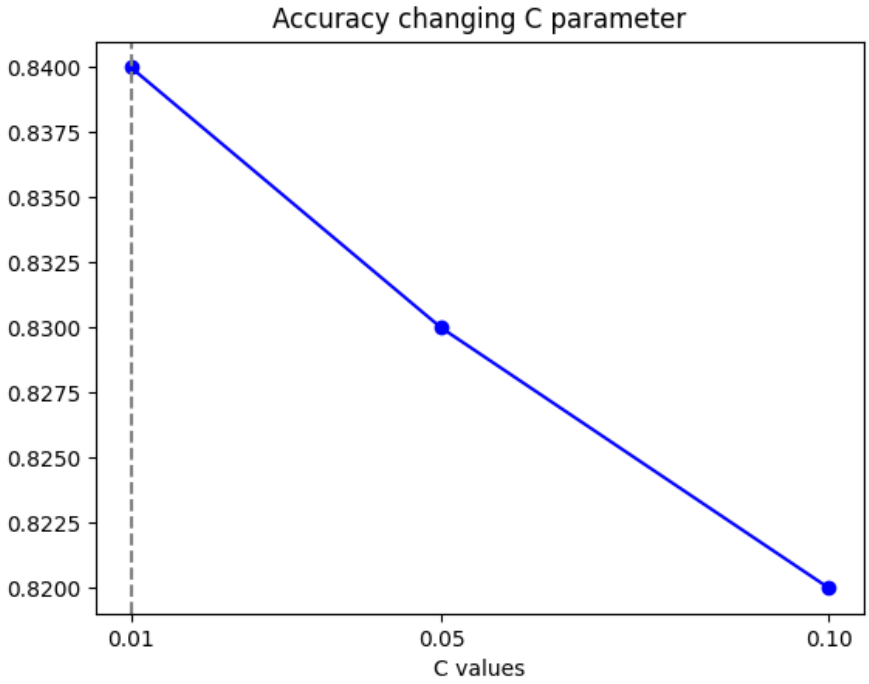
Our model doesn't explain very well the purchasing behaviour of the user, then it is not useful to predict future values. We continue by trying to adjust an alternative model.

SVM

We adjust a Linear SVM model and we obtain just **42% of accuracy**, also the model doesn't converge. As the variance changes a lot between features we apply a **standard scaler** and re-fit the model. The results improve significantly. Now the **accuracy is 82%.**

Then we try to improve our model changing the kernel, the "C" and the "Gamma" values.

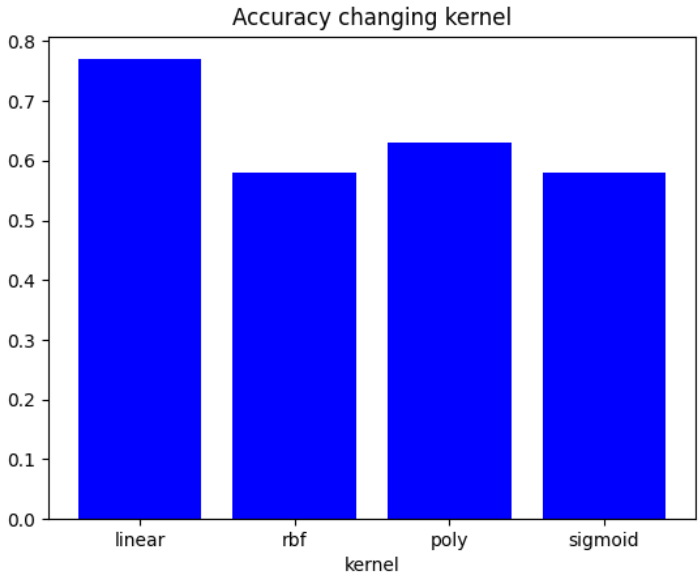
Figure 1



The best value for "C" parameter is 0.01 since accuracy is 84%.

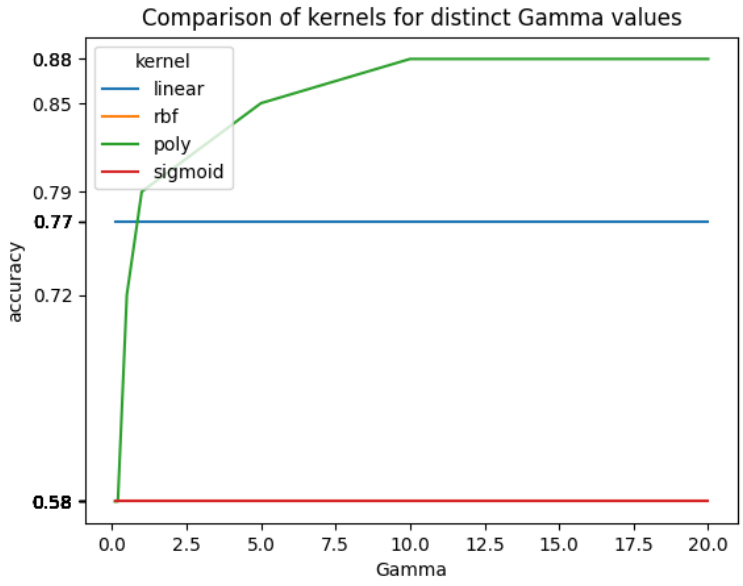
The following step is to analyse the kernel type. As it can be seen in *figure 2* the best kernel is the linear one.

Figure 2



To be sure we fix "C" parameter in 0.01 and we change the "Gamma" value for each kernel. The results are shown in *figure 3.*

Figure 3



This last figure shows that the **best approach** is a polynomial kernel with Gamma=10 and C=0.01. In this case the **accuracy achieved is 88%.**