Data Science Portofolio

Yodi Ramadhani Alfariz

Education

dibimbing.id
Jun 2022 - Nov 2022
Non-formal education: Data Science Bootcamp

Universitas Komputer Indonesia
Jun 2013 - Aug 2017
Major : Computer Science - Informatics Engineering

Working Experience

Indah Logistik Cargo Ahmad Yani Bandung
Exp. Leader

Jan 2019 - Present

PDAM Tirta Raharja Kab. Bandung
Internship

Aug 2016 - Oct 2016

BPS Porvinsi Jawa Barat
Oct 2016 - Dec 2016

Input Correction / Data Entry

Oct 2016 - Dec 2016

Skills and Proficiency

- SQL Database
- Data Visualization
- Machine Learning
- Python Coding
- Dashboarding

Data Science Project

Customer Segmentation

Product Recommendation

Tableau Dashboarding

USA Sales Performance Data Studio Dashboarding

SQL in BigQuery

Online Retail Product Recommendation

Introduction

The following datasets from kaggle is Online Retail Transaction From 2009-2011, but I have cleaned before in my last portofolio Customer Segmentation, and I have split on 2011 only. The cleaned dateset you can get here.

Objective

We will build product reccomendation using Alternating Least Square to offering for Sales Improvement.

Dataset Information

371.722 record

19 columns

- InvoiceNo : number code of the transacation
- StockCode : the code of item
- Description : the description of item
- Quantity : count of item buying.
- InvoiceDate : date of customers transcation .
- UnitPrice : the price per item.
- CustomerID : ID of customers
- Country : country of customers live.
- TotalTrx : total transcation (quantity x UnitPricce)
- invDate : date without time
- year : year of transaction
- month : month of transaction
- day : day of transaction
- hour : hours of transaction
- cluster : cluster of customerID
- class : class of customerID (cluster description)
- encust : encoding customerID
- encprod : encoding StockCode
- qtylog : qty after log transformation

Methodolgy

Data Understanding

Features Engineering

Grouping By Agregate (encust, encprod, (sum qty), (sum totaltrx))

Data Transformation

Finding Rating with K-Means

Modelling

Split Data (Train, Validation, Test)

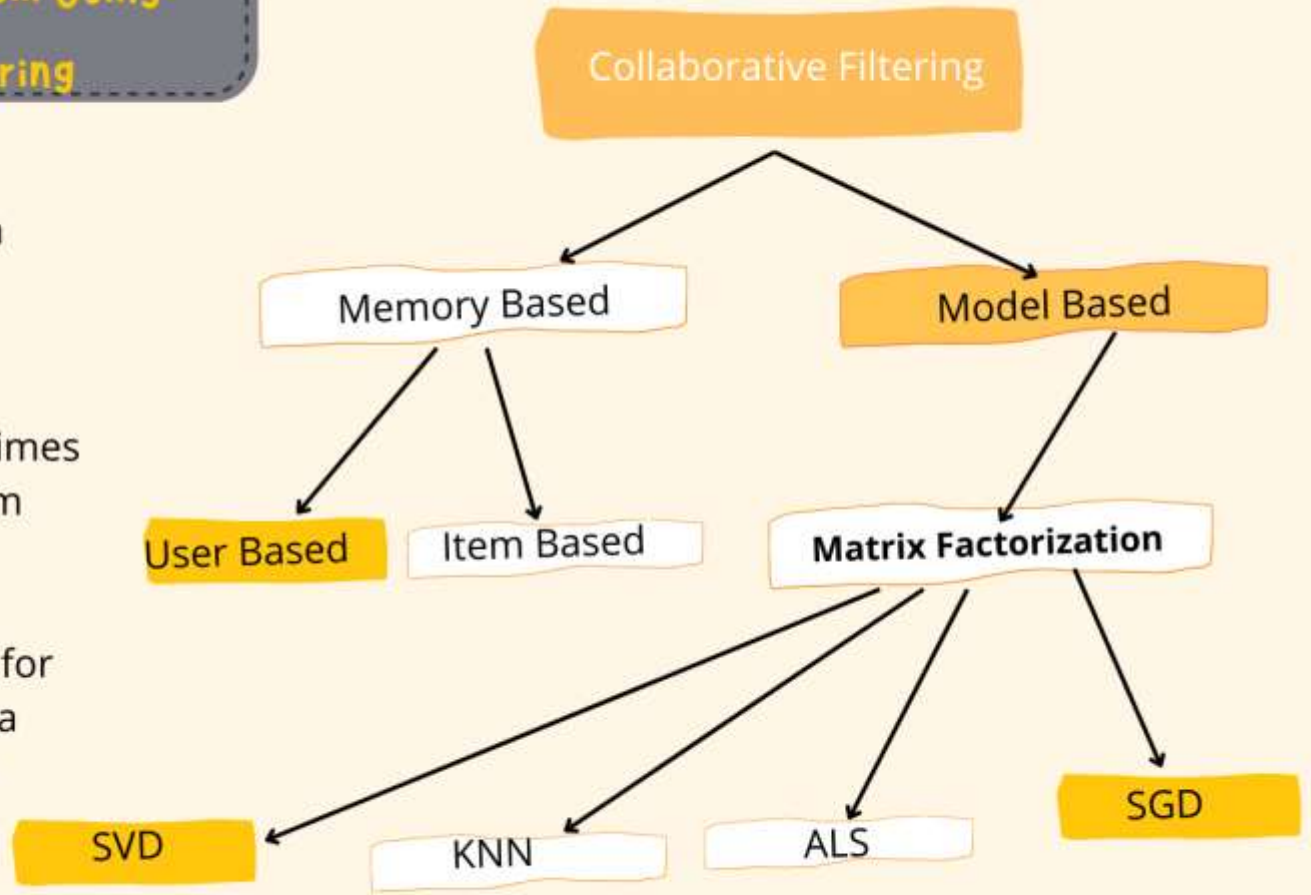
Tunning Hyperparameter

Modelling With Best Parameter

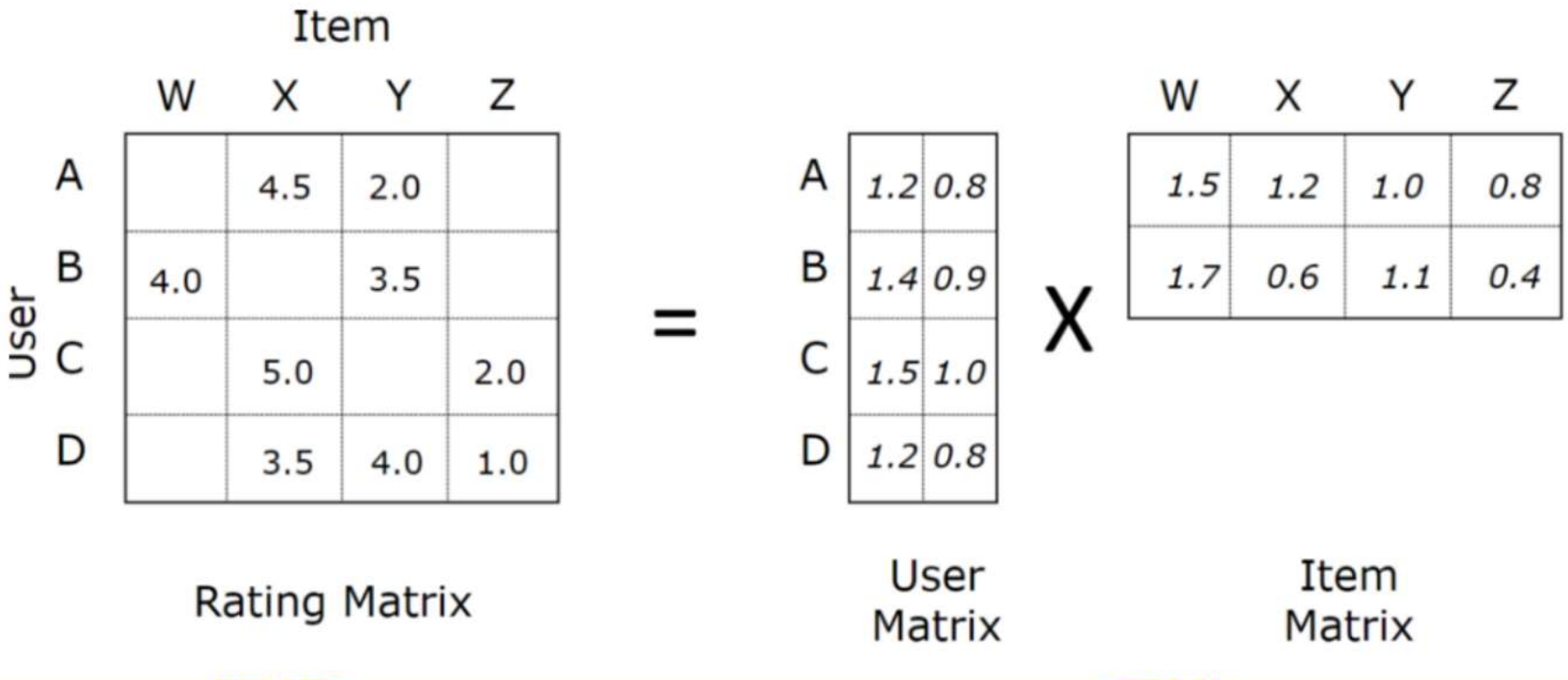
Product Recommendation

Recommendation System Using Collaborative Filtering

Before we going through to main topic, I will explain first about reccomendation system.
A recommender system, or a recommendation system (sometimes replacing 'system' with a synonym such as platform or engine), is a subclass of information filtering system that provide suggestions for items that are most pertinent to a particular user. source Wikipedia



Alternating Least Square



Source here

Alternating Least Square (ALS) is also a matrix factorization algorithm and it runs itself in a parallel fashion. ALS is implemented in Apache Spark ML and built for a larges-scale collaborative filtering problems. ALS is doing a pretty good job at solving scalability and sparseness of the Ratings data, and it's simple and scales well to very large datasets.

Feature Engineering

We need rating for Matriks Factorization

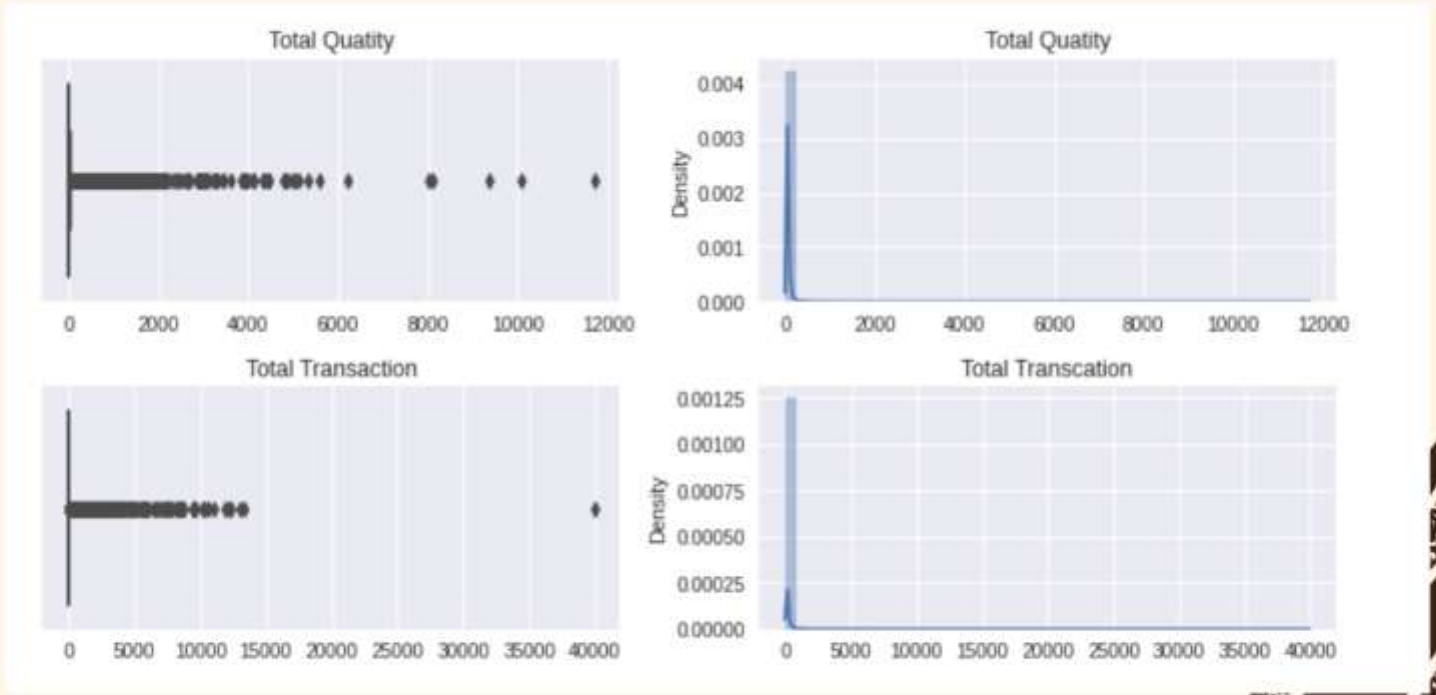
In our datasets didn't have rating, so we will using K-Means for Feature Engineering, and we will using Qty and Total Transaction to get Rating as feature.

encust	encprod	qty	grand
4005	1473	6	9.899999999999999
966	2155	5	4.25

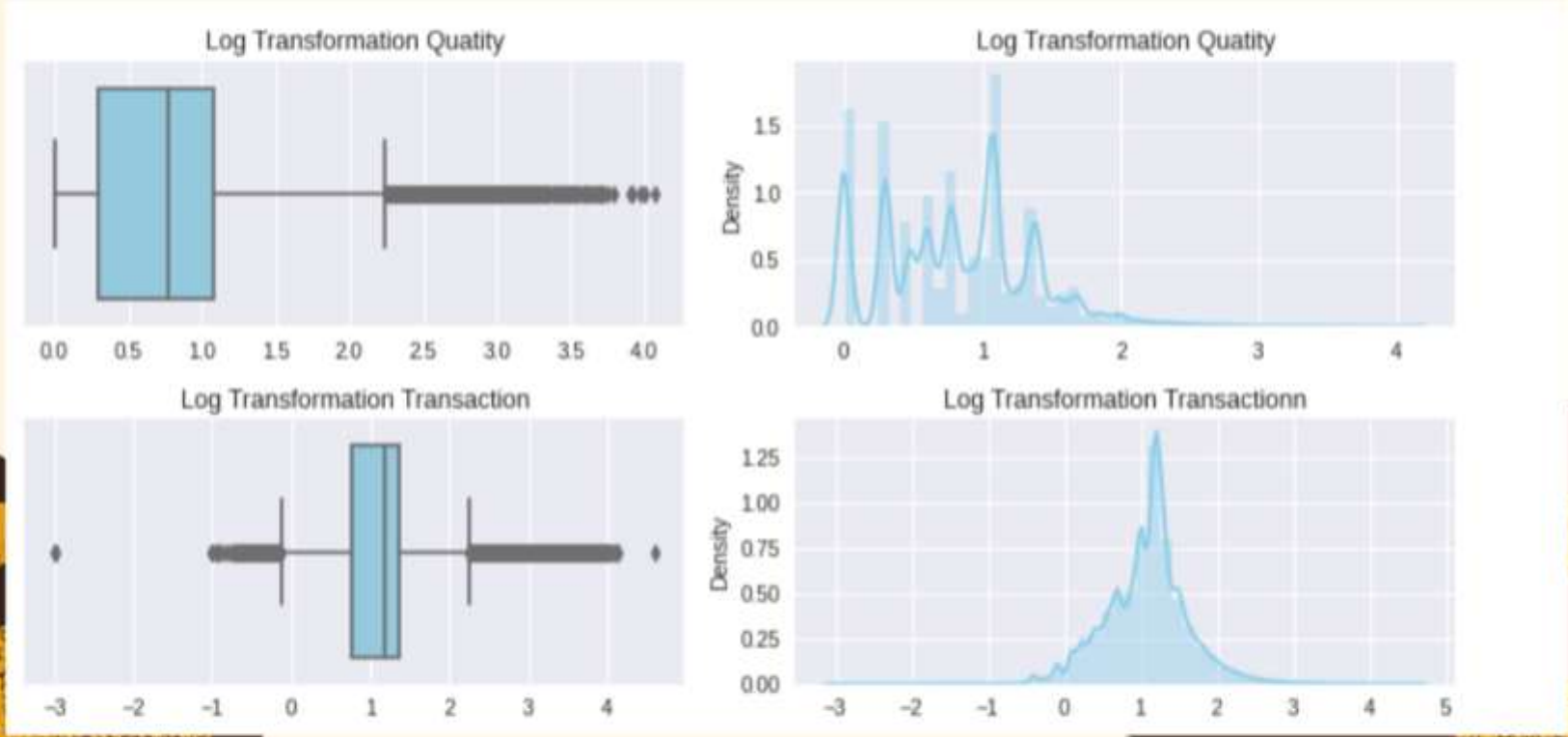
First of all we will grouping by customer, and product with aggregate sum of qty and sum of total transaction

But as you can see, the data didn't normal distribution

We have to normalized using log transformation

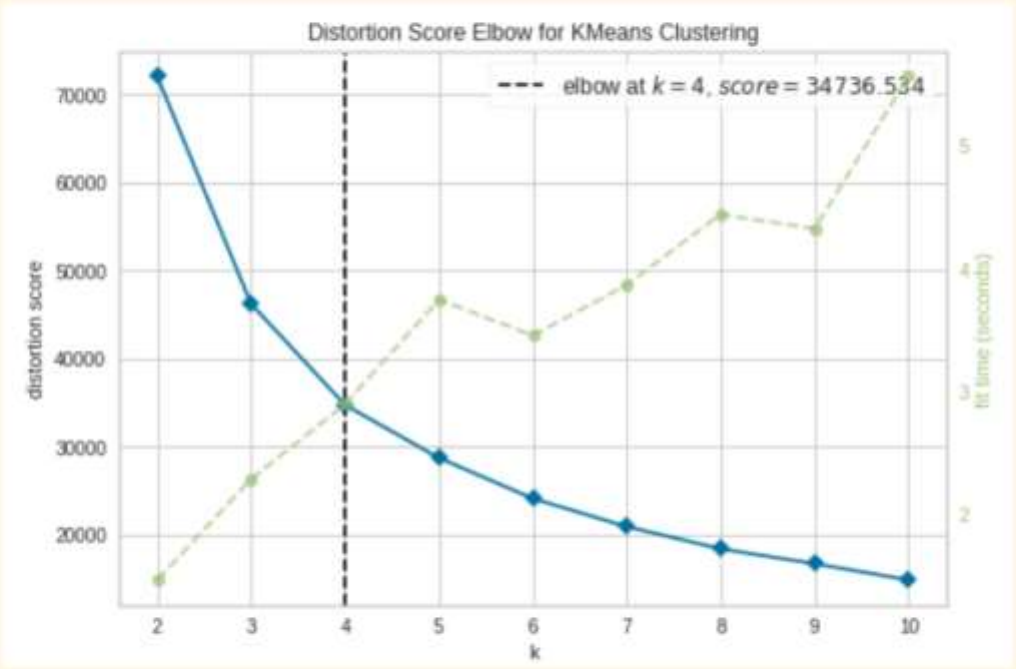


Data distribution after Transformation



It's better than before

we will find rating using k-means, and first we will find k potensial.



And we got 4 for k potensial, then we will fitting into dataset.

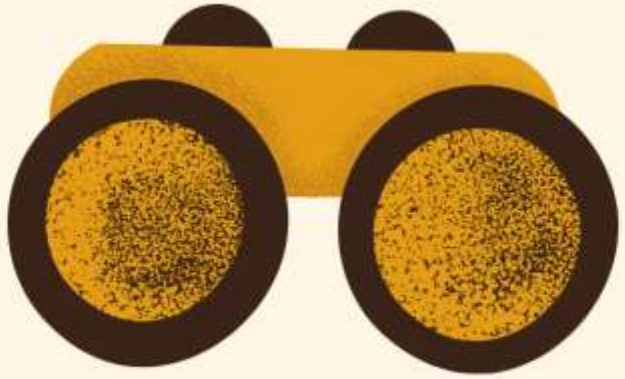
And we got rating like this

encust	encprod	logqty	logtrx	features	prediction
4005	1473	0.7781512503836436	0.9956351945975499	[0.77815125038364...	2
966	2155	0.6989700043360189	0.6283889300503115	[0.69897000433601...	2
966	677	1.255272505103306	1.5774917998372253	[1.25527250510330...	3

but we only need rating(prediction), encust, and encprod for features. We will getting column like this

encust	encprod	prediction
4005	1473	2
966	2155	2

only showing top 2 rows



We will split data into 3 segment, train, validation and test. with 60% train, 20% validation, and 20% test

Modellling

Try Hyperparameter Tunning to find Best Paramaeter

With train and validation data

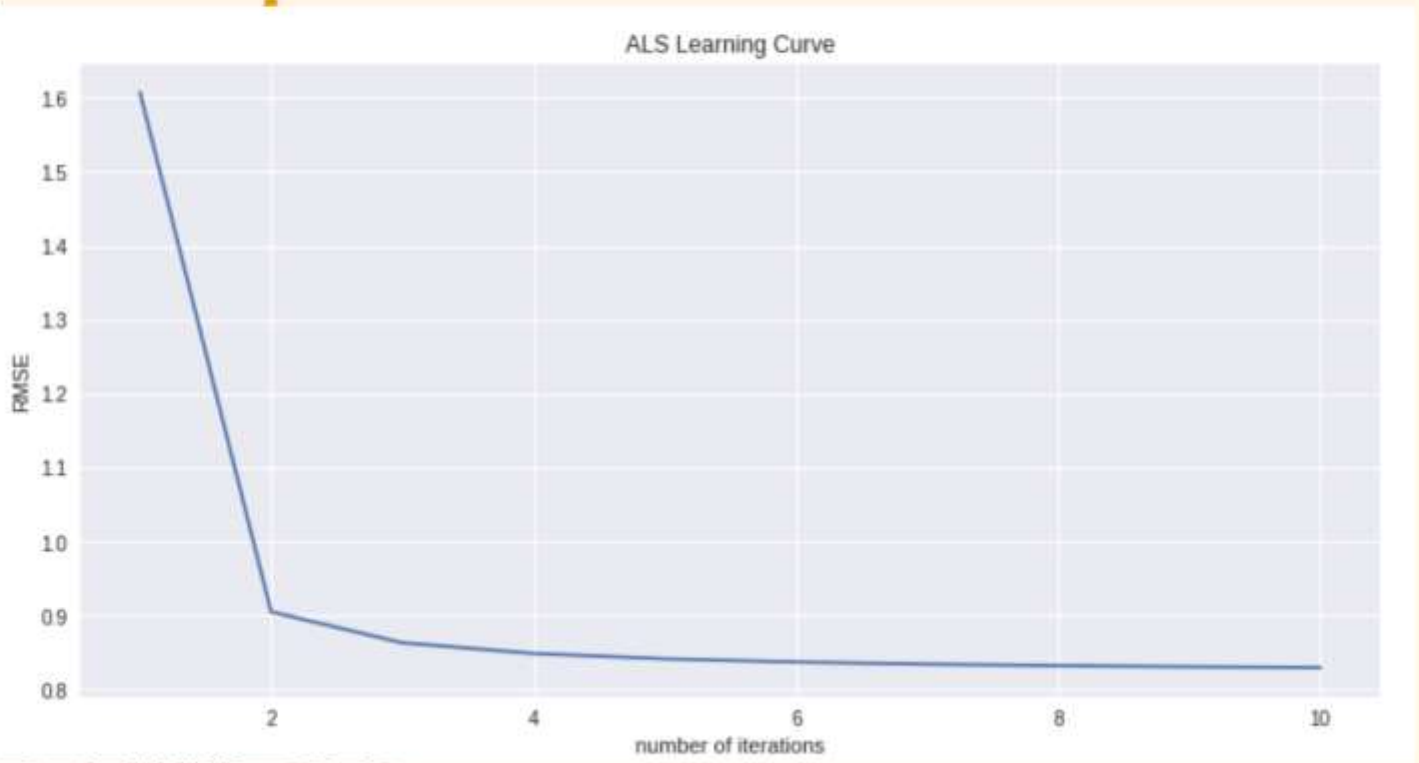
```
20 latent factors and regularization = 0.01: validation RMSE is 1.106612668230327
20 latent factors and regularization = 0.05: validation RMSE is 0.8884571735723852
20 latent factors and regularization = 0.1: validation RMSE is 0.8348978928649736
20 latent factors and regularization = 0.2: validation RMSE is 0.8357110913866176
50 latent factors and regularization = 0.01: validation RMSE is 1.0103233409673376
50 latent factors and regularization = 0.05: validation RMSE is 0.8601728532689794
50 latent factors and regularization = 0.1: validation RMSE is 0.829443568380284
50 latent factors and regularization = 0.2: validation RMSE is 0.8357865788921208

The best model has 50 latent factors and regularization = 0.1
Total Runtime: 389.21 seconds
```

We getting rank = 50 and lambda = 0.1 for the lowest rmse

We will see learning curve

After 10 itteration for learning we can see more better for rmse.



And we try with test data

```
The out-of-sample RMSE of rating predictions is 0.8289320259848187
```

Try to predict Product Recommendation

```
Recommendations for JUMBO BAG ALPHABET:
1: JUMBO BAG PINK VINTAGE PAISLEY
2: CRYSTAL STUD EARRINGS ASSORTED COL
3: JUMBO BAG SCANDINAVIAN PAISLEY
4: JUMBO STORAGE BAG SUKI
5: HAIRCLIPS FORTIES FABRIC ASSORTED
6: ANTIQUE GLASS PLACE SETTING
7: JUMBO BAG APPLES
8: JUMBO BAG OWLS
9: WHITE FRANGIPANI HAIR CLIP
10: JUMBO BAG PINK POLKADOT
```

Notebook can be accessed here

bit.ly/3Fa0fmV


Thank You

CONTACTS

Github : https://github.com/yodialfa/

LinkedIn : https://www.linkedin.com/in/yodialfariz/

Email : yodialfariz@gmail.com

Phone : 082218293933

Twitter : @yodiumh