

Deep Learning (Audio Book)

1. Introduction

Audiobook data is extracted to work on the marketing model. The objective is to identify the probability that a customer will buy again or not. This model will help to improve sales and profitability. The dataset has the following columns:

- ID
- Book length overall in minutes (0)
- Book length average in minutes (1)
- Price overall in \$ (2)
- Price average in \$ (3)
- Review (4)
- Review out of 10 (5)
- Minutes listened (6)
- Completion (7)
- Support requests (8)
- Last visited (9)
- Target

2 years of observation window and an additional 6 months of performance window. Target = 0 indicates the customer did not purchase and target = 1 indicates the customer purchased.

Dataset: Audiobooks_data.csv

2. Data Processing

There are three parts to data processing. In the first part, the data is balanced so that there are equal percentages of target = 0 and target = 1. In the second part, the data is divided into train, validation, and test datasets. In the third part data is scaled using standardization.

Part 1

Initial ratio	Data processing	Final ratio
11,847 observations where target = 0	Over-sampling of minor classes is done. The observations where the target = 1 are replicated 4 times and the observations where the target = 0 are not replicated.	11,847 observations where target = 0
2,237 observations where target = 1		11,185 observations where target = 1

Part 2

The data is divided into 70% training, 10% testing and 20% validation.

Train	Test	Validation
Count of observations = 16,122	Count of observations = 2,303	Count of observations = 4,607
Propensity ratio = 48.57%	Propensity ratio = 48.55%	Propensity ratio = 48.56%

Part 3

The standard scalar is trained on the training dataset and then applied on train, testing, and validation. The univariate analysis of all the datasets combined:

	0	1	2	3	4	5	6	7	8	9
count	23,032	23,032	23,032	23,032	23,032	23,032	23,032	23,032	23,032	23,032
min	-2.715	-1.869	-0.650	-0.692	-0.445	-11.473	-0.385	-0.438	-0.187	-0.755
mean	-0.005	-0.001	0.001	0.001	-0.001	0.001	0.004	0.000	0.000	0.004
max	1.199	6.010	24.316	18.225	2.250	1.571	4.679	7.548	67.522	4.272
std	1.002	1.002	1.015	1.004	0.999	0.990	1.004	1.000	1.067	1.002

3. Ensemble Models

There are three parts to ensemble models. In the first part, 3 models are created. In the second part, their test accuracy is noted. In the third part, the outputs of these models are combined.

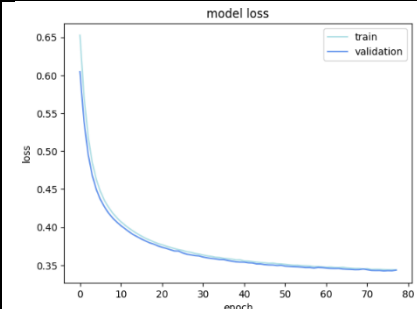
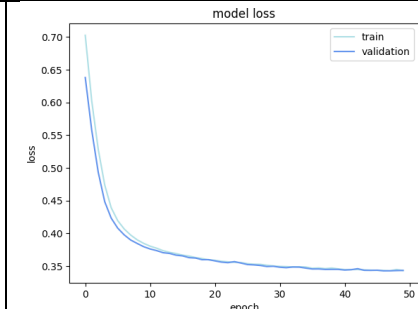
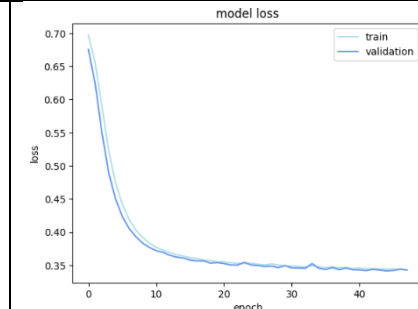
Part 1

Three models are created and each of these models has a separate configuration. Each model is run 50 times and the test accuracy is noted. The average and standard deviation of the test accuracy is published.

Model 1	Model 2	Model 3
Only 1 hidden layer Number of nodes in each hidden layer = 30 Activation function in hidden layer = RELU	2 hidden layers The number of nodes in each hidden layer = 15 Activation function in hidden layer = RELU	3 hidden layers The number of nodes in each hidden layer = 10 Activation function in hidden layer = RELU
The model is run 50 times The average test accuracy = 0.834 The standard deviation of test accuracy = 0.003	The model is run 50 times The average test accuracy = 0.832 The standard deviation of test accuracy = 0.003	The model is run 50 times The average test accuracy = 0.830 The standard deviation of test accuracy = 0.005

Part 2

In this part, the candidate models are trained only once. and their test accuracy is noted.

Model 1	Model 2	Model 3
Only 1 hidden layer Number of nodes in each hidden layer = 30 Activation function in hidden layer = RELU	2 hidden layers The number of nodes in each hidden layer = 15 Activation function in hidden layer = RELU	3 hidden layers The number of nodes in each hidden layer = 10 Activation function in hidden layer = RELU
Model is run for 1 time with random seed of 42 The test accuracy = 83.24% The test accuracy is within a 95% confidence interval. Hence this iteration is suitable	Model is run for 1 time with random seed of 108 The test accuracy = 83.28% The test accuracy is within a 95% confidence interval. Hence this iteration is suitable	Model is run for 1 time with random seed of 108 The test accuracy = 83.50% The test accuracy is within a 95% confidence interval. Hence this iteration is suitable
		

Part 3

The output from the three models is combined using a simple average. The test accuracy of the ensemble model is 83.37%