HBC RECOMMENDATION MODEL

BY: UTTASARGA SINGH

CUSTOMER PURCHASE PREDICTION

- This Model was developed using PySpark on a Windows Machine and the Aim for Model Development was to predict whether a consumer will re-purchase a product that they have already shopped before and provide a list of recommendation of various products to each customer.
- PySpark is a Python-Integrated Apache Spark API, which is used in large-scale, distributed Computing and Machine Learning.
- There are various libraries which are a part of this API and we are using ALS (Alternating Least Squares) to build the Model.
- Recommendation System is a type of Information Filtering tool, which helps to find out the products that a user would like and recommends the user on the basis of this prediction.

- This Recommendation System works based on Utility matrix; where there are users on X axis and Items on Y axis.
- There is a rating assigned to every User-Item relationship and the Highly rated Item for a User is recommended to them. This Rating is calculated using a Machine Learning Model and based on the rating present in the Data feeded into the Model.
- The 2 approaches used widely for this System are Content-based Filtering and Collaborative Filtering.
- These approaches are different from each other as CF uses past behavior of the Users while CBF uses similar contents shared between group of users and recommend them those Items.
- It aggregates on the Past behavior of all users present in the Data and select the similarities of products between two users and recommends item to each of these users.
- This User-to-User approach have 2 different ways, out of which I am trying a Model based Matrix Factorization in which both the User and Item profiles are learned by the model and the RMSE value is being minimized after each iterations of predicted values.

- Alternating Least Squares or ALS Matrix Factorization estimates rating matrix R as the product of two lower rank matrices; X and Y.
- In each Iteration, one of the Factor matrices are held constant while other one is solved using least squares. The newly solved matrix is then held constant while solving for the other matrix and hence the Model continues till the Iterations or Custom Hyperparameters feeded by the user.
- Explicit Feedback is a term which means that the User has self-assigned ratings to the Products according to his/her preferences.
- These ratings may not be present in every data that we encounter. Suppose we thought of refining our built recommendation model with the help of other data such as Views, Clicks, Purchases, Number of Shares and Likes etc. ; those numbers then play a major role in deciding user preferences rather than a default rating given to each product.

- Cold Start Strategy is a technique used widely while making predictions using a ALS Model.
- I have used Cross-Validation in order to split my training set and test set and these random splits increase the chances that a User is encountered in the Test Set which is not present in the training set and hence the Model is likely to fail to predict accurately.
- This Strategy helps in a way that it assigns a NAN prediction during the stage of Model building when a User/ Item is not present in the Model. Then, when the Model is in production and a new User is encountered; System makes the predictions based on the Historical Data of the Content which are being shared by the new user and any user before.
- We cannot control the User Surge, but we can have necessary predictions based on the Item or brand that a previous User is viewing; and recommend the New User these similar brands.
- Then, we can monitor the New User behavior and predict the other Items that he might be interested in; based on the approach of Implicit Feedback.

- I have built two separate Models based on the Dataset and I have predicted the recommended Items that a User Can purchase with the help of his past purchases of Products and the various brands, of which those products are a part of.
- The First Model “HBCALSModel2” is based on the explicit feedback, which are the frequencies of various products purchased by various Users while “HBCALSMODEL4” is based on the implicit feedback of the Quantities of Products purchased as well.
- RMSE Value for the 1st Model was ~0.75 while it was ~1.50 for the 2nd Model.
- The PySpark Notebooks are attached with the mail and you can access the same through my GitHub Link as well. Please feel free to check the below links for your reference:
- GitHub: https://github.com/uttasarga9067
- R-Shiny Application: https://uttasarga.shinyapps.io/BayesianHierarchicalModel/
- Medium