Machine Learning Made Easy

Candidates: GROUP 3
Lecturer: Prof. Jeff Edmonds

AIMS Cameroon (African Institute for Mathematical Sciences)

March 19, 2022

Group Members I

- Aloys
- Brenda
- Christabel
- Danielle
- Nathalie
- Roland

Contents

1. Introduction
2. Objective
3. Description of the Dataset
4. Data exploration
5. Description of the ANN model
6. Training, validation, loss and accuracy
7. Comparison of the different loss
8. Conclusion
9. References

Introduction

A bank is a financial institution licensed to receive deposits and make loans. Banks may also provide financial services such as wealth management. These institutions are responsible for operating a payment system, providing loans, taking deposits, and helping with investments we intend to explore the strategies of a portoguese banking institution via phone calls on convincing their client on subscribing to a term deposit.

A few marketing strategies is : segmentation of the population, Distribution channel to reach the customer's place, Price and Promotional Strategy

Objective

Our objective is to build an ANN and check the performance of this network by testing whether a client will subscribe to a term deposit or not.

Description of the Dataset

The dataset is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.The classification goal of this dataset is to predict if the client or the customer of polish banking institution will subscribe a term deposit product of the bank or not. The dataset consists on 45211 rows for the training data and 4521 rows for the testining data with 17 columns which correspond to the features and they are the following :

variables	features
categorical	job, marital, education, contact, month, poutcome
numerical	age, balance, day_of_week, duration, campaign, pdays
binary	default, housing, loan, y

Data exploration I

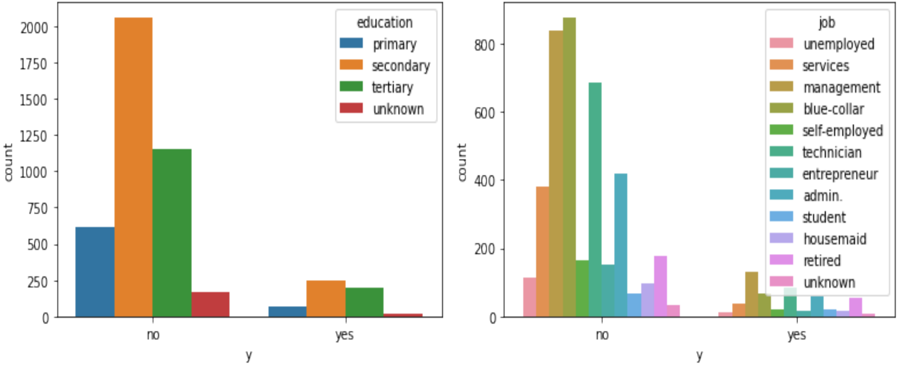

Figure: Education and job Histogram

Data exploration I

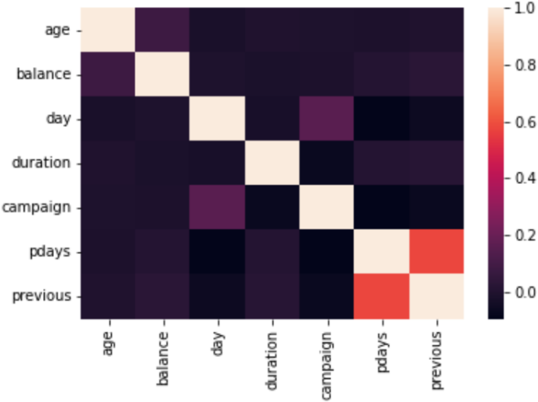

Figure: Correlation between numerical variables

Data pre-processing I

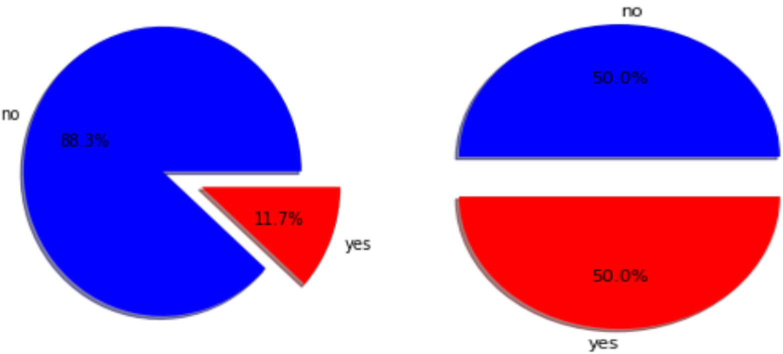

Figure: Imbalanced and Balanced

Description of the ANN model I

- Since we are dealing with a very large dataset, we perform a model with two hidden layers,
- In each hidden layers we have 500 units with the 'relu' activation function
- One unit for the output layer with the sigmoid activation function
- we compile our model using 'adam' as optimizer, 'binary crossentropy' as loss and 'accuracy' as metrics
- We train the model with 50 epochs and we get the following results

- We build our ANN model using train set that we have split into the train and the validation set about of 20% of for the validation set and the remain 80% for the new train data. For our ANN we have choose 500 inputs layers and 2 hidden layers. For the hidden layers, we have choosen the activation function "RELU" and for the output layer, we took the Sigmoid activation function After build it we see that it doesn't fit well so we have tried to introduce some regularizer to ameliorate the ANN model.
- **The L2 regularizer:**
  This regularizer is related to the ridge regression and help to reduce the amount of shrinkage.
- **The L1 regularizer:**
  This regularizer is related to the Lasso regression and help to reduce the amount of sparsity.

Description of the ANN model I

Loss	Validation loss	Accuracy	Validation accuracy
0.071	2.5721	0.9691	0.7174

Table: Differents loss after training the model with 50 epochs

Loss	Validation loss	Accuracy	Validation accuracy
0.1574	0.7104	0.9414	0.7201

Table: Differents loss after training the model using L1-regularizer, with 50 epochs.

Description of the ANN model II

Loss	Validation loss	Accuracy	Validation accuracy
0.1489	0.6751	0.9415	0.7227

Table: Differents loss after training the model using L2-regularizer, with 50 epochs.

Loss	Validation loss	Accuracy	Validation accuracy
0.1503	1.1615	0.9327	0.6842

Table: Differents loss after training the model using Drpout layers, with 50 epochs.

Comparison of the different loss I

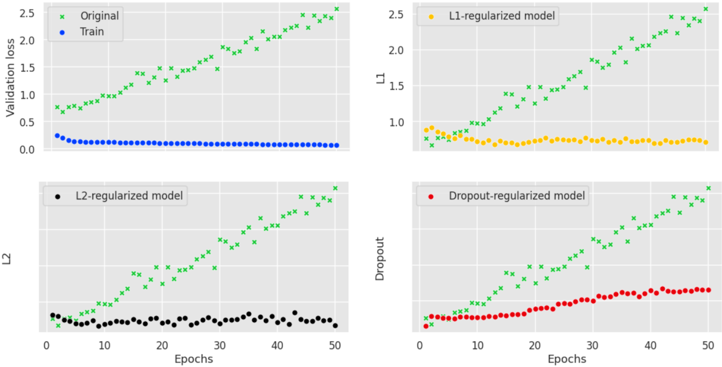

Figure: Comparison of the loss for the original model with the L1-regularizer, L2-regularizer and Droupout layers

Conclusion

Based on our findings and results, a propose concept with which we can infer the amount of influence of an input feature to the target value is checking the correlation coefficients between each input features and the target value. The feature which influences most the output is the one having the highest correlation coefficient. Furthermore from our objective and through building our network and checking the losses we got the loss closed to zero indicating a good performance of our model. Thus we can conclude that the client will subscribe to our term deposit.

- https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets
- Lab and lectures notes
- https://github.com/sinhabishal77/Predicting-whether-the-customer-will-subscribe-to-Term-Deposits