An AI Based Recommendation Engine Case Study

Objective:

Analyse, design and build "An Artificial Intelligent Recommendation Engine" based on "Pearson Correlation Coefficient" machine learning algorithm.

Background:

Have you ever browsed Amazon or similar and noticed the "*Customers That Bought This Item Also Bought These Items*" list, or after you've selected a movie on Netflix, recommendations for similar flicks?

Would you like to add this capability to your e-commerce web site to increase cross-sell opportunities?

The ability to find trends in the browsing habits and choices of users has become a must for many customer facing websites. Collaborative Filters are the magic that makes these features possible.

Recommendation systems are the bread and butter of large e-commerce web sites. Potential customers are presented with highly relevant product recommendations based on their demonstrated likes; thus, increasing the probability of purchase. Customers purchasing one or more items are presented with like-items increasing the likelihood of cross-selling. And returning customers are retained by the quality of recommendations based on their previous purchases. In general, the goal of any recommendation system is to present users with a highly relevant set of items.

In this case study, we'll explore ***"Pearson Correlation Coefficient",*** one of the popular implementation of a recommendation system, how it works, and how to incorporate it into your project.

Pearson Correlation Coefficient – The algorithm

In statistics, the Pearson correlation coefficient (PCC), also referred to as the Pearson's r or Pearson product-moment correlation coefficient (PPMCC), is a measure of the linear dependence (correlation) between two variables X and Y. It has a value between +1 and −1 inclusive, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is widely used in the sciences. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s. Early work on the distribution of the sample correlation coefficient was carried out by Anil Kumar Gain and R. A. Fisher

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. – **By Wikipedia**

For more information visit:
`https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`

Formula: How to Apply?

To determine how strong, the relationship is between two variables, a formula must be followed to produce what is referred to as the **coefficient value**. The coefficient value can range between -1.00 and 1.00. If the coefficient value is in the negative range, then that means the relationship between the variables is **negatively correlated**, or as one value increases, the other decreases. If the value is in the positive range, then that means the relationship between the variables is **positively correlated**, or both values increase or decrease together.

Let's look at the formula for conducting the Pearson correlation coefficient value.

Step one: Make a chart with your own data for two variables, labelling the variables (*x*) and (*y*), and add three more columns labelled (*xy*), (*x*^2), and (*y*^2).

A simple data chart might look like this:

Person	Age (*x*)	Score (*y*)	(*xy*)	(*x*^2)	(*y*^2)
1					
2					
3					

More data would be needed, but only three samples are shown for purpose of example.

Step two: Complete the chart using basic multiplication of the variable values.

Person	Age (*x*)	Score (*y*)	(*xy*)	(*x*^2)	(*y*^2)
1	20	30	600	400	900
2	24	20	480	576	400
3	17	27	459	289	729

Step three: After you have multiplied all the values to complete the chart, add up all the columns from top to bottom.

Person	Age (*x*)	Score (*y*)	(*xy*)	(*x*^2)	(*y*^2)
1	20	30	600	400	900
2	24	20	480	576	400
3	17	27	459	289	729
Total	61	77	1539	1265	2029

Step four: Use this formula to find the Pearson correlation coefficient value.

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- N = number of pairs of scores
- Σxy = sum of the products of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

Pearson Correlation Coefficient Formula

N = 3

Sum(XY) = 1539

Sum(X) = 61

Sum(Y) = 77

Sum(X^2) = 1265

Sum(Y^2) = 2029

R1 = (N * Sum(XY)) – (Sum(X) * Sum(Y))

R2 = ((N * Sum(X^2)) – (Sum(X) * Sum(X))) * ((N * Sum(Y^2)) – (Sum(Y) * Sum(Y))

R = R1 / Sqrt(R2)

R = -0.739853246

Step five: Once you complete the formula above by plugging in all the correct values, the result is your coefficient value! If the value is a negative number, then there is a negative correlation of relationship strength, and if the value is a positive number, then there is a positive correlation of relationship strength.

Note: The above examples only show data for three people, but the ideal sample size to calculate a Pearson correlation coefficient should be more than ten people.

The Problem

The SkillAssure Online Book Store is an eCommerce web site, selling millions of books online. To stay in the competition and increase the cross sales, web site needs to add cutting edge artificial intelligent feature to help its potential customers to provide highly recommended books based on their previous activity on the website.

Sample Data

We have collected the sample data to model the recommendation engine.

Note: Download the sample data at

`http://www2.informatik.uni-freiburg.de/~cziegler/BX/`

Book Details

ISBN	Book Title	Book Author	Year-Of-Publication	Publisher
0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press
0002005018	Clara Callan	Richard Bruce Wright	2001	Harper Flamingo Canada
0060973129	Decision in Normandy	Carlo D'Este	1991	Harper Perennial

…	…			

Customer Details

User ID	Location	Age
1	New York	NULL
2	California	18
3	Victoria	25
…	…	…

Book Ratings Detail

User ID	ISBN	Rating (1 to 10)
1	0195153448	9
1	0002005018	6
2	0002005018	7
…		

Solution Required

To help individuals, identify other books they may be interested in, the goal of the case study is to build a comprehensive recommendation system for books based on user ratings. We also consider user's location and age to improve relevance of recommendation.

We take dataset consisting of quantifiable book rating values given by users who have provided their age and where they reside. We pre-calculate pair wise books similarity for different age groups and location.

A user is asked to choose one of the favourite book from our dataset and based on the user's location and age, our recommendation system recommends other books from our dataset that matches user's interest based on ratings based similarity.

Recruitment

For Example:

When a customer selects a book of ISBN 0195153448, show the top 10 recommended book list for cross-selling using Pearson Correlation Coefficient AI algorithm.

Solution Approach

Guided by Mentor