Recommendation for Online Retail Data

Sumit Patel

Springboard Capstone #2

Abstract

Recommendation Systems are widely used to recommend products and services to customers and clients. This system allows for the prediction of the rating or preference a user would give to an item. The dataset is for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. The dataset is comprised of 8 attributes consisting of InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID and Country. The dataset is composed of 532,618 rows. Through exploratory data analysis and using graphlab we will identify how to make a suitable recommender systems for the dataset. Appropriately matching customers to items that they may be more inclined to purchase would increase the likelihood of another purchase using recommender systems approach.

UCI Machine Learning Repository Dataset can be found at:

https://archive.ics.uci.edu/ml/datasets/Online+Retail

Recommendation for Online Retail

Recommender systems provide recommendations through collaborative filtering, content-based filtering or hybrid approaches. Collaborative filtering approaches build a model from a user's past behavior as well as similar decisions made by other users which is then used to predict items that the user may also have interests in. Content-based filtering approaches utilize use a series of discrete characteristics of an item in order to recommend additional items with similar properties. These approaches are often combined in hybrid systems. We will train and test various recommender systems to find one that is most suitable for this dataset application.

Method

Data

Dataset is a UCI Learning Repository dataset that consists of 532,618 rows of online retail data and 8 columns consisting of attributes InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID and Country. The dataset can be found at: https://archive.ics.uci.edu/ml/datasets/Online+Retail. Our recommender system will be based on the features of CustomerID, StockCode, Description and Quantity to provide our recommendations.

Extract, Transform and Load (ETL)

The dataset exists in xslx format which we converted to csv for easier processing. After converting the columns to the necessary data types we created and SFrame. When converting to SFrame some of the columns required data type manipulation once again. We also concatenated StockCode and Description to a column named Items for simplicity. We formatted the SFrame to only have our columns of interest: CustomerID, Items and Quantity. We then restructured the

SFrame to show unique instances of CustomerID so that there are no duplicitous results that would create errors in the SFrame. This action resulted in reducing the SFrame to 4,340 rows from 532,618 rows of data. Since multiple customers are able to to order similar items we wanted to identify the total number of unique items which happens to be 1,616.

Exploratory Data Analysis (EDA)

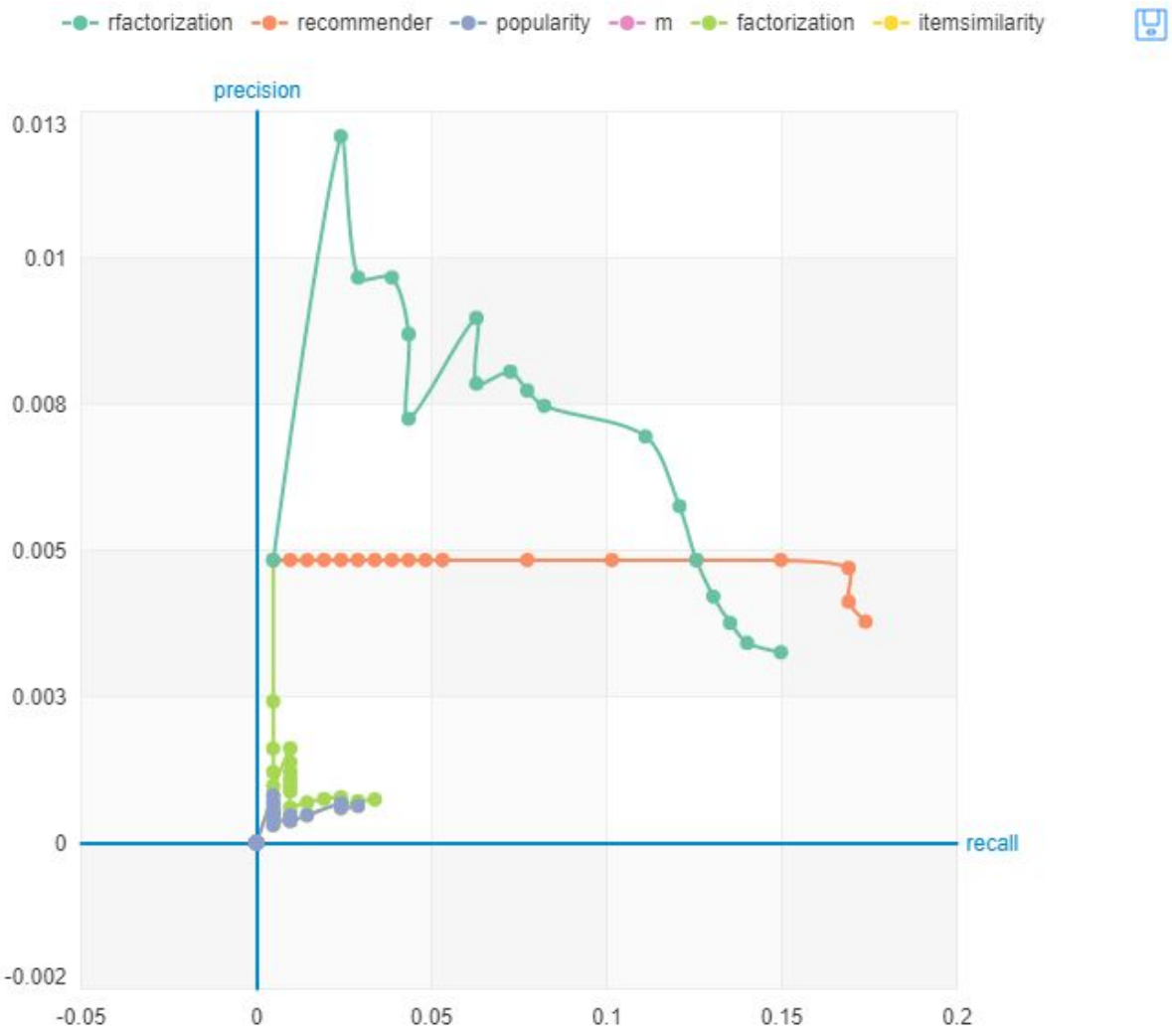
Using protocols and functions available in graphlab we were able to apply recommender create to our SFrame to find out which recommender system model would be suggested, the result was ranking factorization model. This is a collaborative recommender that learns latent factors for each user and item and uses them to rank recommended items according to the likelihood of observing those (user,item) pairs. This result is what we desire as collaborative filtering would be the ideal recommender system for our dataset. We will now test the model performance against other recommender systems and evaluating them with root mean-square error (RMSE) metric. RMSE is the measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. We will be comparing factorization, ranking factorization, popularity and item similarity recommender systems for our observations.

Figure #1 RMSE of Models on Original Dataset

Recommender System	RMSE Overall
Recommender	1087.9851481898477
Factorization	1047.0362041393714
Ranking Factorization	1151.9368663756113
Popularity	980.1790947421174
Item Similarity	1151.9472778226755

This table shows the RMSE overall values of the models using the original dataset. We can see that these are not ideal values and that from this observation the popularity recommender system would bet the most ideal model given that is has the lowest RMSE overall value. We will try normalizing the data to see if this improves, as it stands this is not an ideal result.

Figure #2 Performance of Models on Original Dataset



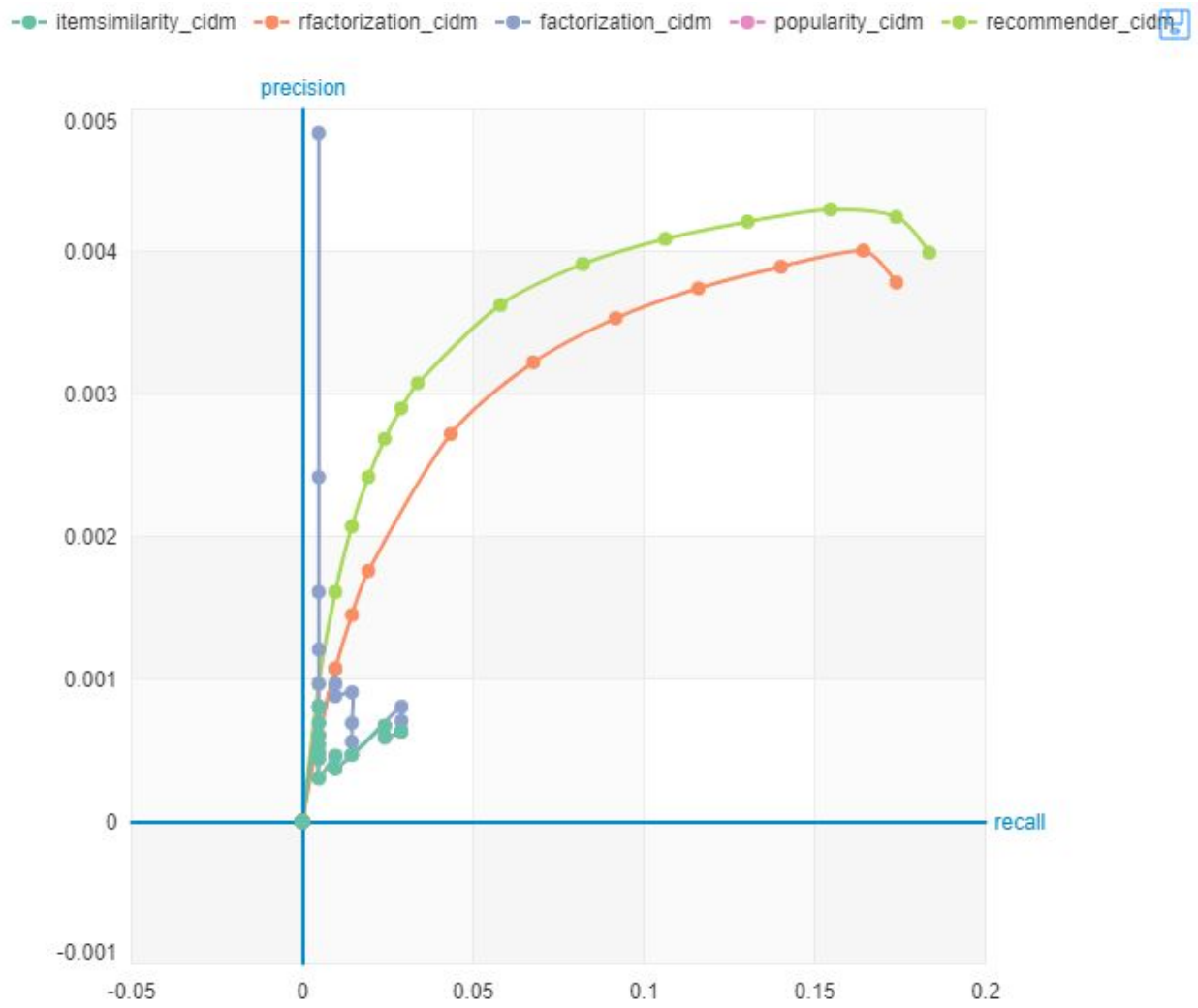
From the above figure we can see that ranking factorization is the most ideal model. We will try observing these results with normalized data to see if there is an observable difference in performance. This does not support our RMSE observations which is understandable given that the performance declines in precision. This gives us further incentive to normalize our data.

Figure #3 RMSE of Models on Normalized Data using mean

Recommender System	RMSE Overall
Recommender	432.71770582163055
Factorization	128.92590232348368
Ranking Factorization	133.84279408718817
Popularity	980.1790947421174
Item Similarity	1151.9472778226755

This table shows the RMSE overall values of the models using the normalized data with mean as the target variable instead of quantity. Using the mean of the quantities we are able to scale the quantities observations. We will observe this data using standard deviation and then using the coefficient of variance to see which provides the most optimum results.

Figure #4 Performance of Models on Normalized Data using mean



The performance of the models greatly improved by scaling the quantity using mean. Recommender suggested that ranking factorization was the recommended model and we can see that that is indeed the case as the next best performing model is ranking factorization after recommender. This is a vast improvement from figure 2 observations.

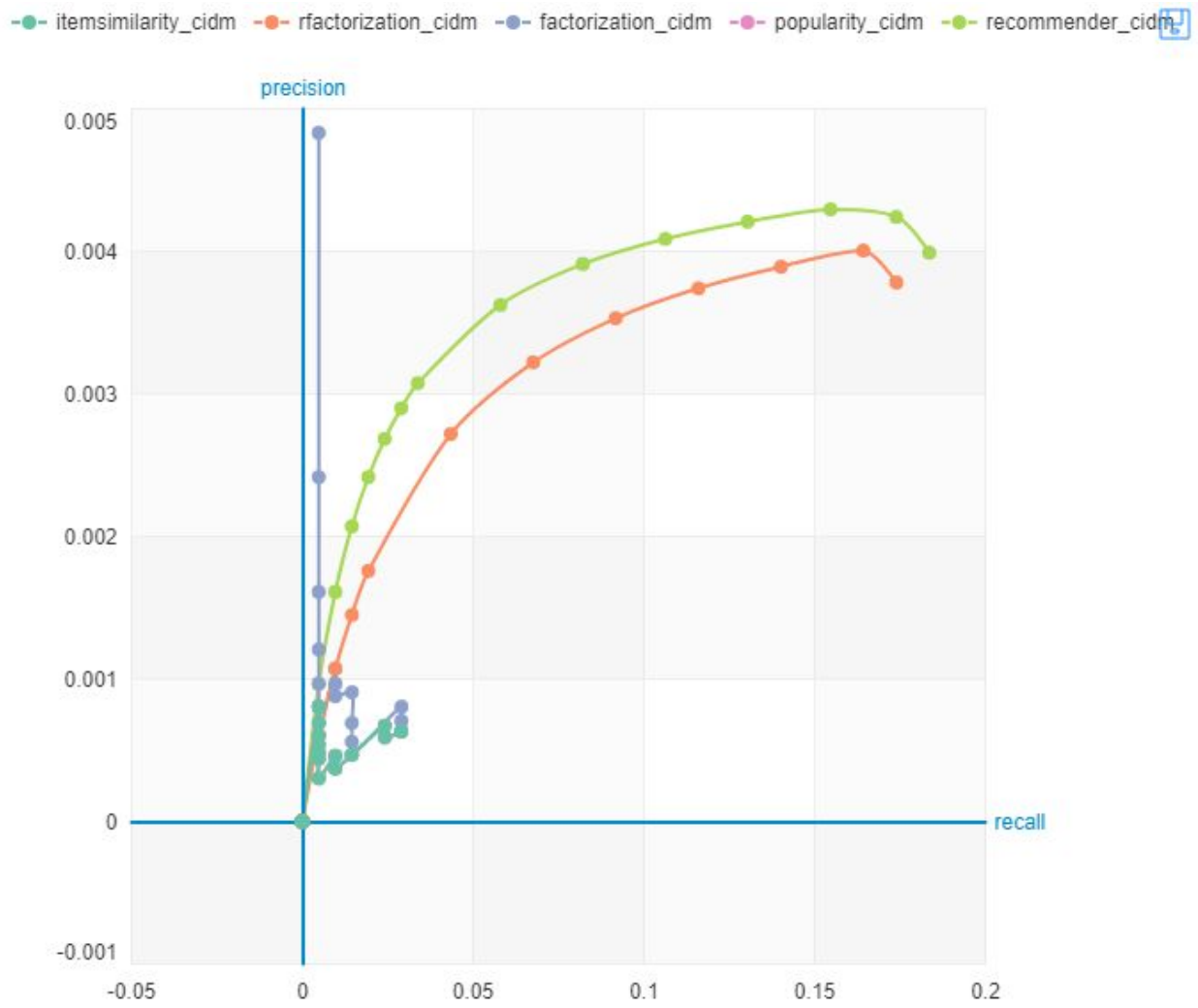
Figure #5 RMSE of Models on Normalized Data using standard deviation

Recommender System	**RMSE Overall**
Recommender	0.10022084767375772

Factorization	0.0002243989586135821
Ranking Factorization	0.006480047999389945
Popularity	0.0
Item Similarity	0.0

Since many of our values are equal to the mean we have instances of standard deviation equalling zero. We can see using the standard deviations that our RMSE vastly improved and that Factorization is the most ideal. We do see instances of zero values generated for popularity and item similarity RMSE values which is a possible error. We will try our observations with coefficient of variance to see if this is remedied.

Figure #6 Performance of Models on Normalized Data using standard deviation



The performance of the models using standard deviation is identical to figure 4 when we used mean to scale. Contrary to the results of the RMSE values ranking factorization has better overall performance. Factorization does yield the highest precision but terrible recall performance. We will now make observations using the coefficient of variance.

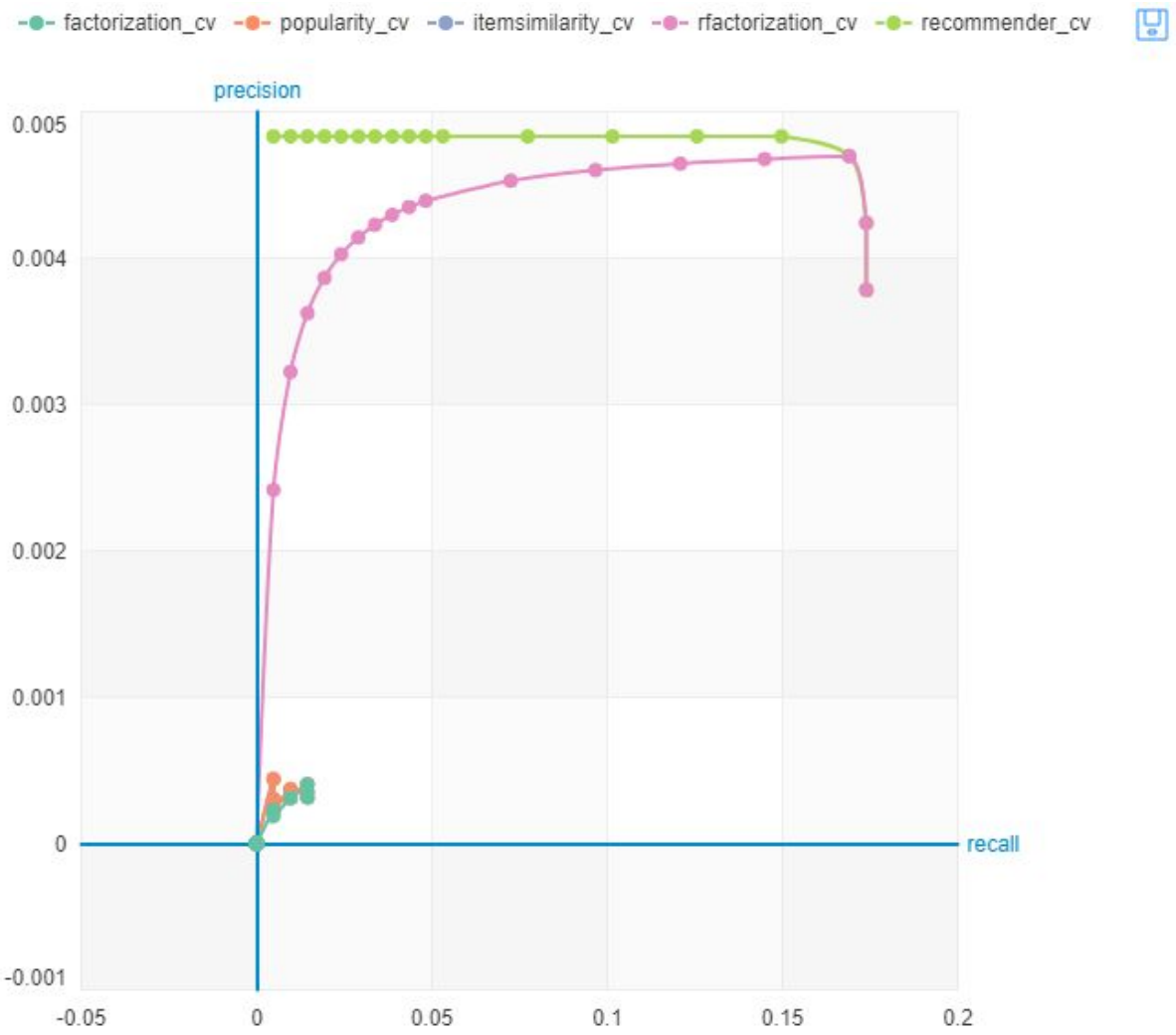
Figure #7 RMSE of Models on Normalized Data using coefficient of variance

Recommender System	RMSE Overall
Recommender	0.002729177844296901

Factorization	0.0001998868975889664
Ranking Factorization	0.16444060972566002
Popularity	0.0
Item Similarity	0.0

Using coefficient of variance to scale our data we see that the RMSE is improved on all models except for ranking factorization. Popularity and item similarity still maintain a zero output. Let's see what the performance of the models looks like with coefficient of variance.

Figure #8 Performance of Models on Normalized Data using standard deviation



We can clearly see that ranking factorization has been greatly improved using coefficient of variance as the target variable. The recommender model has the highest precision recall thus far and degrades ever so slightly at the end in both the recommender and ranking factorization models. Clearly ranking factorization is the ideal model choice given all of our tests.

Normalization of Data

Our observations from the original dataset showed that normalization of data may be required to scale our data to provide a normal distribution. Since multiple customers are able to

purchase the same item this would cause such errors. We will attempt to remedy this by using the mean and standard deviations to scale our data with.

Model Selection

After ETL and EDA we continued to scale our data for our observations. We created and implemented training and test sets to test our various models and their performance. The metrics used to measure model performance were RMSE. We then scaled our data to normalize our results and observations using mean, standard deviation and coefficient of variance. Our observations reflected improvement from original data observations which means that our data benefited from scaling for normalization. Ranking factorization was the recommender systems recommended by the graphlab recommender function. Factorization yielded the better RMSE score but on the performance plot for precision recall we saw that ranking factorization outperformed it vastly.

Conclusion

Evaluation of the models primarily utilized the RMSE metric and the graphical performance of the precision recall. Our results and calculations did provide that many of the mean values were identical to the quantity values which resulted in zeros for standard deviations. This did cause some unique instances on the graphical performance of the precision recall curves. The RMSE values observed in the model comparison clearly shows that ranking factorization is the best choice. Ranking factorization provided for the most ideal RMSE scores which reflected the most minimal error rate less than 1% which is very ideal. Overall RMSE values provided that scaling was necessary to reduce the error rate and that was well reflected in the results.

References

Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197â€“208, 2012 (Published online before print: 27 August 2012. doi: 10.1057/dbm.2012.17).

http://archive.ics.uci.edu/ml/datasets/online+retail