CUSTOMER PROFILING TO IMPROVE SERVICE AND MANAGEMENT OF MOBILITY ON DEMAND (MOD) SYSTEMS

Kumarage Tharindu Sandaruwan Kumarage

(188009A)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2019

CUSTOMER PROFILING TO IMPROVE SERVICE AND MANAGEMENT OF MOBILITY ON DEMAND (MOD) SYSTEMS

Kumarage Tharindu Sandaruwan Kumarage

(188009A)

Thesis/Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Computer Science and Engineering

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2019

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters Dissertation under my supervision.

Name of the Supervisor: Dr. Charith Chithraranjan

Signature of the Supervisor: Date:

Name of the Supervisor: Dr. Amal Shehan Perera

Signature of the Supervisor: Date:

ACKNOWLEDGMENTS

First and foremost, I would like to extend my deepest gratitude to my MSc supervisors, Dr. Charith Chithraranjan and Dr. Amal Shehan Perera, for the invaluable guidance, understanding, and patience offered to me during the MSc project. Their presence as mentors was paramount for me to gain a wider breadth of research experience. Moreover, I would like to thank all the members of DataSEARCH, the knowledge I have managed to gather through their insightful feedback is immense and working with the DataSEARCH team has been truly inspirational.

I would also like to convey my special gratitude to my progress review panel, Prof. Sanath Jayasena, and Dr. Suba Fernando. Their valuable insights and supervision received at the progress reviews helped me to improve the quality of my research work extensively.

I would like to thank the entire staff of the Department of Computer Science and Engineering, both academic and non-academic for all their help during the course of this work and for providing me with any resources necessary to conduct my research.

This work was funded by Digital Mobility Solutions (Pvt) Ltd. a.k.a PickMe.

I would also like to thank PickMe for providing the unique and resourceful opportunity to collaborate with their system throughout this project for the evaluation of the proposed methodology. In addition, I am highly thankful for the support given by Mr. Asanka Perera of PickMe data science team, in laying the technical foundation to familiarize with the data of their system and continuous feedback on system changes. Without your dedicated involvement in the process of generating the required data, it would have been impossible to complete the project on time.

Finally, I am indebted to my family and friends for all their immense support offered in numerous ways during the period of my MSc.

ABSTRACT

A system which caters the mobility requirements/travel needs in real time with user demand is known as Mobility on Demand system (MoDS). Global companies like Uber, Lyft, and local company like PickMe can be considered as examples for a Mobility on Demand systems. With the prevailing rapid growth of these MoDS, there is an explosion in system data where massive amounts of information related to customer rides are gathered on a daily basis. Due to this enormous volume of data, there is a potential for exploiting data mining and machine learning technologies to make the service smart and improve the management functionalities of the system.

Even though there is a vast amount of data at hand, the lack of systematic modelling techniques in MoDS is delaying the businesses from achieving smart systems with improved and personalized services. However, when considering similar E-commerce systems, user profiling and segmentation can be identified as the foundation towards smart improved service and management. Hence it is crucial to form the necessary framework towards user profiling and segmentation in MoDS. Research work found in our work is two-fold. First, we introduce a systematic aggregated and anonymous analysis schemes towards user profiling and segmentation in MoDS. Starting from the feature extraction specific to the MoDS, a detailed methodology for building the profile vectors is defined by this work in the following sections.

Then consequently, we extended the methodology towards enabling recommender systems in MoDS in order to improve the service. Moreover, under the recommender system methodology, a novel deep Collaborative Filtering method is introduced, and evaluation results show that the new model is capable of outperforming the current state-of-the-art techniques for Collaborative Filtering. The outcome under the recommender system for MoD is a hybrid system which incorporates all the profile vectors built in the customer profiling phase. Evaluation of the overall recommender system with historical data shows a significant improvement in recommendations related to MoD services.

Keywords: Mobility on Demand systems (MoDS), Recommender Systems (RS), Collaborative Filtering (CF)

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
MoDS	Mobility on Demand System
RS	Recommender System
BDA	Big Data Analytics
RFM	Recency Frequency Monetary
POI	Point of Interest
CF	Collaborative Filtering
MF	Matrix Factorization
AE	Autoencoder
AAE	Adversarial Autoencoder
VAE	Variational Autoencoder

TABLE OF CONTENTS

Chapter 1

1 INTRODUCTION

With the trends of E-commerce, there is rapid growth in customer interactions towards various smart systems which provides them the services and products which fulfil their day to day work. Moreover, new smartphone technologies enable these systems to be at the user's fingertips. Within the last decade, a similar type of service had rapid growth in the global arena, which is Mobility on Demand systems (MoDS). A system which caters the mobility requirements/travel needs in real time with user demand is known as a Mobility on Demand system (MoDS). When considering the urban cities day to day mobility needs are quite extensive. Therefore, the mobility demand is increasing gradually which leads to high frequent utilization of MoD systems for the mobility needs of the urban population.

1.1 Problem Formation

Rapid growth in MoDS results in an explosion in human mobility/travel related data. MoDS collect their customer travel data on a daily basis thus creates a huge volume of data and consequently creates the potential of utilizing the collected data to improve the service and management. Collected data in MoDS hold information related to user trips such as pick up location, drop location, start time, end time, duration and fare that are captured daily for each trip handled through the system. The research problem we are addressing in this work is to identify different segments/groups of users in the customer base of a MoDS in terms of their mobility patterns and recommend services for such groups of users to suit their mobility needs.

In similar internet-based systems, big data analytics (BDA) is massively utilized in order to reinforce the customer sales and service utilization [1]. However, in the domain of MoDS, the problem of addressing service improvement through customer travel data is still poorly-analysed and yet to be explored in-depth as a research problem and lay the foundation for the future research on MoD BDA.

1.2 Research Objectives

The research objectives of this project can be outlined as below,

Prepare comprehensive algorithms and techniques to model the customer related data of mobility on demand systems in order to profile customers to support predictive management and service enhancement.

- Systematic modelling techniques to identify profile vectors of the users of MoDS based on attributes such as demographics and behavioural.
- Prepare customer segmentation models based on created user profile vectors.
- A comprehensive methodology for a recommender system in MoDS in order to enable personalized service recommendation

Aforementioned objectives can be further described as below.

1. Profiling of users - Know your customer

- As shown in Figure 1.1, we need to identify different attributes of users that can describe each user in the system. These attributes, will govern all the aggregated analytical schemes that will be executed on the data of MoDS.

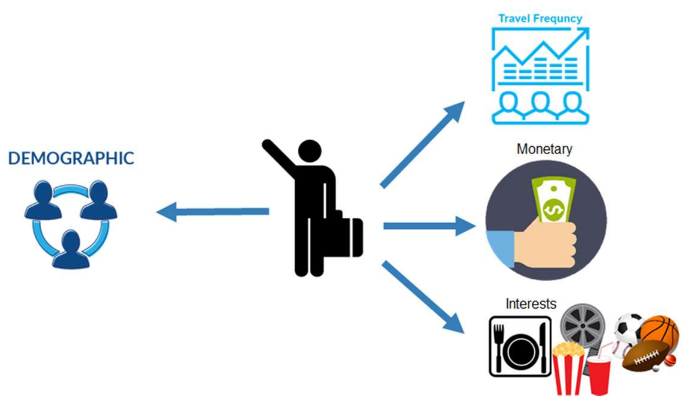

Figure 1.1: Know your customer

2. Segmentation of users - Know the group

- The objective is to identify various segments of the customer base as shown in Figure 1.2. A segment is not just an ordinary cluster of users but a cluster of users in the system, which adds a business advantage. Users are described in terms of profile vectors based on the selected attributes. Using these vectors, comprehensive models were deployed in order to identify the necessary segments within the user-base.

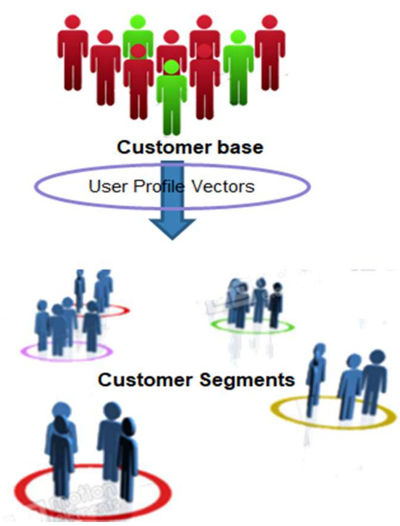

Figure 1.2: Customer segmentation

3. Recommendation of personalized services

- The objective is to lay the groundwork for personalized service recommendations in MoDS, which will help improve the user experience. Therefore a proper recommender system methodology for a MoDS based on the user profile vectors and segments is needed.

1.3 Project Contributions

Following are the contributions of this project towards the growth of the research community in the domain of Mobility on Demand systems.

- A comprehensive methodology for user profiling and segmentation in MoDS
- Novel deep Collaborative Filtering model for personalized service recommendation in MoDS

1.4 Project Scope and Limitations

1. This research work does not consider real-time streaming data processing and analytics, only a static pseudonymous aggregated analysis.
2. Recommender system evaluation is based on historical data collected at the system, live evaluation of the methodology could not be executed due to practical business constraints.

1.5 Organization

The upcoming sections of this thesis are organized as follows. Chapter 2 discusses related works in similar E-commerce systems and current research directions of the MoDS. Then the proposed systematic methodology is described in a detailed manner in Chapter 3 followed by the model evaluations and results on Chapter 4 along with result analysis and discussions. Finally, chapter 5 presents the conclusion of the thesis and future work and potential research directions in MoDS.

Chapter 2

2 LITERATURE SURVEY

With the trends of Big data, there is massive starvation for knowledge in E-commerce applications in various domains. Hence, more and more companies and service providers depend on machine learning techniques to gather valuable information which supports the strategic decision-making process. In this section, we discuss similar research work in various domains, including MoDS related to the research problem at hand. First, we discuss the related work on customer profiling and segmentation, which help understand the need for proper modelling techniques for enabling profiling and the segmentation in MoDS. Next, we discuss similar research work related to the problem of how to allow personalized services. In this subsection, we discuss the Recommender System methodologies thoroughly, which can be seen as the primary technique used by various other domains for enabling personalized services.

2.1 Customer Profiling and Segmentation

When considering the applications of data mining and machine learning in E-commerce applications and digital marketing customer profiling and segmentation is considered as the fundamental component [1, 2]. Various customer profiling based research has been carried out in digital marketing and E-commerce applications in several domains.

Trusov et al. [2] propose a user profiling approach that identifies individual user profiles from online system interaction data and allows online businesses to take predictive decisions and profile predictions when there's only implicit user information is available. Moreover, they have shown that these profiles later support the application of personalized advertising recommendations in online businesses. In recent work, Farias and Li [3] extend this idea of utilizing user side information in order to create generative models to generate user preference so that a personalized service and product recommendations are available in an E-commerce environment.

Another recent work on customer profiling can be seen in bank telemarketing by Palaniappan et al. [4]. In this work, they exploit the use of classification models for predicting customer profiles based on customer personal characteristics and spending behaviors and consequently decide the customers with the highest probability to accept the sales or offers provided via telemarketing. Classifications models such as Naïve Bayes classifier, Random Forest classifier, and Decision Tree classifier have shown good accuracy values in customer profile prediction and increasing telemarketing sales profit.

Kanagasabai et al. [5] incorporated deep learning models for customer profiling in telecommunication companies. They use mobile weblog data collected from the telecommunication companies in order to parameterize the profile vectors and then extract the aggregated insights on customer web behaviours which have the potential to be used for various industrial applications not limited to the telecommunication industry. Dursun and Caber [6] show that traditional RFM model can be exploited in order to profile hotel customer and decide the profitability order. The RFM model found in various business analytics platforms consists of three main user features which are namely Recency, frequency, and monetary aspects.

2.1.1 Clustering Models for Customer-base

Kashwan and Velu [7] present a tool which is capable of automatically identify the segments of the customer-base by using unsupervised clustering on sales data records of supermarkets in real time. In this research, they have used the k-means clustering technique to develop this online system for the supermarkets in order to make predictions on sales in various annual seasonal cycles.

Güçdemir et al. [8] exploit the clustering and multi-criteria decision making (MCDM) concepts in order to identify the customer segments in businesses. There proposed method is two folds as below.

1. Identify proper segmentation variables: They have extended on the RFM model and added five novel segmentation variables into the model.
2. Grouping the customers: This is the segmentation phase where they utilize three hierarchical clustering algorithms (Ward's method, single linkage, and

complete linkage) and one partition clustering algorithm (k-means) to identify the customer segments.

These clustering based user segmentation can be seen in the fashion industry as well. Brito et al. [9] present a clustering and rule-based techniques to identify the subgroups of the customer base in order to predict customer preferences and make the products more customized.

2.1.2 Demographic Features based Customer Classification

Apart from the user system interaction and behavioural attributes demographic features are also often taken into account in various domains in order to identify different groups of customers based on demographics. However, customers often show a tendency of not to share their accurate demographics such as age and gender with the system. Thus it is quite important to have predictive models for customer demographics prediction.

Research on demographic prediction can be seen in various fields of applications. Duc et al. [10] propose a classification model for predicting the gender Demographic of customers based on their catalog viewing data on e-commerce systems, such as the date and time of access, the products viewed, etc. They have gained about balanced accuracy of 81% by utilizing BayesNet SVM classifiers with added supporting techniques such as cost-sensitive learning, sampling and boosting techniques.

Another similar work is done on real-world mobile data where user mobile phone usage patterns were exploited in to predict demographic characteristics such as age, gender, marital status [11]. KNN, Radial Basis Function Network, Random Forest techniques have been used for the classification task and 83.33% accuracy for gender classification using Random Forest and 59.49% for age classification was recorded. Shrestha et al. [12] show that the age and gender prediction of health forum users can be done by creating an author profiling based on forum posts replies etc. and then utilizing a Logistic regression model on top of the profile data with data resampling to rectify the class imbalance problem. They have recorded a 65.59% accuracy for age classification and 88.41% accuracy for the gender classification through classification model they propose.

2.1.3 Travel Data based Customer Classification

To the best of our knowledge, this work is the first comprehensive study on all the steps of user profiling and segmentation in MoDS, However, there are various work on user profiling based on collected user travel data. A comprehensive study can be found on user profiling and segmentation of London's public transport users, using an extensive database of Oyster Card transactions. They have extracted numerous travel behavioural attributes related to the user's temporal and spatial variability of the mobility, activity, demographic features, and mode choices in order to identify homogeneous clusters and segments of users. In recent work, Zhang and Cheng [13] have presented a comprehensive list of features which can be extracted from London Oyster card data to measure the spatial, temporal, and mode choice behaviours, which could give a better understanding of the long-term travel patterns of the users.

One of the state-of-the-art activity mining models was introduced by Phithakkitnukoon et al. [15] on their work on Activity-Aware Map creation. They capture the static activity distribution of different areas based on landmarks and Points-of-Interest (POIs) and then by aggregating human mobility in these areas, a model is created for activity probabilities. This model is then utilized to capture the individual daily activity pattern and analyse the correlations among different people's work and living areas.

2.2 Recommender Systems

After the user profile vectors and segments are created the next task is for the personalized service and product management in the system. Recommender systems are a vital component in enabling personalized service in any E-commerce system.

One of the main objectives of an RS is to solve the customer over-choice problem [16]. Customer over choice can occur in any business or commercial system if there are an extensive amount of available item choices. Nowadays RSs are utilized in many global systems, for example, video hubs such as YouTube, E-commerce companies [17] like Amazon and eBay, online movie streaming giants like Netflix and Music service providers like Spotify, etc.

Current RS methods can be classified into three main categories namely Collaborative Filtering (CF), Content-Base (CB) and Demographic recommender systems [18]. Out of the aforementioned methods, CF-based RSs have shown outstanding performance in many recommender tasks [19, 20]. Recently in most of the implementations of recommender systems, these three methodologies are overlapping each other and unified in order to create hybrid RSs.

Again the CF-based RSs can be categorized into memory-based and model-based systems. Both these systems utilize the user-item ranking/interaction matrix which is considered as the foundation for CF models. Recorded rankings or interactions between users and items can be either explicit such as direct user preference or rating or it can be implicit such as recorded historical usage of a given item by the customers. Here memory-based CF models utilize the user-item matrix in order to establish the neighbourhood of a given user and based on the neighbourhood the recommendation and prediction are performed [21]. Even though memory-based CFs are easy to implement in large scale systems loading user-item matrix can be high time and memory consuming. Therefore, many CF based RSs tend to be model-based systems where linear algebraic and machine learning models are created beforehand based on the user-item matrix and then the created model is used for real-time prediction of ratings/interactions and provide recommendations [20].

There are various memory-based CF models can be found in the current literature. Matrix Factorization based algorithms are considered as a set of state-of-the-art memory-based CF models [22, 23]. Apart from the matrix factorization-based CF models, various machine learning based models have been applied in model-based CF as well. Clustering based CF [24], Bayesian network based CF [25], and deep CF [16] which exploit deep learning models are few of the examples for machine learning memory-based CF.

2.2.1 Deep Collaborative Filtering Models

Deep learning based CF models, which is also known as Deep CF is one of the prominent research areas in the RS domain [16]. The main idea behind utilizing deep learning for CF is to learn the latent model accurately and predict the unseen user-item ratings or interactions [26]. Given a large scale dataset of user-item ratings/interactions deep learning based models have shown outstanding prediction accuracy and also these models tend to perform well in the case of sparse user-item matrices.

In the early stages, Restricted Boltzmann Machine was used to create a Matrix Factorization model ensembles for a hybrid CF model and this model has performed excellently with Netflix data and present recommendations [27]. CNN and RNNs were also used for deep latent feature learning in later research work related to deep CF [28, 29]. When considering the latest deep CF research work, applications of deep generative networks is quite significant.

The earliest form of deep generative networks used for CF was Autoencoders [16, 26]. Utilizing autoencoders generative capabilities, prediction of unseen user-item ratings was made easy. However, by utilizing denoising criteria on autoencoders more robust CF models have been created for CF [30]. With the introduction of Variational Autoencoders (VAE) more accurate stochastic latent representations of the user-item matrix were learned and as a result, VAE tends to outperform the traditional Autoencoder models at the task of CF [31]. Generative Adversarial Networks (GAN) has also been utilized in deep CF. GAN based CF model has performed well in a personalized citation recommender system where it was employed in order to generate a heterogeneous bibliographic latent representation [32].

2.2.2 Adversarial Learning into Collaborative Filtering

Introduction of adversarial learning for deep generative networks revolutionized the solution models in various problem domains such as image recognition, voice recognition, and anomaly detection. The main reason behind this high performance was the precision added to the generative power of these networks by adversarial learning. Thus, adversarial learning is a viable solution for deep CF and also address the problem of data sparsity [16].

He et al. work shows that Adversarial learning has been incorporated into Bayesian models and gained high performance in personalized ranking for recommender tasks [33]. Moreover, GAN based CF model was exploited in order to successfully recommend heterogeneous citations to users [32]. Furthermore, adversarial learning has been utilized in recommender systems which worktop of streaming data from social media and E-commerce platforms [34].

However, there is one of the emerging generative networks known as Adversarial Autoencoders (AAE) yet be exploited for deep CF. This model unifies adversarial learning into the VAE model [35], therefore, forming a generative network which is capable of learning the latent model more accurately than the VAE itself.

2.3 Summary

When considering the customer profiling need of the MoDS, it is clear that the traditional RFM model found in other E-commerce platforms is not adequate to explain the mobility patterns of the customers. Therefore, we need to create a comprehensive feature-set by integrating features under temporal, spatial, and mode preference categories and also human travel activity features that can be found in mobility pattern mining research [13, 15]. Moreover, incorporating demographic features such as age and gender is also essential for the completeness of the profiling in MoDS. Supervised classification models such as SVM and Boosting algorithms are performing well in different domains for the demographic prediction of the system users [10].

When looking into the customer segmentation, clustering techniques such as K-Means is quite often applied by E-commerce platforms. However, when considering the research domain of data clustering models, most of the improved methods such as model-based clustering and also density-based clustering is yet to be employed in customer segmentation [7].

After a thorough analysis, Recommender Systems (RS) were identified as the key component in enabling personalized services to the customer. Under the RS research domain, Collaborative Filtering (CF) based recommender systems are performing well than the Content-based and Demographic-based recommender systems [22]. Out of the prediction models used in CF recommender systems, deep CF models can be identified as the new frontier. However, the deep Collaborative Filtering models are still open for improvement by employing novel deep generative models which combine adversarial learning. In recent research under the Recommender System problem domain, hybrid Collaborative Filtering models, which includes the side information related to the content-based and demographic-based recommender systems, are highly valued due to their high performance [26].

Chapter 3

3 METHODOLOGY

This chapter provides a comprehensive step-by-step description of the proposed methodology. The subsections are divided according to the following order. Section 3.1 discusses the user profiling and profile vector building in MoDS followed by user segmentation models in subsection 3.2. Finally, subsection 3.3 presents the methodology for enabling the personalized services in MoDS.

3.1 User Profiling

3.1.1 Overview

In this research, we identified a set of common profiling criteria that can be utilized in building the profile vectors for users of any general MoDS.

1. System Interaction Profile - Build a vector V_I for each user in the system where V_I is defined as below.
    - For each user U_j there exist vector $V_Ij = \{i_1, i_2 \ldots i_n\}$ where $i_1, i_2 \ldots i_n$ are attributes that can be extracted from users trips in the system.
2. Activity Profile - Build a vector V_A for each user in the system where V_D is defined as below.
    - For each user U_j there exist vector $V_Aj = \{a_1, a_2 \ldots a_n\}$ where $a_1, a_2 \ldots a_n$ are various activities a user might be travelling to complete.
3. Demographic Profile - Build a vector V_D for each user in the system where V_D is defined as below.
    - For each user U_j there exist vector $V_Dj = \{d_1, d_2 \ldots d_n\}$ where $d_1, d_2 \ldots d_n$ are demographic attributes.

3.1.2 System Interaction Profile

In MoDS, all the details related to the user trips are recorded in the system. By aggregating all these details we can build a profile for each user which can indicate important qualities of their mobility behaviour. However, it is quite important to identify what are the most valuable set of features that should come under this vector.

A standard state-of-the-art model used in E-commerce systems is the RFM model.

- Recency – Indicates how recently did the customer purchase
- Frequency – Indicates how often do they purchase
- Monetary Value – Indicates how much do they spend

These three features can be exploited in the MoD system domain as well.

Recency – Average time between two trips of the user

Frequency – Number of trips per given time period of the user

Monetary Value – Average trip expense

However, the RFM model doesn't express the travel behaviour of the users. Therefore, more travel related features are extracted under the following categories as shown in Table 3.1.

1. Temporal
2. Spatial
3. Travel Mode

Table 3.1: System interaction profile features

Feature type	**Profile parameters**
Temporal	Total average travel frequency (Trips per day)
	Average travel duration (hours)
	Average duration difference between trips (in days)
	Proportion of trips on weekdays

	Proportion of trips on weekends
	Proportion of trips during morning peak (7:00 am-10:00 am)
	Proportion of trips during lunch hours (12 pm – 2 pm)
	Proportion of trips during evening peak (4:00 pm-7:00 pm)
	Proportion of trips during night peak (7:00 pm-9:00 pm)
	Proportion of trips during midnight (11:00 pm-2:00 am)
Spatial	Average travel distance (per trip)
Travel Mode	Proportion of each number of trips in different travel modes

3.1.3 Activity Profile

Users of the MoD is utilizing the service in order to fulfil their travel needs. However, the travel purpose is unknown to the system. When profiling the users travel purpose or activity is one of the important profile vectors.

In this work, we propose an activity vector which indicates each user's probability of traveling to a set of predefined activities by utilizing the MoD. We exploit the static activity map methodology introduced by Phithakkitnukoon et al. [15] in order to acquire the activity probabilities for each user. Following are the proposed steps for activity probability mining.

1. The geographical region of mobility was divided into small cells as shown in Figure 3.1

Figure 3.1: Activity map cell division

2. Above defined grid correspond to a set of cells where a given cell is denoted by a cell id (C_i). The next step is to calculate the activity probability vector for each cell given the different POI categories and count that resides in the cell and the overall region.

   $C_i = [\alpha_i(1), \alpha_i(2), \ldots, \alpha_i(m)]$

   α = activity value function;

   Where i = 1, 2, 3... N, N = Number of cells and

   Total number of different activities = m

   Here the activity value function is defined as below equation 3.1.

$$\alpha_i(a) = w1 \times \frac{POI_a}{C_{i_{POI}}} + w2 \times \frac{POI_a}{T_{POI_a}} \tag{3.1}$$

   i = cell id

   a = activity number

   w1, w2 weight values

   POI_a= Number of POIs within the cell i related to activity a

   Ci_{POI}= Total number of POIs within the cell i

T_{POIa} = Total number of POIs within the overall region related to activity a

As shown in above equation activity probability calculation has two parts where the 1st part indicates the proportion of the POI corresponding to the activity type ‘a’ within the cell and the second part indicates the overall proportion of the POI type ‘a’ compared to the region. In our research, we selected the set of activities mentioned in Table 3.2 for activity map creation. Table 3.3 shows the corresponding POI categories tags provided by Google maps for each activity type.

Table 3.2: Predefined activities

Activity		**Activity Number**
Dining		1
Shopping		2
Educational		3
Entertainment	Cinema	4_1
	Sports	4_2
	Clubs	4_3
	Other	4_4
Sports fitness		5
Other		6

Table 3.3: POI categories for activity types

Activity		**POIs**
Dining		‘bakery', 'cafe', 'restaurant', 'meal_takeaway', 'food'
Shopping		'clothing_store', 'convenience_store', 'book_store', 'bicycle_store', 'car_dealer', 'department_store', 'electronics_store', 'florist', 'furniture_store', 'hardware_store', 'home_goods_store','jewelry_store', 'liquor_store', 'pet_store', 'shoe_store', 'shopping_mall', 'store', 'supermarket'
Educational		'library', 'school', 'university'
Entertainment	Cinema	'movie_rental', 'movie_theater'
	Sports	'stadium'
	Clubs	'bar', 'pub'
	Other	'amusement_park', 'aquarium', 'museum', 'art_gallery', 'bowling_alley', 'campground', 'casino', 'spa', 'zoo'
Other		All the other tags

3. After creating the activity map, for each cell there exists an activity probability vector $V_a = <a_1, a_2 \ldots a_n>$. Then for each user activity probability values are calculated based on their travel locations.
   If a user U_j travel to a location which resides in cell C_i that cells activity vectors are added to user’s activity vector. And after executing this for every trip of

the user, the activity vector of the user is normalized by the number of trips he/she had as shown in the below equation 3.2.

$$V_Aj = \frac{\sum_T Cell\ V_a}{n_T} \tag{3.2}$$

Here, T is the set of all Trips and n_T is the total number of trips.

4. Based on the application of the activity probability vector following hyperparameters should be tuned. (Our hyperparameter values will be shown in the evaluation section)

- Cell level - Size of the cell where the geographical region is covered
- w1, w2 - Weight values of the activity value function

3.1.4 Demographic Profile

Another important profiling required is demographic based profiling. One of the main problems in many E-commerce systems is that the users tend not to release their actual demographics such as age and gender. But it is quite important for the system to have a demographic vector correspond to every user, in order to enable personalized services and products.

When considering the MoDS, it is compulsory for the users to provide their demographics to the system. However some people tend to input their demographics while registering to the system. Therefore we have an accurately tagged user with their demographics in the system which leads to a supervised machine learning model for classifying the rest of the users of the system. In our work, we present two classification models for predicting the age and gender demographics of the users in MoD systems.

1. Data set creation

In order to create and evaluate the demographic classification models, training and testing datasets are needed. Since we have already created two profile vectors namely user system interaction and user activity profiles, demographic classification can be done on top of those features. However, it is necessary to employ feature engineering techniques in order to fine-tune the classification models.

2. Model Selection

Selected the following models for the demographic prediction based on the previous work on similar E-commerce environments.

- SVM
- Gradient Boost
- Neural Network Classifier

3. Feature Engineering

We employed a set of derived features in order to acquire the best accuracy values out of the prediction models.

- Principal Component Analysis

Selected the derived principal components of the dataset by employing the PCA technique. Selected 6 components by plotting the variance covered by the number of components on the dataset as shown in Figure 3.2.

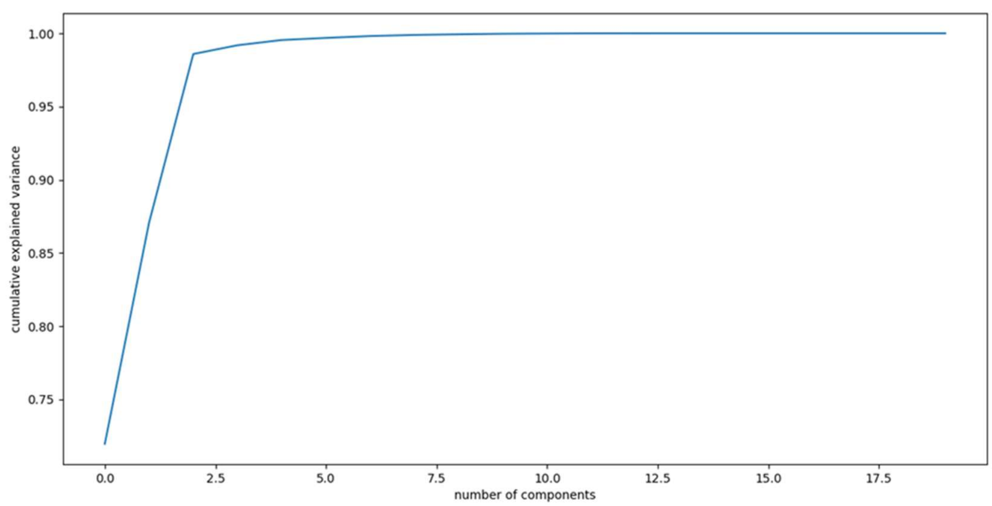

Figure 3.2: Cumulative variance explained by PCA components

- Latent Codes

Latent factors, also known as the hidden labels of the data set was extracted by employing an autoencoder and getting the hidden layer features as shown in Figure

3.3. These 'hidden' features within the data give a semantically relevant 'aggregates' of the observed features which consequently leads to a more accurate classification hypothesis.

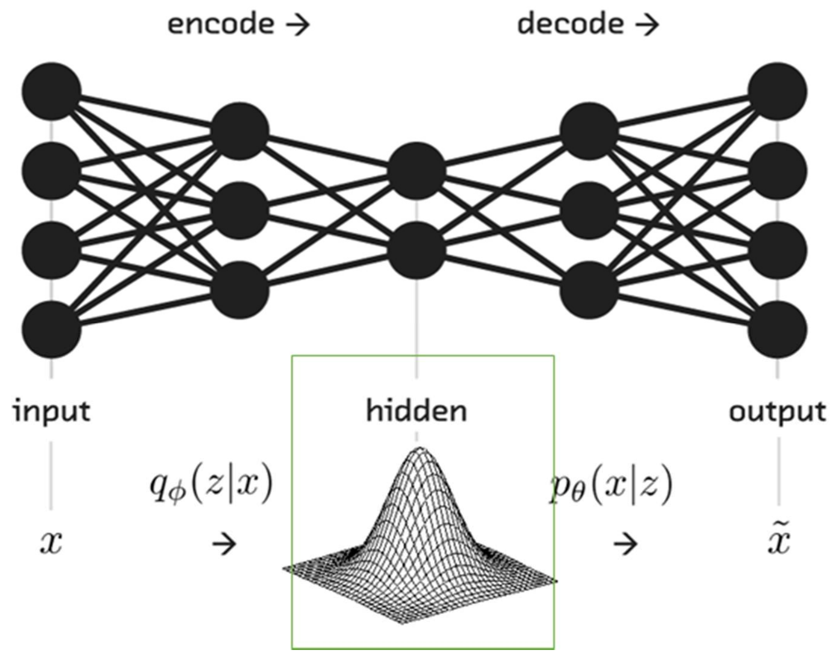

Figure 3.3: Latent model extraction by employing an Autoencoder

4. Feature Selection

We employed the state-of-the-art Recursive Feature Elimination (RFE) to select the most important features for the classification models. This technique proved to be better since this selects the most important features based on the machine learning technique used for prediction. The RFE technique selects features by recursively considering smaller and smaller sets of features. At first, a set of weights are assigned to the features and the weights are updated after training the model. Then the features with the smallest weights are pruned and again another model is trained. This ranks the features and the desired number of features can be selected.

We performed recursive feature elimination for each model that was selected and chose the best features based on the model. The features that were selected using recursive feature elimination based on the selected models are shown in Table 3.4.

Table 3.4: Features selected by Recursive Feature Elimination (RFE)

Model	**Age**	**Gender**
SVM	Total average travel frequency Recency Proportion of trips on weekends Educational Activity Val PCA components	Total average travel frequency Mode Preference Trip expenses
Gradient Boost	Total average travel frequency Proportion of trips on weekends	Total average travel frequency Mode Preference Trip expenses
Neural Network	Total average travel frequency Recency Proportion of trips on weekends	Total average travel frequency Mode Preference PCA components

5. Class imbalanced problem

One of the frequent problems that can occur is the class imbalanced problem when it comes to demographic prediction in MoDS. Table 3.5 shows the number of samples for each age group class we employ for the age classification.

Table 3.5: Label propagation in age classes

Class (label)	**Age Category**	**Number of samples**
1	15-24	22339
2	25-34	30541
3	35-44	14211
4	45-54	6848
5	55-64	3345
6	65+	950

Figure 3.4 shows the imbalance classes in our dataset for gender classification.

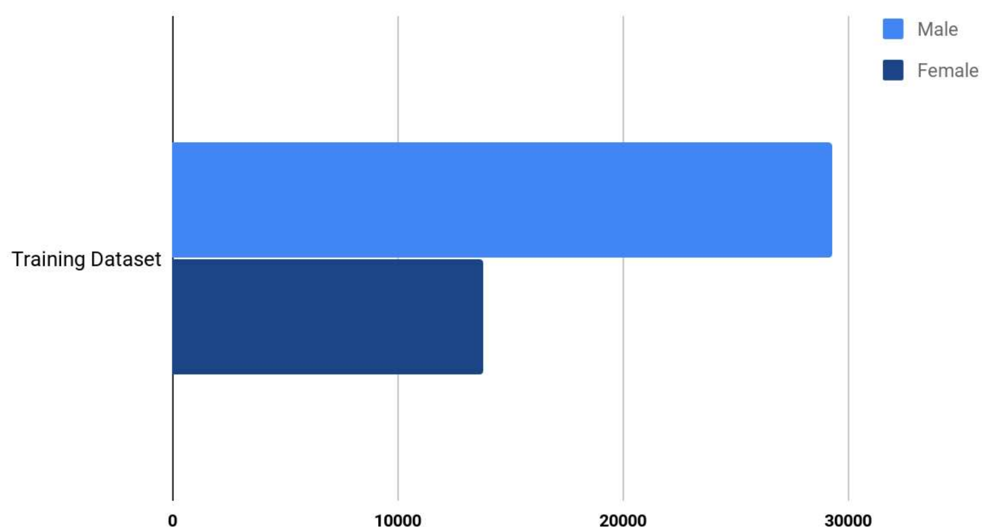

Figure 3.4: Gender classification training dataset classes

In our work, we proposed the following solutions for the class imbalance problem.

- SVM weighted class leaner (Cost-sensitive learning) - In this model, we can assign a weight value for each label that can mitigate the effect of imbalance number of training samples for the classes.

- Boosting algorithms - One of the recommended methods for overcome class imbalance problem is to utilize boosting algorithms. We employ gradient boosting classifier to evaluate the boosting effect on demographic prediction.
- Resampling - Resampling was also employed in order to make the prediction classes balanced, however, the problem with resampling was the overfitting of the classification models.

3.2 User Segmentation

After creating the user profiles another important step in our analysis is to create customer segments. We employed unsupervised clustering on the dataset in order to identify the natural clusters resides in the user base and consequently parameterized and extract the segments.

3.2.1 Unsupervised Clustering

Related work on customer segmentation models shows that distance based K-Means clustering is quite effective in finding natural clusters in the customer base. However, there are much more sophisticated and generalized clustering techniques currently used for cluster data such as GMM instead of the traditional K-Means. We employed user profile vector features as the features for the dataset for the unsupervised clustering using the GMM model.

GMM model also requires a number of clusters pre-defined similar to the K-Means. However, there is a standard method for deciding the number of clusters components in GMM based on Akaike information criterion (AIC) and Bayesian information criterion (BIC). As shown in Figure 3.5 AIC and BIC values are plotted against the number of cluster components of GMM. The number of clusters is selected such that the AIC and BIC becomes a minimum. In the case of a continuously decreasing curve, the Elbow method is employed to decide the cut-off point.

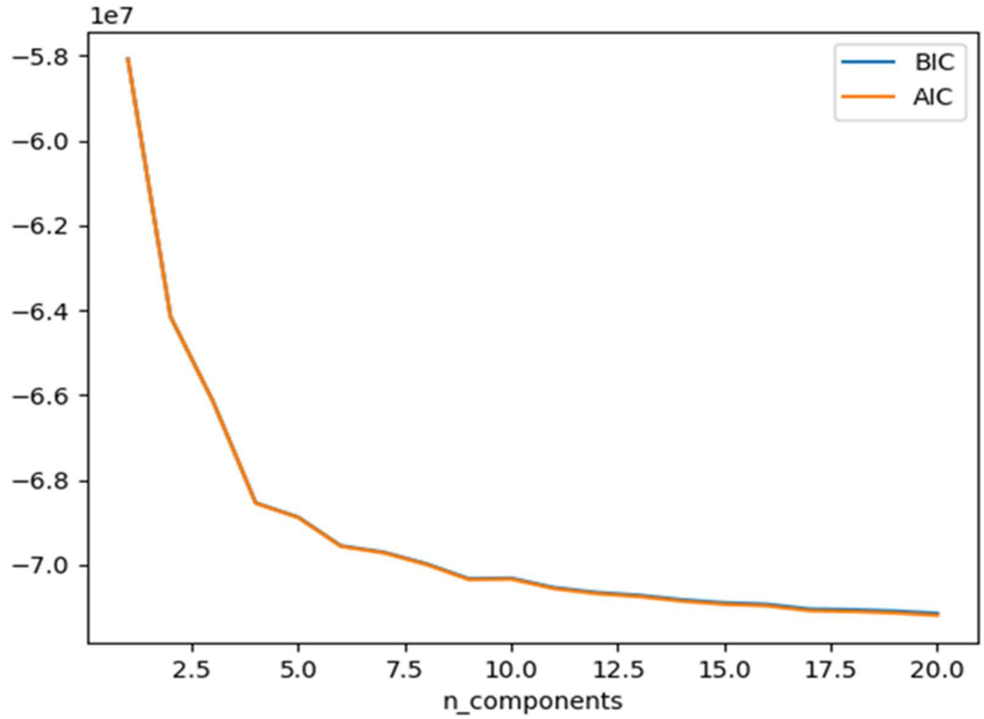

Figure 3.5: AIC and BIC against GMM cluster components

After the clustering Silhouette coefficient as defined by below equation set 3.3 is used to measure the cluster quality.

$$
\begin{aligned}
a(\boldsymbol{o}) &= \frac{\sum_{\boldsymbol{o}' \in C_i, \boldsymbol{o} \neq \boldsymbol{o}'} dist(\boldsymbol{o}, \boldsymbol{o}')}{|C_i| - 1} \\
b(\boldsymbol{o}) &= \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{\boldsymbol{o}' \in C_j} dist(\boldsymbol{o}, \boldsymbol{o}')}{|C_j|} \right\} \\
s(\boldsymbol{o}) &= \frac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{\max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}
\end{aligned}
\tag{3.3}
$$

Here silhouette coefficient s(o) is between -1 and +1. A value closer to +1 implies a good quality cluster.

However, density-based clustering models are recommended over the distance based clustering models. Density-based clustering models can form the natural clusters in the data in spite of the anomalous data points. We employed DBSCAN and OPTICS clustering on the data and acquired natural user clusters for segmentation purpose.

DBSCAN is one of the state-of-the-art density-based clustering techniques which is capable of capturing arbitrarily shaped clusters. Moreover, this has the concept of noise built into the clustering technique and is robust against noise since the samples which are noise are not considered for clusters. OPTICS is an optimized model on top of the DBSCAN algorithm based on the reachability definition.

3.2.2 Segmentation

After extracting the natural clusters in the data special analysis should be employed in order to acquire the segments. A segment in the E-commerce domain is a cluster of customers which has a business value. Thus, in our analysis on MoDS user clusters, it is quite necessary to extract user segments in order to enable the management of the service takes strategic business decisions.

We propose an aggregated analysis based on the profile vectors in order to enable the segmentation in MoDS. For each natural cluster, the aggregated average of each user profile vector is calculated. And based on the average value vector V_p as defined by below equation 3.4, the clusters are given a context and a business value.

$$V_{p_{C_i}} = \frac{\sum_{u \in C_i} V_u}{|C_i|} \tag{3.4}$$

Here, C_i is a given cluster of the user base and V_u is the aggregated profile vector of any user in cluster C_i.

3.3 Personalized services in MoDS

Next step of our work is to apply the formed user profiles to improve the service and management of MoDS. Service improvement in E-commerce domain is achieved by making the services and products customer oriented and personalized. Thus open the next part of the methodology which discusses a set of comprehensive models particularly employed towards the goal of personalized services in MoDS. Recommender Systems (RS) are the models which enable personalization in various other domains similar to MoDS. Therefore it is essential to lay the groundwork for RSs in MoDS in order eventually improve the service.

3.3.1 Problem Formation

The main product /service of MoD systems is the trip/ride provided to the user for their mobility requirement. However, when considering the ride options provided by a certain MoD is limited and the user can easily select a ride that fits their personal needs. Nevertheless, one of the prominent business models of these services is to provide customers with promotions in order to complete their day to day mobility needs. However, these promotions can be put in various categories such as dining promotions (promotions given if the customer traveling to a particular restaurant), evening promotions and weekend promotions, etc. Thus in order to make the given promotions more personalized and relevant to a particular user, a systematic method of promotion recommendation is required. This is what differentiates the normal recommender systems criteria with the recommender system needs in the MoD system because in other domains recommender systems are there to recommend the user base the different categories of the main product or service. Our main goal is to propose a personalized systematic recommender system for promotion matching, which is an indirect product of the system. Based on the profile vectors of the MoD we propose a hybrid recommender system for promotion recommendations as shown in Figure 3.6.

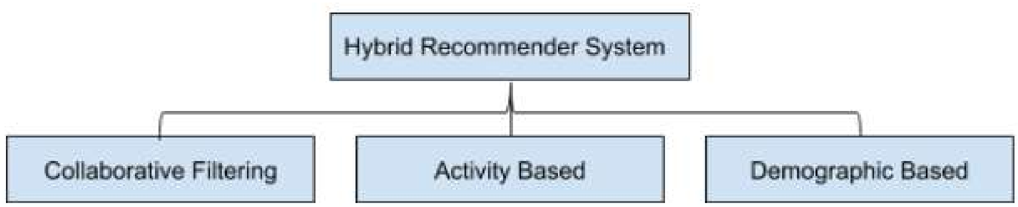

Figure 3.6: Recommender system methodology

The proposed RS for the MoD consists of three components.

1. Collaborative Filtering
2. Activity-based
3. Demographics-based

We employed a generative model for enabling a hybrid recommender system by combining the Collaborative Filtering model with output information from the other two recommender system models.

3.3.2 Collaborative Filtering (CF) Recommender System

There are registered user u_1, u_2... u_n and for each of these users, there are recorded trips in the systems which indicates whether a given trip t, was completed using a promotion given by the service provider or not. Moreover, there is a record which indicates what type of promotion was utilized by each of the user trips. Therefore for each user, we can build an integrated promotion usage matrix as shown in Figure 3.7.

	P_1	P_2	.	.	.	P_m
U_1						
U_2						
.						
.						
.						
U_n						

Figure 3.7: User promotion usage matrix

Here each cell of interaction matrix I_{ij} is calculated by equation 3.5.

$$Iij = T_{Pj} / T_T \tag{3.5}$$

Here T_Pj is the trip count completed using promotions in category P_j by user i and T_T is the total trip count of the user i. This user-promotion usage matrix is analogous to the user-item interaction matrix normally found in RS criteria.

3.3.2.1 Recommender System Overview

The objective of the purposed CF RS is to match the most relevant and personalized promotions to the user based on the previous promotion usage. This RS can be divided into two sub-components based on the functionality.

1. Prediction model: We employed generative network based models to predict the unseen interactions between the users and promotions.
2. Recommendation model: Top N-recommendation for each user based on completed user-promotion interaction matrix.

As defined by the above equation user-promotion interaction matrix was calculated using aggregating the number of trips where a particular promotion was utilized. However, the problem here is that these interactions can be quite limited. For example, user U_i might have used only promotions which comes under P_j category. Thus for the user U_i other interactions with promotion category set P ={P_i; i=0,1,..n ,and i != j} is missing in our dataset. Therefore, predicting these unseen interactions is the most important part in enabling the recommender system for personalized promotion distribution.

Generative Network Models

Autoencoder (AE)

Autoencoder is a feed-forward neural network with Encoder and Decoder model architecture as shown in Figure 3.8,

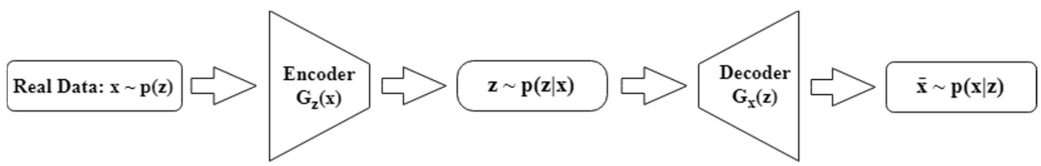

Figure 3.8: Structure of Autoencoder (AE)

Following are the main components found in AE

1. Encoder - This network encodes the provided input samples x onto latent model distribution z.
2. Decoder - This network decodes the stochastic latent model z onto the output of $\bar{x}$.

Variational Autoencoder (VAE)

VAE has the same structure as a normal Autoencoder with optimization in the latent model (Hidden layer) as shown in Figure 3.9. The Loss function as given in equation 3.6, corresponds to the optimized latent model formation.

$$L(i) = E_{q_{\phi}(z)}\left[\ln\left(p\left(x|z,\theta\right)\right)\right] - KL(q_{\phi}(z|x)||p_{\theta}(z)) \tag{3.6}$$

Here, the first part of the equation corresponds to the normal Autoencoder objective of minimizing the reconstruction error between the output and the original input. The Kullback-Leibler (KL) divergence enforce a prior distribution $p_{\vartheta}(z)$ on the hidden layer approximate posterior $q_{\emptyset}(z|x)$, which makes the VAEs hidden variables to take a stochastic from than being deterministic as in the normal Autoencoder.

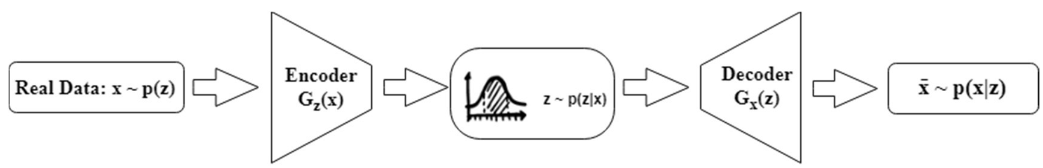

Figure 3.9: Structure of Variational Autoencoder (VAE)

3.3.2.2 Adversarial Autoencoder (AAE)

AAE can be considered as a model where normal Autoencoder (AE) is reinforced by adversarial learning. Hence we can see the AE architecture plus another model to impose adversarial learning as shown in Figure 3.10.

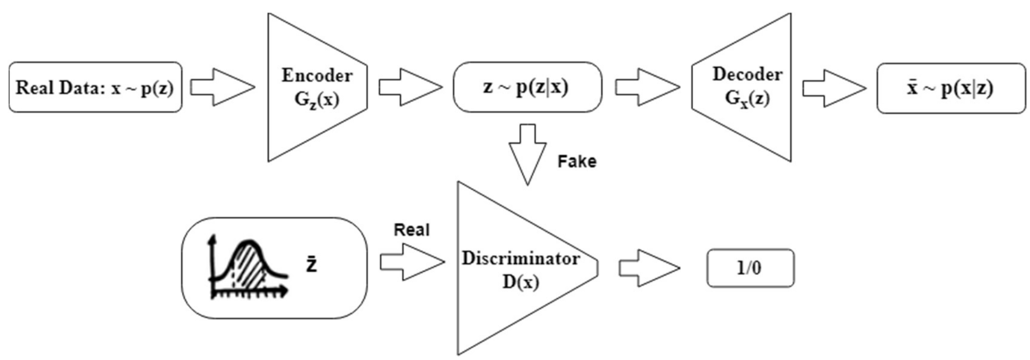

Figure 3.10: Structure of Adversarial Autoencoder (AAE)

Discriminator network outputs the probability of a generated latent model z (by Encoder) belonging to the real distribution $\underline{z}$ we need to impose on the latent model. When considering the AAE model, it can be perceived as a combination of AE and a GAN. Thus when it comes to the training of AAE it has two phases.

- Reconstruction Phase: This corresponds to the training of the AE part where Encoder and Decoder models are trained in order to reconstruct a given data sample x with a minimum error. The objective function for this phase is defined as equation 3.7.

$$\arg\min \|X - \bar{X}\|^2 \tag{3.7}$$

- Adversarial Phase: This corresponds to the training of the GAN part where Encoder (Generator G in GAN terminology) and Discriminator (D) are trained in a manner which conforms to a two-player minimax game. The main goal of this learning phase is to impose a prior distribution on the latent model (z) which is generated by the encoder. The objective function of this phase is defined by the equation 3.8.

$$\min_{G} \max_{D} V(D.G) = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \tag{3.8}$$

3.3.2.3 User-Promotion Interaction prediction

After the completion of the training process, generative models have learned the latent model of the user-promotion interaction matrix. This latent model is similar to a hidden code of the matrix where it consists of the base units which defines the promotion interactions of the users. After the original user-matrix interaction data X is fed into the trained Encoder it will generate the latent code for each row of the matrix and then the Decoder will map the created latent codes into the output data $\underline{X}$which is the predicted interactions for each user of the system. This procedure is shown in Figure 3.11.

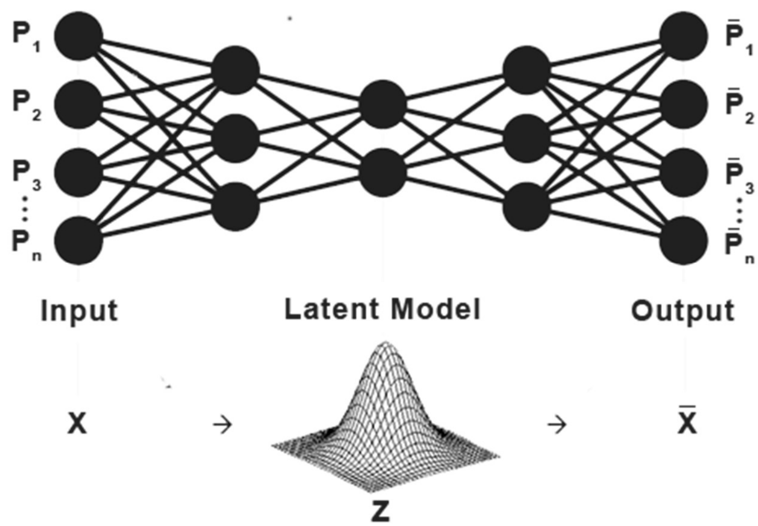

Figure 3.11: User-promotion interaction prediction

Personalized Recommendations

After the user-promotion interactions are predicted by the generative model, a list of ranked promotion categories are generated for each user based on these predicted interactions. Then top N-recommendation is filtered from the ranked list of promotion categories.

3.3.2.4 Dealing with Data Sparsity

The user-promotion matrix is normally a sparse matrix since there are an extensive number of promotion categories and each user interaction with these promotions are limited. This is one of the main problems in RS implementations. Due to the aforementioned sparsity learning models tend to predict the trivial solution when it comes to the unseen examples.

Importance of adversarial learning of AAE model is that using the Discriminator model we can impose an arbitrary prior distribution over the latent model rather than learning the latent model just from the data itself like in normal AE structure. Thus, even though the data is sparse, the latent model of AAE tend to cover the full space of the prior distribution we select. This enables the Decoder model of the AAE to generate meaningful predictions for the unseen samples from any part of the prior space we imposed on the latent model.

Another improvement commonly seen in deep generative networks is to make the model robust by utilizing the denoising criteria. Denoising criteria are a special set of random corruption done to the inputs of the network in order to make the generator susceptible to noisy inputs. In other words, the model is trained on training samples where random noise is added before the training starts. By utilizing Gaussian denoising on AAE model it was made more robust and prediction accuracy was increased in the case of sparse data.

3.3.3 Activity-based Recommender System

The activity-based recommender system is enabled by employing the user activity profile vector (V_Aj) mentioned in the above sections. For each promotion category, we have to define an activity indicator vector (V_{AI}) which infer the purpose of the promotion related to each of the activities. For example, if the management's purpose to give a promotion p to the user base is to promote their trips for dining, then in the activity indicator vector dinning activity has 1 and all other activities are 0. Then for each user, the similarity between their activity profile vector and the promotion activity indicator vector is calculated using the cosine similarity as shown in the below equation 3.9.

$$\text{Cosine Similarity } (U_i, P_j) = \frac{v_{A_i}.V_{AI_j}}{\left\|v_{A_i}\right\|\left\|V_{AI_j}\right\|} \tag{3.9}$$

Personalized Recommendations

After the user-promotion activity similarities are calculated, a list of ranked promotion categories is generated for each user based on these predicted activity similarities. Then top N-recommendation is filtered from the ranked list of promotion categories.

3.3.4 Demographic-based Recommender System

A demographic-based recommender system is enabled by employing the user demographic profile vector (V_Dj) mentioned in the above sections. For each promotion category, we have to define a demographic indicator vector (V_{DI}) which indicates the relevant demographic group where the promotion is allocated. For example, if the management's purpose to give a promotion p to the user base, is to promote people in the age group of 25-35 to travel more in the MoD, then in the demographic indicator vector age group, a demographic label should be 2 which indicate the age group 25-35. Then for each user, the similarity between their demographic profile vector and the promotion demographic indicator vector is calculated using the cosine similarity as shown in equation 3.10.

$$\text{Cosine Similarity } (U_i, P_j) = \frac{v_{A_i}.V_{DI_j}}{\left\|v_{A_i}\right\|\left\|V_{DI_j}\right\|} \tag{3.10}$$

Personalized Recommendations

After the user-promotion demographic similarities are calculated, a list of ranked promotion categories is generated for each user based on these predicted demographic similarities. Then top N-recommendation is filtered from the ranked list of promotion categories.

3.3.5 Hybrid Recommender System

In our work, we propose a hybrid recommender system methodology by combining deep generative model based CF with the side information from user activities and demographics. Model architecture is as shown below in Figure 3.12.

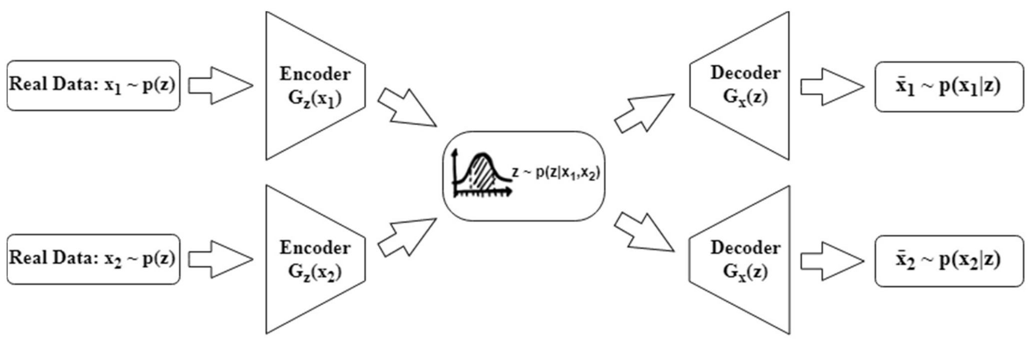

Figure 3.12: Hybrid latent code learning model

Here x_1 is the user-promotion interaction data as in normal CF model. However, the difference in this model is that the latent model z is shared by two networks where x_2 -> $\underline{x}_2$ network corresponds to the side information about user activity and demographics. By making normal CF latent model shared by the side information generative model we are enabling a hybrid model which utilize both user-promotion interaction and user activity, demographic qualities in predicting the interactions thus producing a more comprehensive recommendation.

3.4 Summary

In this section, we discussed in detail the proposed methodology of our research on the Mobility on Demand System domain. Our methodology comprises a step-by-step guide for a systematic analysis starting from customer profiling and then the customer segmentation and eventually, the application of personalized services in MoDS.

Under this section, we presented three user profile models for the task of customer profiling in MoDS.

1. System Interaction Profile
2. Activity Profile
3. Demographic Profile

System interaction profile contains the feature set directly extracted from the system, whereas the activity profile includes the calculated user activity probability levels by analysing the travel locations. The demographic profile is the profiling model which employs above mentioned two profiles to generate user demographic features.

Next step in our methodology was to extract segments from the customer base. We proposed two clustering techniques for the task at hand; one model is a generalization of the state-of-the-art customer segmentation model K-Means, the Gaussian Mixture Model (GMM) clustering, and the other model was a prominent density-based clustering technique OPTICS.

Final subsections cover the comprehensive methodology we propose for the personalized service in MoDS. We introduced a novel Adversarial Autoencoder prediction model based Collaborative Filtering for individualized promotion recommendation in MoDS. Furthermore, we extended the aforementioned model into a hybrid Collaborative Filtering recommender system which incorporates side information from the activity and demographic profiles from the user profiling section.

Chapter 4

4 EVALUATION

This chapter unrolls the experiments employed on a real-world Mobility on Demand system. Subsection 4.1 discusses the system we analysed and the dataset utilized for the evaluation. Experiment formation section provides a comprehensive step-by-step description of the planned experiments on the proposed methodology. Following points are addressed under the experimental evaluation;

1. User profiling model performance by evaluating the Demographic classification models.
2. Performance of the segmentation models by evaluating the clustering models.
3. Performance of the proposed recommender system on the historical data in MoDS.

Subsection 4.3 covers the evaluation results and discussion on the evaluations experiments mentioned above.

4.1 Experiment System and the Dataset

Our experiment was based on a local on-demand vehicle service which caters the public mobility needs in 3 provinces. A Mobile app is used by the users in order to demand vehicles at their convenience. The system provides the following 6 main modes of travel for their users.

1. Tuk
2. Micro Cars
3. Mini Cars
4. Cars
5. Van
6. VIP

In this system, user travel transactions are recorded on a daily basis, where the system keeps track of the basic details such as trip starting timestamp, trip ending timestamp, pickup location, drop location, travel distance, travel fare and whether the trip was promotion based (Promo code was utilized). For our analysis, we collaborated with the service provider and acquired a pseudonymous user transaction dataset which consists of the day to day data gathered travel data from more than 300000 users.

4.2 Experiment Formation

4.2.1 Overview

The experiments of our research can be broken down into 3 main categories and under each of the main categories, following sub-experiments can be found.

1. User profiling
    a. Activity Mining Model
    b. Demographic Prediction Model
2. User segmentation
    a. Unsupervised clustering models
    b. Cluster analysis and segmentation
3. Recommender system
    a. User-promotion interaction prediction
    b. Hybrid recommender system

4.2.2 Dataset Setup

From the user travel transaction, raw data following aggregated datasets in Table 4.1 are created for different experiment model evaluations.

Table 4.1: Aggregated datasets for evaluation

Dataset	**Description**	**Samples**	**Features**
User system interactions	This dataset corresponds to system profiling	374000	17 - temporal features (10) - spatial features (1) - travel mode (6)
User activity dataset	This dataset corresponds to the user activity profiling	374000	8 -dining (1) - shopping (1) - educational (1) - entertainment (4)

			- other (1)
User demographic classification	This dataset is used to evaluate the age demographic classifiers	111763	28 - users system interactions (17) Derived features - PCA features (6) - latent features (4) - demographic label (1)
	This dataset is used to evaluate the gender demographic classifiers	61527	
User promotion interactions	This dataset is used to evaluate the CF model	374000	30 promotion categories

For the activity mining model, an external POI dataset was extracted from Open Street Maps APIs which is shown in Table 4.2. This experiment was done for the western province of Sri Lanka.

Table 4.2: POI data for activity types

Activity		**Number of POIs**
Dining		7742
Shopping		10984
Educational		3555
Entertainment	Cinema	105
	Sports	344
	Clubs	246
	Other	286
Other		47355

For the evaluation of proposed models, we employed the standard criteria of dividing the overall dataset into train, test and validate datasets (70%:20%:10% ratio).

4.3 Evaluation and Results Analysis

In this section, all the experiment results are presented in the order mentioned above in experiment formation section. All the proposed models and components are evaluated utilizing the datasets mentioned in Table 4.1.

4.3.1 Evaluation of the User Profiling

First two profile vectors we defined under the methodology, namely 'System Interaction Profile' and 'Activity Profile' contain explicitly derived features from the raw dataset. Aforementioned two profile vectors were used to derive the 'Demographic Profile' for each user by employing supervised classification techniques.

Age and gender demographics were selected as the experiment demographics due to the availability of labelled dataset for training and evaluating a supervised prediction model. Age prediction problem was a multiclass classification, where the following age groups were selected as the categories needed to add each user into shown in Table 4.3.

Table 4.3: Age Classification Classes

Age Category	**Class label**
15-24	0
25-34	1
35-44	2
45-54	3
55-64	4
65+	5

Results

Following are the accuracy results we got from the age and gender demographic classifications. Here the baseline accuracy is from the zero-r classifier where accuracy value is recorded by assigning the most frequent label to every prediction.

1. Age Group Classification

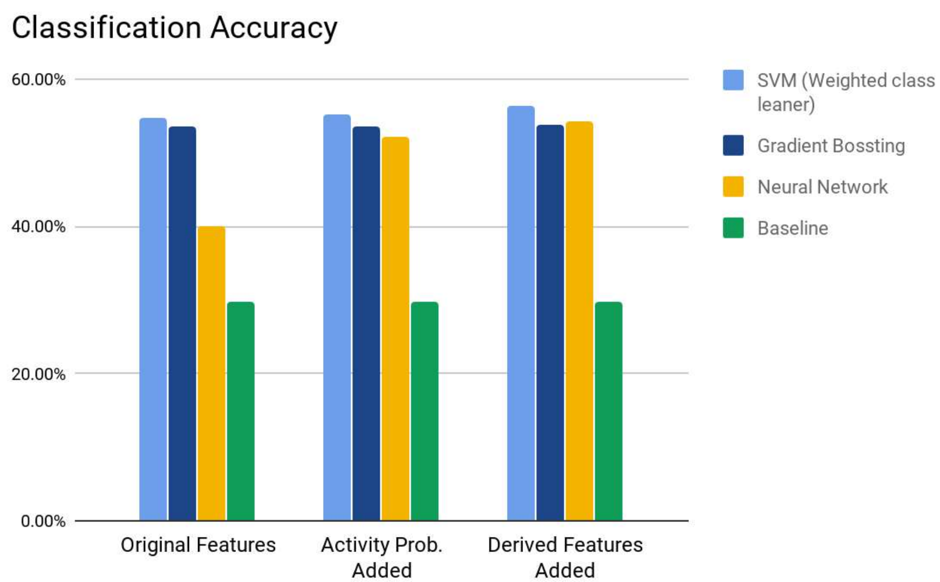

Figure 4.1: Age group classification accuracy

Out of the age classifiers, we can see in Figure 4.1 that SVM weighted class leaner has the highest accuracy. This is due to the class imbalanced problem we earlier discussed. SVM weighted class classifier tends to treat each class according to the assigned weights, consequently forming a more generalized hypothesis which gives good accuracy values. Moreover, there is an increase in the accuracy values by adding the activity profile into the model. Further increase in accuracy was achieved by adding the derived features based on PCA and latent code transformations. Even though, the overall accuracy values are lower we can see that all the models tend to surpass the baseline in a great margin.

2. Gender Classification

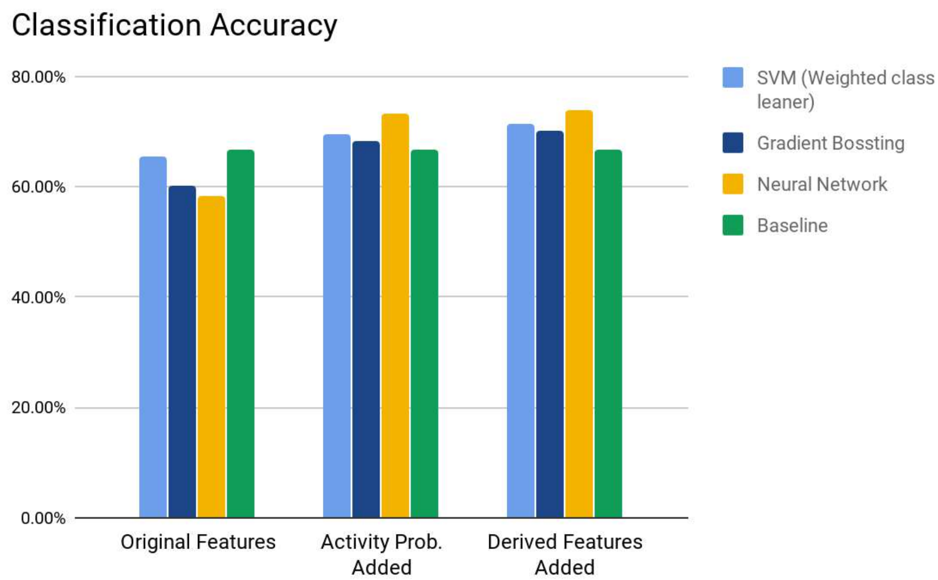

Figure 4.2: Gender classification accuracy

From the gender demographic classification results which are shown in Figure 4.2, we can see that neural network classifier tend to perform better than the other proposed methods. However, when comparing the features sets used for the classification it is clearly visible that user system interaction features are not playing an important role in predicting the user's gender. However, by adding the activity profile and the derived features, accuracy of the models can be increased above the baseline. This can be explained by the fact that the activities of each user can be governed by their gender demographic. Therefore by employing an activity profile, we are getting good prediction accuracy.

4.3.2 Evaluation of the User Segmentation

4.3.2.1 Unsupervised Clustering

Based on the previous work on customer segmentation we employed the following two clustering models to extract the natural user clusters in the MoDS.

1. GMM Clustering
2. OPTICS Clustering

Parameters

1. Number of Clusters

    a. GMM - Number of clusters was 4 which was based on Akaike information criterion (AIC) and Bayesian information criterion (BIC) as shown in Figure 4.3.

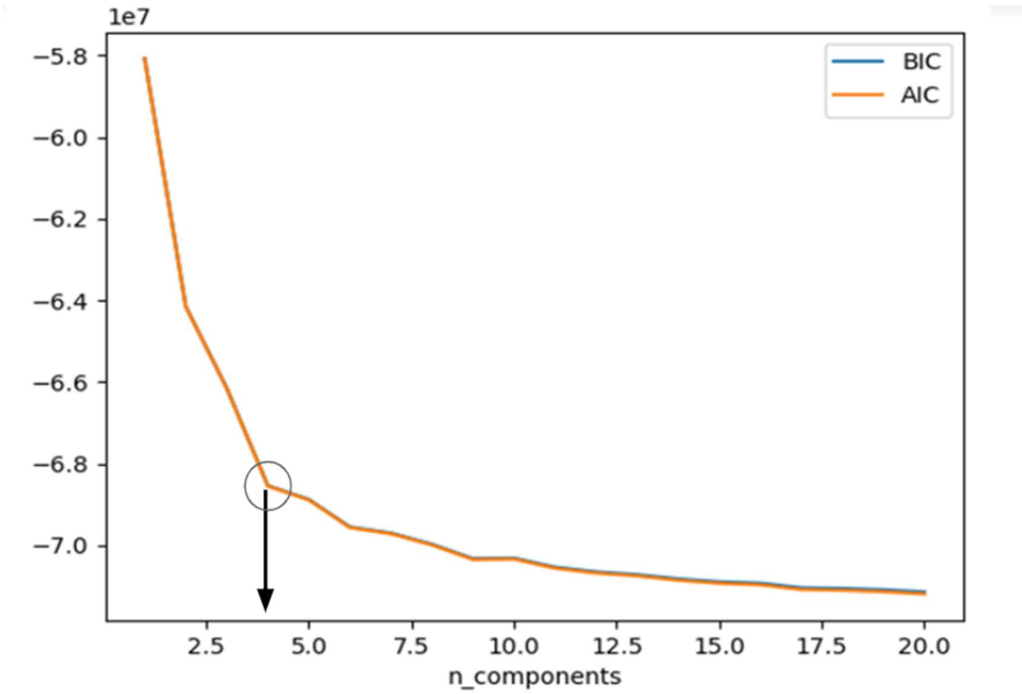

Figure 4.3: Number of component selection for GMM

    b. OPTICS - No need to define the number of clusters

2. Hyperparameters

    a. OPTICS - Following Table 4.4 contains the hyperparameters tuned for OPTICS clustering model.

Table 4.4: OPTICS clustering hyperparameters

Hyperparameter	**Value**
min_samples	10
xi	0.1

Results

When analysing the cluster results present in Table 4.5, GMM has a higher Silhouette coefficient than the density based OPTICS clustering. This result doesn't particularly emphasize that the GMM is better than the OPTICS for the clustering task at hand. Reason for the OPTICS to show a lower Silhouette coefficient value is due to the number of clusters that are formed by it. As shown in Table 4.5, GMM clustering forms 4 clusters whereas by OPTICS, the number of clusters is 48. So the OPTICS clustering can reach more deeply into the data hierarchy than the GMM, consequently making the Silhouette coefficient lower because of the closeness of the clusters.

Table 4.5: Unsupervised clustering results

Model	**No. Clusters**	**Silhouette coefficient**
GMM	4	0.68
OPTICS	48	0.26

4.3.2.2 Cluster Analysis (Segmentation)

As mentioned in the methodology section, user clusters and user segments differ due to the business value. According to the definition, the user segment is a user cluster which has a business value to the MoDS. Therefore, we employed a cluster analysis based on user system interactions in order to identify the user segments.

1. GMM Clustering

By analysing the four clusters given by GMM as shown in Table 4.6, following segments were identified.

- High frequent user segment
- Infrequent user segment
- One time user segment
- Promo seeker user segment

Table 4.6: GMM cluster analysis

Segment		One time	Infrequent	Promo seeker	High frequent
Total frequency (per 6 months)	Mean	**1**	5	12	**37**
Avg. total frequency (Trips per day)	Mean	**1**	0.21	0.43	0.31
Recency (Avg. days between trips)	Mean	**0**	16	17.	5
Trip expense (Per trip)	Mean	606.24	492.05	452.50	416.40
Discount amount (Per trip)	Mean	4.76	3.7	66.77	5.8
Proportion of Promo Trips (%)		2.9	4.5	**16**	7.6
Cluster Size		16806 (8%)	41516 (19%)	23856 (11%)	130863 (62%)

2. OPTICS Clustering

By analysing the four clusters given by OPTICS, following segments were identified as shown in Table 4.7.

- The high frequent user segment
  - Cooperate
  - General public
- One time user segment

Table 4.7: OPTICS cluster analysis

Segment		**General frequent**	**Cooperate frequent**	**One time**
Total frequency	Mean	**70**	**74**	**1**
Avg. total frequency (Trips per day)	Mean	0.52	0.54	1
Recency (Avg. days between trips)	Mean	2	3	0
Trip expense (Per trip)	Mean	399.82	588.85	606.24
Discount amount (Per trip)	Mean	4.51	480	4.76
Proportion of Promo Trips (%)		6	2.9	0.04
Proportion of Cooperate Trips (%)		0.0024	**73.8**	0
Cluster Size		45777 (22%)	2026 (0.95%)	16294 (7.6%)

Here the difference between GMM and OPTICS clustering is the number of clusters. GMM has only 4 clusters and OPTICS creates 47 clusters where the promo seeker segment found in GMM is divided into 10 small clusters in OPTICS. Moreover, OPTICS was able to recognize another sub-segment 'cooperate frequent users' which was not recognized under GMM segments. When comparing these segmentation results we can conclude that OPTICS has the power to form more deep and valuable segments which are hidden from the GMM.

4.3.3 Evaluation of the Recommender System

When considering the Recommender System proposed in our work, there exist two components where independent evaluations are needed in order to check the performance.

1. Prediction model: We employed generative network-based models to predict the unseen interactions between the users and promotions.
2. Recommendation model: Top N-recommendation for each user based on completed user-promotion interaction matrix.

4.3.3.1 Prediction Model Evaluation

The prediction model is evaluated using the RMSE criteria as shown in equation 4.1.

$$\mathrm{RMSE}_i = \sqrt{\sum_{j=0}^{m} \left\| I_{ij} - \bar{I}_{ij} \right\|^2} \tag{4.1}$$

Here I_{ij} is the original user-promotion interaction matrix and $\bar{I}_{ij}$ is the predicted interactions matrix by the model. And m is the number of promotion categories in the MoDS. Using the above equation RMSE for user i is calculated and mean RMSE for all the users in the system is recorded as the model performance.

For the AAE training, complete user-promotion interaction data set is divided into training, validation, and testing datasets. The sample sizes and dimensions of the corresponding datasets are recorded in Table 4.8

Table 4.8: Experimented dataset description

Dataset	Number of Samples	Features
Training	260000	30
Testing	76000	30
Validation	38000	30

We employed a random predictor as our baseline score and state-of-the-art matrix factorization model is utilized to emphasize the generative model performance in CF. Recorded results can be seen in Table 4.9.

Results

Table 4.9: RMSE score for user-promotion interaction prediction model

Training Size	MF	AE		VAE		AAE	
		Vanilla	Denoising	Vanilla	Denoising	Vanilla	Denoising
50%	0.78	0.62	0.56	0.53	0.46	0.38	0.31
75%	0.64	0.45	0.39	0.38	0.31	0.25	0.17
100%	0.56	0.31	0.15	0.18	0.13	0.14	0.06

The prediction model evaluation experiment has the following three main goals.

1. Evaluating our model AAE for deep CF - AAE model performance was compared with the other well - established generative models such as AE and VAE and also with the state of the art MF model.
2. Evaluating the importance of denoising for generative networks - Gaussian denoising was implemented for all the generative networks based CF models (AE, VAE, AAE) and calculated the RMSE scores for prediction.
3. Evaluating the performance variations of the CF predicting models with the training data size - Training the models on different proportions of the overall training dataset and calculated the aggregated RMSE score.

According to the results obtained from the experiment, it can be clearly seen that the AAE model tend to outperform all the baseline models we have selected. AAE outperforming the MF model in a large margin is moderately intuitive because of the extra sparsity of our experiment user-promotion interaction dataset. As explained in the above sections AAE is more robust with sparse data due to the completeness of the latent model than the MF which extract the latent codes only based on the data itself.

Another significant observation in the experiment results is the performance increase in the generative model due to the denoising. As explained in the methodology section, denoising makes the generative models more robust and perform well when the input data is noisy. Consequently, even in the case of user-promotion interaction matrix is sparse and has less information in the inputs, the denoising helps models learn the meaningful latent models which lead to accurate predictions.

When considering the model performances with the size of the training data there is a clear increase. However, in contrast to the MF model curve, the deep CF model curves showing a steep gradient which implies that increasing the training data size has a more impact on the deep CF than the MF model. When there is more data autoencoder based deep CF models tend to learn good latent representation which leads to high performing predictive models. However, when comparing the AAE curve with both VAE and AE the increase in accuracy is moderately uniform. This is due to the fact even though there are less training samples, AAE has a complete latent model due to adversarial learning.

4.3.3.2 Hybrid Recommender System Evaluation

Evaluation of the proposed Hybrid recommender system was executed on the previous user promotion data. Dataset preparation for the overall recommender system evaluation is based on the promotion usage time. Following are the steps for validation of the recommender system model.

1. Divide the overall user transactions into training and testing data. (Ex: first nine month transactions for the training and last three months transactions for the testing)
2. Aggregate the transactions for each user in the training period and train the hybrid collaborative filtering model for interaction prediction.

3. Use the training set for predicting the user-promotion interaction via the trained hybrid model.
4. For each user in test data, sort promotions based on the predicted interaction value. This will be the final recommendation list of promotions for that user.
5. Out of the recommendations in the recommendation list, select k recommendations, and evaluate the performance using the Mean Average Precision (MAP) as defined by equation 4.2.

$$MAP = \frac{\sum_{u=1}^{m} AveP(u)}{m} \tag{4.2}$$

Here u = 1, 2, 3… m are the users in the test data. AveP(u) is the average recommendation precision for the user u. AveP() is defined as below.

$$AveP(u) = \frac{\sum_{p=1}^{n} \frac{1}{index(p)}}{n} \tag{4.3}$$

p = 1, 2, 3… n is the sorted promotion recommendation list for the user u. Here the index function returns the index which the promotion category p found in the sorted promotion usage list of the user u.

We evaluated the hybrid recommender system performance against the state-of-the-art matrix factorization Collaborative Filtering model and the novel AAE based Collaborative Filtering model. Recorded results can be seen in Table 4.10.

Results

Table 4.10: Mean Average Precision score for recommender systems

Training Size	MF Collaborative Filtering RS	AAE Collaborative Filtering RS	Hybrid Collaborative Filtering RS
50%	0.53	0.54	0.52
75%	0.55	0.60	0.59
100%	0.58	0.63	0.66

The recommender system evaluation experiment has the following main goals.

1. Evaluating the recommender system performance - Hybrid model performance was compared with the AAE CF and also with the state-of-the-art MF CF model.
2. Evaluating the performance variations of the Hybrid recommender system model with the training data size - Training the models on different proportions of the overall training dataset and calculated the aggregated MAP score.

When evaluating the MAP scores of recommender systems, hybrid recommender system tend to outperform the other models when the complete training dataset is used. Thus, we can emphasize the importance of employing side information towards a better recommender model. However, the Hybrid model performance tends to decrease when the training dataset size is lowered. When we consider the attribute complexity, the hybrid model is in the top because of the additional side information related to user activity and demographics. Consequently, when there are fewer training samples due to the high complexity of features, the latent model of the hybrid model become less expressive and exhibits poor performance in promotion recommendations.

4.4 Summary

In this section, all the evaluations of the proposed techniques and methods were presented. First, we discussed the review of the user profiling methods introduced by our research. User profiling based demographic classification models performed well, especially the age group prediction when compared with the baseline accuracy values.

Next step in the evaluation was to evaluate the customer segmentation modelling. Under this evaluation, GMM and OPTICS clustering technique performances were compared and with a proper cluster analysis OPTICS clustering was selected as the best method for customer segmentation in MoDS.

The final step in the evaluation was to evaluate the recommender system methodologies for personalized services in MoDS. First, the novel Adversarial Autoencoder based Collaborative Filtering (CF) model performance was evaluated against well-known deep CF models and the state-of-the-art Matrix Multiplication technique. It was evident that the novel model is performing quite well and outperforms other comparable models in the task of predicting user-promotion interactions. Then we evaluated the overall recommender system methodology by utilizing previous promotion usages. We introduced a Hybrid model that incorporates side information such as user activity profiles and demographic profiles alongside with the standard AAE based CF architecture. Through this evaluation, it was evident that the hybrid recommender system can outperform the other individual CF models.

Chapter 5

5 CONCLUSION

This research presents a comprehensive methodology for customer profiling and segmentation in Mobility on Demand systems. To the best of our knowledge, this work is the first systematic research on how to improve service and management in MoDS by employing customer profiling and segmentation.

We propose three distinguish profile vectors which can be developed in any MoDS. First profile vector was based on straightforward user-system interactions, where state-of-the-art RFM model and other mobility attributes were aggregated. Second profiling model was based on inferred activity probabilities of the users based on their travel locations. Third profile vector was derived from the first two profiling models, which is based on user demographics. Under the demographic profiling, we addressed the problem of predicting user demographics in MoDS. After setting the framework for user profiling in MoDS, we proposed a methodology to utilize profile vectors for user segmentation. GMM and OPTICS clustering models showed good performance in forming user segments in MoDS.

Our work was further extended into enabling personalized services in MoDS which help to augment the service. This work was demonstrated by creating a working model of a recommender system for promotion recommendation in MoDS. We analysed the new frontier of deep recommender systems, which utilize deep generative networks and was able to propose a novel Collaborative Filtering model based on Adversarial Autoencoders. And with the experiments done on a local MoD service, we have shown that the novel AAE based deep CF model tends to outperform other state-of-the-art deep CF models based on AE and VAE. Furthermore, deep CF models on MoDS promotion recommendation task outperforms the state-of-the-art Matrix Factorization model by a large margin due to the extra sparcity, which can be found in the user-promotion interactions. In addition, we propose a hybrid mechanism which integrates the CF models with other valuable information such as user activities and demographics. By evaluating the hybrid RS on historical data we can see an extensive overlap in the usage set and the recommended set of promotions which is implied by the higher mean average precision values.

5.1 Future Works

As future work related to this research we try to model more profile vectors which in fact acts as fundamental building blocks for smart MoDS. Demographic profiling can be further improved by identifying more demographic features such as living and working conditions of the user. Activity profiling is also open for improvements by adding more inferred activities into the dataset.

Furthermore, the evaluation done on the Hybrid RS is currently based on the historical data of the system. Our plan is to extend the proposed model to the real-time live recommendations and evaluate the performance on live user feedback.

References

[1] Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26(2), 173-194.

[2] Trusov, M., Ma, L., & Jamal, Z. (2016). Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science*, *35*(3), 405-426.

[3] Farias, V. F., & Li, A. A. (2019). Learning preferences with side information. *Management Science*.

[4] Palaniappan, S., Mustapha, A., Foozy, C. F. M., & Atan, R. (2017). Customer Profiling using Classification Approach for Bank Telemarketing. *JOIV: International Journal on Informatics Visualization, 1*(4-2), 214-217.

[5] Kanagasabai, R., Veeramani, A., Shangfeng, H., Sangaralingam, K., & Manai, G. (2016, December). Classification of massive mobile web log URLs for customer profiling & analytics. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 1609-1614). IEEE.

[6] Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism management perspectives, 18*, 153-160.

[7] Kashwan, K. R., & Velu, C. M. (2013). Customer segmentation using clustering and data mining techniques. *International Journal of Computer Theory and Engineering, 5*(6), 856.

[8] Güçdemir, H., & Selim, H. (2015). Integrating multi-criteria decision making and clustering for business customer segmentation. *Industrial Management & Data Systems, 115*(6), 1022-1040.

[9] Brito, P. Q., Soares, C., Almeida, S., Monte, A., & Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing, 36*, 93-100.

[10] Duc, D. T., Son, P. B., & Hanh, T. (2017). A Resampling Approach for Customer Gender Prediction Based on E-Commerce Data. *Journal of Science and Technology: Issue on Information and Communications Technology, 3*(1), 76-81.

[11] Brdar, S., Culibrk, D., & Crnojevic, V. (2012, June). Demographic attributes prediction on the real-world mobile data. In *Proc. Mobile Data Challenge by Nokia Workshop*, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK.

[12] Shrestha, P., Rey-Villamizar, N., Sadeque, F., Pedersen, T., Bethard, S., & Solorio, T. (2016, May). *Age and Gender Prediction on Health Forum Data. In LREC*.

[13] Ortega-Tong, M. A. (2013). Classification of London's public transport users using smart card data *(Doctoral dissertation, Massachusetts Institute of Technology)*.

[14] Zhang, Y., & Cheng, T. (2017, January). Feature Extraction for Long-term Travel Pattern Analysis. In *Proc. 25th GISRUK Conference* (pp. 39-42).

[15] Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., & Ratti, C. (2010, August). Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *International Workshop on Human Behavior Understanding* (pp. 14-25). Springer.

[16] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR), 52*(1), 5.

[17] Huang, Z., Zeng, D., & Chen, H. (2007). A comparison of collaborative-filtering recommendation algorithms for e-commerce. *IEEE Intelligent Systems, 22*(5), 68-78.

[18] Bobadilla, Jesús, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. "Recommender systems survey." *Knowledge-based systems 46 (2013):* 109-132.

[19] Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). Springer, Boston, MA.

[20] Yang, Z., Wu, B., Zheng, K., Wang, X., & Lei, L. (2016). A survey of collaborative filtering-based recommender systems for mobile Internet applications. *IEEE Access*, 4, 3273-3287.

[21] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering,* (6), 734-749.

[22] Gogna, A., & Majumdar, A. (2015). A comprehensive recommender system model: Improving accuracy for both warm and cold start users. *IEEE Access, 3*, 2803-2813.

[23] Bokde, D., Girase, S., & Mukhopadhyay, D. (2015). Matrix factorization model in collaborative filtering algorithms: A survey. *Procedia Computer Science, 49*, 136-146.

[24] Ungar, L. H., & Foster, D. P. (1998, July). Clustering methods for collaborative filtering. In *AAAI workshop on recommendation systems* (Vol. 1, pp. 114-129).

[25] Lian, D., Ge, Y., Yuan, N. J., Xie, X., & Xiong, H. (2016, July). Sparse Bayesian Content-Aware Collaborative Filtering for Implicit Feedback. In *IJCAI* (pp. 1732-1738).

[26] Dong, X., Yu, L., Wu, Z., Sun, Y., Yuan, L., & Zhang, F. (2017, February). A hybrid collaborative filtering model with deep structure for recommender systems. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[27] Mnih, A., & Salakhutdinov, R. R. (2008). Probabilistic matrix factorization. In *Advances in neural information processing systems* (pp. 1257-1264).

[28] Chu, W. T., & Tsai, Y. L. (2017). A hybrid recommendation system considering visual information for predicting favorite restaurants. *World Wide Web, 20*(6), 1313-1331.

[29] Bansal, T., Belanger, D., & McCallum, A. (2016, September). Ask the gru: Multi-task learning for deep text recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 107-114). ACM.

[30] Li, S., Kawale, J., & Fu, Y. (2015, October). Deep collaborative filtering via marginalized denoising auto-encoder. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 811-820). ACM.

[31] Liang, D., Krishnan, R. G., Hoffman, M. D., & Jebara, T. (2018, April). Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 689-698). International World Wide Web Conferences Steering Committee.

[32] Cai, X., Han, J., & Yang, L. (2018, April). Generative adversarial network based heterogeneous bibliographic network representation for personalized citation recommendation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[33] He, X., He, Z., Du, X., & Chua, T. S. (2018, June). Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 355-364). ACM.

[34] Wang, Q., Yin, H., Hu, Z., Lian, D., Wang, H., & Huang, Z. (2018, July). Neural memory streaming recommender networks with adversarial training. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2467-2475). ACM.

[35] Mescheder, L., Nowozin, S., & Geiger, A. (2017, August). Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 2391-2400). JMLR. org.