


ASSIGNMENT III ON "MACHINE LEARNING MADE EASY"		
Student's Code		Deadline
GROUP 3		19.03.2022, 01:00 pm
March 19, 2022		Ac. Year: 2021 - 2022
Lecturer(s): "Prof.Jeff Edmonds, Chester Wyke, Göktug Alkan, James Jr. Njong"		

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 portuguese banking institution via phone calls on convincing their client on subscribing to a term deposit. Marketing is a process by which companies create value for customers and build strong customer relationships in order to capture value from customers in return. Marketing campaigns are characterized by focusing on the customer needs and their overall satisfaction. Nevertheless, there are different variables that determine whether a marketing campaign will be successful or not. There are certain variables that we need to take into consideration when making a marketing campaign.

Marketing strategies

Marketing strategies are based on the following principle :

- **Segment of the Population:** To which segment of the population is the marketing campaign going to address and why this aspect of the marketing campaign is extremely important since it will tell to which part of the population should most likely receive the message of the marketing campaign.
- **Distribution channel to reach the customer's place:** Implementing the most effective strategy in order to get the most out of this marketing campaign. What segment of the population should we address? Which instrument should we use to get our message out? (Ex: Telephones, Radio, TV, Social Media Etc.)
- **Price:** What is the best price to offer to potential clients? (In the case of the bank's marketing campaign this is not necessary since the main interest for the bank is for potential clients to open deposit accounts in order to make the operative activities of the bank to keep on running.)
- **Promotional Strategy:** This is the way the strategy is going to be implemented and how are potential clients going to be address. This should be the last part of the marketing campaign analysis since there has to be an indepth analysis of previous campaigns (If possible) in order to learn from previous mistakes and to determine how to make the marketing campaign much more effective.

Problem Statement

Term deposits are a major source of income for a bank. A term deposit is a cash investment held at a financial institution. Your money is invested for an agreed rate of interest over a fixed amount of time, or term. The bank has various outreach plans to sell term deposits to their customers such as email marketing, advertisements, telephonic marketing and digital marketing. Telephonic marketing campaigns still remain one of the most effective way to reach out to people. However, they require huge investment as large call centers are hired to actually execute these campaigns. Hence, it is crucial to identify the customers most likely to convert beforehand so that they can be specifically targeted via call.

1 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.

Dataset Attributes : Our dataset consists on 45211 rows for the training data and 4521 rows for the testing data with 17 columns which correspond to the features and they are the following:

- age : Age of the client (numeric variable)
- job : Type of job (categorical variable)
- marital : Marital status of the client (categorical variable)
- education : Education level (categorical variable)
- default : Credit in default (binary variable)
- balance : average yearly balance, in euros (numerical variable)
- housing : has housing loan (binary variable)
- loan : has personal loan (binary variable)
- contact : contact communication type (categorical variable)
- month : last contact month of year (categorical variable)
- day_of_week : last contact day of the month (numerical variable)

- **duration** : last contact duration, in seconds (numerical variable)
- **campaign** : number of contacts performed during this campaign and for this client (numerical variable)
- **pdays** : number of days that passed by after the client was last contacted from a previous campaign (numerical variable)
- **previous** : number of contacts performed before this campaign and for this client (numerical variable)
- **poutcome** : outcome of the previous marketing campaign (categorical variable)
- **Marital statut** is categorical variable which has 3 categories : married, single and divorced as shown in the picture below. Our output y which evaluate if the costumer takes the term deposit or not and it consist on two categories which are yes or no. When we look at the figure, we see that among the clients who answer yes to the cash deposit, the category married has a larger number of people, and among the clients who answer no to the cash deposit, the category married still has the largest number of people.
- **Housing** is categorical variable which evaluate if a costumer has already a house and the possible answers are yes or no. When we look at the figure, we see that among the clients who answer yes to the cash deposit, the category of people who do not have a house has higher number of people, and among the clients who answer no to the cash deposit, the category of people who had a house have a high number of people.
- **Education** is categorical variable with 04 categories which evaluate if the costumer has ended his studies at the primary, secondary, tertiary or unknown. When we look at the figure, we see that among the clients who answer yes to the cash deposit, the category of people who ended their study in the secondary school have higher number of people, and among the clients who answer no to the cash deposit, the category of people who ended their study at the secondary school have a high number of people.
- **Default** is categorical variable with 02 categories which evaluate if the costumer has a credit in default or not. So the categories are yes or no. When we look at the figure, we see that among the clients who answer yes to the cash deposit, the category of people who did not have credit in default have higher number of people, and among the clients who answer no to the cash deposit, the category of people who did not have credit in default have a high number of people.
- **Loan** is categorical variable with 02 categories which evaluate if the costumer has personal loan or not. So the categories are yes or no. When we look at the figure, we see that among the clients who answer yes to the cash deposit, the category of people who have personal loan have lowest number of people, and among the clients who answer no to the cash deposit, the category of people who do not have a loan have a high number of people.
- **Contact** is categorical variable with 03 categories which evaluate the type of contact communication that costumers used. So the categories are cellular, telephone and unknown. When we look at the figure, we see that among the clients who answer yes to the

cash deposit, the category of people who used cellular have the higher number of people, and among the clients who answer no to the cash deposit, the category of people who used cellular has the higher number of people.

- **Month** is categorical variable with 12 categories which are the different months of a year. So we see that the months where costumers accepted to take the credit are may, july and november while the month where costumers do not want to take the credit is january.
- **Job** is categorical variable with 12 categories which are unemployed, services, management, blue-collar, self-employed, technician, entrepreneur, admin, student, housemaid, retired or unknown. So we see that we have the costumer with management as job that have the higher number of people wile those with self-employed, blue-collar and technician who said no to the cash deposit has also the higher representation .
- **poutcome** is categorical variable with 04 categories which are failure, other, success and unknown. we see that among the clients who answer yes to the cash deposit, the category unknown have the higher number of people, and among the clients who answer no to the cash deposit, the category unknown has the higher number of people.

2 Methodology

3 Methods and results

3.1 Methods

3.2 Results

Data pre-processing

- **Imbalanced data:** We notice that our dataset is imbalanced because looking at this figure bellow, there is 88.3% of clients who did not subscribe to a term deposit, which is more than the 11.7% of the one who subscribe.

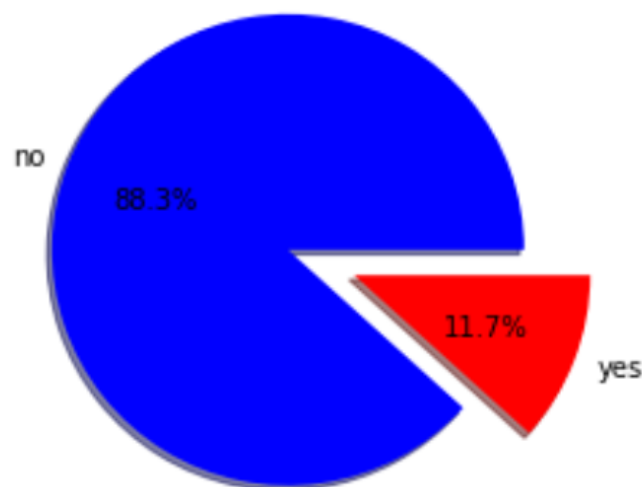


Figure 1: Initial histogram of the output 'y' in the dataset

This difference in percentages is very large and to perform a model which will have a good ability to predict we need to work with a balanced dataset.

- **Balanced the dataset:** We modify our dataset in such a way that the output 'y' has 50% of clients who subscribed to a term deposit or not. The result can be shown in this figure bellow:

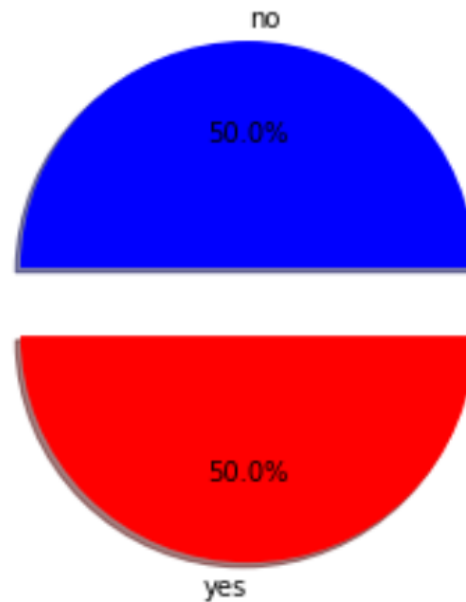


Figure 2: Histogram of the predicted values 'y' with a balanced data

- **Selection of features:** while plotting the correlation matrix between the numerical features of the dataset, we obtain the following figure:

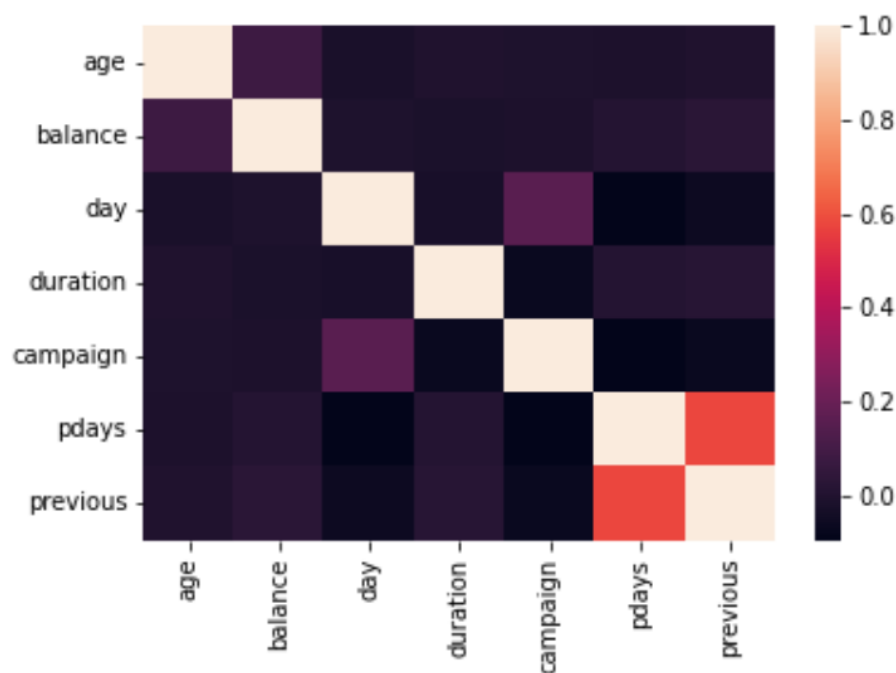


Figure 3: Correlation between the numerical features of the dataset

We observe that the all the features are fairly correlated, except **pdays** and **previous**

which are moderately correlated. Since there is not a strong correlation between those variables, (we can't say that there is a relationship between them) we decide to keep all of them.

- **Statistics of the training data before standardization of the numerical features**

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

Figure 4: summary table of the training data for the numerical variables before standardization

In this figure above, we can observe that the model we will train using this training data, will consider the clients who is at least 18 years old and a maximum age of 95 years old. We can easily observe that those features have a very high variance reason why it's really important that we standardize our train data and also the test data.

- **Statistics of the training data after standardizing the numerical features**

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	0.297873	0.085171	0.493547	0.052494	0.028449	0.047245	0.002110
std	0.137906	0.027643	0.277416	0.052364	0.049968	0.114827	0.008376
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.194805	0.073457	0.233333	0.020943	0.000000	0.000000	0.000000
50%	0.272727	0.076871	0.500000	0.036600	0.016129	0.000000	0.000000
75%	0.389610	0.085768	0.666667	0.064864	0.032258	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 5: Summary table of the training data after standardization

We can now observe that those numerical features have now variance close to 1. We recall that this is only for the numerical features, and in order to have a uniform dataset we have converted the categorical variables into numerical variables and since the dataset is too large we can present the final result in this document but, we give in this figure below a few presentation on how our some categorical variables look after encoding.

marital_divorced marital_married marital_single

0.0	1.0	0.0
0.0	0.0	1.0
0.0	1.0	0.0
0.0	1.0	0.0
0.0	0.0	1.0

Figure 6: Categorical variable 'marital status' after encoding

Description of the ANN model

- 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

Loss	Validation loss	Accuracy	Validation accuracy
0.071	2.5721	0.9691	0.7174

Table 1: Differents loss after training the model with 50 epochs

Training and validation loss

With the architecture of our model we get



Figure 7: Training and validation loss

The training loss show a constant loss closed to 0 ,following the number of iterations while the validation loss keeps fluctuating.

Training and validation accuracy

With the same architecture, we get

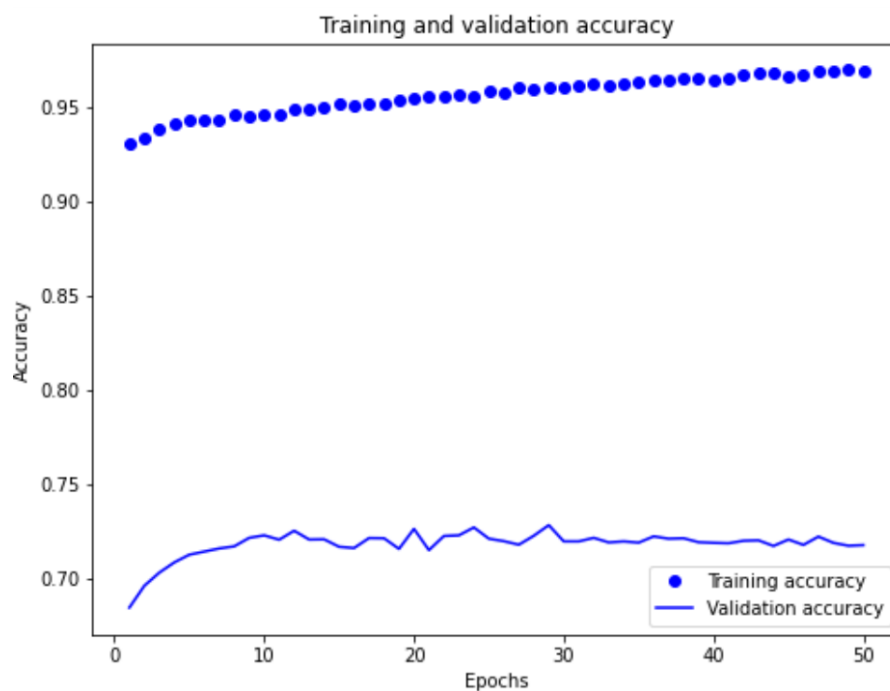


Figure 8: Training and validation accuracy

In this figure above, we observe that the network has a very high accuracy which is keep increasing and it's very higher than the validation accuracy we can't conclude on the performance of the model.

In order to mitigate overfitting, we tried to perform a model in which the distribution of the weight values is more regular. To do that, we used two methods of weight regularization which consist to add to the loss function of the network a cost associated with having large weight, which can be proportional to the absolute value of the weight (**L1-regularizer**), or proportional to the square of the value of the weight coefficients (**L2-regularizer**).

Adding L1-regularizer

We have built an L1-regularizer model with

- Two hidden layers with 20 units in each of them, using relu activation function
- One output layer with the sigmoid activation function
- We compile the model with 'rmsprop' as optimizer, 'binary-crossentropy' as loss and 'accuracy' as metric.

We train the model with 50 epochs and we get

Loss	Validation loss	Accuracy	Validation accuracy
0.1574	0.7104	0.9414	0.7201

Table 2: Different loss after training the model using L1-regularizer, with 50 epochs.

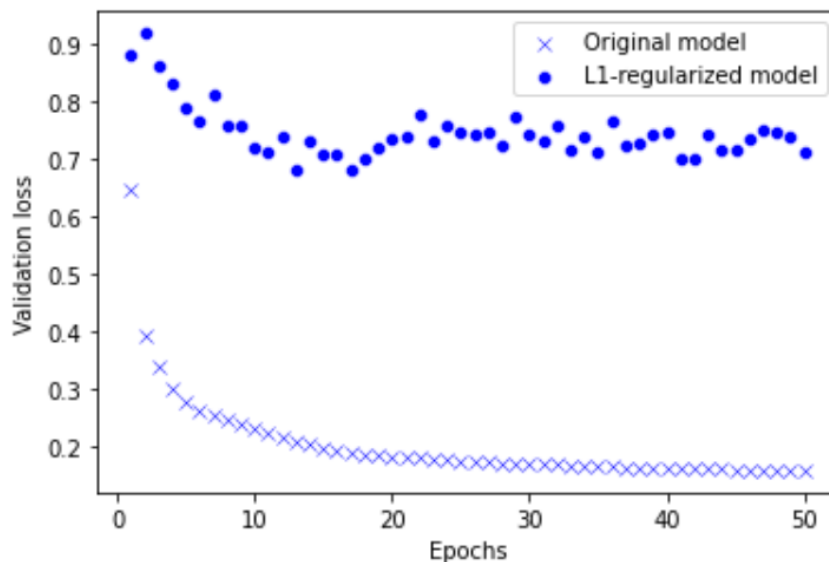


Figure 9: Training loss and validation loss using L1-regularizer

As we can see in this figure above, the model using L1-regularizer tend to become resistant to overfit compared to the original model.

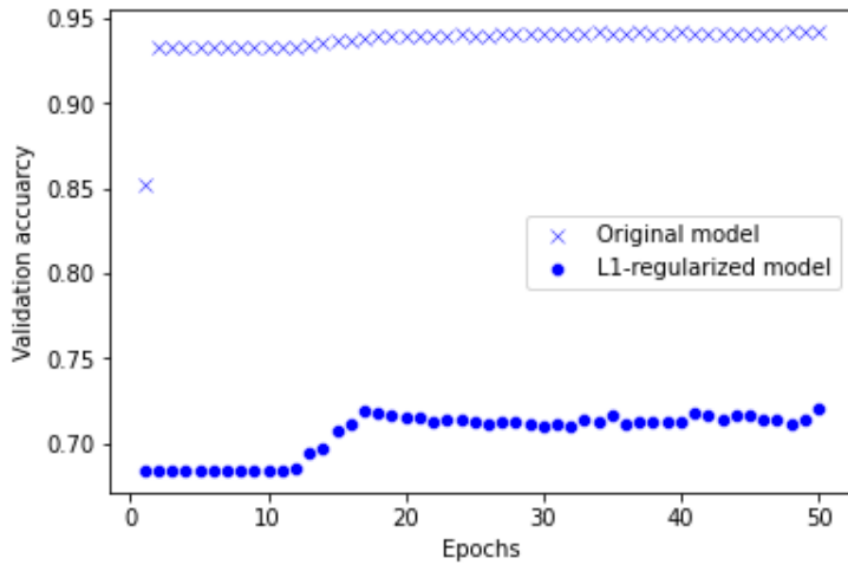


Figure 10: Validation accuracy using L1-regularizer

With L1-regularizer there is not a big difference with the original model.

Adding L2-regularizer

We have built an L2-regularizer model with

- Two hidden layers with 20 units in each of them, using relu activation function
- One output layer with the sigmoid activation function
- We compile the model with 'rmsprop' as optimizer, 'binary-crossentropy' as loss and 'accuracy' as metric.

We train the model with 50 epochs and we get

Loss	Validation loss	Accuracy	Validation accuracy
0.1489	0.6751	0.9415	0.7227

Table 3: Different loss after training the model using L2-regularizer, with 50 epochs.

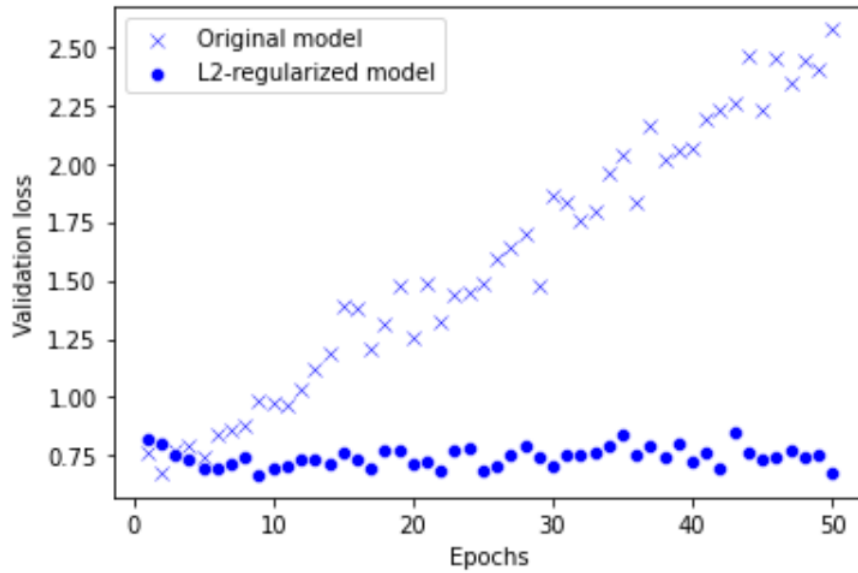


Figure 11: Validation loss using L2-regularizer

We observe that the model with L2-regularizer becomes more resistant to overfitting than the original model and overfitting start after around 10 epochs.

Adding Dropout layers

We add dropout layers model with the same architecture as L1-regularizer and L2-regularizer and we train the model with 50 epochs and we get the results bellow

Loss	Validation loss	Accuracy	Validation accuracy
0.1503	1.1615	0.9327	0.6842

Table 4: Differents loss after training the model using Drpout layers, with 50 epochs.

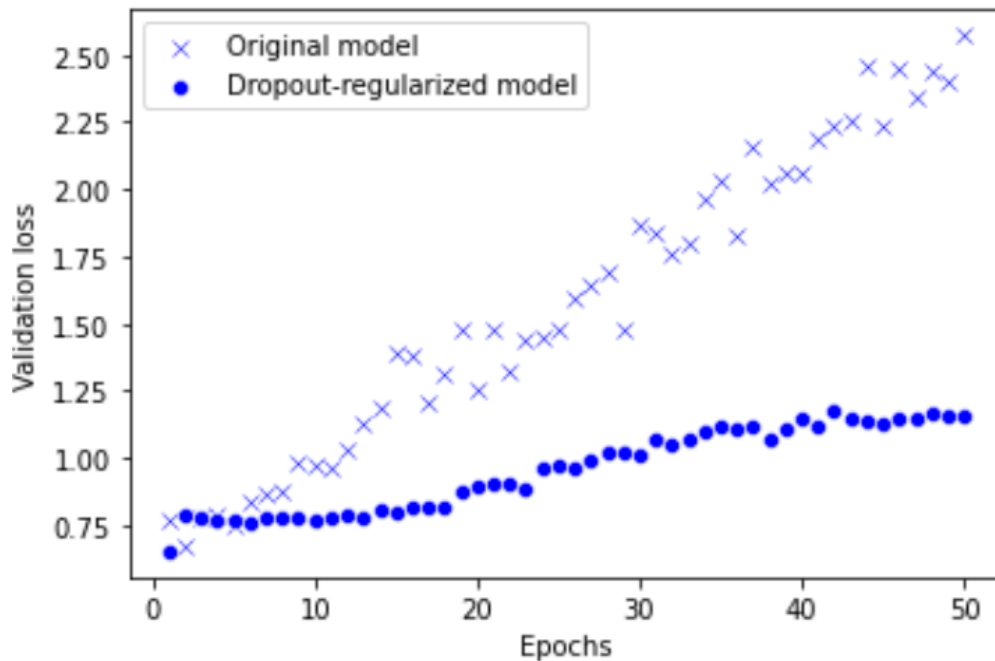


Figure 12: Validation loss using dropout regularized model

We observe that overfitting has significantly reduce when using a model with dropout layers, compared to the original model. Overfitting start after around 12 epochs.

Comparison of the different loss

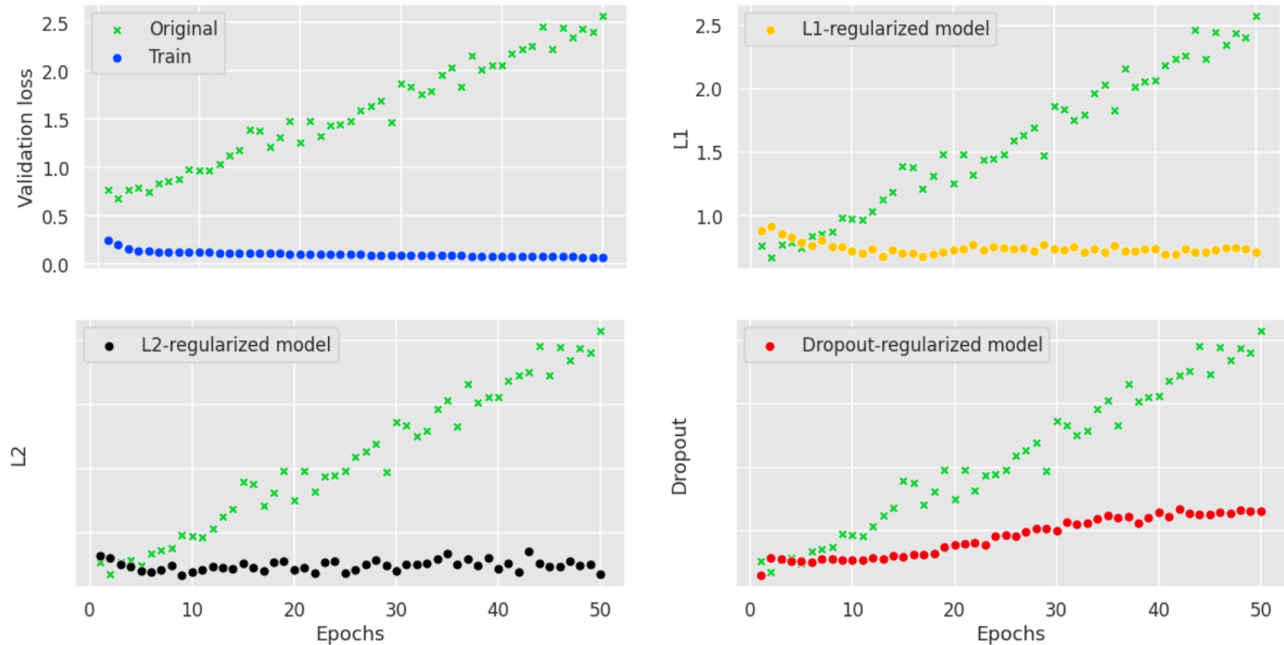


Figure 13: Comparison of the loss for the original model with the L1-regularizer, L2-regularizer and Dropout layers

The above figures shows the different ways we used to prevent overfitting in our network and we observe that the best method is the dropout method, since the curves tends to be more closer than the other graphs .

Interpretations of the results

Conclusion

References

- <https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets>
- Lab and lectures notes