Amazon Co-Purchasing Network Link Prediction

Submitted by: Jainabou Barry Danfa

SI 608 Networks

Final Project Report

Fall 2019

Abstract

This paper focuses on the methods of link prediction in a Amazon Co-purchasing Network. Using product data such as, product category similarity and network attributes such as Jaccard coefficient, we were able to test various machine learning model's ability to predict if products are co-purchased together or not. We found that most models that were tested had an average accuracy score of around 85% of a link existing between two products, with Gradient Boosting Classifier performing the best of all the models tested. The most important features in predicting a link was the category similarity of the two products and the Resource Allocation Index. Further analysis can add more attributes to improve accuracy score and use on different networks.

Keywords: Amazon, link prediction, machine learning

Amazon Co-Purchasing Network Link Prediction

Research Question and Motivation

This project focused on the analyzing the Amazon co-purchasing network of books, CD's, DVD's, and VHS tapes. Amazon is one of the world's biggest online marketplaces and data from this network can give us great insight into co-purchasing networks. Using the attributes of the nodes and the co-purchases, we can recommend products based on that history. This area of research is of great importance to companies since they aim to increase revenue and targeted recommendations of co-purchased products can increase the average spend of customers. Suggesting the "right" product to buy is challenging, especially in large networks like Amazon, due to the volume of similar products. This environment provides the motivation for this project.

Related Work

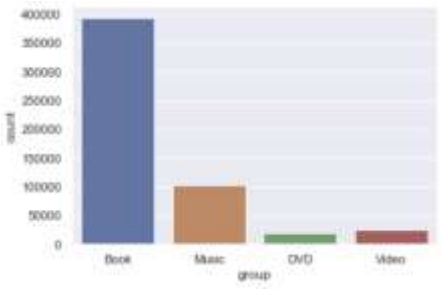
This project is related to link prediction. Link prediction uses a network structure to predict the most likely links to form in the network, ie. the next product to be purchased with that product. Other works have highlighted the different similarity measures and link prediction functions that can be used to predict links in a network. Many previous works have shown the increased predictive power of linkage through network topology. In the paper by Dr. David and Dr. Jon [2], we see how measures to predict node proximity outperformed more direct measures in predicting the existence of an edge.

Data

This project uses the Stanford Network Analysis Project Datasets for Amazon metadata from summer 2006[1]. The meta-data contains 542,663 products with the following information for each product:

Label	Information
Id	*Numerical id for each product in dataset*
ASIN	*Amazon unique identifier for each item*
Title	*Item name*
SalesRank	*Amazon number that captures the items popularity in each category. Amazon algorithm calculates this number based on several factors. The most important factor being the period between items being sold.*
Group	*The overall group the item belongs to*
Categories	*The different tags for the item*
Rating	*The overall rating for the item based on user responses*
Reviews	*The number of written reviews for the item*
Copurchased	*The list of other items "frequently" with the item*

The data was contained in a text file. After downloading the meta-data, significant pre-processing was needed to get the data into the proper format for analysis. In figure 1, we see that the distribution of products is mostly books, with fewer items belonging to music, DVD's and Videos.



Methods

Data Processing

After parsing through the text file, a pandas dataframe was used to contain all the data for each node. Then a separate dataframe was created for the co-purchasing network. This was developed from the asin id of each node and co-purchasing products for each of the nodes.

Network Development

NetworkX was used to create a undirected graph of co-purchases. This graph edges were weighted with the category similarity of the two nodes. Node attributes of the Group, SalesRank, and Average Rating were also represented in the undirected graph. Our network contained 366,997 nodes and 987,942 edges. Any co-purchased product from the metadata that did not have metadata associated with it was not included in the graph, hence the reduction in products represented in our network graph.

Model Development

Since we were developing the model for link prediction, our current graph only contained edges that represent a co-purchase between two products. In order to train the model with a balanced representation of co-purchases and non-co-purchases, a random list 500,000 of non-co-purchases was generated. These pairs were checked to not be have been included in our existing set already and that they were not the same product. Both the co-purchases and non-co-purchases were stored in separate dataframes. The following measures were added to each of the dataframes before they were combined:

Category Similarity

- *This was computed for each pair of items based on the number of words that intersect between the two category descriptions over the union of all words in the category descriptions. This was used as the weight of the co-purchasing graph. The similarity was calculated for the non-co-purchased products as well.*

SalesRank

- *This value was provided by in the metadata. It is an indicator of the popularity of the item in the amazon purchasing network. This metric is calculated by Amazon. For the network, the sales rank of the two items were added together for both co-purchased and non-co-purchased pairs.*

Jaccard Coefficient

- *This is a measure of the neighborhood similarity of two nodes. It is an indicator of the relationships that may exist because they share the same neighborhoods. This metric was calculated for the co-purchasing pairs and assumed zero for non-co-purchased pairs.*

Resource Allocation Index

- *This is a measure of common neighbors the two nodes have and the fraction of co-purchases that can occur using the common neighbors. This metric was calculated for the co-purchasing pairs and assumed zero for non-co-purchased pairs, since they cannot have common neighbors if no edge exist between the two nodes.*

Preferential Attachment

- *This is a measure of the degree correlation of the two nodes. It assumes that the a connection will be more likely given both nodes have higher degrees. This was calculated for both co-purchased and non-co-purchased pairs.*

Once the metrics were calculated for each dataframe. The two were combined into one. A 80/20 stratified test-train split was performed on the set. The stratified split was used to ensure that the classes of link and no link were equally represented in the split.

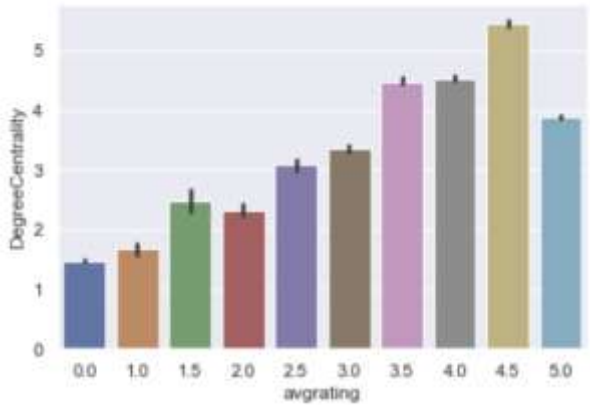
Three machine learning algorithms were evaluated on our data, Logistic Regression, Gradient Boosting Classifier, and Random Forest. These three models all have their strengths and weaknesses and were compared using accuracy and Area under the ROC curve. These two measures let us see how accurate the model is and how well it deals with positive and negative values.

The baseline selected for this project will be logistic regression using only SalesRank. This is a good indicator of how the model performs when we only want to test a connection based on the amount sold. Adding the other features will give us the actual improvement of prediction based on those features.

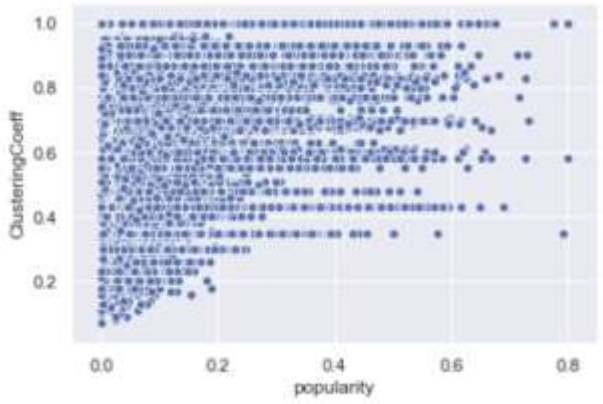
Results

Co-purchasing Network Analysis

The structure of the undirected graph has nodes that are the ASIN number, edge being the existence of a co-purchase, and the weight being category similarity of the two nodes. The density of the network is 0.0000146, indicating that the network does not have many ties between items and is not that dense. Looking at the relationship between degree centrality and average rating, we can see the higher the average rating, the higher the degree centrality. This is indicative of how having a higher rating does not necessarily mean the item is purchased



with other products. The local clustering coefficient was also calculated for each node, which provides us with an indication of cliques in the network.



As we can see, most nodes with a high clustering coefficient can have a range of popularity but nodes with low clustering coefficients also have low popularity.

Link Prediction Analysis

When performing the link prediction on the various machine learning models, we found adding the network features mentioned in the methods.

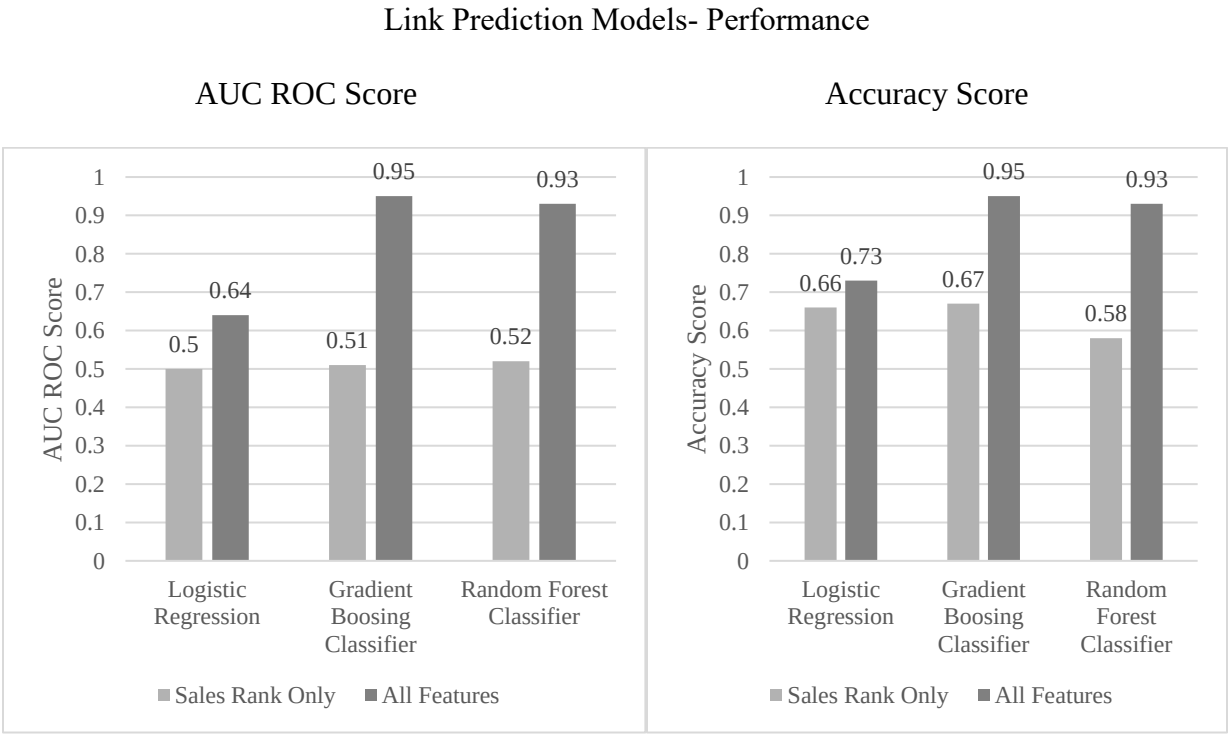


Figure 1. These graphs highlight the improvement of both the AUC ROC and accuracy scores over the three models when the network features were added to the model.

As we can see in the graph above our baseline of Logistic Regression with only SalesRank as a predictive feature was beat significantly when the other features were added. Gradient boosting provided the highest accuracy and ROC score after adding in network features. When determining which features provided the highest importance when predicting a link. Category similarity and resource allocation index provided the highest predictive importance in our model.

Feature	Importance
Category Similarity	0.378
Resource Allocation	0.247
Jaccard Coefficient	0.161
SalesRank	0.147
Preferential Attachment	0.066

This gives us insight into the co-purchasing network, highlighting that common neighbors and common categories are the most indicative features of two products being co-purchased together. SalesRank and Preferential attachment had the lowest importance, indicating that a nodes degree or popularity is not a big indicator of a link between two nodes.

Challenges

In this project, several challenges presented themselves, especially in data cleaning and preparation. Due to the format of the Amazon metadata, ample time was spent placing the data into the correct format in a memory-efficient was since the dataset was very large.

Due to co-purchasing network only representing items that have been purchased together, another function had to produce the nodes that were not purchased together and represent them in the train and test set. I was not expecting this and it caused some delay.

Conclusion

In this project, we saw the predictive power that network attributes can add to a machine learning model. When using the 4 network attributes; Category Similarity, Jaccard Coefficient, Preferential Attachment, and Resource Allocation, we found that most models that were tested had an average accuracy score of around 85% of a link existing between two products. The most important features in predicting a link was the category similarity of the two products and the Resource Allocation Index. Further analysis can add more attributes to improve accuracy score and use on different networks.

References

[1] : https://snap.stanford.edu/data/amazon-meta.html

[2] https://www.cs.cornell.edu/home/kleinber/link-pred.pdf