

# Project Final Report

## RECOMMENDATION ENGINE FOR CAMERA RENTAL SERVICES (KLACHAK)

Group No. 7

Jayaprakash Nallathambi  
Abhishek Ramachandan  
Harish Ganesan  
Santhosh Murali

PGDBA-BI – Jan 2017

Mentored By  
Dr. Monica Mittal



# Project Report

## Recommendation Engine for Camera Rental Services (Klachak)

Submitted towards partial fulfillment of the criteria for award  
of PGPBABI

**Submitted By**

<b>Jayaprakash Nallathambi</b>	BACJAN17022
<b>Abhishek Ramachandran</b>	BACJAN17002
<b>Harish Ganesan</b>	BACJAN17020
<b>Santhosh Murali</b>	BACJAN17062

Course: **Post Graduate Program in Business Analytics and Business Intelligence [PGP-BABI]**

Mentor : **Dr.Monica Mittal**



East Coast Road,  
Manamai Village,  
Thirukazhukundram Taluk,  
Kancheepuram District,  
Manamai,  
Tamil Nadu 603102  
Phone: 044 3080 9000



## Acknowledgements

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them. We are highly indebted to Dr Monica Mittal for her guidance and constant supervision as well as for providing necessary information regarding the project & also for the support in completing the project. Her experience and support guided us to make the project development process look simple. She provided valuable inputs at every step of the project.

We would also like to express our gratitude towards Mr. Ganesh Siddhamalli, Co-Founder of Klachaks, Velachery, Chennai , for his offer and co-operation which helped us in completion of this project.

Last but not the least we wish to thank Dr. Prof. P. K. Viswanathan , our course Director, for constant supervision, guidance and for being a source of inspiration in helping us to work on this project.

Date : Nov 15, 2017

Place : Chennai

Jayaprakash Nallathambi

Abhishek Ramachandran

Harish Ganesan

Santhosh Murali

## Certificate of Completion

I hereby certify that the project titled “Recommendation Engine for Camera Rental Services (Klachak)” was undertaken and completed under my supervision by Jayaprakash Nallathambi, Abhishek Ramachandran, Harish Ganesan & Santhosh Murali all the four students of Postgraduate Program in Business Analytics and Business Intelligence (PGPBABIJAN2017).

Date : Nov 15, 2017

(Dr. Monica Mittal)

Place : Chennai

Mentor

## Table of Contents

Acknowledgements.....	4
Certificate of Completion.....	5
Table of Contents.....	6
List of Figures .....	8
List of Tables .....	8
Executive Summary.....	9
Introduction .....	10
Camera Rental Services Industry .....	10
Klachak.com .....	12
Equipment Rentals.....	12
Photography Services.....	13
Problem/Opportunity Statement .....	13
Analysis Methodology.....	14
Methodology.....	14
Data Source .....	15
Tools & Techniques.....	16
Limitations .....	16
Literature Review.....	18
Recommender Systems .....	18
Content Based Filtering:.....	18
Collaborative Filtering Algorithms: .....	18
Hybrid collaborative Filtering:.....	18
Exploratory Data Analysis .....	19
Sales Yearly Trend. ....	19
Seasonal Trend.....	20
Percentage of Revenue .....	21
Cohort Analysis .....	23
Model Development .....	25
Data Extraction and Data Preparation. ....	25
How RFM value is segmented.....	25

Mapping Cheat Sheet got Rating Derivation .....	26
Web Scrapping – Using Scrapy.....	26
Scrapy Data Flow.....	26
Model Building and Evaluation .....	28
Content Based Filtering.....	28
Collaborative Filtering.....	31
Why Hybrid Approach? .....	37
Implementation .....	38
Overall Architecture.....	38
Post Implementation .....	38
Recommendation Systems in Action .....	39
Conclusion.....	40
Future Scope .....	40
References .....	41



## List of Figures

Figure 1 .....	11
Figure 2 .....	14
Figure 3 .....	19
Figure 4 .....	20
Figure 5 .....	21
Figure 6 .....	22
Figure 7 .....	23
Figure 8 .....	24
Figure 9 .....	25
Figure 10 .....	26
Figure 11 .....	26
Figure 12 .....	27
Figure 13 .....	28
Figure 14 .....	30
Figure 15 .....	33
Figure 16 .....	34
Figure 17 .....	35
Figure 18 .....	35
Figure 19 .....	36
Figure 20 .....	36
Figure 21 .....	38

## List of Tables

1 .....	10
Table 2.....	15
Table 3.....	29
Table 4.....	30

## Executive Summary

Klachak.com is an online camera and accessories rental company, enabling photographers to rent cameras, accessories or photography services online. The company is positioned itself as a marketplace for amateur photographers to professionals across Tamil Nadu. Klachak is the first e-commerce website in Tamil Nadu operated in Hybrid model of Managed Market place and Inventory model for Camera Rentals. Initially company wanted to find out its state of market position and its some performance measures through exploratory data analysis, which later expanded its scope to build a Recommendation System for its website. This will help customer to analyze various other options before renting the camera, as well as for the company to up-rent and cross-rent of products.

The project takes on following sequence to complete, starting from understanding the business through several meetings with Klachak, then deep diving into transaction data of klachak. Through several exploratory analysis and visualization understanding the data well. Exploiting and preparing the data for Developing Recommendation Engine through Content and Collaborative Filtering based hybrid model. Evaluate and fine tune the model. Finally deploy the Program in Klachak website.

The duration of Data Extracted was from Feb-2014 to Oct-2017 for performing Exploratory data analysis. This gave enough space for understanding customer behavior and inventory performance. For building Recommendation engine same duration of data is used but with limited fields. The completed model is evaluated based on Root Mean Square Error between Actual and Predicted. Post implementation, data will be streaming near Realtime and model will become functional.

Once it is deployed in Production, it should improve Customer Satisfaction in addition to Customer Experience. As well as, Improved inventory performance as a auxiliary results.

## Introduction

### Camera Rental Services Industry

This is a budding industry with only very few active players. So, what is the business? Let's assume that Someone called "John doe" is a excellent Photographer and would not mind crossing any limits to get best click. He wanders around the globe to capture the perfect moments. Every photographer has their own gears which they used carry all around with them ever. {gear: Camera Body, Lenses, Filters, Tripod, Lights etc). John doe also has his own set of gears, he has Canon 5D Mark II Body, Canon 70-200 IS2 and Canon 16-35 L lens, Tripod and Travel bag. His total gear value is provided below.

<b>Canon 5D M2</b>	<b>₹ 70,000.00</b>
<b>Canon 70-200 IS2</b>	₹ 120,000.00
<b>Canon 16-35 L</b>	₹ 110,000.00
<b>Tripod</b>	₹ 27,000.00
<b>Bag</b>	₹ 3,000.00
<b>Total</b>	<b>₹ 330,000.00</b>

1

With his limited existing set of gears, he limits his creativity. With Canon 70-200 IS2, he could capture long range objects but still not a very extensive telescopic range and f-stops are not wide enough. On the other hand, he has 16-35mm wide angle lenses which will cover larger area and minimal depth. So, whatever photograph he takes it will either long distance and not too long with moderate depth information or a wide angle. Suppose if he wants to take a portrait his gear pack will not help him. He has to have either "Canon EF 35mm f/1.4L II USM" or "Canon EF 50mm F/1.2L USM". It will cost him additional ₹ 200,000 or ₹ 100,000 respectively. If he wants to go for macro photography, he has to invest ~₹ 100,000 additionally. John Doe can afford to buy these new lenses. But, there are Hobbyists, cost sensitive photography professionals, who might not invest such a large amount. Also, not all lenses are used frequently and portable. Application of Camera and lenses are very dependent on time and situation(event). Not everyone can afford to buy all kinds of cameras and lenses. This a costly profession.

That's where the rental services industry identified the opportunity. These companies will have an inventory of all kinds of cameras, lenses and its accessories from almost all brands, (it can be also based on Managed market model). They rent these gears to customers and the rent will be charged based on daily basis and also rent is dependent on the cost of the gear. This way it becomes win-win

deal for both the company and Customers. All their inventory has now started minting money while Customers(Photographers) need not invest whooping money for purchasing lenses, instead spend very little fraction of the cost of the gear as rent and also gain access to wide range of options.

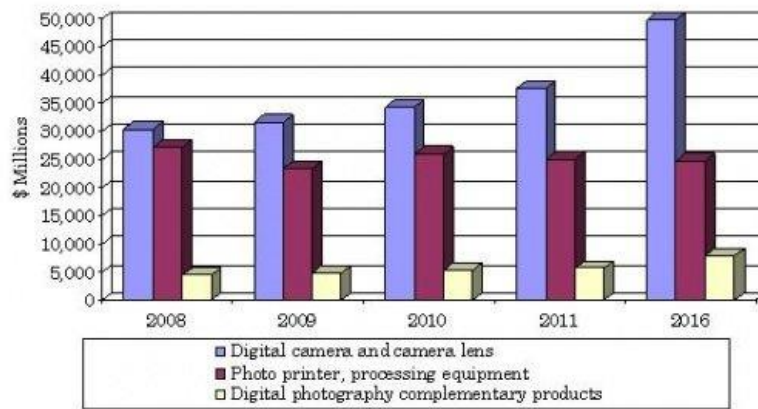


Figure 1

BBC Research has released a new report stating that the digital photography industry has an annual growth rate of 3.8%. Valued at \$68.4 billion last year, the global market will reach an estimated value of \$82.5 billion by 2016. The study defined the market as a combination of camera equipment, printing equipment, and complementary products. While the photo printing industry is predicted to struggle and lose \$300 million between now and 2016, digital cameras and lenses will reportedly do just fine: they have a healthy annual growth rate of 5.8%.

Though it seems to be very attractive, this industry is not an exception from challenges from market and competition.

**Top 5 challenges are listed below.**

- Inventory cost
  - Average lens cost comes out to be ₹135000.
  - If the Rental company wants to keep all the lenses available in market, it will go beyond \$100 M.
- Narrow customer segment
  - Unlike other rentals products like Car, garments etc, where anyone can be a prospective customer. In Camera rental services, it is mostly the Photography aspirants and professionals. Their population is negligible.

- Services and handling
  - Services and handling charges for these gears are very costly.
- Securing the devices
  - These are very costly devices . It requires at most care and make sure the gears that are rented out are safe. Customer's background verification test is also important to avoid gear loses.
- Pricing
  - Not all the camera gears go out for rent in same frequency and not everything is of same price. Some gears are less expensive with aggressive movement, some are very expensive and little movement. Rental pricing should be customized at the gear level to meet early break even.

Though we have such challenges, the industry is growing as more and more people are interested in Photography. Not only that, Social media and media services over the internet are complementing the camera industry for its growth.

#### [Klachak.com](http://Klachak.com)

Klachak.com is a company based out of Chennai, founded by a team of entrepreneurs, who are professionals in their own fields of expertise, yet have a great passion and skill in the art of photography. The company was formed with a vision to make photography accessible to all sincere enthusiasts who would like to create art with light.

It facilitates artists of every level to overcome their barriers to engage with photography, and help inspire creativity and ingenuity in this form of art. Its aim is to provide our discerning clientele services of every kind that is related to the field of photography including education, lens rental, and experiences on field both within and outside the country.

#### [Equipment Rentals](#)

Klachak provides a top of the line service in equipment rentals to assist professionals and amateurs to hire equipment that is usually out of reach financially. Be it lenses, cameras or any other photography related product, Klachak aims to make it accessible at a fair price, while still maintaining the best quality of service. Whether you have a wedding to shoot, a friend's birthday party to capture, or bring home memories of wild animals in the forest.

## Photography Services

Through both in-house photographers, as well as through selected professionals in our network, Klachak provides a range of photography services such as educational workshops, commercial shoots, portfolio development, for both individuals, corporates or advertising agencies. We also hold a high-quality stock photography collection, that can be licensed for a wide range of needs.

## Problem/Opportunity Statement

*“Primary Objective is to help users in finding the cameras, lenses or any other camera related gadget they would like to rent by predicting similarity score or a list of top 5 or 10 recommended items for the given users based on the transaction history, by using Recommender System Algorithms”.* To Reach that Goal, we have to perform very extensive Exploratory Data Analysis, which will produce Key Reports and Dashboards. This also gives opportunity to Perform Cohort Analysis to capture loyal customer and RFM Analysis as well for customer segmentation and base the recommendation on its weightage , It can recommend the Items to users through offline channels also.

Exploratory reports will further help Klachaks in,

1. Identifying all the factors contributing for more sales (in our case rentals) and revisit the Business Strategy accordingly.
2. How can Klachak take advantage of seasonality in rental activities.

## Analysis Methodology

### Methodology

This project will follow CRISP-DM model as a framework and methodology.

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<b>Determine Business Objectives</b> Background Business Objectives Business Success Criteria  <b>Assess Situation</b> Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits  <b>Determine Data Mining Goals</b> Data Mining Goals Data Mining Success Criteria  <b>Produce Project Plan</b> Project Plan Initial Assessment of Tools and Techniques	<b>Collect Initial Data</b> Initial Data Collection Report  <b>Describe Data</b> Data Description Report  <b>Explore Data</b> Data Exploration Report  <b>Verify Data Quality</b> Data Quality Report	<b>Select Data</b> Rationale for Inclusion/Exclusion  <b>Clean Data</b> Data Cleaning Report  <b>Construct Data</b> Derived Attributes Generated Records  <b>Integrate Data</b> Merged Data  <b>Format Data</b> Reformatted Data  Dataset Dataset Description	<b>Select Modeling Techniques</b> Modeling Technique Modeling Assumptions  <b>Generate Test Design</b> Test Design  <b>Build Model</b> Parameter Settings Models Model Descriptions  <b>Assess Model</b> Model Assessment Revised Parameter Settings	<b>Evaluate Results</b> Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models  <b>Review Process</b> Review of Process  <b>Determine Next Steps</b> List of Possible Actions Decision	<b>Plan Deployment</b> Deployment Plan  <b>Plan Monitoring and Maintenance</b> Monitoring and Maintenance Plan  <b>Produce Final Report</b> Final Report Final Presentation  <b>Review Project</b> Experience Documentation

Figure 2

Understanding the business and data are paramount for this project. Once we have data, We will proceed with preparing the data for further Statistical and descriptive analysis. This data preparation will include data cleaning, masking, missing value treatment, removing bias by applying various sampling techniques, etc., Before even proceeding with the project, initial study of data has been conducted as proof of concept and measure the feasibility.

Data preparation will be very critical. Wherever necessary we will apply transformation logics, like converting categorical variable to set of dummy variables, clubbing to variables to form a new variable, segmenting continuous variable to convert them to categorical variable, etc., As we have data from 3 data sources, it is important to establish a relationship between them through some key variables.

Once the data set is prepared, Descriptive Statistics will be applied again on the prepared data to understand the data and ensure the data is normal. Descriptive Stats includes studies like, central tendency, frequency distribution, variance study etc.,

Once the required data is prepared, formatted, it will be subjected to model building based on Recommendation Systems Algorithm. In Our case, we are using a hybrid approach of using content based and Collaborative Filtering based algorithms, Specifically Memory based.

Once the Model is Built, it is evaluated based on Root Mean Square Error Value for its prediction accuracy, Later it will be deployed in Klachak production system.

## Data Source

### Primary Data Source

All the transaction details and customer information are pulled from Klachak's Primary Database which is in MySQL. The data will be pulled using MySQL Connectors and exported to Excel which will be used for Exploratory Data Analysis, Building RFM, Building Recommendation Systems and Evaluation.

### Secondary Data Sources

We will access <https://www.dpreview.com> website to scrap, customers review and expert review of all lenses and cameras. Which will be used for sentimental analysis, that will later fed to Model for deriving User-Item Ratings.

Data is also pulled from "flickr" database, for capturing equipment usage ranking across globe that will be published to user post recommendation..

Primary Data Source	Secondary Data Source
Sales/Transaction Data from Klachaks	Detailed Product Information from <a href="https://www.dpreview.com">https://www.dpreview.com</a>  Detailed Application of Product is obtained from <a href="https://www.flickr.com/">https://www.flickr.com/</a>

Table 2

This data will help us in improving the accuracy of recommendation based on Situation.



## Tools & Techniques

We have conducted our analysis using Python 3.6.\* in Anaconda Distribution. Data visualization are done using Tableau 13 and few of them in Excel. Transaction information is stitched along with the Klachak's content management system tables and data bases. It was extremely difficult to identify the tables and key columns. Later We have to perform multiple joins and extract meaningful and correct data. On the other hand, dpreview.com does not expose any API's, so we had to perform Web scraping using scrapy in python. Data from flickr is obtained through API call.

Recommendation System Algorithms are memory and process intensive. Especially for predicting and scoring the rating for  $n$  items will require  $N \times N$  Matrix and each cell will have to go through Cosine Distance calculation of 2 vectors. Number items will keep growing along the time, so we need a data base which is scalable. We opted for NoSQL database, MongoDB.

Just not limited to the above tools, we have used following libraries and techniques.

1. Data Integration using Python
2. Data Profiling using Python and Tableau
3. Data Treatment using Python libraries using Scikit learn, numpy and pandas
4. Visualization in python using matplotlib
5. Collaborative Filtering using Pandas and numpy

## Limitations

Building a recommendation engine for Klachaks has following limitations.

1. Data Sparsity :

The reported matrix of user-item ratings is usually very sparse (up to 99%) due to users' lack of knowledge or incentives to rate items. In addition, for the new users or new items, in general, they report or receive only a few or no ratings. Both issues will prevent the CF from providing effective recommendations, because users' preference is hard to extract. Here it is ~97%

2. As the underlying data is transaction based data, probability of a product which is in the inventory for long duration being rented out will be more than the one which is new.

3. Probability of any new product added to the system being picked for recommendation might be low. Which we have tried to address using cold start mechanism. Using Hybrid approach.
4. User Rating for the products are not enabled or effectible used in Klachak's website. whereas Content bases filtering or collaborative filtering uses only user rating for recommendation. As workaround, we have derived the user-item rating using various factors, implicitly. Which is explained in model building section.
5. Model Evaluation: there are very limited techniques available for evaluating the accuracy and performance for Recommender Systems Algorithm.

## Literature Review

### Recommender Systems

This is how “Joonseok Lee” puts it, Any software system which is actively suggesting an item or a group of items to buy, subscribe, invest or to rent can be referred as a Recommender System. Any campaign or advertisement can also be considered as a recommendation. But, we mainly consider, however, a narrower definition of "personalized" recommendation system that base recommendation on user and item specific information, on Realtime streaming information.

Recommender System algorithm can be broadly can be classified as 2 categories

- Content Based Filtering
- Collaborative Filtering

Collaborative filtering can further be classified as “Memory based” and “Model based” each has its pros and cons and also highly dependent on the type of underlying data, number of users, number of items, how sparse the data is, etc.,

#### Content Based Filtering:

This is explicitly based on domain knowledge concerning the users and item. Both User profile and item profile plays a key role in this recommendation system. Further, domain knowledge may correspond to user information such as Age, Gender, Occupation, Geographical location etc. In case of Item, it may be like, product features, like, lens type, max ISO, fstops, Shutter Speed, Brand, Price Etc.,

#### Collaborative Filtering Algorithms:

Collaborative filtering does not use user or item information with the exception of a partially observed rating matrix. The rating matrix holds ratings of items (columns) by users (rows) and is typically binary, for example like vs. do not like, or ordinal, for example, one to five stars recommendation. The rating matrix may also be gathered implicitly based on user activity. Similar to our case.

#### Hybrid collaborative Filtering:

Hybrid collaborative and content-based filtering strategies combine the two approaches above, using both the rating matrix, and user and item information. Such systems typically obtain improved prediction accuracy over content-based filtering systems and over collaborative filtering systems.

## Exploratory Data Analysis

Before dwelling into preparing for development of model, we have to understand the data well. As Klachak does not have any explicit rating mechanism, an alternative method is used to derive best possible rating based on the observations. In order to identify Key fields that can be used for deriving the rating the following study is conducted. This study will also help in understanding the Klachaks business position and its performance.

### Sales Yearly Trend.

Yearly Trend

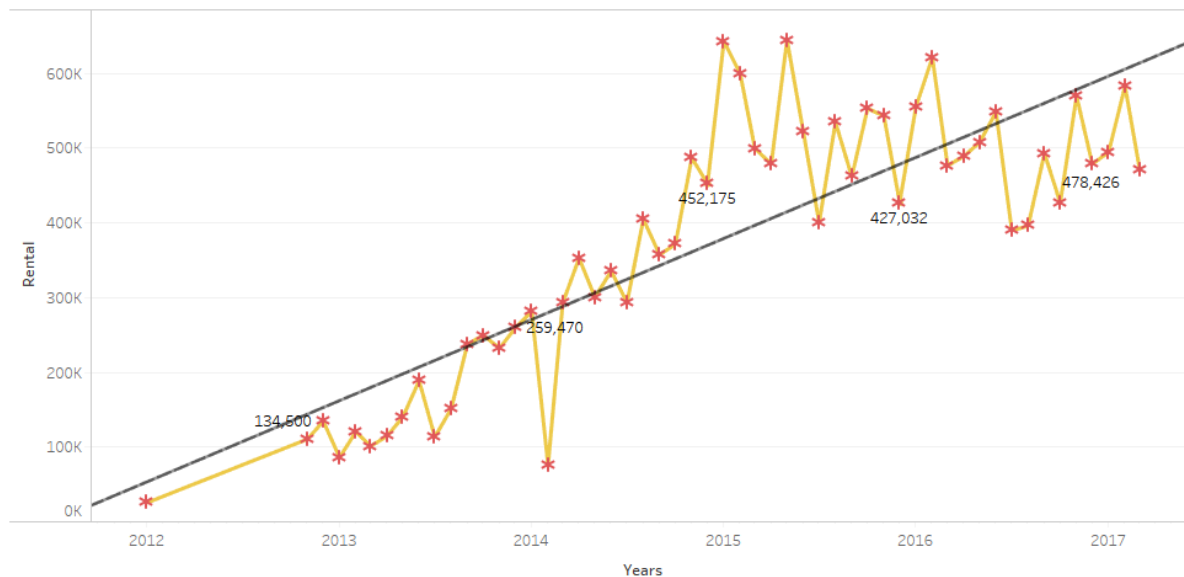


Figure 3

Figure 3, shows the Sales (Revenue from Rentals) over the period of 5 years. We can clearly see that overall sales have increased significantly. But, if we can just look at the past 3 years we can see that the trend is almost horizontal. Considering if this trend continues, this will have an impact on profit and it becomes very important for the company to find new customers, hold on to existing loyal customer and also where ever possible, perform up-sell or cross-rent the product.

## Seasonal Trend

Seasonal Trend

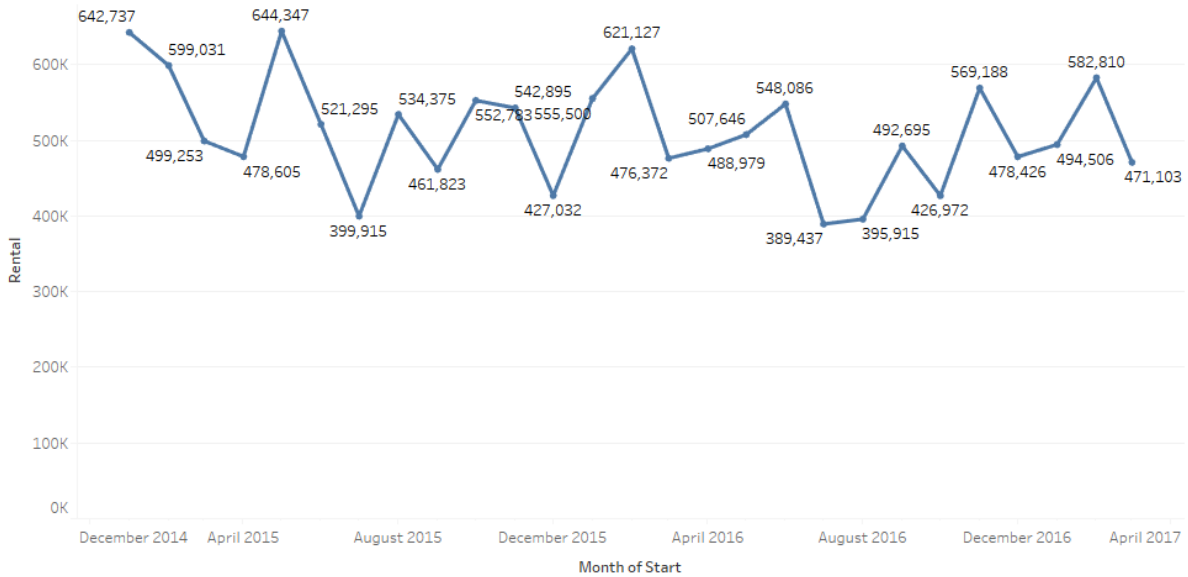


Figure 4

**Inference:** Three seasonal peaks per year in the months as January/February, July/August and October/November.

## Percentage of Revenue

% of Revenue

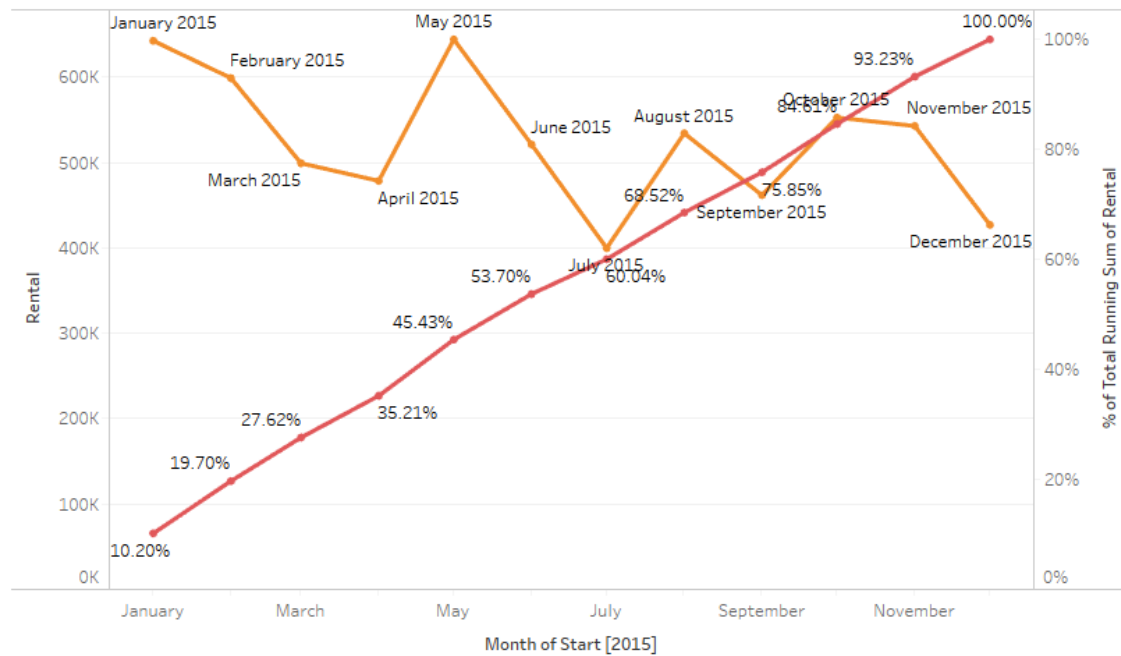


Figure 5

**Inference:** For the Year 2015, 60% of revenue achieved in first six months. This is a pattern that can be observed year over year. This correlates very well with previous observation on seasonality. There are 2 seasons before third Quarter of every year, these contributes heavily on sales.

### Gear Taken > 40 in a Month

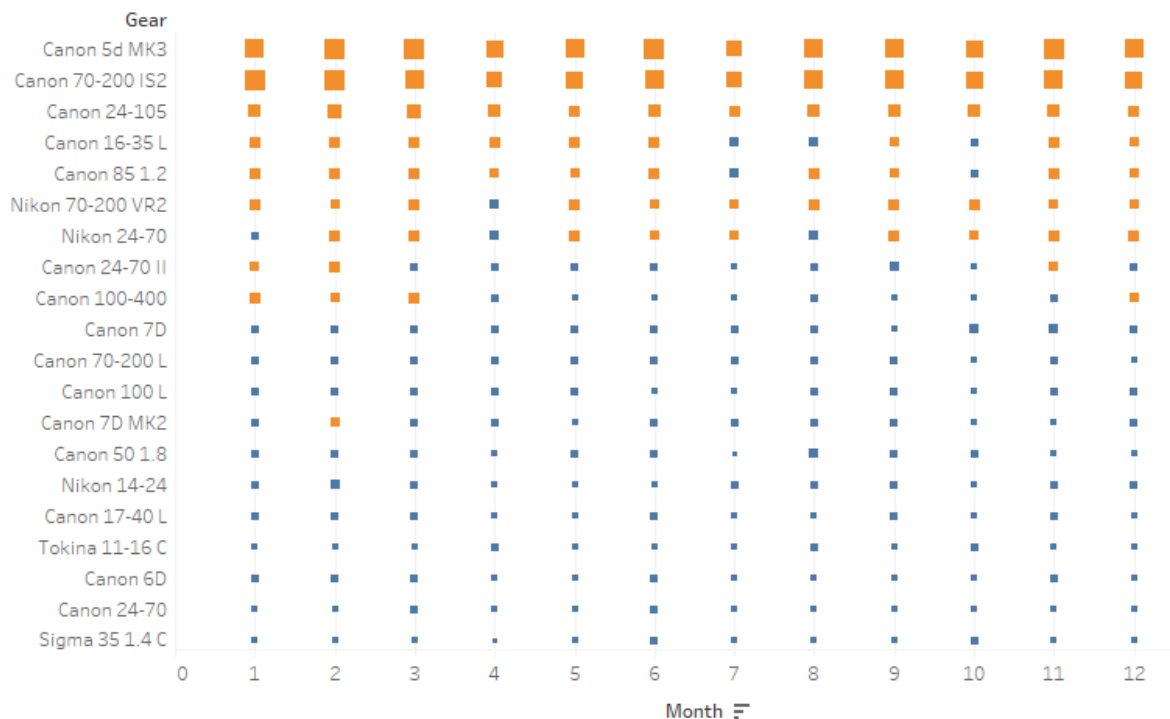


Figure 6

Seasonality and type of cameras rented confirms that most of the transactions were related to marriage events or other festivals. Also, Canon is contributing more for sales. Top5 are Canon Brand.

### Recommendation:

1. Replace cameras rented very few times like Nikon 200-400 VR2, Nikon 60 F4, Canon 1D MK4, etc. (extensive list provided separately) with frequently rented cameras like Cannon 5d MK3, Canon 70-200 IS2, Canon 24-105 (preferred mostly for marriage events) to increase revenue. This purely based on the Recommender that is to be built which will also consider the purpose of the rent.
2. Promotional exercise in the later part of the year between August to December will increase revenue during off-season

## Cohort Analysis

### Cohort Sales

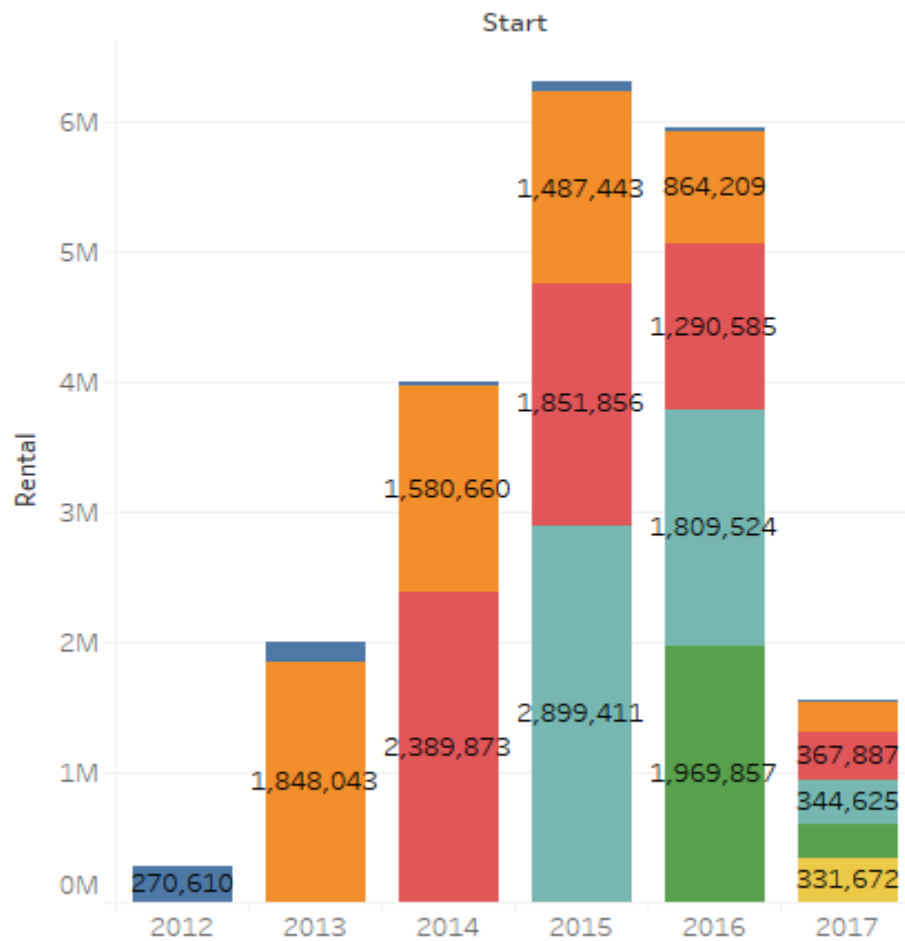


Figure 7

Inference: Customer who rented in 2013 or 2014 rented frequently and they are most loyal ones, cohort on loyalty will explain more.



## Cohort Loyalty

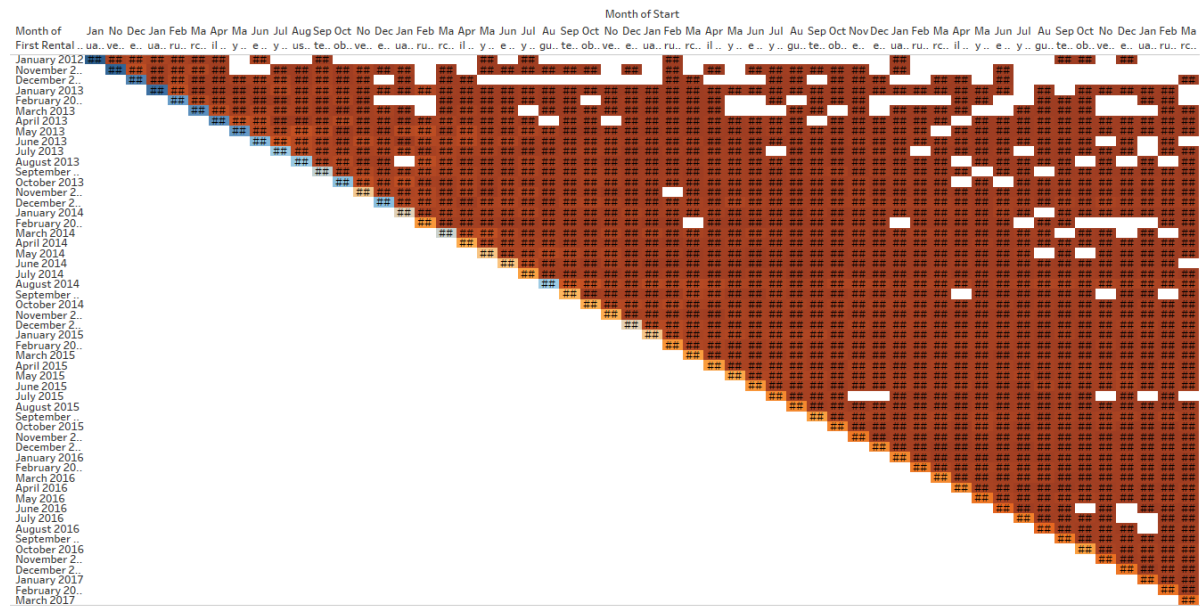


Figure 8

**Inference:** Very few new customers are renting in the last two years and high share of revenue generated by customers from initial years. Does that mean, people who become customers of Klachaks always stays as customer because of the value they see in it? It can be true, what can also be true is, Kalachaks is a near to monopoly in Chennai when it comes to Camera Rentals.

### Recommendation

Recommending any advertisement/promotional exercises to bring new customers which is not done till now by Klachak. Also, Klachacks needs to be proactive and focus on retaining its customer base and also aggressively look for adding new customers to its Loyalty list.

Cohort and Inventory flow visualization helps us to understand the strength of our customer base as well as fast running product. We will basically use customer loyalty score as well as product inventory performance for deriving the initial rating of user-item.

This can be further enhanced by performing RFM analysis, Segment the Users and assign weightage for deriving the rating based on segmentation. Before getting into RFM, we will prepare the data, understand the model and understand how RFM contributes for Collaborative filtering model.

## Model Development

Developing the model for collaborative filtering involves 3 Stage process as listed below.

1. Data Extraction and Data Preparation
2. Model Building and Evaluation
3. Deploying in Live system.

### Data Extraction and Data Preparation.

Input for this model comes from 2 different data sources. Primarily one from Klachaks Native Data base and other from a website dpreview. For building content based filtering we need features of the product, which will be obtained from dpreiew. As it was mentioned earlier that klachak.com does not capture the user rating for each item, we have to derive those as Collaborative filtering is designed based on user rating. Apart from user rating, collaborative filtering algorithm just needs only the user id/name and product id/name, which is directly pulled form Klachak's database.

Based on the RFM Score and number of time a particular product is rented by the same customer, probability weightage of rating scale will be applied to randomly derive the rating. Rating will be of scale 1 to 5, 1 being low an d5 being high,

### How RFM value is segmented.

RFM Score	Segmentation Type
444,445,455,555,554,544	Very Loyal and High Value Customers (VLHV)
4*5,5*4	New Customers and High Value (NCHV)
*45,*54	Very Frequent and High Value Customers (VFHV)
441,442,443,451,452,453,551,552,553,541,542,543	Very Loyal and Low Value Customers (VLLV)
Rest	Onetime Visitors.

Figure 9

Based on the RFM Segmentation Type and Number of times an user has rented a particular product we will assign probability and the rating randomly

### Mapping Cheat Sheet got Rating Derivation

RFM Segment	Number of times Rented	Probability				
		1	2	3	4	5
Very Loyal and High Value Customers (VLHV)	1	0.1	0.2	0.3	0.3	0.1
New Customers and High Value (NCHV)	1	0.1	0.1	0.35	0.25	0.2
Very Frequent and High Value Customers (VFHV)	1	0	0.15	0.25	0.4	0.2
Very Loyal and Low Value Customers (VLLV)	1	0.3	0.3	0.2	0.1	0.1
Onetime Visitors.	1	0.2	0.2	0.2	0.2	0.2
Very Loyal and High Value Customers (VLHV)	2	0.1	0.1	0.3	0.3	0.2
New Customers and High Value (NCHV)	2	0.1	0.15	0.35	0.25	0.15
Very Frequent and High Value Customers (VFHV)	2	0	0.1	0.25	0.4	0.25
Very Loyal and Low Value Customers (VLLV)	2	0.3	0.2	0.3	0.1	0.1
Onetime Visitors.	2	0.2	0.2	0.2	0.2	0.2
Very Loyal and High Value Customers (VLHV)	greater than or equal to 3	0.1	0.1	0.2	0.3	0.3
New Customers and High Value (NCHV)	greater than or equal to 3	0.1	0.1	0.3	0.25	0.25
Very Frequent and High Value Customers (VFHV)	greater than or equal to 3	0	0.1	0.2	0.4	0.3
Very Loyal and Low Value Customers (VLLV)	greater than or equal to 3	0.25	0.15	0.3	0.2	0.1
Onetime Visitors.	greater than or equal to 3	0.2	0.2	0.2	0.2	0.2

Figure 10

Now that we have Rating scale derived from RFM and Product performance per user, we will regularize it by comparing against the sentiment score and global rating from dpreview.

### Web Scrapping – Using Scrapy

Klachak inventory sources very limited products metadata information. Hence, we scrapped dpreview.com for full product specifications, attribute scorings, products reviews by dpreview.com and users. Scrapped Product specifications and attribute scorings were used for building recommendation model and reviews are converted into scored using sentiment analysis.

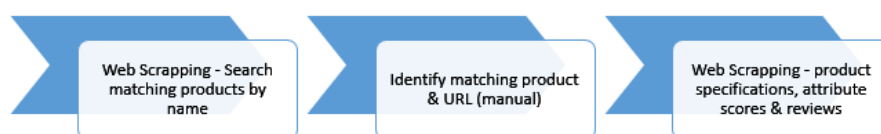


Figure 11

### Scrapy Data Flow

Data flow in Scrapy is controlled by the execution engine, and the steps are

1. The Engine gets the initial Requests to crawl from the Spider.
2. The Engine schedules the Requests in the Scheduler and asks for the next Requests to crawl.
3. The Scheduler returns the next Requests to the Engine.

4. The Engine sends the Requests to the Downloader, passing through the Downloader Middlewares.
5. Once the page finishes downloading, the Downloader generates a Response (with that page) and sends it to the Engine, passing through the Downloader Middlewares.
6. The Engine receives the Response from the Downloader and sends it to the Spider for processing, passing through the Spider Middleware.
7. The Spider processes the Response and returns scraped items and new Requests (to follow) to the Engine, passing through the Spider Middleware.
8. The Engine sends processed items to Item Pipelines, then send processed Requests to the Scheduler and asks for possible next Requests to crawl.
9. The process repeats (from step 1) until there are no more requests from the Scheduler.

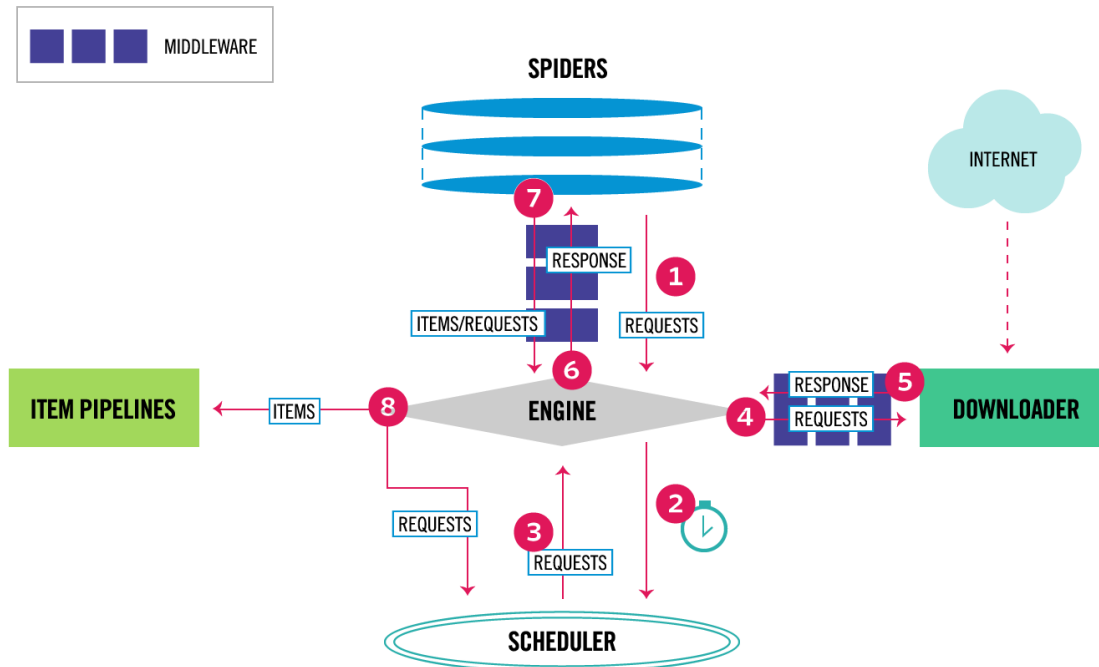


Figure 12

## Model Building and Evaluation

Why are we putting hybrid approach of recommendation system instead of specifically content based or collaborative filter based? We will come to that shortly after explaining the how model works.

### Content Based Filtering

This also can be called as Feature based filtering. Where we will match the similarity of the feature with the product selected by the user. In Klachak's as we are renting the camera's and accessories, we will have to find the similarities of the features of such product and recommend accordingly.

Following features are identified,

<b>Technical Features</b>	Brand	Canon
		Nikon
		Sony
	Min ISO	50 - 120
		greater than 120
	Max ISO	3200 - 12800
		12800 and greater
	Min Shutter Speed	15 sec
		60 Sec
	Max shutter Speed	1/1000 sec to 1/4000 sec
		1/4000 sec to 1/16000 sec
	Apperture	f 1.8
		f 2
		f 2.8
		f 3.5
		f 4
<b>Genre</b>	Aerial	
	Architectural	
	Candid	
	Documentary	
	Fashion	
	Food	
	Landscape	
	Photojournalism	
	Portraiture	
	Street photography	
	Sports	
	Wildlife	

Figure 13

The Technical feature list is obtained from dpreview for each product. Genre information is extracted from flickr. Now that we have transaction data from klachak database, based on the user transaction we update the feature table for all the existing users. Suppose, a customer has rented a camera once or more than once, that has a features like,

“Canon”, “50-120” min iso, 12800 max iso, “15 sec” min shutter speed, “1/1000 sec to 1/4000 sec” maximum shutter speed, “f 2.8” Aperture, Genres like, “Candid”, “Fashion”, “Portraiture”.

Then it will be scored like this.

Feature	Score
Canon	1
Nikon	0
Sony	0
50 - 120	1
greater than 120	0
3200 - 12800	0
12800 and greater	1
15 sec	1
60 Sec	0
1/1000 sec to 1/4000 sec	1
1/4000 sec to 1/16000 sec	0
f 1.8	0
f 2	0
f 2.8	1
f 3.5	0
f 4	0
Aerial	0
Architectural	0
Candid	1
Documentary	0
Fashion	1
Food	0
Landscape	0
Photojournalism	0
Portraiture	1
Street photography	0
Sports	0
Wildlife	0

Table 3

Based on all transaction, Content Based Item- Feature score for all items are scored and later will be used for feature similarity matrix.

Feature	Score 1	Score 2	...	Score n-1	Score n
Canon	1	0	...	0	1
Nikon	0	1	...	0	0
Sony	0	0	...	1	0
50 - 120	1	0	...	1	0
greater than 120	0	1	...	0	1
3200 - 12800	0	0	...	0	0
12800 and greater	1	1	...	1	1
15 sec	1	1	...	0	1
60 Sec	0	0	...	1	0
1/1000 sec to 1/4000 sec	1	0	...	0	1
1/4000 sec to 1/16000 sec	0	1	...	1	0
f 1.8	0	0	...	0	1
f 2	0	1	...	0	0
f 2.8	1	0	...	0	0
f 3.5	0	0	...	1	0
f 4	0	0	...	0	0
Aerial	0	0	...	0	0
Architectural	0	0	...	1	1
Candid	1	0	...	0	1
Documentary	0	1	...	0	0
Fashion	1	1	...	0	1
Food	0	0	...	1	0
Landscape	0	0	...	0	0
Photojournalism	0	1	...	0	1
Portraiture	1	0	...	0	1
Street photography	0	0	...	0	0
Sports	0	1	...	0	0
Wildlife	0	1	...	1	1

Table 4

Then similarity is computed based on Cosine Vector distance formula provided below.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Figure 14

Where A is a Feature Score vector of the first product, and B is a Feature score vector of second Product.

Like, every product is compared with the rest of the product in N dimensional space to arrive at similarity matrix based on product features. This matrix is later used for obtaining top 5 items similar to the selected items through recommendation engine.

Note: Here we are not measuring the similarity between the products and between the users, we are deriving the similarity between only features of the products.

### Collaborative Filtering

In Badrul Sarwar's words "The goal of a collaborative filtering algorithm is to suggest new items or to predict the utility of a certain item for a particular user based on the user's previous likings and the opinions of other like-minded users. In a typical CF scenario, there is a list of m users  $U = \{u_1, u_2, \dots, u_m\}$  and a list of n items  $I = \{i_1, i_2, \dots, i_n\}$ . Each user  $U_i$  has a list of items  $I_{ui}$ , which the user has expressed his/her opinions about. Opinions can be explicitly given by the user as a rating score, generally within a certain numerical scale, or can be implicitly derived from purchase records"

There exists a distinguished user  $u_a \in U$  called the active user for whom the task of a collaborative filtering algorithm is to find an item likeliness that can be of two forms, Prediction and Recommendation.

Prediction is a numerical value,  $P_{a,j}$ , expressing the predicted likeliness of item  $i_j \notin I_{u_a}$  for the active user  $U_a$ . This predicted value is within the same scale (e.g., from 1 to 5) as the opinion values provided by  $U_a$ .

Recommendation is a list of N items,  $I_r \subset I$ , that the active user will like the most. Note that the recommended list must be on items not already purchased by the active user, i.e.  $I_r \cap I_{u_a} = \emptyset$ . This interface of CF algorithms is also known as Top-N recommendation.

Though there are 2 types of algorithm for CF, one is user based and the other is item based, given the former one has limitations like scalability and sparsity issues, we are using item based Collaborative Filtering.

### Item Based Collaborative Filtering

Item-based approach looks into the set of items the target user has rated and computes how similar they are to the target item i and then selects k most similar items  $\{i_1, i_2, \dots, i_k\}$ . At the same



time their corresponding similarities  $\{si_1, si_2, \dots, si_k\}$  are also computed. Once the most similar items are found, the prediction is then computed by taking a weighted average of the target user's ratings on these similar items. We describe these two aspects namely,

the similarity computation

and, the prediction generation in details here.

Before getting to the next step, let's understand our data. We do not have User Opinion or rating for each item, we only have following information, with that we will implicitly derive the Rating for each product each user.

Here we use the sales transaction data obtained from Klachaks DB.

Variable	Type	Description
ProductID	Numeric	Unique identifier of the product
ProductName	String	Name of the Product
Cust_Id	Numeric	Unique Identifier of the Customer
Customer Name	String	Name of the customer

#### First Input for Deriving the Rating – Sales Data

Aggregate the Transaction Count bases on Customer and Product. This will result in number of times a customer rented the given product. Based on the count we can randomly assign the rating in way that, if the count is more give more weightage for Rating Score 4 and 5, if the count is one or two, flip the weightage for rating 1 and 2, any number between the maximum and minimum count adjust the Probability weight of random number relatively.

#### Second Input for Deriving the Rating – RFM Segments

To make the Rating Derivation more accurate, We will use the RFM Segment information also clubbed with the Sales data in assigning the probability weightage for Rating Score. Please [refer figure 10](#) to understand more.

#### Third Input for Deriving the Rating – Sentiment Score

To make it more precise, This system will perform web scraping of dpreview and obtain the Sentimental score. Based on the Sentiment Score this rating will be again adjusted. Then we arrive at the final rating value from Scale 1 to 5 for each item that user has rented.

### Sparsity Test

There are 3 Things that will affect the Prediction and Recommendation accuracy of CF algorithm.

1. Number of Customers
2. Number of Products
3. Rating Sparsity.

We have around 2118 Unique Customers, with 176 unique products to choose from. In other words we have 372, 768 combination of possible user item ratings, of which we have 12982 combinations rated. Which means, out of all possible ways of capturing rating by every user for every product, we have only 3.49% rated or ratings derived. We still have to Predict the Rating for remaining 96.51% of combinations or in other words, 359,786 ratings needs be predicted.

Generally 98% Sparsity is considered sufficient for Building Recommendation System, we have 96.51%, which holds good for us to proceed.

When looking at Rating sparsity of each product, which means, 2118 Unique customers, how many of them rated for each product, we get the following graphs.

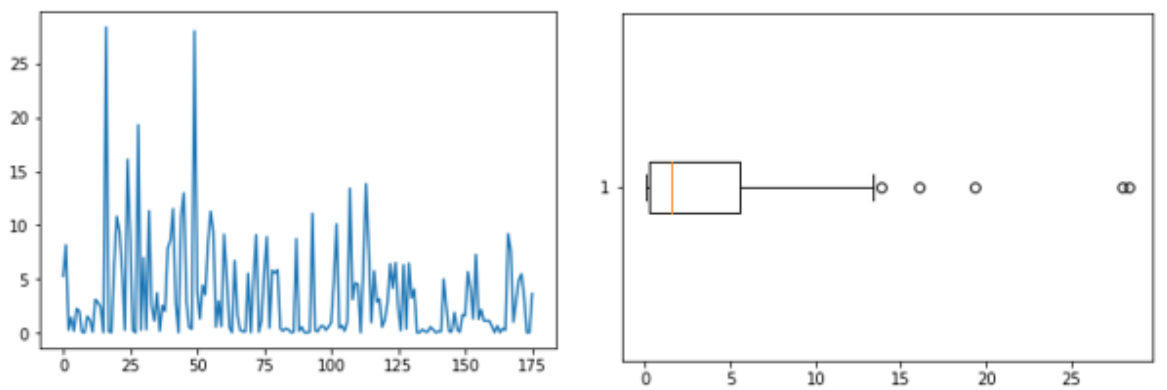


Figure 15

Nearly 28% of Products have dense Ratings, while remaining 72 percentage very weak density with respect to rating. This might impact the similarity score calculation, to overcome it, we are centering the data based on Item rating average.

### Item Similarity Computation

One critical step in the item-based collaborative filtering algorithm is to compute the similarity between items and then to select the most similar items. The basic idea in similarity computation

between two items  $i$  and  $j$  is to first isolate the users who have rated both of these items and then to apply a similarity computation technique to determine the similarity  $s_{ij}$ . Figure 2 illustrates this process, here the matrix rows represent users and the columns represent items. There are a number of different ways to compute the similarity between items. Here we present three such methods. These are cosine-based similarity, correlation-based similarity and adjusted-cosine similarity.

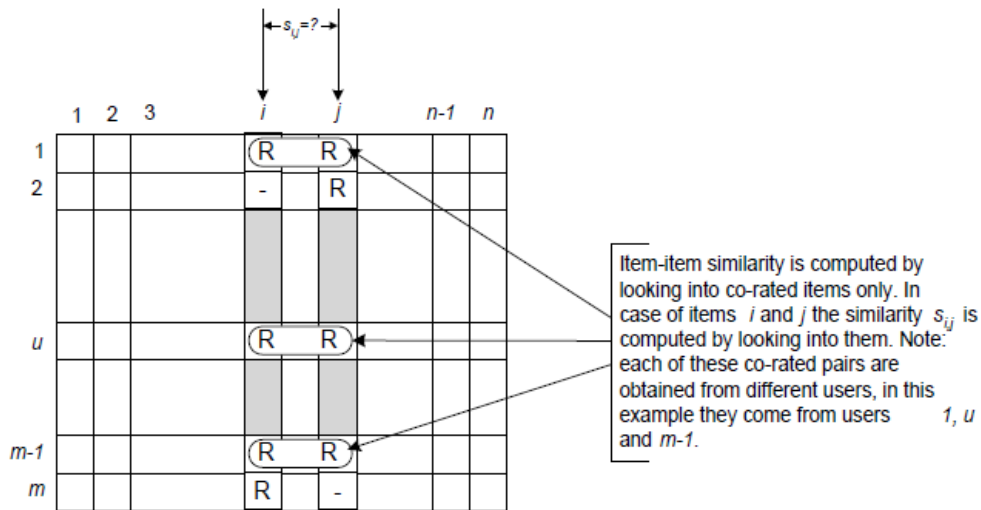


Figure 16

In our case we have taken “adjusted-cosine similarity” for calculating the Similarity Score,

$$sim(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_u)^2}}$$

Here  $\bar{R}_u$  is the average of the  $u$ -th user's ratings.

## Sample Output from Item-Item Similarity Matrix

click to scroll output; double click to hide

	3 feet Slider	4 feet Slider	5 in 1 Reflector Disc	77mm Graduated ND Filter	77mm Graduated blue Color Filter	Aperture DSLR Shoulder Rig V2	Aperture 7 inch HD SScreen	Arm for GoPro
3 feet Slider	1.000000	0.015006	0.000000	0.007627	0.000000	-0.021777	0.001647	-0.069846
4 feet Slider	0.015006	1.000000	-0.000007	0.065154	-0.003165	0.064558	-0.020661	-0.102720
5 in 1 Reflector Disc	0.000000	-0.000007	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
77mm Graduated ND Filter	0.007627	0.065154	0.000000	1.000000	0.021193	0.051220	-0.007613	0.000000
77mm Graduated blue Color Filter	0.000000	-0.003165	0.000000	0.021193	1.000000	0.000000	0.000000	0.000000

Figure 17

## Prediction Generation

There are couple of ways we can generate the Rating prediction for those remaining 359,786 possible combinations,

1. Using Weighted Sum
2. Using Regression.

This project used Weighted Sum to Predict the Ratings.

This method computes the prediction on an item  $i$  for a user  $u$  by computing the sum of the ratings given by the user on the items similar to  $i$ . Each ratings is weighted by the corresponding similarity  $s_{i,j}$  between items  $i$  and  $j$

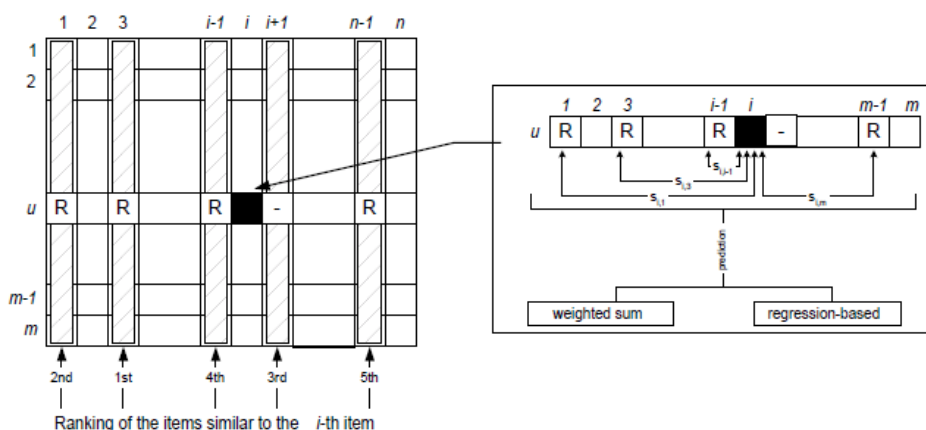


Figure 18

Formula is,

$$P_{u,i} = \frac{\sum_{\text{all similar items, } N} (s_{i,N} * R_{u,N})}{\sum_{\text{all similar items, } N} (|s_{i,N}|)}$$

Figure 19

Predicted Output (Sample)

	PredictedRating	SourceRating	customer	product
0	5	0	PARTHIBAN	3 feet Slider
1	2	0	A	3 feet Slider
2	4	0	A J Aravind	3 feet Slider
3	1	0	A P S	3 feet Slider
4	3	0	A.G	3 feet Slider
5	2	0	A.R	3 feet Slider
6	3	0	A.R.Naazar	3 feet Slider
7	4	0	A.X.P Leo	3 feet Slider
8	3	0	ABDUL	3 feet Slider
9	2	2	ABIN	3 feet Slider

Figure 20

#### Collaborative Filtering Model Evaluation

There is very limited methodology to evaluate the model.

#### Root Mean Square Error (RMSE)

Root mean square error computes the mean value of all the differences squared between the true and the predicted ratings and then proceeds to calculate the square root out of the result [5]. As a consequence, large errors may dramatically affect the RMSE rating, rendering the RMSE metric most valuable when significantly large errors are unwanted. The root mean square error between the true ratings and predicted ratings is given by

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (d_i - \hat{d}_i)^2}$$

$d_i$  is the actual rating

$\hat{d}_i$  is the predicted rating

$n$  is the amount of ratings

We took the following approach.

Randomly Divide the Original Rated Data into Training and Test in the proportion 60% and 40%

Total number of Items Originally rated : 12,982

Training Set : 7789

Test Set: 5163

Overall sparsity of Original Set : 64%

Sparsity Ratio for Training set : 60%

Run the same algorithm on test data for which we already know the Original rating.

Arrive RMSE value based on original rating and predicted rating.

**RMSE : 0.35 [ Which says Model is fit and good for predicting the rating with overall deviation of .35 from the actual rating]**

### Why Hybrid Approach?

Going back to the Question of Why Hybrid approach and why not either of them, is because, of the following reason.

As not every product are having equal number of transaction, we might not be able to recommend similar product for those products as there won't be any intersection of ratings, also, any new product will not be entering into the systems as it will not have any transactions or ratings.

Hence for such Products that have no transaction or only one transaction we are using content based filtering as it will match based on feature similarity and not rating similarity. For those products that gas several transactions, we will use rating similarity.

## Implementation

### Overall Architecture

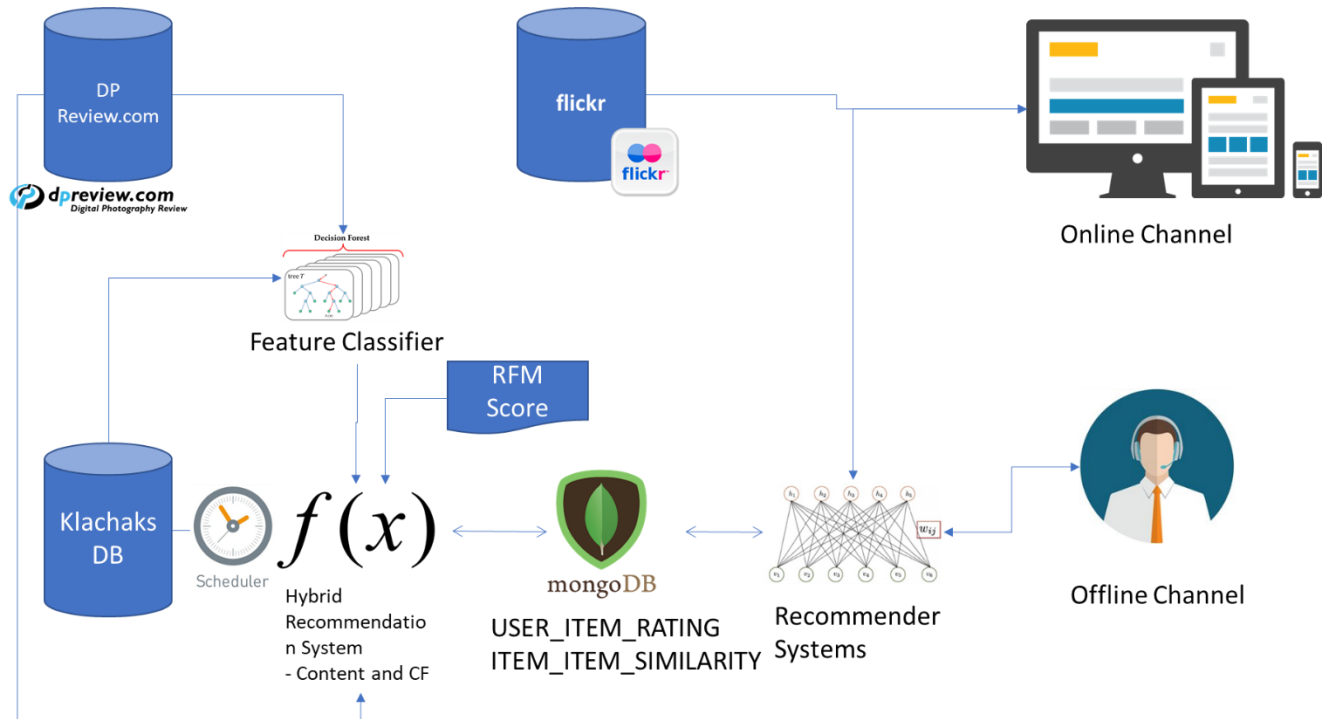
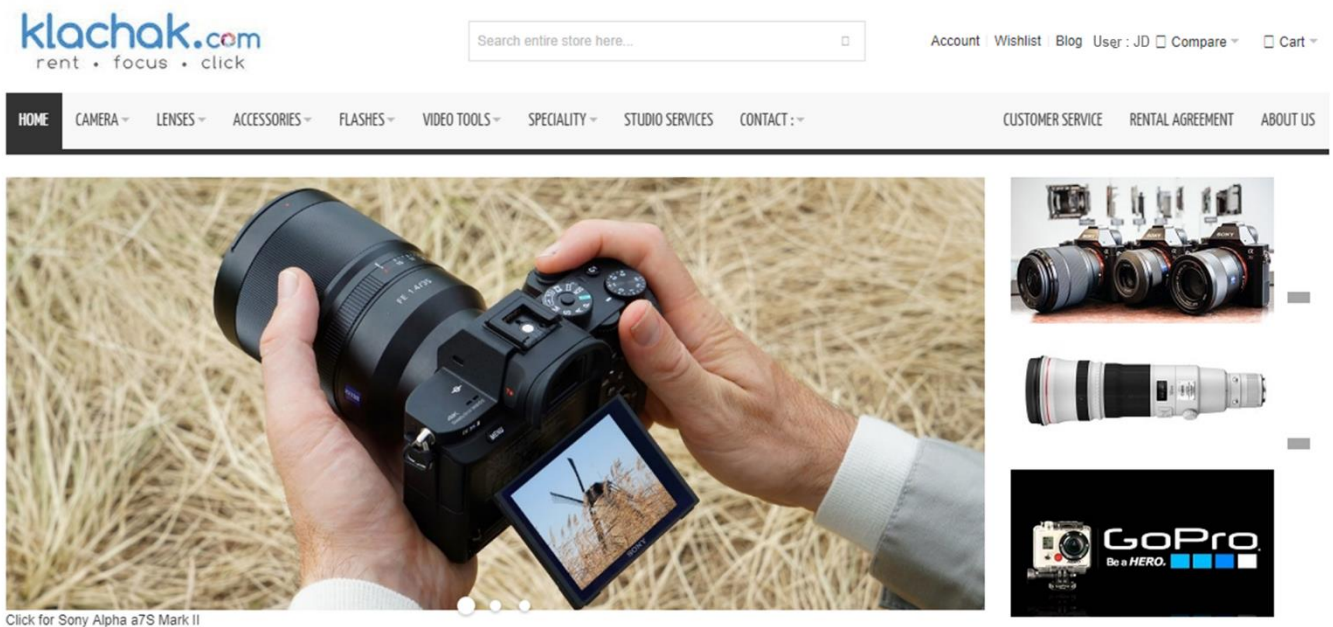


Figure 21

## Post Implementation



## Recommendation Systems in Action

Contact : +91 8939148844

**klachak.com**  
rent • focus • click

Search entire store here...


Account | Wishlist | Blog | User : JD | Compare | Cart

HOME CAMERA LENSES ACCESSORIES FLASHES VIDEO TOOLS SPECIALITY STUDIO SERVICES CONTACT


CUSTOMER SERVICE RENTAL AGREEMENT ABOUT US

Home → Camera → Canon


**On Selecting this...**




Canon EOS 5D MK III  
1 Day: ₹ 1,625.00  
2 Days: ₹ 3,200.00  
3 Days: ₹ 4,775.00  
Qty: 1  
Rent



Canon EOS 7D  
1 Day: ₹ 890.00  
2 Days: ₹ 1,730.00  
3 Days: ₹ 2,570.00  
Qty: 1  
Rent



Canon EOS 5D MK IV  
1 Day: ₹ 2,150.00  
2 Days: ₹ 4,250.00  
3 Days: ₹ 6,350.00  
Qty: 1  
Rent



Canon EOS 1Dx DSLR Camera  
1 Day: ₹ 3,200.00  
Not available for selected dates  
Qty: 1  
Rent

**You may also be Interested to...**

**Canon EF 500mm f/4L IS II USM**  
1 Day: ₹ 3,200.00  
2 Days: ₹ 6,350.00  
3 Days: ₹ 9,500.00  
Rent

**Manfrotto Video Tripod**  
1 Day: ₹ 470.00  
2 Days: ₹ 940.00  
Rent

**Canon 400mm f5.6 USM L**  
1 Day: ₹ 1,100.00  
2 Days: ₹ 2,045.00  
3 Days: ₹ 3,042.50  
Rent

**Recommend these**

32

NA

12

**Sentiment Analysis**



## Conclusion

The Hybrid Model, based on Content based and Collaborative filtering model has shown effective results in recommending the products for the user based on feature similarity as well as item similarity. This will also help Klatchaks to upsell and cross sell the items effectively. As RFM being one of the byproduct of this project, it will help Klatchaks in perform targeted campaigns as well.

## Future Scope

1. Use Analytical Hierarchy Processing for Enhancing the User Decision making process post recommendation.
2. Use Conjoint Analysis for deriving Part Worth Utility, and use it as input for Deriving user-item rating in more precise way.
3. Real time in-memory computing using spark so that it will be scalable going forward as and when the size of the item matrix and user matrix grows.

## References

[1] Item-based Collaborative Filtering Recommendation Algorithms.

- Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl

[2] A Comparative Study of Collaborative Filtering Algorithms.

-Joonseok Lee, Mingxuan Sun, Guy Lebanon

[3] Mining Massive Datasets – Recommender Systems

- Leskovec, Rajaraman and Ullman