Books’ Rating Prediction Using Just Neural Network

Alaa Mazen Maghari, Iman Ali Al-Najjar, Said Jamil Al-Iaqta, Samy S. Abu-Naser

Department of Information Technology,
Faculty of Engineering and Information Technology,
Al-Azhar University, Gaza, Palestine

Abstract: The aim behind analyzing the Goodreads dataset is to get a fair idea about the relationships between the multiple attributes a book might have, such as: the aggregate rating of each book, the trend of the authors over the years and books with numerous languages. With over a hundred thousand ratings, there are books which just tend to become popular as each day seems to pass. We proposed an Artificial Neural Network (ANN) model for predicting the overall rating of books. The prediction is based on these features (bookID, title, authors, isbn, language_code, isbn13, # num_pages, ratings_count, text_reviews_count), which were used as input variables and (average_rating) as output variable for our ANN model. Our model were created, trained, and validated using data set in JNN environment, which its title is “Goodreads-books”. Model evaluation showed that the ANN model is able to predict correctly 99.78% of the validation samples.

Keywords: Predictive Analysis, Artificial Neural Networks, Books Rates, Goodreads.

Introduction

In this study, we will analyze the dataset which contains various information of the books on the website of the world’s largest book archive and book proposal site GoodReads. This dataset also includes information on the names, writers and spelling languages of books, as well as the rating and total score based on the votes given by various users. Artificial neural networks (ANNs) will be used for the analysis. Artificial neural networks are like biological neural networks and offer a technique, which solves the problem of prediction [3]. Neural networks contain input, hidden and output layers. Hidden layers convert the input into usable thing to the output layer [5]. The ANN Model goes through training and validation on a dataset. Training in which that the network is trained is done on a dataset. Then a configuration is done to the weights of the connections between neurons. Validation in which that the network is validated to determine the prediction of a new dataset [6]. In this study, we used about 33% of the dataset instances for network validation, the remaining 67% for training. ANN Architecture is shown in figure 1.

1. Literature Review

Artificial Neural Networks have been used many fields. In Education such as: Predicting Student Performance in the Faculty of Engineering and Information Technology using ANN[5], Prediction of the Academic Warning of Students in the Faculty of Engineering and Information Technology in Al-Azhar University-Gaza using ANN[5], Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach[6].

In the field of Health such as: Parkinson’s Disease Prediction [7], Classification Prediction of SBRCTs Cancers Using ANN [7], Predicting Medical Expenses Using ANN[8], Predicting Antibiotic Susceptibility Using Artificial Neural Network[8], Predicting Liver Patients using Artificial Neural Network[7], Blood Donation Prediction using Artificial Neural Network[9], Predicting DNA Lung Cancer using Artificial Neural Network[10], Diagnosis of Hepatitis Virus Using Artificial Neural Network[10], COVID-19 Detection using Artificial Intelligence[11].

www.ijeais.org/ijeais
In the field of Agriculture: Plant Seedlings Classification Using Deep Learning [12], Prediction of Whether Mushroom is Edible or Poisonous Using Back-propagation Neural Network[15], Analyzing Types of Cherry Using Deep Learning[21], Banana Classification Using Deep Learning[13], Mango Classification Using Deep Learning[14], Type of Grapefruit Classification Using Deep Learning[7], Grape Type Classification Using Deep Learning[3], Classifying Nuts Types Using Convolutional Neural Network[2], Potato Classification Using Deep Learning[3], Age and Gender Prediction and Validation Through Single User Images Using CNN[5].

In other fields such as : Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods [14], Predicting Overall Car Performance Using Artificial Neural Network [8], Glass Classification Using Artificial Neural Network [9], Tic-Tac-Toe Learning Using Artificial Neural Networks[14], Energy Efficiency Predicting using Artificial Neural Network[15], Predicting Titanic Survivors using Artificial Neural Network[14]. Classification of Software Risks with Discriminant Analysis Techniques in Software planning Development Process[13], Handwritten Signature Verification using Deep Learning[12], Email Classification Using Artificial Neural Network[14], Predicting Temperature and Humidity in the Surrounding Environment Using Artificial Neural Network[12], English Alphabet Prediction Using Artificial Neural Networks[9].

2. Methodology

We downloaded a data set from kaggle that contains books information from goodreads application/website. This dataset created by the user Soumik [19]. We did some preprocessing on the data, and then we trained our ANN model and validated it.

3. Original Dataset Description

Table 1: Original Dataset Description

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bookID</td>
<td>A unique identification number for each book.</td>
<td>Integer</td>
</tr>
<tr>
<td>2</td>
<td>title</td>
<td>The name under which the book was published.</td>
<td>String</td>
</tr>
<tr>
<td>3</td>
<td>authors</td>
<td>Names of the authors of the book. Multiple authors are delimited with -.</td>
<td>String</td>
</tr>
<tr>
<td>4</td>
<td>average_rating</td>
<td>The average rating of the book received in total.</td>
<td>Real</td>
</tr>
<tr>
<td>5</td>
<td>isbn</td>
<td>Another unique number to identify the book, the International Standard Book Number.</td>
<td>Long</td>
</tr>
<tr>
<td>6</td>
<td>language_code</td>
<td>Helps understand what is the primary language of the book. For instance, eng is standard for English.</td>
<td>String</td>
</tr>
<tr>
<td>7</td>
<td>isbn13</td>
<td>A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN.</td>
<td>Long</td>
</tr>
<tr>
<td>8</td>
<td># num_pages</td>
<td>Number of pages the book contains.</td>
<td>Integer</td>
</tr>
<tr>
<td>9</td>
<td>ratings_count</td>
<td>Total number of ratings the book received.</td>
<td>Integer</td>
</tr>
<tr>
<td>10</td>
<td>text_reviews_count</td>
<td>Total number of written text reviews the book received.</td>
<td>Integer</td>
</tr>
</tbody>
</table>

3.1 Dataset Preprocessing

We wanted to use this dataset to build an ANN model to predict the overall rating of the books (attribute number 4). The first thing we had to do, is choose a suitable factors for this prediction, and delete the unnecessary ones, we chose these factors to be our input to the predictive model: #num_pages, rating_count, text_reviews_count, language_code. Moreover, the dataset contain 11128 instances. After preprocessing it becomes 11122 which is a large a number to a neural network to deal with, so, we divided these samples to 7451 training instances, and 3670 validation instances. In addition, because of the integer numbers of the inputs are too large comparing with the real rate values, we did a normalization to them so all the data are real.

Normalization formula was:
While checking the instances, it has been noticed that there is a conflict between some instances; which means, there are at least two books with the same input values but different rates, we excluded the secondary ones. Moreover, there were validation instances that are out of range, we converted them to training. Now, the dataset is ready for training and validation.

3.2 Our ANN Model

The resulted predictive ANN model is shown in Figure 2 and Figure 6.

3.3 Validation

Our ANN model was able to predict the books’ overall rate with 99.78% accuracy, with about 0.005 errors as seen in Figure (3). Furthermore, The Model showed that the most effective factor in a book’s rate is the rating_count. More details are shown in Figure (4).

![Our ANN Model](image_url)
Figure 3: Validation and Errors

Figure 4: Attributes Importance

Figure 5: imported pre-processed Dataset
4. Conclusion

A predictive Artificial Neural Network Model for predicting books’ rating was developed. The Model trained and validated using a dataset from the goodreads application/website. We did some preprocessing on the dataset to make it suitable as input to our ANN model. Validation showed that the model is 99.78% accurate.

References


20. EasyNN Tool