National College of Ireland

Customer Behaviour Prediction Using Recommender Systems

MSc Research Project
Data Analytics

Ifeoma Delphine Onyeka
Student ID: x20189231

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ifeoma Delphine Onyeka
Student ID:	x20189231
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Vikas Sahni
Submission Due Date:	15/08/2022
Project Title:	Customer Behaviour Prediction Using Recommender Systems
Word Count:	6721
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Ifeoma Delphine O
Date:	13th August 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Customer Behaviour Prediction Using Recommender Systems

Ifeoma Delphine Onyeka
x20189231

Abstract

It is possible to improve business decisions by gaining insight into consumer behavior through predictive research.The goal is to predict customer purchasing habits and recommend items based on their behavioral data. It is possible to improve business decisions by gaining insight into consumer behavior through predictive research. This study compares the statistical approach to data mining in predicting customer behavior with Logistic regression, Random Forest, Support Vector Machine and K-Nearest Neighbour. As a result, Logistic regression achieved 80% accuracy, An accuracy of 82% was gotten from Random Forest while K-Nearest Neighbour gave 81% and finally For Support Vector Machine, 81% accuracy was obtained. Random Forest gave a higher accuracy.

1 Introduction

For several decades, businesses have invested heavily in predicting customer behaviour. During the process, they adopt numerous data prediction models to provide detailed insights which guide the business operations and processes. Consumer behaviour involves analyzing the consumers, their preferred processes, their consumption choices, and their disposal of products. In online marketing, predicting consumer behaviour is important as it describes the eCommerce journey of consumers. Once online marketing businesses predict the behaviour of their customers, they can quickly establish ways of leading them to the business's end goals. Various goods and services are available via online platforms due to recent technological advances, such as e-commerce. Many products are available in online marketplaces, so it is difficult for customers to find the one best suited to their needs. By analyzing their needs or behaviour, recommender systems can reduce the number of alternatives individuals may prefer by analyzing their preferences. For brands to survive in the era where customers can browse and visit multiple platforms online, they need to have detailed insights into their customers' behavioural and social influences.

1.1 Background

A recommender system is all about generating recommendations for products and items that may interest a collection of users. A recommendation system is critical in assisting e-commerce customers in overcoming information overload. The recommendation system has come a long way over the years. To effectively assist customers, it is critical to identify their specific needs/requirements and suggest a tailored buying list. Researchers

have proposed various approaches to achieving this goal, but the proposed framework is the most effective. The recommender system will be able to learn about a customer's preferences and then recommend an item based on past purchases and necessitates. This technique helps enterprises to provide multiple options to their customers. An adequately shaped recommender system can add significance to an e-commerce company by providing data about a given item, increasing repurchase agreements for a customer, and sustaining a customer's loyalty by conveying appropriate services. It is necessary to investigate the critical factors that may influence analyzing the favourable outcome of the recommender system in digital marketing.

1.2 Motivation

Recommender systems were developed primarily to enhance search results. They help individuals find the desired items on the Internet more efficiently. The main aim of RS is to recommend personalized items (products or services) to the users and is evolving to be more user-centric. Personalized recommendations based on the history of a user's actions are suggested by a sound recommender system that employs active and selective information-filtering techniques on the data collected. Many types of user-provided data can be incorporated into recommendation systems, including expressive data like favourites and rankings, clickstream data, and indirect information like logs that describe the customer's behavioural traits. Understanding a user's preferences through different data mining approaches is essential before recommending items based on their explicit and implicit data. The service must filter the numerous item details offered to make recommendations according to consumers' preferences or requirements using data analysis outcomes. Online marketing using recommender systems is an efficient way to achieve client-centricity. Online stores increasingly incorporate recommender systems into their websites to provide personalized recommendations to shoppers, enhancing the user experience and driving sales.

Over 4.5 months, real-world data was collected from e-commerce websites to aid customers in making better decisions over the long run by providing them with better recommender systems through several global channels. Several models are compared in the study, such as Logistic Regression, support vector machine, K-nearest Neighbor, Random Forest, and Neural Network Regression. To recommend users better products, a comparison between these models is conducted. In this study, we look at the impact of customer behaviour and product similarity and examine how predictive models work and how they affect consumer behaviour overall. Various studies on recommender systems in online marketing have been conducted, but little research has been done on determining the correlation between the progress of recommender systems and customer behaviour.

1.3 Reseach Question

- The extent to which machine learning model can be used to predict Customer purchasing power using a recommender systems.

- Comparison of different Machine learning models in customer behaviour prediction using recommender systems on historical data.

2 Related Work

2.1 Recommender Systems

E-commerce platforms and other data-centric online services today have a problem of information overload that recommender systems (RS) are designed to eliminate. With the help of implicit information from internal e-commerce systems and interactions with users, they assist users in exploring and exploiting the system's information environment. In research done by Yu et al. (2018), they proposed an aesthetic-based clothing recommendation method based on a paired matrix and tensor factorization model. They extracted image features and aesthetic features using CNN models. Based on experiments on real-world datasets, the suggested technique can grasp customers' tastes. Wang et al. (2017) investigated the effects of graphical representations on point-of-interest (POI) or location recommendation. They used CNN models to retrieve image features and suggested a POI recommender system with visual content enhancements (VPOI). They investigated the interrelations between visual content, latent user factor, and latent location factor using probabilistic matrix factorization (PMF).

Chen et al. (2017)described an intelligent search tool for internet purchases that made use of the Amazon dataset and two CNN models, VGG and AlexNet. Related to product images, they utilized neural networks to give the nearest product. The results show that the method's accuracy has improved. They likewise utilized Jaccard similarity to compute the similarity mark. Nevertheless, because of small computational resources, their study needs scalability, and they are content with about 0.3 percent of the dataset images. Tuinhof et al. (2018) presented an image-based recommendation system that trains a CNN model for solving image classification tasks using a fashion dataset. They fed the ranking system with the trained model acting as a visually aware feature extractor. In addition, they utilized the ball tree search to conquer memory resource constraints, a unique implementation of the K-NN algorithm for simple ranking. They concluded that the technique could be useful in other fields, such as music.

Abdul Hussien et al. (2021) focus on developing a recommendation system to address the problems of cold start, scalability etc. and obtain greater reasonable prediction results. This is accomplished by creating the system depending on customer behaviour and collaborating with statistical analysis to assist in decision making, to be used on an e-commerce site and to enhance its effectiveness. The test findings using precision, recall, F-function, mean absolute error (MAE), and root mean square error (RMSE) metrics, that are being used to analyze system performance, demonstrate the project's contribution. The results demonstrated that applying statistical methods enhance decision-making, which is used to improve the accuracy of recommendation lists recommended to customers.

By providing better and more accurate marketing strategies, marketing decision-makers can benefit from deep learning techniques. According to Salehinejad and Rahnamayan (2016) research, recurrent neural networks are used to predict customer behavior based on variables such as client loyalty number (CLN), recency, frequency, and monetary (RFM). According to the experiment results, RNNs can predict RFM values of customers efficiently. A recommender system based on this model can then be used to manage loyalty programs and exclusive promotional offers. According to (IbukunT et al., 2016) research, the majority of studies investigated customer retention prediction, and administrative datasets were applied more for prediction than other types of datasets. Furthermore,

by comparing the statistical method to data mining in customer behaviour prediction, this study demonstrated that data mining is largely exploited for analytical purposes. In comparison, the Artificial Neural Network is by far the most popular data mining method for predicting consumer habits. The research study concludes that tasks are performed to predict consumer behavior, which is critical to the company. Many of the studies discussed customer retention modeling, which was done primarily by using organizational data as a predictive dataset.

Wang et al. (2009) made a proposal for how wireless network companies could make use of recommender systems to understand also evade customer churn. The aim of this research was to obtain an accurate analysis by analysing the data of over 60,000 transactions which belong to over 4000 members by using the decision tree algorithm. This data contains a period of three months, with its first nine weeks being used as training data, while the final month was used as testing data. The paper got results which proved to be very useful for making strategic recommendations to retain customers. According to Wakil et al. (2019), a structural modelling technique can be employed for use in validating the success of the recommender systems in the e-commerce industry based on the customer history, criteria of prices and the classification of products. The paper discussed the impact the above criteria had on the success of the recommender system. The results obtained from the experiment showed the efficiency of the proposed model in terms its success in the e-commerce industry.

The study conducted by Zhou and Hirasawa (2019) investigated how the genetic network programming (GNP) and the ant colony optimization (ACO) were implemented in solving the sequential rule mining problem in transactional databases that are time related. The paper also discussed how the rapid development of technology has led to an explosion of information and excellent recommender systems are now able to capture a customer's interests or potential needs accurately and in a proactive manner. The paper made use of a methodology which analyzed the customer database of a real-world online supermarket. The recommender systems have also been used in other fields asides for customer preference purposes. Adomavicius and Tuzhilin (2011) talked about how relevant contextual information is vital when making use of recommender systems. It argued that a number of research work in the past have majorly focused of merely recommending the most relevant items to users, but do not usually consider key additional information like time and location. The article then discussed how key context can be modelled into recommender systems. It came up with three algorithmic paradigms which include contextual prefiltering, post-filtering and modelling. The paper discussed how possible it is to combine multiple context-aware recommendation techniques into one unifying approach. The classification and review of some recent research in recommender systems used in the Computer Science and Information Systems field was discussed by Jannach et al. (2012). This research work was particularly aimed at recognizing open issues and probable directions for future research works. 330 papers on recommender systems which were published in high profile journals and conferences between the year 2006 – 2011 were used as the basis for this analysis.

2.2 Customer Behaviour

Abid et al. (2018) researched logistic regression to predict the customer behaviour in defaulting loan payments across commercial banks in Tunisia. The research was conducted when loan defaulting was common in Tunisia, and most banks allocated credits to

bad borrowers. To predict future trends in consumer behaviour regarding loan defaults, the banks opted to predict the behaviour of customers and borrowers and establish the optimal loan limit to be issued. The researchers utilized logistic regression and discriminant analysis to predict customer behaviour. The researchers then compared the results obtained through logistic regression and discriminant analysis. The results revealed that logistic regression was more effective in predicting customer behaviour than discriminate analysis. The logistic regression (LR) model yielded excellent results, which were 99% good in classifying the customers and predicting their behaviour. The LR model portrayed only a 1% error rate, significantly lower than the one recorded by the discriminate approach (DA) model . The DA model yielded a classification rate of 68.49% with an error rate of 31.51%. The research indicated that the LR model was far better than the DA model in predicting customers' behaviour while borrowing loans from Tunisian commercial banks.

Juan et al. (2017)researched the factors that influence the consumer behavioural activities on the purchasing of green building premiums. The research was motivated by the increasing cases of environmental degradation, which led to an energy crisis. Encouraging customers to purchase green buildings was essential as it could effectively achieve environmental sustainability. Even though greenhouses are valued at higher prices than ordinary buildings, most customers believed their higher prices were justified since the houses promote environmental sustainability. To determine and analyze customer behaviour, the researchers utilized the Howard-Sheth model of identifying customer behaviour. They then created an artificial neural network (ANN) to create a pricing model that could assist them in predicting the price premiums. The research revealed that ANN was approximately 94% effective in predicting customer behaviour. The researchers then compared the results of the ANN with those obtained by the multiple regression analysis. They then concluded that ANN was an effective model in predicting customer behaviour towards green building pricing and developing effective market strategies.

De Caigny et al. (2018) developed a hybrid algorithm that could predict customer behaviour. They utilized the logistic regression model and the decision tree model in their model. According to the researchers, the logistic regression model and decision trees are very effective methods for predicting and comprehending customer behaviour. However, the researchers noted that decision trees were less effective than the logistic model while handling linear relationships between variables. Similarly, logistic regression struggled to predict the interaction between variables. They identified a logit leaf model (LLM) as a better predictor than the other two models since it is constructed from the segments of the dataset rather than the actual data. Prediction of customer behaviour is essential among retail companies as it enables them to estimate the sales of various products from their product category. Javed Awan et al. (2021) utilized a big data approach to predict customer behaviour based on the sales volume recorded during black Fridays. The study's primary objective was to assist retail companies in developing a personalized structure for promoting their products to their customers even during times of uncertainty due to pandemics. The researchers utilized the sales data extracted from the Kaggle website and conducted qualitative and quantitative studies. They used linear regression to predict the behaviour and compared the results with the random forest model. They first did not utilize the Spark framework in predicting the data. Results indicated that the linear regression model was 68% effective in predicting customer behaviour while the random forest model was 74% accurate. Afterwards, the researchers used the Spark framework, whereby the accuracy of their prediction models changed. The linear regression model

recorded an accuracy level of 72%, whereas the random forest model recorded an 81% accuracy level.

Khodabandehlou and Rahman (2017) utilized supervised machine learning to analyze and predict customer behaviour. The research focused on creating a predictive framework that could be utilized in business and followed all the prediction stages. During the research, behavioural data were collected from the participants before the analysis variables were formed. Data training and testing were then conducted, and the prediction model used in the research was identified. The prediction was conducted using discriminant analysis and other predictive models. The results revealed that the ANN model was the most accurate, followed by logistic regression. The decision trees model emerged as the least accurate. Analyzing data drawn from Google trends is one of the most effective ways of predicting customer behaviour. Silva et al. (2019) researched the fashion trends using Google data. The researchers explored diverse predictive models to forecast fashion trends. The analysis revealed that single univariate models are not good predictors of customer behaviour on fashion trends. Examples of single univariate models include the ARIMA model, neural network autoregression (NNAR), and exponential smoothing. The results portrayed NNAR as the least effective model for predicting customer behaviour.

Mansouri et al. (2016) evaluated the effectiveness of the ANN model and logistics regression in predicting the customer behaviour that could drive bans into bankruptcy. They used the Tehran stock exchange data from three consecutive years to analyze the consumer behaviour trend. The research involved an analytical mathematical approach with over seven independent variables to test the hypothesis. The research results revealed that both models could predict customer behaviour and, in turn, the causes of bankruptcy in the Tehran stock exchange market. However, the effectiveness and accuracy of the two techniques were different since ANN was more accurate and efficient than linear regression. ANN was more effective in bankruptcy prediction than linear regression, even though both techniques proved effective in analyzing and predicting the overall consumer behaviour.

With technological advancement and e-commerce, businesses and customers have many business opportunities. Technology has facilitated the development and advancement of marketing behaviour. It has made it easier for customers to shop online, and businesses can easily record more sales via eCommerce. Ecommerce has helped businesses broaden their brands; it is more convenient to customers, has increased the reach of businesses, is scalable, and has presented numerous marketing opportunities. However, entrepreneurs need to predict their customer's behaviour to reduce the risk of unexpected changes and plan effectively for the business expansion to cater for their emerging needs.Gordini and Veglio (2017) researched how businesses can predict customer behaviour and establish retention strategies in the e-commerce industry. The researchers utilized the newly developed support vector machine (SVM) to predict customer behaviour. The results revealed that the SVM is critical and accurate in behaviour prediction. Its results are more accurate when applied to non-linear marketing data, and the technique outperforms other predictive models.

Over the past several decades, businesses have resorted to predicting customer behaviours. Businesses are heavily investing in diverse techniques to predict customer behaviour. The primary motivation behind the prediction is to obtain detailed customer insights, which guide almost all business processes. Accurate prediction is key to effective marketing and business sustainability as it enables businesses to reduce uncertainty risks. Consequently, businesses that conduct more accurate predictions of customer data

are likely to record increased revenue, better their brand, and reduce uncertainty-related losses. Zuo et al. (2016) utilized machine learning techniques to predict consumer behaviour in a grocery business. Most businesses have, over the years, embraced linear equations such as logistic regression and discriminant analysis to predict customer behaviour. As more technologies emerge and consumer data becomes more complex, linear models become insufficient prediction techniques. According to the research conducted by Zuo et al. (2016), machine learning techniques such as SVM and the Bayes classifier are more robust and effective in predicting consumer behaviour than the linear equation techniques. Customer behaviour analysis is essential in financial institutions as it enables them to avoid losses and determine the creditworthiness of borrowers. Financial institutions globally have resorted to analyzing and predicting consumer behaviour to quantify and lower the risk they are exposed to. When performing customer analysis and prediction, financial institutions evaluate the probability of customers defaulting on loan repayment, determining the risk to be experienced if the customers default their loans, and establish their exposure at default. Online marketers also utilize the same technique to determine the risk they are exposed to in case of any change in customer behaviour. Even though there exist several mathematical and statistical techniques that can predict customer behaviour, good classifiers and predictors have not yet been identified. Abedin et al. (2018) predicted the risk of customers defaulting credit using the SVM and a probabilistic neural network (PNN) model. The two models were compared to determine the most effective model for predicting customer behaviour. The empirical study revealed that the PNN model was a better predictor of customer behaviour and more robust than other techniques. Moreover, the research findings revealed that emerging performance evaluation techniques were better predictors than traditional ones.

Numerous factors influence the personality, character, and behaviour of customers. Customers portray diverse characteristics, with individuals portraying diverse motivation levels, occupations, income levels, social status, beliefs, and attitudes. Numerous researchers have applied data mining to sort through large datasets and identify critical patterns and relationships that can help organizations solve existing problems through data analysis and make informed choices. The data mining process involves sophisticated technology stored in the designated database. The data is usually collected without the knowledge of the customers and is used to determine their purchase behaviour. Maheswari and Priya (2017) utilized the SVM classifier to predict customer online marketing and shopping behaviour. The researchers aimed to identify whether the SVM model effectively analyzed mined data and predicted customer behaviour. Commonly, SVM is used in predictive analysis to assign new data elements to a particular data category. The binary classifier assumes two target values from the presented data and is used to optimally discriminate between the two data groups and create a margin between the clusters. The research revealed that SVM is one of the best and most accurate models for predicting customer behaviour. Customer behaviour has a direct impact on their bidding process. To acquire detailed insights on the impacts of customer behaviour in predicting the bidding price, Petruseva et al. (2016) conducted a study using an SVM. The research focused on predicting the bidding prices in the construction industry. This was because budding price, determined by the overall customer behaviour, directly impacts the construction business. The researchers collected and analyzed information from fifty-four members. Out of the collected information, data from only twenty-six participants was used for further study. The researchers then used the SVM to develop a forecasting model. The study results indicate that SVM is helpful and more efficient in developing

prediction models which are useful in analyzing customer behaviour. The SVM obtained an accurate model with a mean absolute percentage error of approximately 2.5%. This implies that SVM is very useful in developing accurate and efficient models that can be used to predict customer behaviour in online marketing and purchasing.

The invasion of the Covid-19 pandemic affected almost all sectors and shook governments globally. Healthcare facilities became flooded with patients, so they could no longer admit more patients. Consumer habits changed, and most customers shifted to purchasing goods and services online. Businesses also shifted their operational model and embraced online marketing. The pandemic greatly influenced consumer buying habits, forcing many to purchase online. With most customers and businesses shifting to online platforms, customer behaviour analysis was essential as it would help businesses reduce risks and adjust their operations based on the emerging customer behaviour. Jamunadevi et al. (2021) researched to identify how the coronavirus pandemic influenced online buying behaviour. The researchers administered questionnaires as a tool for data collection. The collected information was analyzed using the analysis of variance (ANOVA), Chi-square, and ranking methods. The results indicated that the pandemic greatly influenced customer behaviour since most preferred online shopping and delivery of goods over physical marketing. ANOVA and Chi-square proved to be very effective and accurate models in predicting consumer analysis behaviour during and after the pandemic.

Customer relationship management (CRM) is one of the most critical aspects businesses must consider to learn about their customers. Effective CRM helps entrepreneurs understand their customers, their general behaviours, how and why they purchase their products, purchasing patterns, and histories. For businesses to better predict their customer behaviours, they must maintain a positive CRM as it allows them better anticipate their customers' needs and establish ways of satisfying them. Aissa et al. (2018) evaluated the effect of CRM on predicting customer behaviours. Their study was limited to Ooredoo telecommunications and focused on determining the link between CRM and customer behaviour. The researchers collected customer data, utilized an analytical approach, and analyzed the data statistically. They utilized analysis of variance (ANOVA) to establish the effect of interactive communication and the overall customer behaviour. The prediction technique (ANOVA) proved to be an effective forecasting technique that could identify customer behaviour patterns in online marketing. The researchers identified that CRM is crucial in determining the overall customer behaviour, and organizations can obtain a significantly competitive market share when they handle their customers appropriately.

3 Methodology

This chapter concisely describes the methodology used in conducting this research.

3.1 Knowledge Discovery in Databases

The Knowledge Discovery in Databases (KDD) methodology is utilised in the experiments. The KDD methodology is a recursive and sequential process that occurs in different stages and is used to discover information (knowledge) in data. The stages are: Data Selection, Data Preparation, Data Mining etc. The KDD methodology process is shown in the diagram below:

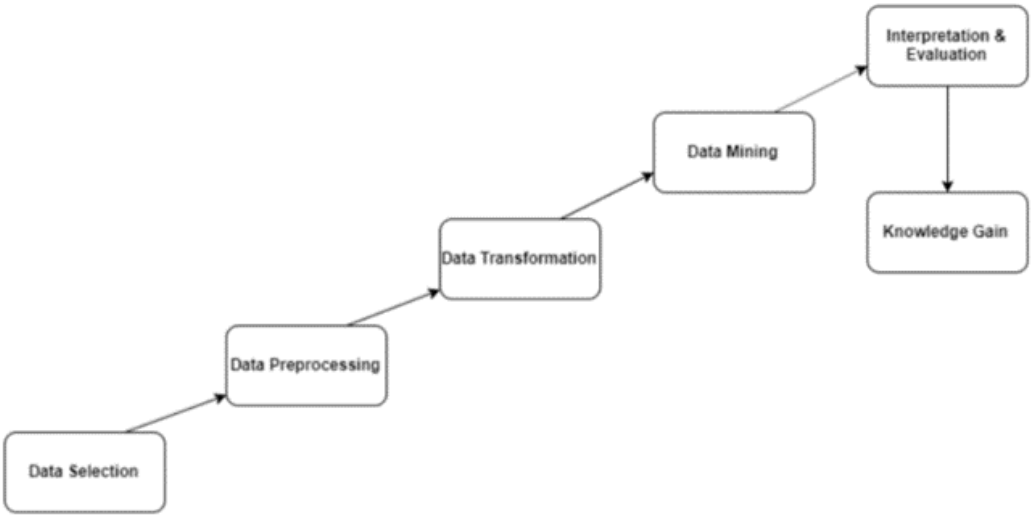

Figure 1: KDD Process

3.2 Dataset Resource and Explanation

Dataset Link: `https://www.kaggle.com/retailrocket/ecommerce-dataset`. The dataset is available from Kaggle and its public data doesn't require permission or authorization. The data is a collection of real-world operations based on 4.5 months of launching a website and analyzing its raw data without content transformation. We aim to build an implicit recommendation system based on implicit feedback and predict customer purchases with this dataset. A hash is applied to all values for reasons of confidentiality. The dataset consists of three files.

- Events.csv : behaviourial data is stored in this file. Events such as "click", add to cart", "view" and "transactions". 69332 are for "add to cart", 2664312 are for "view" and 22457 for "transaction" events. All together making it 2756101 events. These transactions were created by 1407580 special customers. Approximately 90% of events have corresponding properties stored in "item properties.csv. Example "1439694000000,1,view,100," means visitorId = 1, clicked the item with id = 100 at 1439694000000 (Unix timestamp)

- Category.csv – 1669 rows make up the category tree file. A child categoryId is listed for every row in the file, along with its parent categoryId. Example - Line "200, 300" means that categoryid=2 has a parent with categoryid=300. The line "200," It means there is no parent for categoryid in the tree

- Item_properties.csv . In this file, it includes 20275902 rows showing different properties, describing 417 053 unique items. Because of file size limitation, the file was split into two. Due to the fact that an item's property can change over time (e.g., price), each row in the file has its own timestamp. Since the property of an item can vary in time (e.g., price changes over time), every row in the file has corresponding timestamp. As a result, the file consists of weekly snapshots concatenated with the behavior data. Nevertheless, if a property of an item remains constant over time,

the file will only contain one snapshot value. Time stamp columns are present in item properties files because all of their properties are time related, e.g. prices, categories, etc. In the beginning, this file has over 200 million rows of snapshots from every week in the events file. Our form has changed from a snapshot to a change log due to combining consecutive constant property values. The file would therefore contain only one instance of constant values.

The row count has been cut in 10 as an outcome of this move. With the exception of "categoryid" and "available," all property values in item properties.csv are hashed. Item category identifier is contained in the value of the "categoryid" attribute. The "available" property's value indicates whether the item was available; a value of 1 indicates that it was, and a value of 0 otherwise. All numbers were preceded by the "n" char and had a precision of three digits after the decimal point, for example, "5" will become "n5.000" and "-3.67584" will become "n-3.675." All text values were standardized, hashed, and processed numerically as described above. For example, the sentence "Hello world 2017!" will become "24214 44214 n2017.000."

3.3 Data Preparation and Transformation

After retrieving the data from the source, exploratory data analysis is done to identify the relationships between the variables. Figure 2 below displays a pie chart of the target variable in the dataset. There are three classes, View, Add to Cart and Transaction. From the figure, 96 percent of the customers viewed the product, 2.5 percent added the product to cart and finally, 0.8 percent purchased the product. The stage also involves transforming the dataset to make it appropriate for the machine learning models.

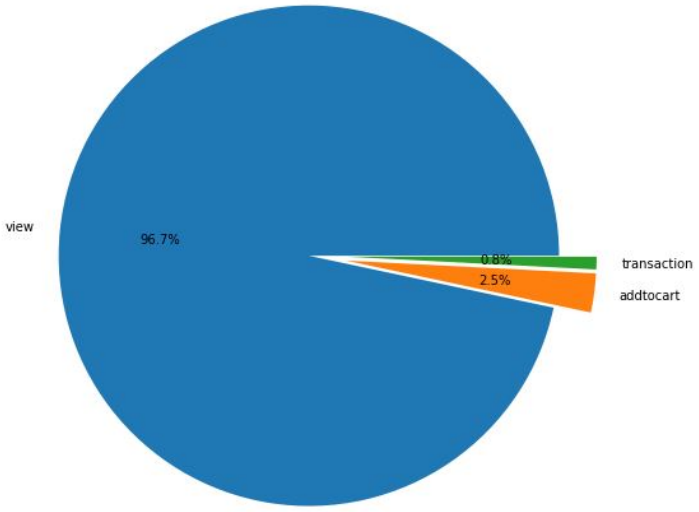

Figure 2: Target Variable

3.4 Data Mining

This research sought to accurately predict customer purchase intention. The literature provides an extensive list of machine learning models to pick from, although the selected

models are chosen because of their occurrence in the literature. The machine learning models are

- Random Forest
- Logistic Regression
- KNN
- Support Vector Machine

3.5 Data Preparation and processing

In order to build an accurate model and make accurate predictions, raw data must be processed, so data pre-processing is crucial to research. By pre-processing data, we mean transforming, structuring, and encoding it in a way that is easily predictable by algorithms.The data is sourced from Kaggle, there are three files: behaviour.csv, category.csv and properties.csv. The dataset used in this research contains both numerical and categorical variables.the data is explored to find the underlying relationships between all the variables in the dataset and the dataset is subsequently merged into one dataframe. The library pandas is used for preprocessing data. It provides tools for processing, analyzing and wrangling data. Numeric data is particularly easily manipulated with pandas.The data is split into testing and training sets with a 70:30 ratio, before this split is done, the data is randomised.

3.6 Data Mining Techniques

There are four experiments performed on the dataset.

1. Logistic Regression: The first experiment uses a Logistic Regression model with default parameters.
2. K-Nearest Neighbour: The second experiment is done using a K Nearest Neighbour model, The number of neighbours is tuned for this experiment with the highest number of neighbours yielding the best accuracy.
3. Support Vector Machine: There two experiments performed using the third model (Support Vector Machines). The first experiment is done with default parameters and the second is done tuning the C parameter to 0.1 and the kernel being linear. However both experiments using the Support Vector Machines produce the same Accuracy.
4. Random Forest: The fourth model is the Random Forest model, it is built using ten as the number of estimators, criterion as entropy, and a random state set to zero.

The literature highlights different evaluation metrics, however the models in this research are evaluated using Accuracy, Receiver Operating Characteristic (ROC) curve and Confusion Matrices. The chosen metrics are described below:

- ROC Curve: This is a pictorial representation of the true positive rate (TPR) against the false positive rate (FPR). The ROC Curve shows the diagnostic performance of binary machine learning classification algorithms with varying threshold.

- Confusion Matrix: Confusion matrices are used to determine the performance of machine learning classification models given a set of test data. The number of the correct and incorrectly predicted instances are broken down by class

- Accuracy: Accuracy is a metric adopted for all machine learning classification models. Accuracy measures the number of correctly classified instances (true positives and true negatives) over the total number of instances in the dataset. The formula is shown below:

3.7 Model

This section describes and explains all applied models

3.8 Model Description

1. Logistic Regression: Logistic regression is popular for binary classification problems. The concept of logistic regression is streched from multiple linear regression. A discontinuous variable is the dependent variable in logistic regression. It is mainly logistic regression that is used to predict the probability of a particular situation. According to Bahrami, Bozkaya, and Balcisoy, logistic is mainly used for variables with two dependent values. "1" and "0" are used to predict the results. A logistic function is also known as a sigmoid function. Weights or coefficient values are linearly combined with inputs(x) to predict result(y). The equation of logistic regression is as follows. y = e$\hat{}$(b0 + b1*x) / (1 + e$\hat{}$(b0 + b1*x)) In the equation y is the output, b0 is bias and b1 is the coefficient of a single value(x).

2. K-Nearest Neighbour: The KNN model is used mainly for non-linear regression, mapping a new point based on existing data points. It will be helpful to categorize a new product using a KNN model. Nearest and distant neighbors will be determined using this model. Based on our dataset, KNN will identify items whose properties are similar and predict items based on these properties.

3. Support Vector Machine: This model is frequently used to establish the outcome of a data point in a multidimensional space as well as for segmentation. It helps with text classification and stock market forecasting. As a vector in a multidimensional space, it displays our data. In our dataset, it will be applied to object attributes of the same categories.

4. Random Forest: For the purpose of prediction, this model employs numerous decision trees.To compute the result, random data points are used. It is frequently used for non-linear Regression and models a number of decision trees in order to produce its forecast. To identify objects with similar qualities and identify predictions, decision trees will be made using random forest.

3.8.1 Model Building

This section explains how models are built. Before creating a model to gain insights from things viewed, added to cart, and sold, it can be useful to group the customers for that new data frame and develop a few features for it. The first step was to add all visitor ids into a single array 14 and sort them ascendingly. There were 1407580 visitors, of whom

11719 purchased something and 1395861 just viewed stuff. The next step is to create a new data frame with the following new features: visitorid, number of items viewed, total number of views, and whether or not the visitor bought anything. The data was plotted using seasborn.

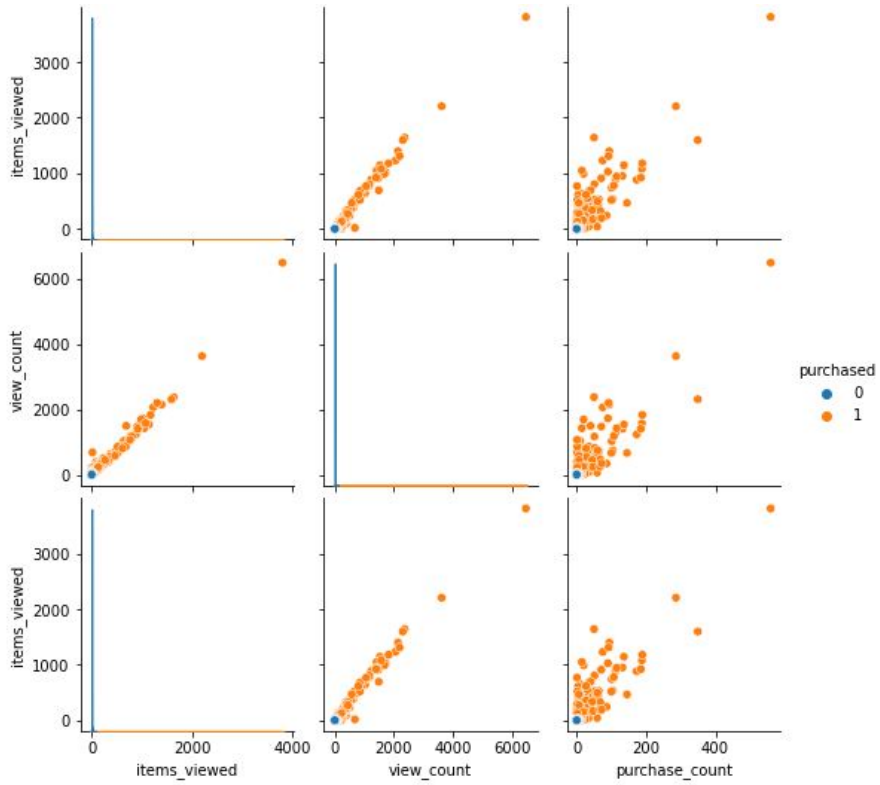

Figure 3: Number of view count and chances of purchase.

According to the plot below, the higher the view count, the greater the chance of a visitor purchasing a product. The next step is to import the libraries used to build the logical regression model. The dataset is loaded into the Pandas library, and an array is created using Numpy. The data is then split using Sklearn into ratio 70:30, the training set and the test set after defining the target variable. The train_split_function is used to split the dataset to set aside the amount of data for training and testing. The training set is applied to train the logistic regression while the test data is for validation. In building the logistic regression, the logistic regression is imported from Sklearn. The next step is to create an instance classifier and fit it into the training set, and predictions must be made on the test data. Furthermore, our model's performance is checked using the confusion matrix. Prediction and accuracy are discussed in the implementation section.

4 Implementation

The proposed solution's implementation is explained in this section. Python 3.6 is used to implement the models, together with pandas, numpy, seaborn, and matplotlib for data processing, graphing, and visualization.

4.1 Logistic Regression

The accuracy of this model's output is evaluated using the AUC ROC curve. The accuracy of this model, which was attained at 80%, indicates that 80% of visitors who made purchases were present in the data. A code snippet of the obtained accuracy is added below.

```
# Let's now use the model to predict the test features
y_pred_class = logreg.predict(X_test)
```

```
print('accuracy = {:7.4f}'.format(metrics.accuracy_score(y_test, y_pred_class)))
```

```
accuracy =  0.8011
```

Figure 4: Logistic Regression Accuracy

```
# Plot the ROC Curve
plt.figure()
lw = 2
plt.plot(fpr, tpr, color='darkorange', lw = lw, label = 'ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color = 'navy', lw = lw, linestyle = '--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc = "lower right")
plt.show()
```

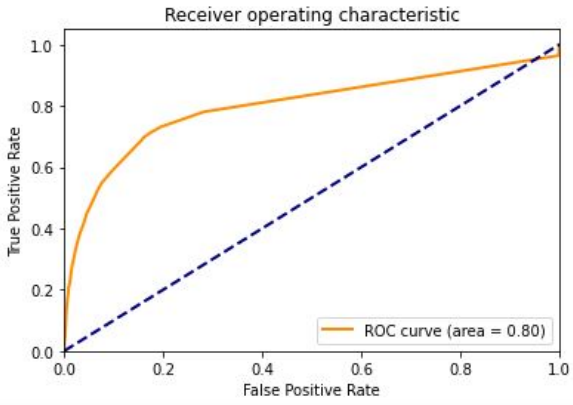

Figure 5: Logistic Regression Accuracy

The above graph shows the accuracy of the binary classification.This means the closer the orange line to the top left corner of the graph shows more accuracy.

4.2 K-Nearest Neighbour

To check the accuracy of the K-Nearest Neighbor model Train AUC and Test AUC is calculated. Achieved accuracy is 80%. It is observed above that the maximum testing accuracy is for k=6. A K-Neighbors Classifier is created with number of neighbors as 6.

```
#Generate plot
plt.title('k-NN Varying number of neighbors')
plt.plot(neighbors, test_accuracy, label='Testing Accuracy')
plt.plot(neighbors, train_accuracy, label='Training accuracy')
plt.legend()
plt.xlabel('Number of neighbors')
plt.ylabel('Accuracy')
plt.show()
```

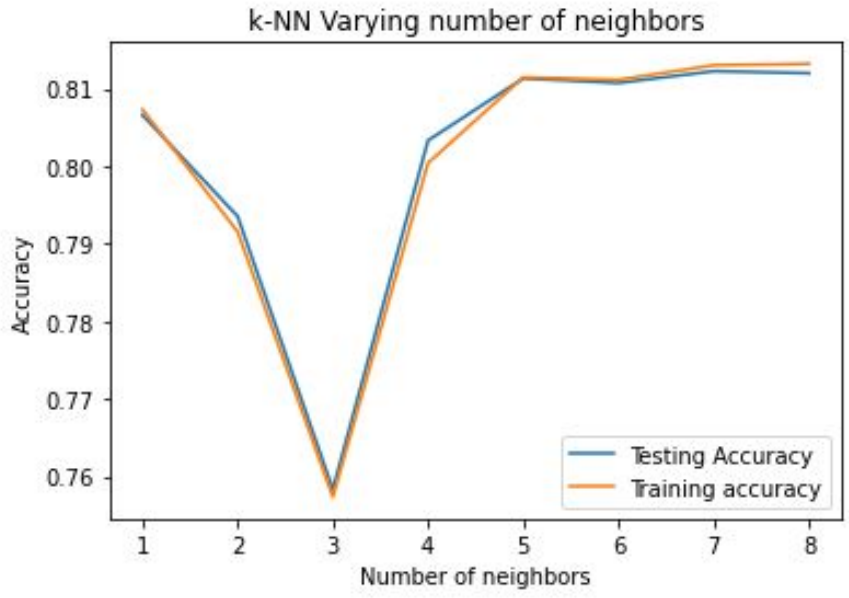

Figure 6: KNN ROC curve

```
#Get accuracy. Note: In case of classification algorithms score method represents accuracy.
knn.score(X_test,y_test)
```

```
0.81200472095768
```

Figure 7: KNN Accuracy

4.3 Support Vector Machine

In our second model, a default parameter was used in support vector machine. Based on our previous splitting of ratio 70:30 for training and testing our model, we used three different types of model in the support vector machine, the first was a linear kernel to show if we will have a different result and also hyperparameter of 0.1 instead of the usual default hyperparameter. with the 0.1, we achieved an accuracy of 81%. For the hyperparameter, we got 80%.When K-fold cross validation is used we obtained different score in each iteration. The reason for this is because of the train_test_split method used.

The dataset is splitted in randomly into training and test set. As a result this result depends on the splitting fashion and samples taken for training and test set. we also used our training data, we fit our training data into the model and made a prediction on our test data.

Running SVM with default hyperparameter

```
from sklearn.svm import SVC
from sklearn import metrics
svc=SVC() #Default hyperparameters
svc.fit(X_train,y_train)
y_pred=svc.predict(X_test)
print('Accuracy Score:')
print(metrics.accuracy_score(y_test,y_pred))
```

```
Accuracy Score:
0.8019726858877086
```

default Linear Kernel

```
svc=SVC(kernel='linear')
svc.fit(X_train,y_train)
y_pred=svc.predict(X_test)
print('Accuracy Score:')
print(metrics.accuracy_score(y_test,y_pred))
```

```
Accuracy Score:
0.8072837632776935
```

Figure 8: Running SVM

SVM by taking hyperparameter C=0.1 and kernel as linear

```
from sklearn.svm import SVC
svc= SVC(kernel='linear',C=0.1)
svc.fit(X_train,y_train)
y_predict=svc.predict(X_test)
accuracy_score= metrics.accuracy_score(y_test,y_predict)
print(accuracy_score)
```

```
0.8105715730905412
```

Figure 9: SVM by taking hyperparameter C=0.1 and kernel as linear

In our code with K-fold cross validation we split the dataset into 10 equal parts hereby accounting for all the data set, which is why we obtained 10 different accuracy score.

With K-fold cross validation(where K=10)

```
from sklearn.model_selection import cross_val_score
svc=SVC(kernel='linear',C=0.1)
scores = cross_val_score(svc, X, y, cv=10, scoring='accuracy')
print(scores)
```

```
[0.80576631 0.80854831 0.80374305 0.80273141 0.79564997 0.79969651
 0.8113303  0.80146687 0.80576631 0.80925879]
```

Figure 10: With K-fold cross validation(where K=10

Putting into consideration all values of C and the accuracy score with linear kernel. The C parameter presents the SVM optimization on the need to avoid misclassifying each training example.

A smaller-margin hyperplane is chosen for a large values of C on the premise that hyperplane gives a better result in classifying the training points correctly. On the other hands, a larger-marging hyperplane is used for a very small value of C even if the hyperplane classified more points incorrectly.

Hence, overfitting of the model can arise for a very large values C and underfitting of the model for a very small value of C thus chosen a value of C that generalised well for data not yet unseen.

4.4 Random Forest

This model was built using 10 as number of estimators, criterion as entropy and random state set to 0.To check accuracy of the random forest model,confusion matrix is used and obtained an accuracy of 82%

```
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)
#predicting the Test set results
y_pred = classifier.predict(X_test)
```

```
from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, y_pred)
```

```
print(cm)

accuracy_score(y_test, y_pred)
```

```
[[7860  574]
 [1535 1893]]

0.8222053616590794
```

Figure 11: Random Forest Accuracy

5 Evaluation

Specifically, this section examines and recommends which model is better at predicting customer behavior in the short- and long-term.

Model Name	Accuracy
Logistic Regression	80%
Support vector machine	81%
K-Nearest Neighbor	81%
Random Forest	82%

The above table shows the evaluation metrics for the models in the experiment. The Logistic Regression model has an Accuracy of 80% , while the K Nearest Neighbours, Support Vector Machine, Random Forest models have accuracy of 81%, 81%, 82% respectively,

The confusion matrix for the K Nearest Neighbours model is shown in table 2. The model accurately predicts 7747 instances correctly in the positive class, and 1941 instances correctly in the negative class.

Table 2 K Nearest Neighbours Confusion Matrix

	Actual	Predicted
Actual	7747	632
Predicted	1542	1941

Figure 12: Logistic Regression Accuracy

The Random Forest confusion matrix is shown in table 3 above. The model accurately predicts 7741 instances in the positive class and 2029 instances in the negative class.

Table 3 Random Forest Confusion Matrix

	Actual	Predicted
Actual	7741	638
Predicted	1454	2029

Figure 13: Logistic Regression Accuracy

5.1 Summary

In conclusion, the preprocessing stage is important to attain a high accuracy score. Whilst the accuracy of the models are high, this can be attributed to overfitting of these models and the class imbalance in the target variable. Improvement to these models can be

made by performing an oversampling of the minority classes or downsampling of the main class. Finally, hyperparameter optimization can also be explored to improve model performance. In summary, Random Forest gave the highest accuracy.

6 Conclusion and Future Work

Over the last decade, E-Commerce (electronic commerce) has gradually become the conventional way for individuals in society to purchase goods and services. This is notably due to the growth and widespread use of the Internet. Individuals are driven to make these purchases online because of countless benefits; for example the absence of time and space constraints, accessibility etc. However, the information available on the internet can become overwhelming, this is known as "information overloading" and this makes decision making difficult. To tackle this issue numerous researchers and companies utilise different tools to understand their customers and anticipate their needs. Furthermore, companies become conscious of their customers' purchasing behaviour and as a result implement appropriate brand and marketing strategies.

This study examines how recommender systems (RS) can be applied to gain an understanding of customer behaviour and their purchase intent in the hopes of reducing customer churn and increasing sales. Four machine learning models: Logistic Regression, Support Vector Machines, K Nearest Neighbours, Random Forest were built using customer behaviour data from Retail Rocket Recommender System dataset. The results from research suggest that the Random Forest model is best suited for classifying customer purchase intention with an Accuracy of 82%.

Although this research shows that Machine Learning and Recommender Systems can be used to understand customer behaviour and purchase intention, the research can be further expanded on by exploring the effect of performance on the type of Recommender Systems i.e Content-Based, Collaborative Filtering and Hybrid and also by implementing deep learning techniques in other to see if there will be a better accuracy.

References

Abdul Hussien, F. T., Rahma, A. M. S. and Abdulwahab, H. B. (2021). An e-commerce recommendation system based on dynamic analysis of customer behavior, *Sustainability* **13**(19): 10786.

Abedin, M. Z., Chi, G., Colombage, S. and Moula, F.-E. (2018). Credit default prediction using a support vector machine and a probabilistic neural network, *Journal of Credit Risk, Forthcoming* .

Abid, L., Masmoudi, A. and Zouari-Ghorbel, S. (2018). The consumer loan's payment default predictive model: an application of the logistic regression and the discriminant analysis in a tunisian commercial bank, *Journal of the Knowledge Economy* **9**(3): 948–962.

Adomavicius, G. and Tuzhilin, A. (2011). Context-aware recommender systems, *Recommender systems handbook*, Springer, pp. 217–253.

Aissa, S. A. H., Thabit, T. and Hadj, H. (2018). The impact of customer relationship management on customer behavior: Case study of ooredoo for telecommunications, *Revue Des Sciences Commerciales* **17**(1).

Chen, L., Yang, F. and Yang, H. (2017). Image-based product recommendation system with convolutional neural networks.

De Caigny, A., Coussement, K. and De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, *European Journal of Operational Research* **269**(2): 760–772.

Gordini, N. and Veglio, V. (2017). Customers churn prediction and marketing retention strategies. an application of support vector machines based on the auc parameter-selection technique in b2b e-commerce industry, *Industrial Marketing Management* **62**: 100–107.

Jamunadevi, C., Deepa, S., Kalaiselvi, K., Suguna, R. and Dharshini, A. (2021). An empirical research on consumer online buying behaviour during the covid-19 pandemic, *IOP Conference Series: Materials Science and Engineering*, Vol. 1055, IOP Publishing, p. 012114.

Jannach, D., Zanker, M., Ge, M. and Gröning, M. (2012). Recommender systems in computer science and information systems–a landscape of research, *International conference on electronic commerce and web technologies*, Springer, pp. 76–87.

Javed Awan, M., Mohd Rahim, M. S., Nobanee, H., Yasin, A. and Khalaf, O. I. (2021). A big data approach to black friday sales, *MJ Awan, M. Shafry, H. Nobanee, A. Yasin, OI Khalaf et al.," A big data approach to black friday sales," Intelligent Automation & Soft Computing* **27**(3): 785–797.

Juan, Y.-K., Hsu, Y.-H. and Xie, X. (2017). Identifying customer behavioral factors and price premiums of green building purchasing, *Industrial Marketing Management* **64**: 36–43.

Khodabandehlou, S. and Rahman, M. Z. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior, *Journal of Systems and Information Technology* .

Maheswari, K. and Priya, P. P. A. (2017). Predicting customer behavior in online shopping using svm classifier, *2017 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)*, IEEE, pp. 1–5.

Mansouri, A., Nazari, A. and Ramazani, M. (2016). A comparison of artificial neural network model and logistics regression in prediction of companies' bankruptcy (a case study of tehran stock exchange), *International Journal of Advanced Computer Research* **6**(24).

Petruseva, S., Sherrod, P., Pancovska, V. Z. and Petrovski, A. (2016). Predicting bidding price in construction using support vector machine, *Tem Journal* **5**(2): 143.

Silva, E. S., Hassani, H., Madsen, D. Ø. and Gee, L. (2019). Googling fashion: forecasting fashion consumer behaviour using google trends, *Social Sciences* **8**(4): 111.

Tuinhof, H., Pirker, C. and Haltmeier, M. (2018). Image-based fashion product recommendation with deep learning, *International conference on machine learning, optimization, and data science*, Springer, pp. 472–481.

Wakil, K., Alyari, F., Ghasvari, M., Lesani, Z. and Rajabion, L. (2019). A new model for assessing the role of customer behavior history, product classification, and prices on the success of the recommender systems in e-commerce, *Kybernetes* .

Wang, S., Wang, Y., Tang, J., Shu, K., Ranganath, S. and Liu, H. (2017). What your images reveal: Exploiting visual contents for point-of-interest recommendation, *Proceedings of the 26th international conference on world wide web*, pp. 391–400.

Wang, Y.-F., Chiang, D.-A., Hsu, M.-H., Lin, C.-J. and Lin, I.-L. (2009). A recommender system to avoid customer churn: A case study, *Expert Systems with Applications* **36**(4): 8071–8075.

Yu, W., Zhang, H., He, X., Chen, X., Xiong, L. and Qin, Z. (2018). Aesthetic-based clothing recommendation, *Proceedings of the 2018 world wide web conference*, pp. 649–658.

Zhou, H. and Hirasawa, K. (2019). Evolving temporal association rules in recommender system, *Neural Computing and Applications* **31**(7): 2605–2619.

Zuo, Y., Yada, K. and Ali, A. S. (2016). Prediction of consumer purchasing in a grocery store using machine learning techniques, *2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, IEEE, pp. 18–25.