Data Driven Ecommerce App Suite

Submitted By

Balram M Mirani

14MCEC03

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Data Driven Ecommerce App Suite

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

Submitted By

Balram M Mirani

(14MCEC03)

Guided By

Dr. Sanjay Garg

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Certificate

This is to certify that the major project entitled **"Data Driven Ecommerce App Suite"** submitted by **Balram M Mirani (Roll No: 14MCEC03)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project part-I, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Dr. Sanjay Garg
Internal Guide,
Head & Associate Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Mr. Amrish Patel
External Guide,
WebLoudSpeaker Private Limited,
Ahmedabad

Dr. Priyanka Sharma
Associate Professor,
Coordinator M.Tech - CSE
Institute of Technology,
Nirma University, Ahmedabad.

Prof. P N Tekwani
Director,
Institute of Technology,
Nirma University, Ahmedabad

WEBLOUDSPEAKER PVT. LTD.
14/A, GULAB PARK,
THALTEJ,
AHMEDABAD-380054.
Mo.: 94264 56129. Email: amrish84@gmail.com.
CIN: U72900GJ2012PTC069974

CERTIFICATE

This is to certify that Mr. **Balram Mirani** student of M.Tech Computer Engineering, Institute of Technology, Nirma University, Ahmedabad-382481, Gujarat, India is doing his full-time training at Webloudspeaker Pvt. Ltd., Ahmedabad. He has sincerely completed his project titled "**Data Driven Ecommerce App Suite**".

It gives us indeed pleasure to highlight that Mr. Balram Mirani has worked hard and with sincerity throughout the project work. We wish him all the best for his bright future.

Please do not hesitate to contact us if you required any further information.

Kind Regards,

amrish

Mr. Amrish Patel

(Founder & Project Guide)

Webloudspeaker Pvt. Ltd., Ahmedabad

Statement of Originality

I, **Balram Mirani**, Roll. No. **14MCEC03**, give undertaking that the Major Project entitled "**Data Driven Ecommerce App Suite**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Dr. Sanjay Garg

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Dr. Sanjay Garg**, Associate Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

A special thank you is expressed wholeheartedly to **Prof. P N Tekwani**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also like to thank **Mr. Amrish Patel** and **Mr. Amit Goyal**, WebLoud-Spekaer Private Limited for the guidance and support provided in my internship & the project work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Balram M Mirani

14MCEC03

Abstract

Electronic Commerce is process of doing business through computer networks. A person sitting on his chair in front of a computer can access all the facilities of the Internet to buy or sell the products.On-line shopping has become the trend and people are more comfortable to buy stuffs on-line instead of going to shop.This has increased the competition among different store owners to show more relevant products to each user in order to make customers life easy by providing recommendation of certain products which he seeks.

Recommender system is one of the applications to predict rating or preference for the items that have not been seen by a user. This system typically produces a list of recommendations. Recommending books, CDs, and other products at amazon.com, news etc.. are examples of such applications to name a few. However, despite these developments, the current generation of recommender systems still requires further improvements to make recommendation methods more accurate and applicable to an even broader range to make customer buying process as simple as ever . Hence, advanced recommendation modelling methods, incorporation of various contextual information into the recommendation process, and measures to determine performance of recommender systems are considered.

The rapid growth of the market in every sector is leading to a bigger subscriber base for service providers. More competitors, new and innovative business models and better services are increasing the cost of customer acquisition. In this environment service providers have realized the importance of the retention of existing customers. Therefore, providers are forced to put more efforts for prediction and prevention of churn.

In this dissertation,we are focusing on two essential tools for an E-commerce store: Recommender System,Churn detection and prevention model.We have proposed an Item based Recommender System which will recommend viewed also viewed products by considering your current interest only and discarding previous history.We have also proposed a churn detection model which is backed by random forest in order to detect the root cause of customer churn.

Abbreviations

CF	Collaborative Filtering.
CB	Content based.
CRM	Client Relationship Model.
CMF	Churn Management Framework.

Contents

List of Figures

Chapter 1

Introduction

E-business has been developing quickly alongside Internet. Its fast development has made both organizations and clients confront another circumstance. While organizations are harder to get by because of more rivalries, the open door for clients to pick among more items has expanded the weight of data handling before they select which items address their issues. Thus, the requirement for new showcasing systems, for example, onetoone advertising and client relationship administration (CRM) has been focused on both from inquires about and also from useful issues. One answer for understand these techniques is customized proposal that offers clients some assistance with finding the items they might want to buy by delivering a rundown of prescribed items for every given customer.[2]
The objective of a Recommender Engine is to create significant suggestions to a gathering of clients for things or items that may intrigue them. Proposals for books on Amazon, or films on Netflix, are some of the good and genuine samples of the operation of enterprise-quality recommender frameworks. The configuration of such suggestion motors relies on upon the space and the specific qualities of the information accessible. For instance, film watchers on Netflix every now and again give appraisals on a size of 1 (despised) to 5 (enjoyed). Such an information source records the nature of communications in the middle of clients and things. Also, the framework may have admittance to client particular and thing particular profile traits, for example, demographics and item depictions separately. Recommender frameworks contrast in the way they investigate these information sources to create ideas of liking in the middle of clients and things which can be utilized to recognize all around coordinated sets. Synergistic Filtering frameworks break down verifiable associations alone, while Content-construct Filtering frameworks are situated

in light of profile traits; and Hybrid strategies endeavor to join both of these plans. The structural engineering of recommender frameworks and their assessment on true issues is a dynamic region of exploration.[1]

The nature of the proposals importantly affects the client's future shopping conduct. Poor suggestions can bring about two sorts of trademark mistakes: false negatives, which are items that are not prescribed, however the client might want them, and false positives, which are items that are suggested, however the client does not care for them. In an e-business environment, the most essential blunders to maintain a strategic distance from are false positives, in light of the fact that these mistakes will prompt furious clients and consequently they will be dissimilar to return to the site (Sarwar et al., 2000). On the off chance that we attempt to discover clients why should likely purchase prescribed items and prescribe items to just them, that could be an answer for maintain a strategic distance from the bogus positives of the poor proposal.[2] In this paper, we have proposed a personalized recommender engine based on web usage mining.Also, CF and similarity models are used to minimize recommendation errors by making recommendation only for customers who are likely to buy recommended products. Optional features includes: Auto Language switcher which will switch the contents language to users native language in order to cherish the user-centric experience. Stock notification will trigger an email to respective user with product availability alert and list of recommended product. Such integration will enhance the user experience and will eventually be helpful to store owner which results in up-sell and cross-sell. We begin by reviewing related contents and Literature survey in Section 2. Section 3 contains literature survey about different recommednation methodologies.It also contains proposed system and their results. Similarly,Section 4 contains the brief overview about churn,churn detction and various methodologies along with proposed methodology to detect as well as prevent churning.Finally in section 5,our contributions and future work are summarized.

1.1 Motivation

The recent hike in e-shops have highlighted the need of cart management, stock management, stock notification, churn prediction, recommendation system etc. apps in order to provide ease of use and manage the customers.

1. Embrace recommendation system Suggest most related products to customers in

order to ease and enhance their viewing/searching experience and ultimately leading to up-sales. For example : It is difficult to track down each client and send them email individually with list of personalized recommended products.

2. Harness Churn prediction (customer abandonment) Client agitate alludes to when a client (player, supporter, client, and so on.) stops his or her association with an organization. Online organizations normally regard a client as stirred once a specific measure of time has slipped by since the client's last cooperation with the webpage or administration. The full cost of client agitate incorporates both lost income and the advertising expenses included with supplanting those clients with new ones. Diminishing client stir is a key business objective of each online business.

   For instance: Send exclusive coupon/offers to customer to keep them from beat.

   The capacity to anticipate that a specific client is at a high danger of stirring, while there is still time to make a move, speaks to an enormous extra potential income hotspot for each online business. Other than the immediate loss of income that outcomes from a client forsaking the business, the expenses of at first getting that client might not have as of now been secured by the client's spending to date. (As it were, obtaining that client may have really been a losing speculation.) Furthermore, it is constantly more troublesome and costly to gain another client than it is to hold a current paying client.[3]

3. Special Affections Certain buyers turns to heroes (frequent buyers).Cherish them with special offers may result to up-sells.

   Various plug-ins and addons are available in the market which can resolve above stated issues, but either they are costly or they aren't provide much integration support. Therefore, we need a better e-commerce suite which can fill the voids and provide an optimum solution at free/low rate.

Chapter 2

Background

This chapter focuses on the general related work covering a few sub-areas of the recommendation systems and churn detection. All related work described in each of the following chapters is discussed in those respective chapters

2.1 Recommender System

E-commerce websites help customers to find interesting items from huge data available over the Internet. These websites use Recommender systems to find relevant items from the large number of choices.

Recommender framework can be seen as one of the applications to predict rating or preference for the items that have not been seen by a user. This framework can also enlist a number of recommendations. The view is not just limited to this but can also include some predictions or forecasts that help users in making appropriate decisions [4].

Recommedner systems are generally classified into three categories as stated below :[4, 5]:

1. Content-based Recommender: The user is recommended items similar to the ones the user preferred in the past.[6]

2. Collaborative Recommender: The user is recommended items that people with similar tastes and preferences liked in the past.[6]

3. Hybrid Recommender: These recommenders combine collaborative and content-based methods. These type of recommenders can also be learnt based on combination of usage and content data.[6]

Comparison between recommender methods:

1. User independence: CF needs other users rating (interest in certain item) in order to deliver the similarity between/among the users and then give the recommendations whereas CB just need to investigate the items and customer's profile for item suggestion.

2. Transparency: CF method provides recommendation to a user just because some other user pertains the same interest in product/taste whereas content based method will let you know why and what feature were considered while recommendation (tagging).

3. cold start: Collaborative filtering requires some sort of rating or weighing data so that it can recommend a product/entity to user whereas via content based method, new items can be recommended before being rated by a certain number of customers.

4. Over-specialization: CB method provides a limited amount of surprises, since it has to compare the features of user profile and items. A perfect CB filtering might suggest stereotype recommendations whereas CF systems doesn't require any sort of information of users or items to be machine-recognizable. A pure CF method utilize only ratings and do not require any additional information about users or items, thus it may comeup with surprise products

2.2 Churn Detection

'Churn' is a word derived from change and turn.Churn means termination of an agreement There exists three types of churn conditions:

1. active(voluntary) - User terminates his agreement and switch to another provider. Probable Reasons: Unsatisfaction with the quality of service,Costly compared to the other similar products of rival companies, no rewards for customer loyalty, improper aftersales support, privacy concerns,no continuity or faultresolution, etc.

2. Incidental(voluntary) - User terminates the agreement without considering the idea of switching to a rival company. Reasons may include conditional changes that makes a user reluctant to use the service, e.g. - financial problems, change of the geographical location of the user where company might not exist.

3. passive - organization terminates agreement itself. Voluntary churns are hard to predict. As rotational churn only explains a minor aspect of overall churn,it is a matter of concern to predict and taking adequate steps to control active churn. In order to control customers' voluntary lapse, it is equally important for the company to know who are the possible churners and why that particular user has chosen to leave the organization.

Menwhile, customer churn can be categorized likewise in three different groups:
totally churned : Agreement is socially terminated;
hidden churn : Agreement is not terminated, but the user is not frequently using store service since a certain span of time;
partially churned - Agreement is not terminated, but the user is not using store services at max possible but utilizing just only few fragments of it, and is rather constantly using rival companies services.

Churn Management Framework

A 5 stage model for creating a customer CMF has been identified [7, 8].

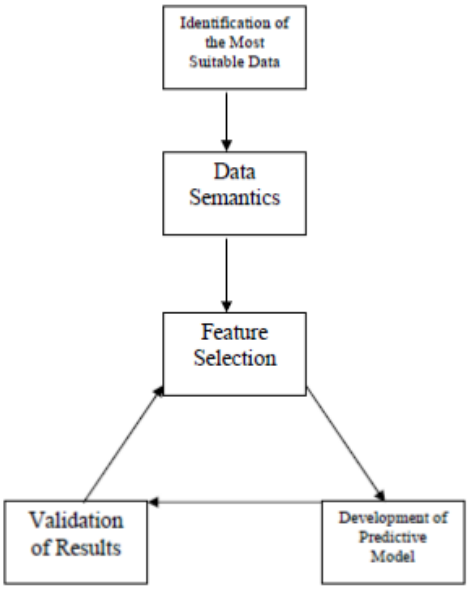


Figure 2.1: The Stages of CMF

The quality of the data necessary for prediction is an important factor. Which data held within the data warehouse would provide greatest accuracy for predicting customer churn? How much of the data should be used; i.e. should all historical data be considered or should the input be based on only a few of the most recent months[7]

Stage 2 is the problem of data semantics. Data semantics has been included in the model shown in Figure 3 because it has a direct relationship to stage 1, identification of the most suitable. In order to identify the most suitable there has to be a complete understanding of the data and the information each variable represents. Data quality is an important problem with many issues being directly related to data misinterpretation. Data semantics also covers representational heterogeneity and ontological heterogeneity. Representational heterogeneity understands the representation of variables. Similar variable names can have different values types associated with them.

Stage 3 covers feature selection. A definition for feature selection has been taken from Chen et al. (2008) who state, Feature selection is about finding helpful (important) features to depict an application area. Selecting applicable and enough features to viably speak to and list the given dataset is a critical errand to tackle the classification and clustering issues shrewdly[9].

Stage 4 is the development of a predictive model. Many models exist for determining the prediction of a desired event including statistical, classification and soft computing approaches.

The final stage involves validating the model to ensure that it is achieving an accurate prediction.

Chapter 3

Recommendation System

3.1 Literature Survey

3.1.1 Collaborative Filtering

CF work by collecting client feedback as evaluation for items in a given area and abusing similitudes in rating conducts amongst a few users and decide how to suggest a product. CF methods can be further sub-divided into neighbourhood-based and model-based approaches. Neighbourhood-based methods are also commonly referred to as memory based approaches.[1]

Neighbourhood-based Collaborative Filtering

In neighborhood-based methods, subsets of clients are picked in view of their comparability to the dynamic client, and a weighted mix of their appraisals is utilized to create expectations for this client. A large portion of these methodologies can be summed up by the calculation outlined in the accompanying steps:

1. Allot a weight to all clients as for comparability with the dynamic client.
2. Select k clients that have the most astounding comparability with the dynamic client normally called the area.
3. Process prediction from a weighted blend of the chose neighbors rodentings.[1]

User based similarity It is a collaborative filtering systems which uses rating similarity metric between users. For example news articles sites, build user profile based on

your past browsing history and map to particular user bucket. After that recommend news articles for you by computing user similarity metrics.[10]

The user-based recommender algorithm comes out of this intuition. Its a process of recommending items to some user. Algorithm structure is like it contains nested for loops in algorithm.The external circle basically considers each known thing (that the client has not as of now ex-squeezed an inclination for) as a contender for suggestion. The inner loop looks to some other client who has communicated an inclination for this competitor thing, and sees what their inclination esteem for it was. At last, the qualities are found the middle value of to think of an weightd average, that is. Every preferred value is weighted in the normal by how comparable that client is to the objective client. The more comparative a client, the all the more vigorously their preference value is weighted. It would be awfully ease back to inspect each thing. Truly, an area of most comparative clients is figured first and just things known not clients are considered: [11]

	Item 1	Item 2	Item 3	Item 4	Item 5
User1	$r_{1,1}$		$r_{1,3}$	$r_{1,4}$	
User2	$r_{2,1}$	$r_{2,2}$	$r_{2,3}$		$r_{2,5}$
User3		$r_{3,2}$		$r_{3,4}$	$r_{3,5}$
User4	$r_{4,1}$				$r_{4,5}$

Table 3.1: Rating of items given by user

The weight $w_{a,u}$ is computes the degree of similarity between the user u and the active user a. A well known method (Pearson correlation) has been used here to compute similarity coefficient between the ratings of the two users, defined below:

$$w_{a,u} = \frac{\sum_{i \epsilon I}(r_{a,i} - \overline{r}_a)(r_{u,i} - \overline{r}_u)}{\sqrt{\sum_{i \epsilon I}(r_{a,i} - \overline{r}_a)^2 \sum_{i \epsilon I}(r_{u,i} \overline{r}_u)^2}}$$

where I is the set of items rated by both users, $r_{u,i}$ is the rating given to item i by user u, and ru is the mean rating given by user u. Predictions are normally computed as the weighted average of deviations from the neighbours mean, as in:

$$p_{a,i} = \overline{r}_a + \frac{\sum_{u \epsilon K}(r_{u,i} - \overline{r}_u) * w_{a,u}}{\sum_{u \epsilon K} w_{a,u}}$$

The primary difference is that similar users are found first, before seeing what those most-similar users are interested in. Those items become the candidates for recommendation. The rest is the same. This is the standard user-based recommender algorithm, and the way its implemented in Mahout. [11]

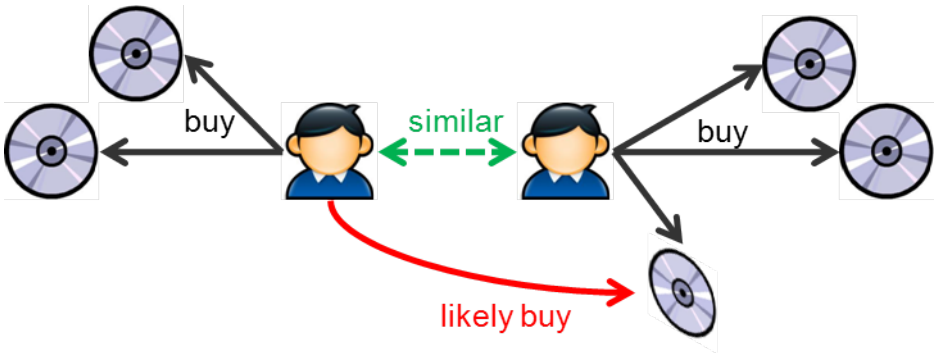


Figure 3.1: User based recommendation

Disadvanatages of User-based CF systems

Client based CF frameworks have been extremely effective in past, yet their broad use has uncovered some genuine difficulties, for example,

- Sparsity.

  Practically speaking, numerous business recommender frameworks are utilized to assess huge thing sets (e.g., Amazon.com prescribes books and CDnow.com suggests music collections). In these frameworks, even dynamic clients may have bought well under 1% of the things (1% of 2 million books is 20,000 books). Appropriately, a recommender framework taking into account nearest-neighbour calculations may be not able make any thing proposals for a specific client. Therefore the exactness of proposals may be poor. [12]

- Scalability.

  Nearest neighbour calculations require calculation that develops with both the quantity of clients and the quantity of things. With a great many clients and things, a run of the million webbased recommender framework running existing calculations will endure genuine versatility issues.[12]

Item based similarity

Itembased recommender is gotten from how comparative items are to items, rather than one user to other user.Suppose there are a larger number of clients than items, every item has a tendency to have more evaluations than every client, so an items normal rating typically doesn't change rapidly.[10][1]

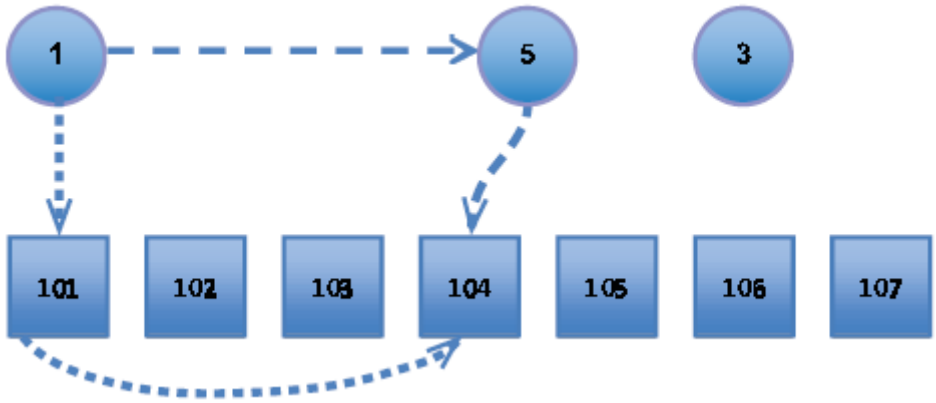


Figure 3.2: User based and item based recommendation

Algorithm indicates how its taking into account item likenesses, not client similitudes as illustrated before. The calculations are comparative, however not so much symmetric. They do have eminently diverse properties. For example, the running time of a item based recommender scales up as the quantity of item increments, while a user based recommenders running time goes up as the quantity of clients increments. [11]

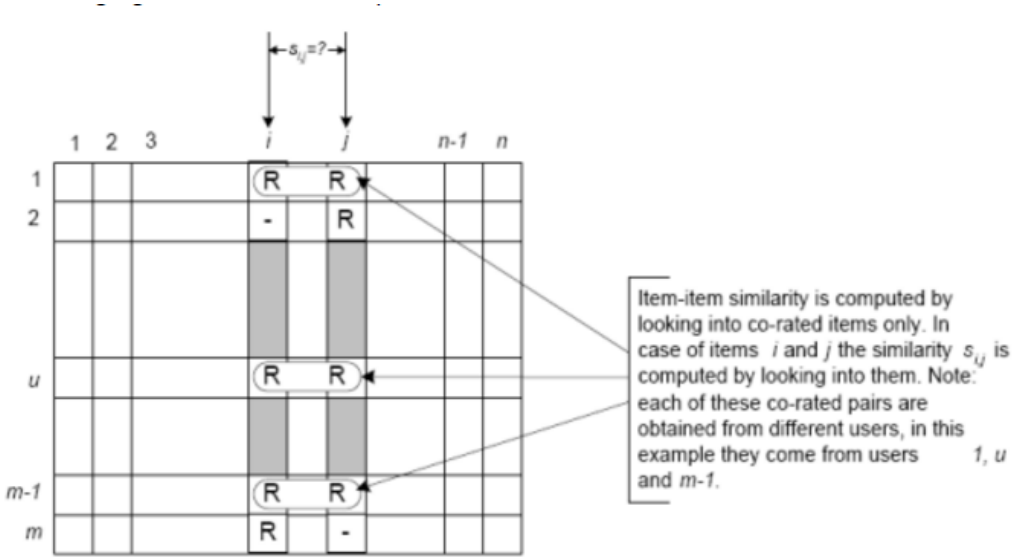


Figure 3.3: item based recommendation matrix

This proposes one reason that you may pick an item based recommender.If the quantity of items is moderately low contrasted with quantity of clients,the execution advantage could be astonishing [11]

Items are regularly less subject to change than clients.At the point when the items will be items,its sensible to expect that after some time,as you secure greater information,assesments of the likenessess among items will merge.There is no motivation to anticipate that them will change profoundly or every now and again.A portion of the same may be said of clients, however clients can change after some time and new learning of clients is liable to come in blasts of new data that must be processed rapidly.

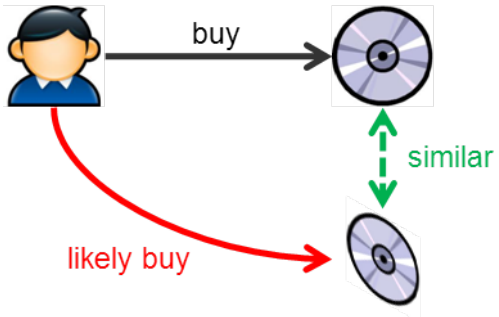


Figure 3.4: item based recommendation

In these methodology similitudes between sets of items i and j are calculated offline utilizing Pearson connection, given by:

$sim(i,j) = \frac{\sum_{u \epsilon U}(R_{u,i} - \overline{R_i})(R_{u,j} - \overline{R_j})}{\sqrt{\sum_{u \epsilon U}(R_{u,i} - \overline{R_i})^2}\sqrt{\sum_{u \epsilon U}(R_{u,j} - \overline{R_j})^2}}$

where U is the set of all users who have rated both items i and j, $r_{u,i}$ is the rating of user u on item i, and ri is the average rating of the ith item across users.

Presently, the rating for item i for client a can be anticipated utilizing a basic weighted normal, as in:

$p_{a,i} = \frac{\sum_{j \epsilon K} r_{a,j} w_{i,j}}{\sum_{j \epsilon K} |w_{i,j}|}$

where K is the area set of the k things appraised by a that are most like i. [1]

This could be alluring in settings where conveying suggestions rapidly at runtime is fundamental consider a news site that must conceivably convey suggestions instantly with every news article view.

The framework executes a model-finding so as to build stage the likeness between all sets of items. This likeness capacity can take numerous structures, for example, relationship between's evaluations alternately cosine of those rating vectors. [13]

For reasons unknown,processing item based CF has more advantage than figuring user-user similarity for the accompanying reasons:

- Number of items observed lesser than number of users
- While clients preferences may change after some time and henceforth the likeness framework should be redesigned more regular, item-item likeness has a tendency to be more steady and requires less upgrade. [14]

Model-based Collaborative Filtering

Model-based procedures give recommendations by assessing parameters of factual models for client evaluations. Model-based recommender frameworks include building a model taking into account the dataset of appraisals. As it were, we remove some data from the dataset, and utilize that as a "model" to make suggestions without using the complete dataset unfailingly. This methodology conceivably offers the advantages of both pace and versatility.[15][1]

Similarity matrix methods

Similarity matrix is utilized for discovering anticipated suggestions of a dynamic client. Similarity Matrix is a measure of likeness between a quantities of information focuses. Every component of the similarity matrix contains a measure of likeness between two of the data points.

	Item1	Item2	Item3	Item4	Similarity with User/Item
User1	$r_{1,1}$		$r_{1,3}$	$r_{1,4}$	X
User2	$r_{2,1}$	$r_{2,2}$	$r_{2,3}$		Y
User3		$r_{3,2}$		$r_{3,4}$	Z
User4	$r_{4,1}$				W

Table 3.2: Similarity values

There will be similarity values for all the items/users (depends on which type you have opt for user-user or item-item) will be getting and the highest/lowest (depends of type of similarity) value will be our 1st prediction.
There are different types of similarities to generate similarity matrix. Some of them are explain as follows.

Pearson Correlation Similarity

It gauges the propensity of the numbers to move together relatively, such that there's a generally direct relationship between the qualities in one arrangement and the other. At the point when this propensity is high, the relationship is near 1. At the point when there has all the earmarks of being little relationship by any stretch of the imagination, the quality is close to 0. At the point when there has all the earmarks of being a contradicting relationshipone arrangement's numbers are high precisely when the other arrangement's numbers are lowthe quality is close 1.
This idea, broadly utilized as a part of insights, can be connected to clients to quantify their closeness. It gauges the propensity of two users' inclination qualities to move togetherto be moderately high, or generally low, on the same items. [11]

	Item1	Item2	Item3	Item4	Correlation with User 1
User1	$r_{1,1}$		$r_{1,3}$	$r_{1,4}$	X
User2	$r_{2,1}$	$r_{2,2}$	$r_{2,3}$		Y
User3		$r_{3,2}$		$r_{3,4}$	Z
User4	$r_{4,1}$				W

Table 3.3: Correlation similarity values

Formula for correlation is [16]:
$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$ A table will be maintained for correlation values like:

Pearson Correlation Similarity weighted

The Pearson correlation doesnt reflect, specifically, the quantity of items over which its registered, and for our reasons, that would be valuable.At the point when taking into account more data,the subsequent relationship would be more dependable result.In order to reflect this, its expected to push positive correlated values toward 1.0 and negative toward 1.0 when the correlation depends on more items. Then again,you could envision the correlation values towards some mean preference values when the correlation depends on less items;the impact would be comparable, however usage would be more tricky because it would require the mean preference value for pairs of users (basically tracking the values). [17]

Euclidean Distance Similarity

This similarity matrix registers the Euclidean distance d between two such user focuses. This quality alone doesn't constitute a legitimate closeness metric, in light of the fact that bigger qualities would mean more far off and in this manner less comparative, users.

The quality ought to be smaller when clients are more comparable. Along these lines, the execution really returns 1/(1+d).when the separation is 0 (clients have indistinguishable preferencces) the outcome is 1, diminishing to 0 as d increments. This simiarity matrix stays away forever a negative quality, yet bigger values still mean more comparability.[11] To discover the separation between two clients taking after equation will be utilized $\sqrt{\sum_{i\epsilon items}(r_{u,i} - r_{v,i})^2}$

Where $r_{u,i}$ is the rating of user u and rv,i is the rating of user v

Tanimoto Coefficient Similarity

There are likewise User-Similarity executive that overlook preference values completely.They couldn't careless whether a client communicates a high or low preferencce for a itemjust that the client communicates an preference by any stretch of the imagination.Tanimoto Coefficient Similarity is one such usage,in light of (amazement) the Tanimoto Coefficient.This wualtiy is otherwise called the Jaccard coefficient,

Its the quality of items that two clients express some inclination for,isolated by the quantity of iems that wither client communicates some preference.

As such,it's the proportion of the extent of the crossing point to the span of the union of their favored things.It has the required properties:when two clients items totally cover,the outcome is 1.When they don't have anything in likeness,its output will be 0.Good thing is,it never results into a negative number.

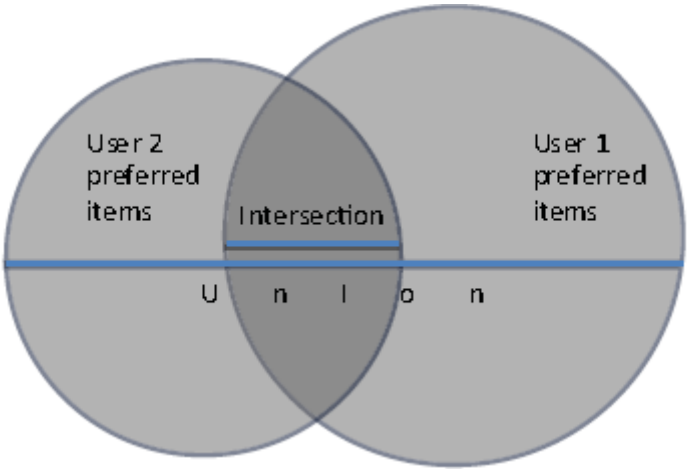


Figure 3.5: tanimoto ven daigram

3.2 Content based[1]

These methodologies suggest items that are comparable in substance to items the client has preferred previously, or coordinated to traits of the client.

Content based recommenders allude to such methodologies that give suggestions by looking at representations of substance portraying a thing to representations of substance that hobbies the client. These methodologies are now and again additionally alluded to as content based separating. This system concentrated on prescribing things with related literary substance, for example, website pages, books, and motion pictures; where the site pages themselves or related substance like portrayals and client surveys are accessible.

In that capacity, a few methodologies have regarded this issue as an Information Retrieval (IR) undertaking, where the content connected with the client's preference is dealt with as an inquiry, and the unrated records are scored with importance/likeness to this question.

Every evaluating classification are changed over into tf-idf word vectors, and afterward arrived at the midpoint of to get a model vector of every class for a client. To order another archive, it is contrasted and every model vector and given an anticipated rating taking into account the cosine likeness to every classification. [1]

3.3 Hybrid Approaches[1]

Keeping in mind the end goal to influence the qualities of content based and CF recommenders, there have been a few crossover methodologies suggested that consolidate the two. One straightforward methodology is to permit both to create separate ranked lists of suggestions, and afterward combine their outcomes to deliver a last rundown.

3.4 Proposed System

We have proposed a system which will consider only few of the previous viewed products in order to produce better matchup with his current mood and interest.

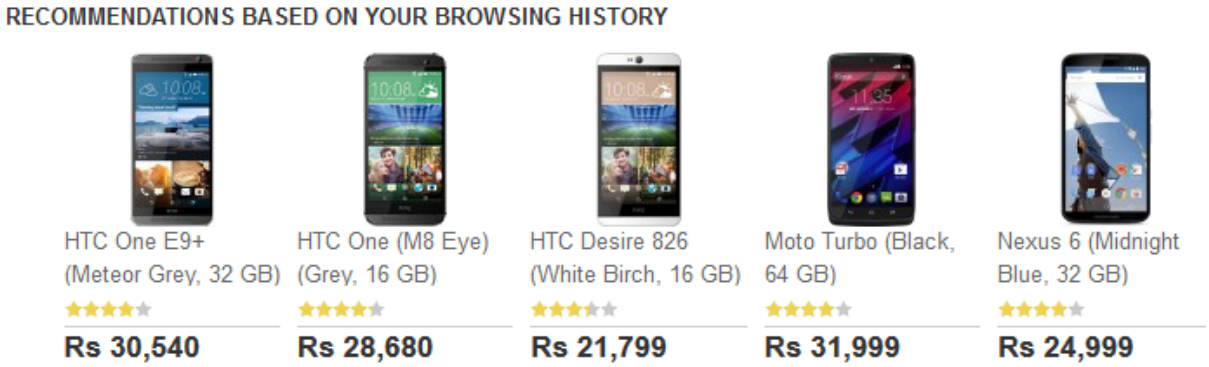


Figure 3.6: Existing System

3.4.1 Example Case

Consider a scenario where a user came to our site/service.For simplicity,let us consider that customer come to buy a blue jeans.

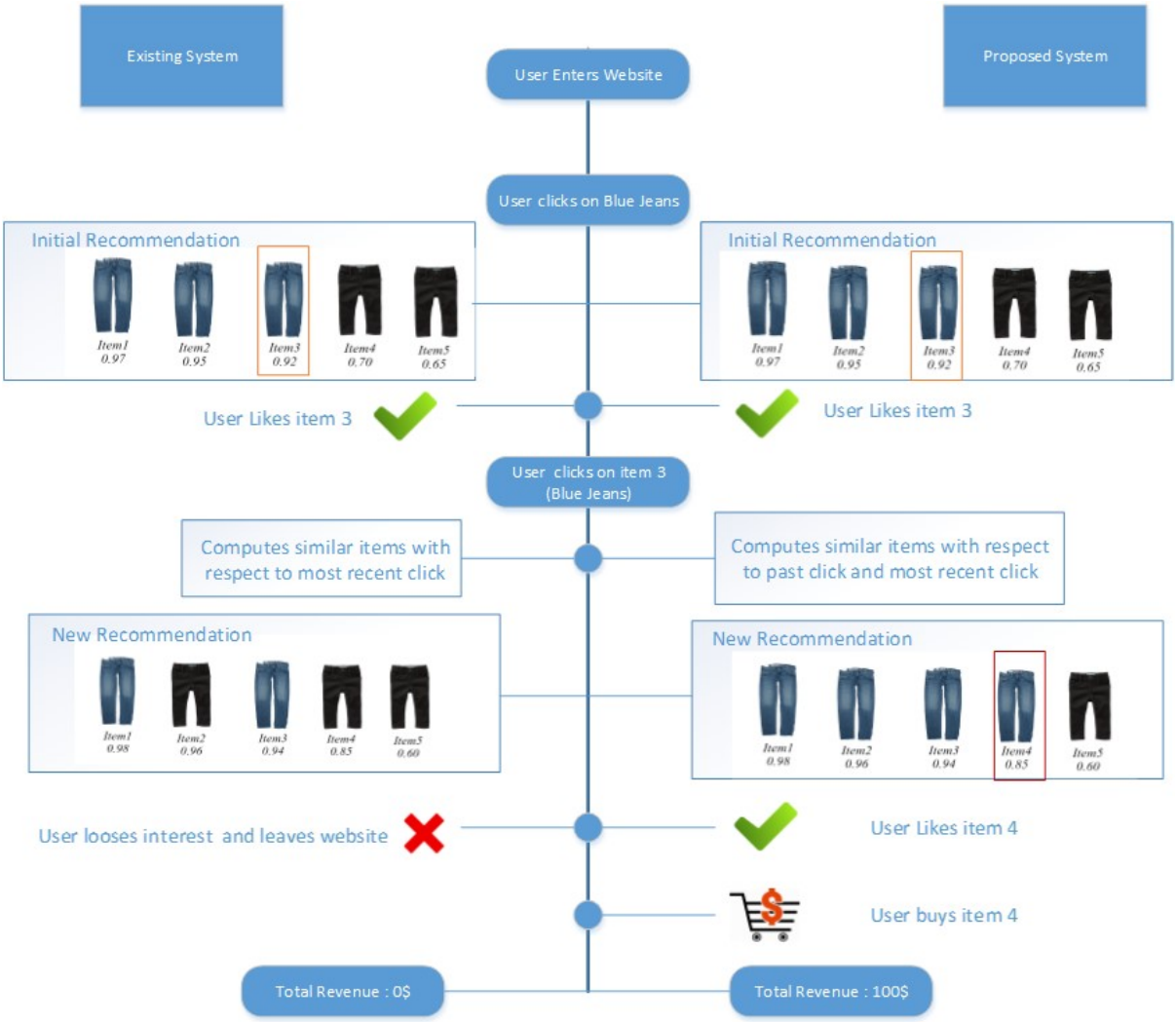


Figure 3.7: Existing vs Proposed System

figure 3.7 shows that a user comes to a site and clicks for blue jeans.Now,due to lack of proper recommendation,user losses interest and leaves without buying anything which sums to a total revenue of 0$.

Whereas in proposed system,user finds the item of his interest at third attempt in recommendation list and buys that item which sums to a total revenue of 100$ and a 1000$ worth customer.

3.4.2 System Approach

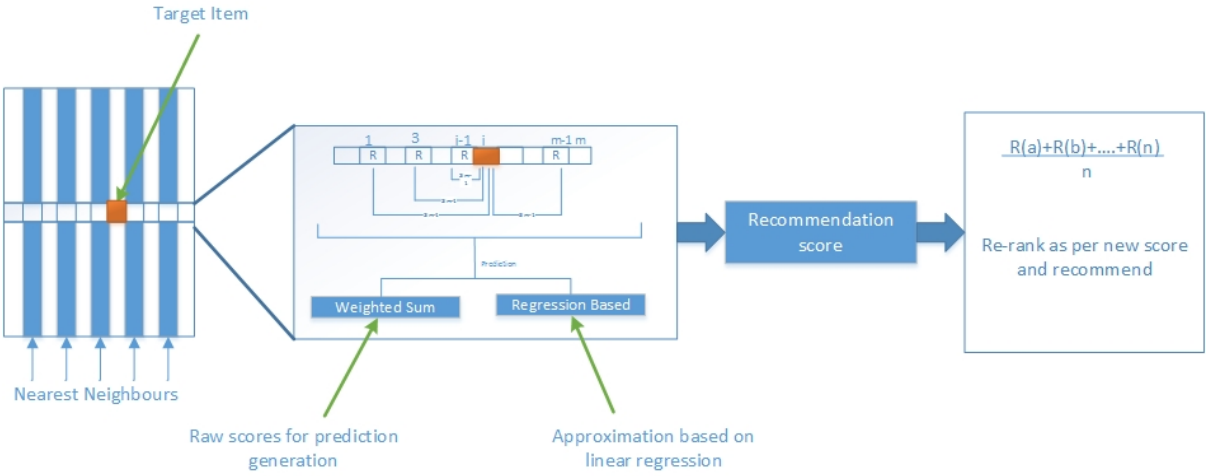


Figure 3.8: Proposed System

figure 3.8 shows the schematic flow of data processed to serve recommendation at user centric level.

Here, the process flow is started from calculating nearest-neighbours of target item.

Nearest Neighbour will draft the most similar items based user-item matrix and will forward it for further computation at similarity matrix generation.

Similarity Matrix will compute for item -item vector and will generate final vector of recommended score (List of nearest neighbours asked in algorithm..,will be discussed later).

At prediction section,linear regression is employed to tune model so as to reduce MAE.

Also,weighted sum is to aid the model to minimize the error and optimize the recommendation support.

After gaining recommended scores for an item,compute with previous Rscore vector (calculating average of both vectors) and produce a fused recommendation list.Repeat till k items (bag size) and deliver the final outcome.Sort the final outcome by their Rscore and populate rows accordingly.

Similarity Computation

One essential distinction between the similarity calculation in client based CF and item CF is that if there should arise an occurrence of client based CF the closeness is figured along the matrix rows however if there should arise an occurrence of the item CF the likeness is registered along the segments i.e., every pair in the co-rated set relates to an alternate client. Processing similarity utilizing fundamental cosine measure as a part of item based case has one imperative downside :the distinction in rating scale between diverse clients are not considered. The adjusted cosine similarity overcomes this disadvantage by subtracting the relating user average from every co-rated pair. Formally, the similarity between items i and j utilizing this plan is given by

$sim(i,j) = \frac{\sum_{u \epsilon U}(R_{u,i} - \overline{R_u})(R_{u,j} - \overline{R_u})}{\sqrt{\sum_{u \epsilon U}(R_{u,i} - \overline{R_u})^2}\sqrt{\sum_{u \epsilon U}(R_{u,j} - \overline{R_u})^2}}$

Here R_u is the average of the u-th users ratings.

prediction computation

When we disengage the arrangement of most comparative items in view of the similarity measures, the following step is to investigate the objective users ratings and utilize a procedure to get forecasts.

Therefore,we are using weighted sum approach.

As the name suggests, this system registers the forecast on a item i for a user u by figuring the total of the evaluations given by the user on the items like i . Each appraisals is weighted by the comparing comparability $s_{i,j}$ between items i and j . We can illustrate the prediction P as

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

Algorithm

Following is the proposed algorithm to implement above stated method.The algorithm returns a vector of newRscore for items which are filled in Bag B (items in which active user showed interest)

Algorithm 1 Proposed Algorithm

Consider a bag B which will handle users most recent interested items.

Suppose there is a user u.

User clicks on item a.

Now, $a \subset B$

Pick nearest 10 neighbours of item a(x1,x2,x3..,x10)

Rscore [size(B)] [n]

NewRscore[n] {Calculate Recommendation score using above mentioned similarity method.}

$j \leftarrow 1 : n$

for all j **do**

  for all item I in Bag B **do**

    Rscore [I][j] = rscore of item j with Item I in bag B;

    NewRscore[j]+=Rscore[I][j];

  end for

  NewRscore[j]=NewRscore[j]/Size(B);

end for

sort(NewRscore)

return *NewRscore*

Challenges

While processing through our proposed approach,we faced a fallback which was degrading the quality of recommendation engine.

1. Adjusted-cosine similarity measurement calculation

   The issue showed itself amid adjusted-cosine similarity computation, for the situation when there was one and only regular client between products. Since we subtract the normal rating for the client, the adjusted-cosine similarity for things with one and only regular client is 1, which is the most elevated conceivable worth. Thus, for such things, which are regular in our product database, the most comparable things wind up being just these things with one basic client. The arrangement we actualized was to indicate a base number of clients (consider 5 users) that two products needed in like manner before they could be called comparative.

3.5 Results

The snapshots of a sample run is shown below . The scenario considered while sample run was:

- shortlistted nearest 25 neighbours but shown only top-10 to user
- User shows interest in item 79
- From top-10 recommendations,user clicks on item 96
- user is looking for item 568
- Dataset used : Large product dataset

1. Existing System

- Existing system results following Recommendation Score
- User losses interest as he didnt find his product of interest.

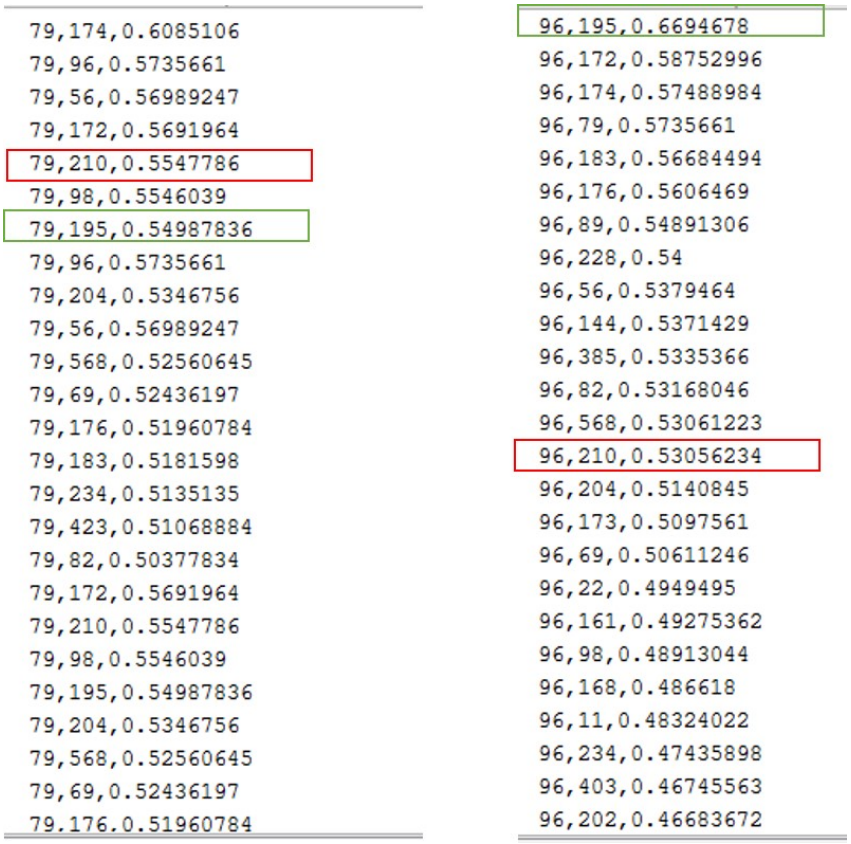


Figure 3.9: Existing System Results

2. Proposed System

- Proposed system computes relative Recommended score of last two items of interest and populates results
- User finds his product of interest at 8th recommendation(within 2 attemps).

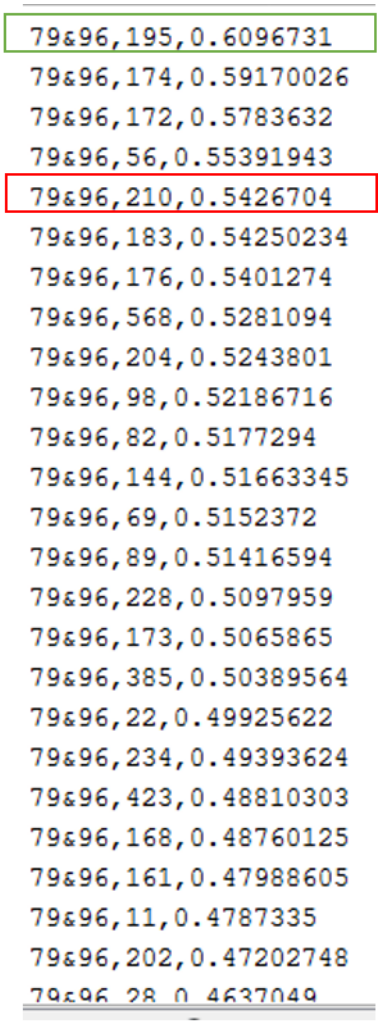


Figure 3.10: New System Results

Chapter 4

Churn Predictions

4.1 Introduction

4.1.1 What is churn?

A Customer, who is not interested in site/services or about to leave site/services.Customer churn alludes to when a client (player, supporter, client, and so on.) brings an end of his orher association with an organization. Online organizations regularly regard a client as churnedonce a specific measure of time has slipped by since the clients last communication withthe website or service.

4.1.2 What is churn prediction?

It consists of detecting customers who are likely to cancel a subscription to a service.[1]One of the easiest ways to keep the existing customers is to predict potential churn early and respond fast. Identify the signs of potential churn, understand customer wants and needs.

4.1.3 Importance of churn prediction

predictionThe capacity to anticipate that a specific client is at a high danger of stirring, while thereis still time to make a move, speaks to a gigantic extra potential revenuesource for each online business.Other than the immediate loss of income that outcomes from acustomer relinquishing the business, the expenses of at first procuring that client may nothave recoverd by the clients spending to date. (As such, acquiringthat client may have really been a losing speculation).Furthermore, it is always been more difficult and costly to procure another client than it is to hold a current one.

4.1.4 Why Customer begins to Churn?

Impossible to guarantee that all customers will stay forever Churn happens for a variety of reasons such as

- Customer can't see/get the worth suggestion of the item anymore
- Product or administration doesn't meets the desire or needs quality or features
- Product is great yet client administration is not
- Customer gains interest in competitor's product; Here, some events and variables are uncontrollable, such customers will be uncontrollable churns. But in our case most of cases we can handle like product price, product quantity, customer services etc.

4.2 Literature Survey

4.2.1 Decision Tree

graphical representation of conceivable answers for a choice based oncertain conditions. It's known as a choice tree since it begins with a solitary box (or root),which then branches off into various arrangements, much the same as a tree. [18]

A choice tree is a graphical representation of conceivable answers for a choice based oncertain conditions. It's known as a choice tree since it begins with a solitary box (or root),which then branches off into various arrangements, much the same as a tree. [18]Decision trees are valuable, not just in light of the fact that they are illustrations that help you s̈eeẅhatyou are considering, additionally on the grounds that to settle on a decision tree requires a deliberate anddocumented thought process.Limitation about our decision making is that we can just selectfrom the known options. Decision trees formalize the conceptualizing procedure sopotential arrangements can be distinguish by us more than ever.[18]

Decision tree will give you an exceptionally effective structure inside which you can lay outoptions and examine the conceivable results of picking those choices. They additionally helpyou to frame an adjusted scene of the dangers and reward connected with each possiblecourse of activity.[19]

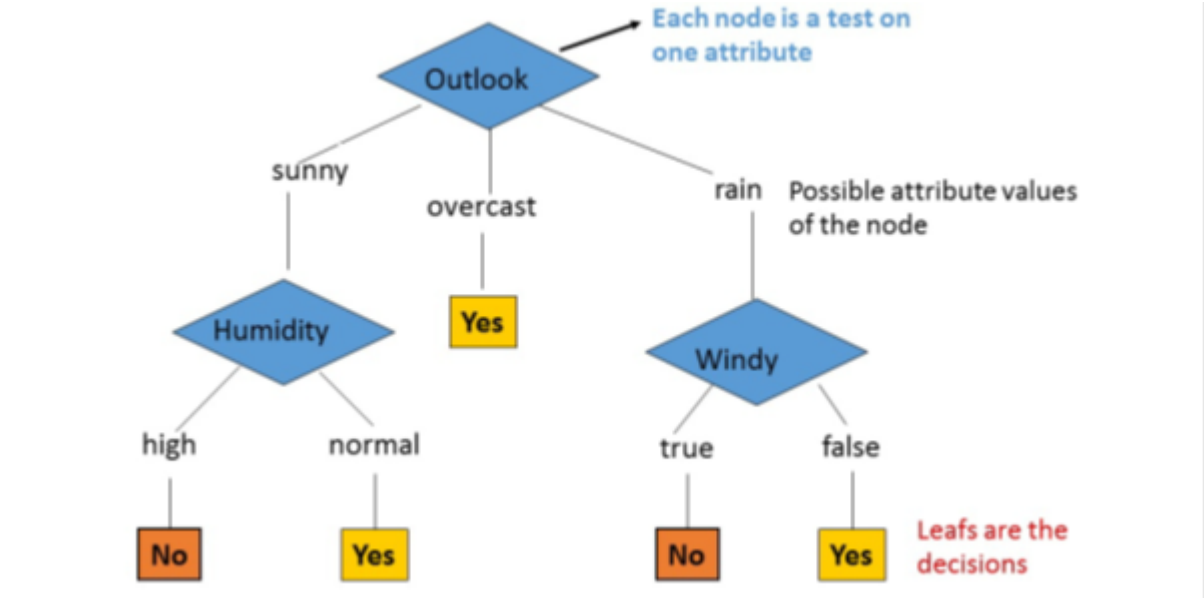
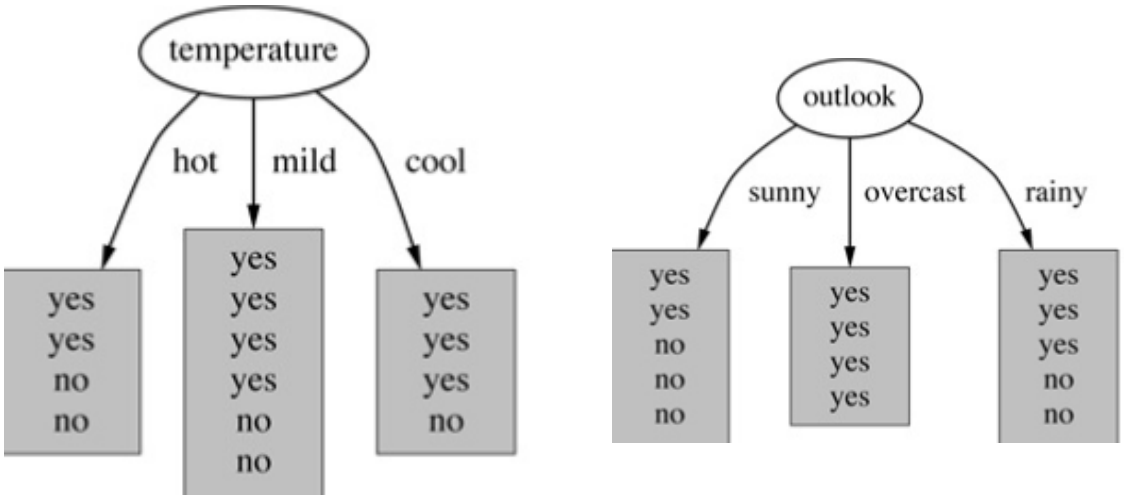
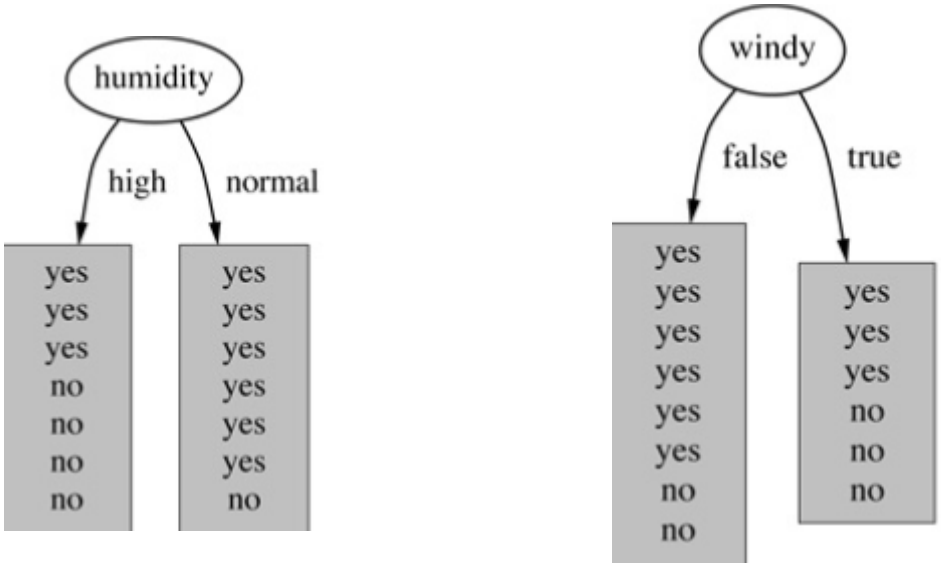


Figure 4.1: decision tree example

Every node in the tree determines a test of some property of the occurrence, and every branch diving from that node compares to one of the conceivable qualities for this trait. A case is characterized by beginning at the root node of the tree, testing the trait indicated by this node, then moving down the tree branchrelating to the estimation of the quality in the given illustration. This procedure is then rehashed for the sub tree established at the new node.[20]

Overall, decision trees speak to a disjunction of conjunctions of limitations on the characteristic estimations of examples.[20]





So as to characterize information gain definitely entropy, that portrays the immaculateness of a discretionary gathering of cases. Given an collectionm S, containing positive and negative case of some objective idea, the entropy of S in respect to this boolean order is

$Entropy(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$

where $p_{\oplus}$, signifies positive examples and $p_{\ominus}$, signifies negative examples.

One interpretation of entropy from information theory is that it specifies the minimum number of bits of information needed to encode the classification of an arbitrary member of S. If target label can take on c different values,then the entropy of S can be defined as

$Entropy(s) = \sum_{i=1}^{c} -p_i \log_2 p_i$

where p_i is the proportion of S belonging to class i.

Given entropy as a measure of the contamination in a gathering of preparing illustrations, we can now characterize a measure of the adequacy of a quality in arranging the training data. The measure we will utilize, called information gain, is basically the normal lessening in entropy brought on by dividing the case as per this trait. All the more unequivocally, the information gain, Gain(S, A) of a property A, with respect to a set of samples in S, is characterized as,

$Gain(S, A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$

the second term is the normal estimation of the entropy after S is divided utilizing property A. The normal entropy portrayed by this second term is essentially the whole of the entropies of every subset S_v, weighted by the part of cases $\frac{|S_v|}{|S|}$ that have a place with S_v. Gain(S, An) is consequently the normal lessening in entropy brought about by knowing

the estimation of attributeA.

The estimation of Gain(S, An) is the quantity of bits spared when encoding the objective estimation of a self-assertive individual from S, by knowing the estimation of A.

4.2.2 Random Forest

To comprehend and utilize the different choices, additional information about how they are figured is helpful. The majority of the alternatives rely on upon two information objects created by random forests.

When the preparation set for the present tree is drawn by testing with substitution, around 33% of the cases are let well enough alone for the example. This oob (out-of-pack) information is utilized to get a running impartial evaluation of the order blunder as trees are added to the forest. It is likewise used to get assessments of variable significance.[21]

After every tree is assembled, the greater part of the information are keep running down the tree, and vicinities are processed for every pair of cases. In the event that two cases possess the same terminal node, their vicinity is expanded by one. Toward the end of the run, the vicinities are standardized by isolating by the quantity of trees. Vicinities are utilized as a part of supplanting missing information, finding exceptions, and delivering lighting up low-dimensional perspectives of the information.[21]

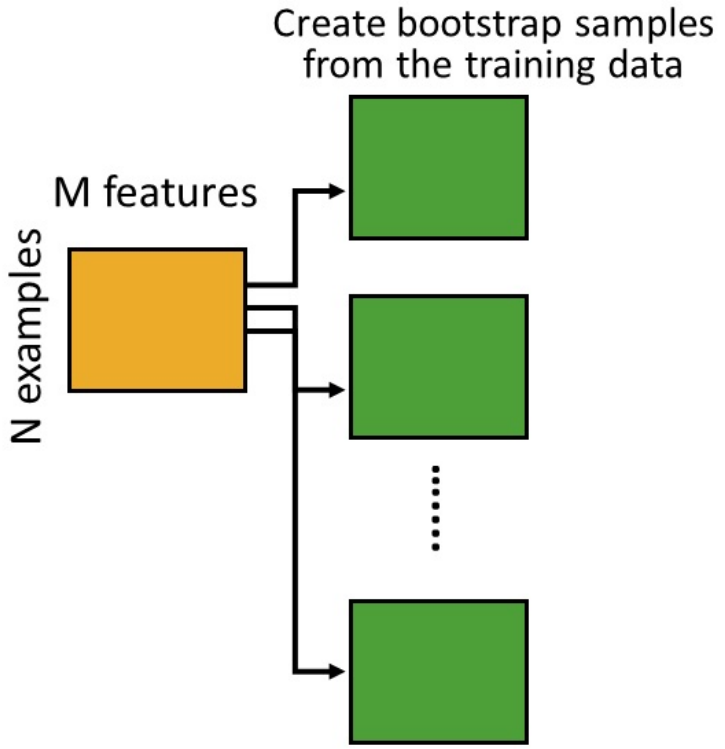
Features of Random Forests

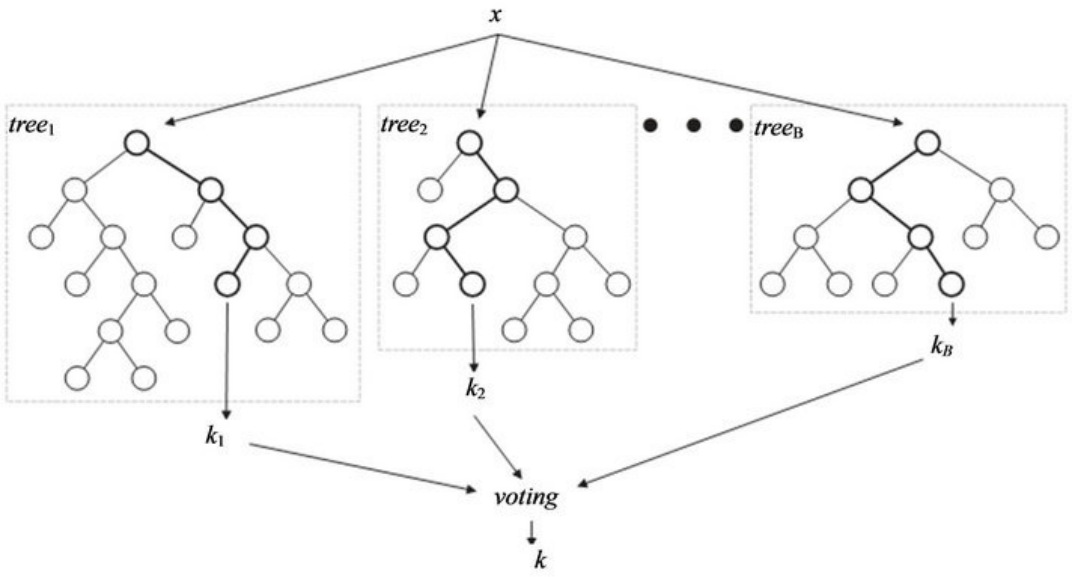
- It is unexcelled in precision among current calculations.
- It runs productively on expansive information bases.
- It can deal with a large number of input variables without variable clearance.
- It gives assessments of what variables are vital in the classification.
- It creates an interior unprejudiced appraisal of the errors as the forest building advances.
- It has a successful strategy for assessing missing information and keeps up exactness when a substantial extent of the information are absent.
- The abilities of the above can be reached out to unlabeled data, prompting unsupervised clustering and anomaly discovery.

- It grants a trial technique for identifying variable interactions.

Bagging

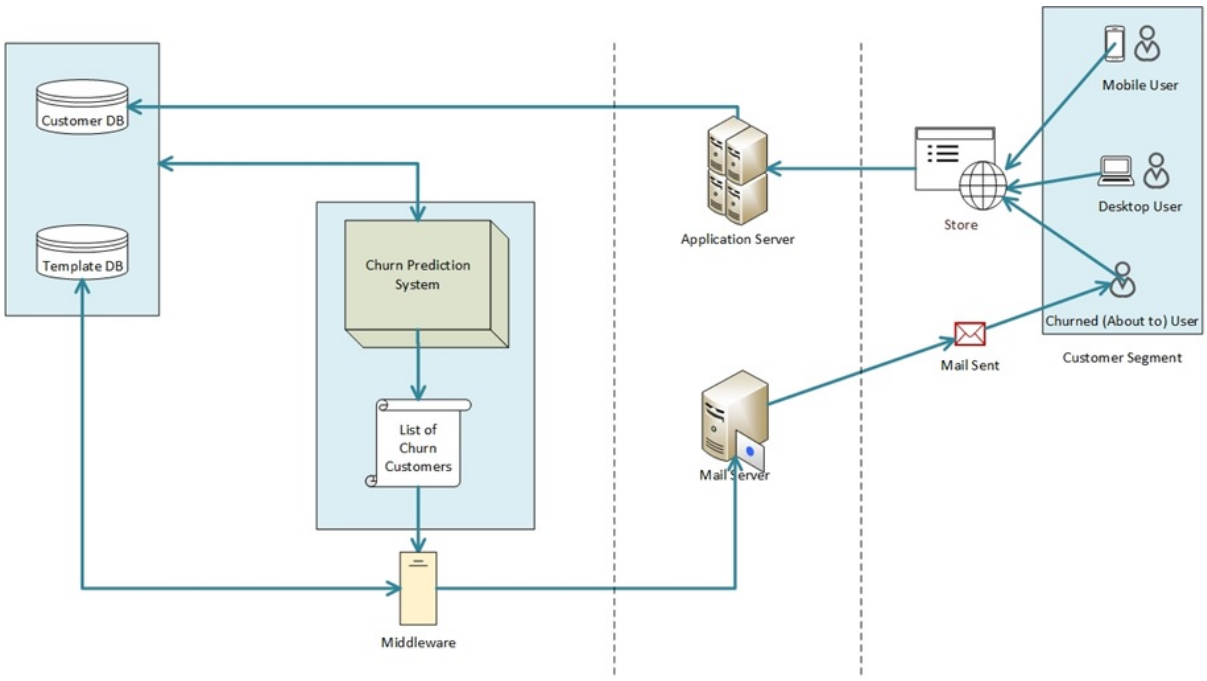
Bagging or bootstrap aggregationa strategy for decreasing the fluctuation of an expected preditionfunction. For classification, from a board of trees,everyone subtree make a choice for the predicted class.





4.3 Proposed System

4.3.1 System Architecture



4.3.2 Approach

The typical approach to solve churn detection is by using a large data set which contains several churning and non-churning customers. Above set is being analysed to develop a churn classifier. Such classifiers are constructed using,

- Regression Analysis
- Decision Trees
- Random Forest

We are considering random forest for churn detection model due to its benefits as stated above in section 4.2.2.

4.3.3 Finetune Model

Here,we are tuning two parameters to improve the prediction and hence accuracy of the model.

1. Number Of Trees

   Test conducted on GCM dataset

Dataset	Number of Trees											
	2	4	8	16	32	64	128	256	512	1024	2048	4096
Accuracy	0.72	0.77	0.83	0.87	0.89	0.91	0.91	0.92	0.92	0.92	0.93	0.93

   We concluded to put Number of Trees = [64,128]

2. Max Features

   Maximum number of features Random Forest is allowed to try in individual tree.

   We have concluded to use **sqrt(Max Features)**

   Above heuristic is based on empirical results [22]

4.3.4 Data Set

Here,we are considering two datasets: one is the purchased by MLVeda and other one is our own dataset(app data).

Dataset 1 (let us call it as GCM) contains 20 features and 20000 observations.Class label signifies a binary classification i.e,,0 or 1. It means if the user is suppose to be churn,its class labe would be 1 else 0.Entire dataset is available at bigML website [23]

Similarly,MLVeda datasset pertains 78 Observations and 8 features.Our app is still in beta phase,so a small number of records are only considered.The class label conatins binary values::Label 1: Churn Customer and Label 0: Not Churn. So,we'll start our analysis on GCM dataset and will finetune model considering GCM dataset only.

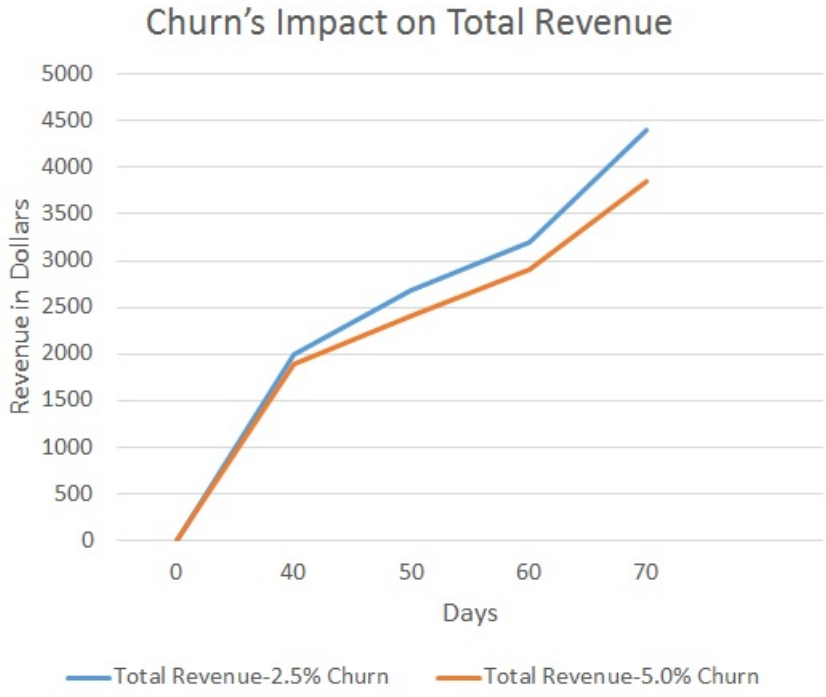
4.4 Results

```
UserNo.0 Label: 1
UserNo.1 Label: 0
UserNo.2 Label: 1
UserNo.3 Label: 0
UserNo.4 Label: 1
UserNo.5 Label: 1
UserNo.6 Label: 0
```

Figure 4.4: Output

	Last 1 Month				Last 3 Months					
Label	Number of Mail Open	Number of Page visit	Number of Added to Cart	Number of Order	Number of Mail Open	Number of Page visit	Number of Added to Cart	Number of Order	Total Order	Last visit
1	0	4	0	0	1	4	1	1	1	07-02-2016
1	0	0	0	0	12	20	2	3	4	09-09-2015
0	2	10	3	1	6	30	9	3	4	22-02-2016
0	4	20	5	3	4	20	5	3	7	20-02-2016
1	0	9	0	0	1	9	10	10	15	14-02-2016
1	0	6	0	0	0	6	2	2	3	04-02-2016
0	4	16	2	1	4	16	2	1	2	01-03-2016
0	5	25	4	1	10	50	8	2	3	29-02-2016
0	3	16	3	2	3	16	3	2	4	10-02-2016
1	0	3	0	0	0	9	1	1	1	09-02-2016
1	0	4	1	0	0	8	2	1	1	10-02-2016
0	3	10	3	1	9	100	12	10	15	01-03-2016

Table 4.1: Test DataSet



Above graph illustrates the comparison between two states of a store.Here,blue line resembles the system without churn control and orange line resembles a system with churn control.

- Total Customers : 78
- Overall churn rate :5% =4customers
- Customer Conversion from churn to non-churn:50% =(2.5%) = 2 customers
- Revenue :$4400 vs $3850
- Therefore, system helped store to generate $550 + retain 2 customers

Chapter 5

Conclusion

E-commerce store do need data analytics tools in order to outgrow his sales and maintain a healthy relation between store and clients. To resolve such issues, Recommender System and churn prediction algortihms are discussed and developed. In order to ensure the application of above methods, the comparison table was shown which indicated the gigantic difference between a system pertaining both tools and a system without them. Summary of work done, conclusions derived therein and possible future work are mentioned herewith in detail.

Recommender system, the items which are viewed/purchased by the user are the labelled examples about the users' preference. However, there are many other items which are not rated by the user. These items form the set of unlabelled examples. One of the major problems with recommender system is that, if it does not have sufficient number of labelled examples for the user (for whom recommendations are to be made), it may not be adequately accurate and useful.

Current Recommendation Systems either recommends considering only single previous item viewed or considering whole browsing history of user which may contain some products in which he/she isnt interested anymore.

Hence, a method is proposed using item-item collaborative filtering to consider previous three products in order to produce relevant recommendations. Eventually, itll produce recommendations as per the users current interest.

This can be improved by considering users interest in current product (measuring how long a user stays on a product page).Results will further enhance the ranking of recommendations subsequently.

Retention of conceivably churning clients has risen to be as vital for administration suppliers as the procurement of new clients. High churn rates and considerable income misfortune because of churning have turned right churn expectation and counteractive action to an imperative business process. In spite of the fact that churn is unavoidable, it can be overseen and kept in satisfactory level.

Hence, a model backed by random forest is proposed. To further improve the performance, model is further tuned by restricting number of trees and introducing Max features (Maximum number of features Random Forest is allowed to try in individual tree).

Further, a mailing system is plugged in to offer churn control. Itll trigger email as per the feedback(s) provided by churn detection model.

Bibliography

[1] P. Melville and V. Sindhwani, "Recommender systems," *IBM T.J. Watson Research Center, Yorktown Heights, NY 10598*.

[2] "A personalized recommender system based on web usage mining and decision tree induction," *Applicat 2002*, pp. 329–342.

[3] M. Solutions, "Customer churn prediction and prevention." http://www.optimove.com/learning-center/customer-churn-prediction-and-prevention, 2014.

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.," *IEEE Transactions*, pp. 734–749, 2005.

[5] M. Balabanovi c and Y. Shoham, "content-based, collaborative recommendation.," *ACM*, 1997.

[6] R. I. B. B. Koutrika, Georgia and H. Garcia-Molina, "Flexible recommendations over rich data," *Proceedings of the 2008 ACM Conference on Recommender Systems - RecSys*, 2008.

[7] M. B. M. D. R. Datta, P. and B. Li, "Automated cellular modeling and prediction on a large scale, issues on the application of data mining," *ACM*, pp. 485–502, 2001.

[8] R. R. J. Hadden, A. Tiwari and D. Ruta, "A computer assisted customer churn management: State-of-the-art and future trends," *Computers and OperationsResearch*, pp. 2902–2917, 2007.

[9] S. T. Chen, W. and T. Hong, "An efficient bit-based feature selection method," *Expert Systems with Applications*, 2008.

[10] H. Krishna, "item similarity vs user similarity." http://self-learning-java-tutorial.blogspot.in/2015/09/item-similarity-vs-user-similarity.html, 2015.

[11] R. A. Ted Dunning, Sean Owen, *Mahuot in action*. Shelter Island, NY 11964: Manning Publications, 2012.

[12] B. M. Sarwar, "Challenges of user-based collaborative filtering algorithms." http://www10.org/cdrom/papers/519/node9.html, 2001.

[13] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the tenth international conference on World Wide Web - WWW 01*, Association for Computing Machinery (ACM), 2001.

[14] R. Ho, "Recommendation engine models." https://dzone.com/articles/recommendation-engine-models.html, 2015.

[15] D. Lew, "Model-based recommendation systems." http://www.cs.carleton.edu/cs_comps/0607/recommend/recommender/modelbased.html, 2007.

[16] W. M. Trochim, "Correlation." http://www.socialresearchmethods.net/kb/statcorr.php, 2006.

[17] A. S. Foundation, "Recommender documentation." https://mahout.apache.org/users/recommender/recommender-documentation.html, 2014.

[18] "Decision tree." http://study.com/academy/lesson/what-is-a-decision-tree-examples-advantages-role-in-management.html.

[19] "Decision tree." https://www.mindtools.com/dectree.html.

[20] T. Michel, *Machine Learning*. 1993.

[21] L. X. E.W.T. Ngai and D. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification,expert system with applications," *Elsevier*, pp. 2592–2602, 2009.

[22] L. X. E.W.T. Ngai and D. Chau, "Marco bertini,alberto del bimbo, andrea ferracan,francesco gelli ,daniele maddaluno,daniele pezzatini," *ACM*, pp. 13–18, 2013.

[23] "Gcm." https://bigml.com/dashboard/dataset/56d933cc200d5a1e8a0104a9.