Melissa LaHoud

Grand Canyon University

DSC 540: Machine Learning

Dr. Aiman Darwiche

5/5/2021

K-means Algorithm

I choose a Loan Dataset to use for clustering from the Kaggle website which I thought would be suited for unsupervised cluster detection.

There are three questions that could be important taken from this data especially if you are trying to receive a loan.

1. For the income I make, how much of a loan can I receive?
2. Does my education effect how much of a loan I can receive?
3. Does having a co-signer allow me to get a larger loan?

These are beneficial for someone that is looking for a loan. If you want a larger loan then we need to know what factors contribute to the amount of the loan.

Let's look at question 1: For the income I make, how much of a loan can I receive? We can see the results I came up with in python using Kmeans cluster analysis. K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided.

First, we can see what the data looks like between Income and Loan amount in Figure 1.

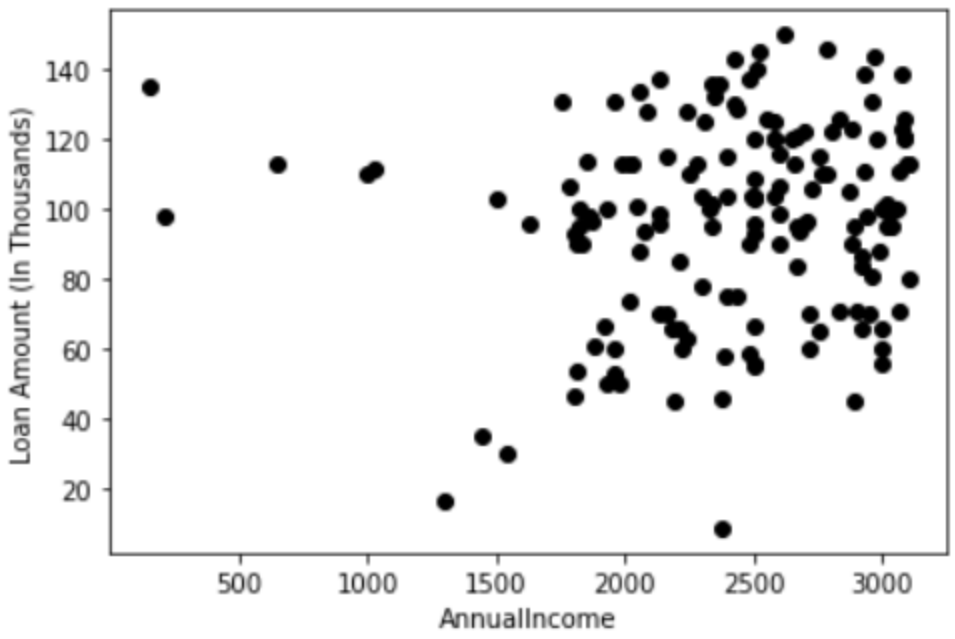

Figure 1

It is what we would expect. As your income increases, the loan amount you are able to get increases as well. I then implemented k means and we can see in Table 1, these are the three cluster details.

	LoanAmount	ApplicantIncome
count	41.000000	41.000000
mean	120.926829	6519.414634
std	25.683059	1085.138654
min	26.000000	5316.000000
25%	115.000000	5746.000000
50%	128.000000	6216.000000
75%	137.000000	7100.000000
max	150.000000	9703.000000

	LoanAmount	ApplicantIncome
count	175.000000	175.000000
mean	96.782857	2493.868571
std	29.012854	564.980510
min	9.000000	150.000000
25%	74.000000	2166.000000
50%	100.000000	2526.000000
75%	120.000000	2952.000000
max	150.000000	3254.000000

	LoanAmount	ApplicantIncome
count	149.000000	149.000000
mean	110.181208	4023.885906
std	26.163481	527.980325
min	25.000000	3273.000000
25%	100.000000	3593.000000
50%	115.000000	3902.000000
75%	129.000000	4408.000000
max	150.000000	5250.000000

Table 1

We can also see it graphically in Figure 2 and Figure 3. Figure 3 gives you a better understanding of the groups and the centroids within each group. These clusters will allow you to look at your income and see the range of loan you could possibly get.

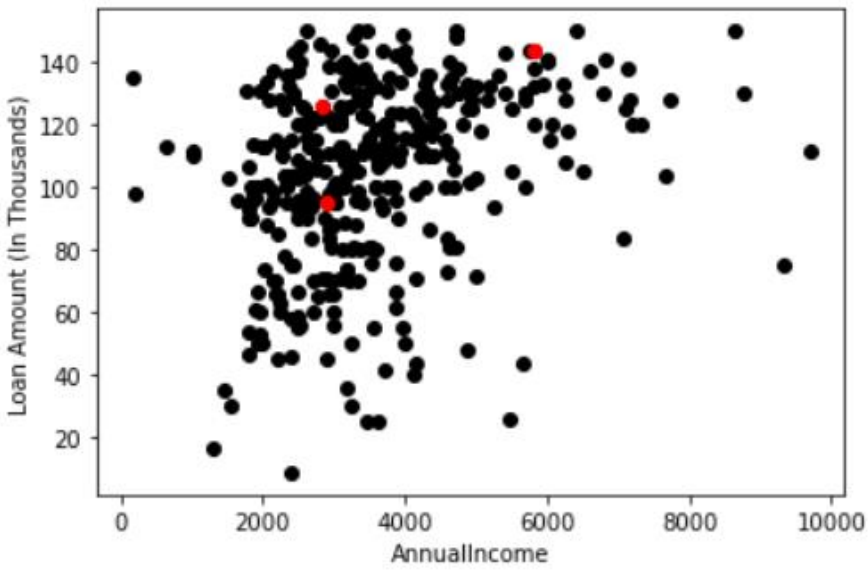

Figure 2

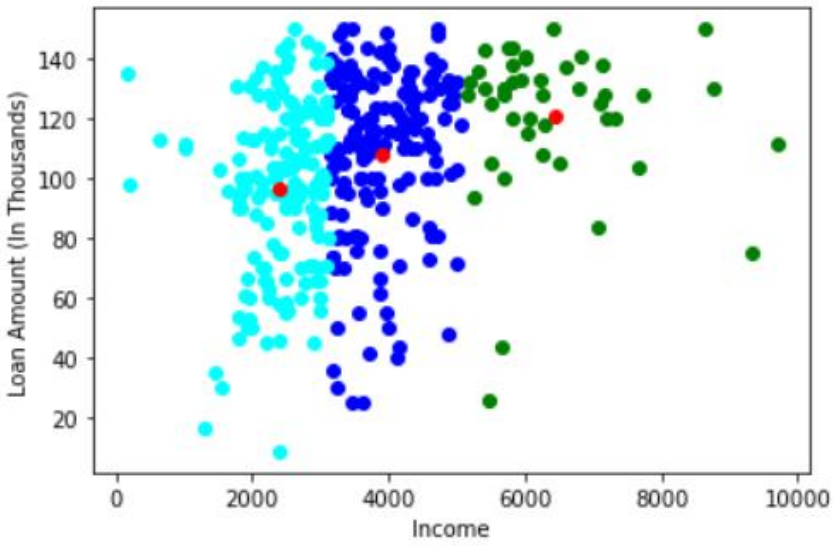

Figure 3

Therefore, using the k means clustering, I would conclude, the more income you make, the more likely you are to receive a large loan amount.

For question two, “Does my education effect how much of a loan I can receive?”, I would conclude yes. Although your education level doesn't actually determine your career opportunities or your level of pay, it is more likely that your career and financial success will increase as your education increases. Therefore, there would be two clusters, “graduated” and “Not graduated” with the Loan amount being higher for the Graduated cluster.

For question three, “Does having a co-signer allow me to get a larger loan?”, I would conclude yes also. This is going to be the same for question one because if you have a co-signer, there income is added to the application. This means there will be more income when you qualify, meaning your more likely to receive are larger loan because there is two incomes instead of one that will be able to pay off the loan.

Ethical Aspects of Data Management

“Data privacy (or information privacy or data protection) is about access, use and collection of data, and the data subject’s legal right to the data. This refers to:

- Freedom from unauthorized access to private data
- Inappropriate use of data
- Accuracy and completeness when collecting data about a person or persons (corporations included) by technology
- Availability of data content, and the data subject’s legal right to access; ownership
- The rights to inspect, update or correct these data” (Wanbil, L., 2016)

Protecting data privacy is necessary because of the ubiquity of the technology-driven and information-intensive environment. Technology-driven and information-intensive business operations are typical in contemporary corporations. The benefits of this trend are that, among other things, the marketplace is more transparent, consumers are better informed and trade practices are more fair. The downsides include socio-techno risk, which originates with technology and human users (e.g., identity theft, information warfare, phishing scams, cyberterrorism, extortion), and the creation of more opportunities for organized and sophisticated cybercriminals to exploit. This risk results in information protection being propelled to the top of the corporate management agenda.

"An explanation of the DPPs is provided by the Hong Kong,11 Office of the Privacy Commissioner for Personal Data, and can be summarized as:

1. Data Collection and Purpose Principle:

   Personal data must be collected in a lawful and fair way for a purpose directly related to a function/activity of the data user (i.e., those who collect personal data).

   Data subjects (i.e., individuals from whom personal data are collected) must be notified of the purpose and the classes of persons to whom the data may be transferred.

   Data collected should be necessary, but not excessive.

2. Accuracy and Retention Principle—Personal data must be accurate and should not be kept for a period longer than is necessary to fulfill the purpose for which they are used.

3. Data Use Principle—Personal data must be used for the purpose for which the data are collected or for a directly related purpose, unless voluntary and explicit consent with a new purpose is obtained from the data subject.
4. Data Security Principle—A data user needs to take reasonably practical steps to safeguard personal data from unauthorized or accidental access, processing, erasure, loss or use, while taking into account the harm that would affect the individual should there be a breach.
5. Openness Principle—A data user must make personal data policies and practices known to the public regarding the types of personal data it holds and how the data are used.
6. Data Access and Correction Principle—Data subjects must be given access to their personal data and allowed to make corrections if the data are inaccurate." (Wanbil, L., 2016)

References

Kaggle. (n.d.). Your Machine Learning and Data Science Community. Retrieved from https://www.kaggle.com

Pulkit, S. S. (2020, October 18). K Means Clustering: K Means Clustering Algorithm in Python. Retrieved from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Wanbil, L. W. (2016, December). An Ethical Approach to Data Privacy Protection. Retrieved from https://www.isaca.org/resources/isaca-journal/issues/2016/volume-6/an-ethical-approach-to-data-privacy-protection