PREDICTING THE NEXT EVENT TIME IN GOOGLE ANALYTICS USING MACHINE LEARNING

Sena DERELİ

OCAK 2022

PREDICTING THE NEXT EVENT TIME IN GOOGLE ANALYTICS USING MACHINE LEARNING

A THESIS SUBMITTED TO THE

GRADUATE SCHOOL

OF

BAHÇEŞEHİR UNIVERSITY

BY

SENA DERELI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF MASTER OF ARTIFICICAL INTELLIGENCE

IN THE DEPARTMENT OF ENGINEERING

İSTANBUL, 2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as requiredby these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname :

Signature :

ÖZET

Makine Öğrenimini Kullanarak Google Analytics Üzerinde Kullanıcının Sonraki Etkinlik Zamanını Tahmin Etmek

Dereli, Sena

Yapay Zeka Yüksek Lisans Programı

Tez Danışmanı: Doç. Dr. Ethem Canakoglu

Ocak 2022, 52 sayfa

Web ve mobil uygulamalara ait birçok veri toplanmakta ve işlenmektedir. Birçok kullanıcının bu platformlarda bazı eylemleri vardır ve bu eylemler kuruluşlar için davranışların analiz edilmesini ve tahmin edilmesi açısından önemli bir role sahiptir. Kuruluşlar müşterilere web ve ya uygulamalar aracılığıyla ulaşmaya çalışmaktadır. Gelişen veri işleme ve analitiği işlemleriyle kullanıcılara ait bir çok veri saklanmaya başlandı ve böylece şirketler kullanıcıya dair bir çok fikre sahip oldular. Kullanıcıların demografik bilgileri, websitesi/uygulama üzerindeki tıklama, ziyaret, siteden çıkma, sitede geçirdikleri vakit, hangi ürün/hizmetlere baktıkları, hangi ürünleri sepete ekledikleri, favori ürün/hizmetleri gibi bir çok bilgiye sahip olan şirketler bunları saklamakta ve işlemekte iyi olmalıdırlar. Örneğin, Google Analytics uygulaması bu veriler için iyi bir altyapı hazırlamaktadır. Belirli entegrasyonlar yapıldıktan sonra az önce bahsettiğimiz verileri toplamaya başlar ve uygun bir raporlama sistemi ile şirketlere sunar. Bu verilerle kullanıcının bir sonraki ziyaret tarihi, bir sonraki alacağı ürün, bunu ne zaman alabileceği, websitesine gelen kişiye uygun ürün önerisi sunma, kişinin uygulamayı bırakabileceğinin tahmini gibi bir çok ileriye dönük tahmin çalışması yapılabilir ve uygulamalar bunlar için farklı aksiyonlar alabilir. Bu çalışmada, Google Analytics örnek veri seti kullanılarak kullanıcıların siteyi tekrar ziyaret ederlerse ne zaman olacağı, web ve mobil üzerindeki hareketlerine göre tahmin edilmiştir. Bu çalışmanın diğer çalışmalara göre farklılığı, tahminlemeye bir segmentasyon katmanının eklenmesi ve tahmin özelliklerine kümelerin eklenmiş

olmasıdır. Segmentler bir özellik olarak tahmin kümesine eklendiğinde model perfomansının arttığı gözlemlenmiştir. Farklı makine öğrenimi modelleri kullanılmış, en yüksek performans ölçütlerine sahip XGBoost modeli kullanılmış ve bir tahmin çalışması yapılmıştır.

Anahtar kelime: google analytics, işlem zaman tahmini, makine öğrenmesi

ABSTRACT

PREDICTING THE NEXT EVENT TIME IN GOOGLE ANALYTICS USING MACHINE LEARNING

Dereli, Sena

Artificial Intelligence Department

Thesis Supervisor: Associate Prof. Ethem Canakoglu

January 2022, 52 pages

A lot of data belonging to web and mobile applications are collected and processed. Many users have some actions on these platforms, and these actions play an important role for organizations in analyzing and predicting behavior. Companies that have a lot of information such as demographic information of users, clicks on the website / application, visits, exit from the site, the time they spend on the site, what products / services they look at, which products users add to the cart, users' favorite products / services and they should be good at storing and processing them. After certain integrations are made, it starts to collect the data we just mentioned and presents it to companies with an appropriate reporting system. With this data, many forward-looking predictions such as the next visit date of the user, the next product to buy, the prediction of churn behavior can be made, and the organizations can take different actions for them. In this study, using the Google Analytics sample data set, when users visit the site again, it was estimated based on their movements on the web and mobile. The difference of this study from other studies is that a segmentation layer is added to the prediction and clusters are added to the prediction features. As the segments were added to the prediction set as a feature, the performance increase is observed. Different machine learning models were used, the XGBoost model with the highest performance criteria was used and a prediction study was made.

Keywords: google analytics, event time prediction, machine learning

To my family who always supported me and raised me with great devotion, to my husband who made me live the best days of my life..

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my supervisor Associate Prof. ETHEM CANAKOGLU for his guidance, advice, criticism, encouragements and insight throughout the research.

TABLE OF CONTENTS

ETHICAL CONDUCT ii
ÖZET iii
ABSTRACT v
DEDICATION vi
ACKNOWLEDGMENTS vii
TABLE OF CONTENTS viii
LIST OF TABLES xi
LIST OF FIGURES xii
LIST OF ABBREVIATIONS xiii
Chapter 1: Introduction 1
1.1 Theoretical Framework 2
1.2 Statement of the Problem 3
1.3 Purpose of the Study 3
1.4 Significance of the Study 4
Chapter 2: Literature Review 5
Chapter 3: Web Analytics and User & Customer Behavior 6
3.1 Understanding User & Customer Behavior 6
3.1.1 Collecting User & Customer Data 7
3.1.2 Processing User & Customer Data 9
3.1.2.1 Databases – Traditional & Cloud Systems 9
3.1.2.2 Google Analytics – Google Bigquery & Data Studio 10
3.1.2.2 Open-Source Programming Languages 10
3.1.2.3 Software 10
3.2 Data in Marketing 10
3.2.1 Structured Data 11
3.2.1.1 Transactions 11
3.2.1.2 Customer Data 11
3.2.1.3 Other Structured Data Examples 12
3.2.2 Unstructured Data 12
3.2.2.1 Surveys 12

3.2.2.2 Images, Videos, Audios 12
3.2.2.3 Social Media 12
3.2.2.4 Website 13
Chapter 4: Google Analytics 14
4.1 Previous Studies with Google Analytics 14
4.2 Google Analytics Dimensions and Metrics 15
4.3 Google Analytics Data Collection & Data Processing 22
4.4 Google Analytics Reporting 23
4.4.1 Real-time 23
4.4.2 Audience 24
4.4.3 Acquisition 24
4.4.4 Behavior 25
4.4.5 Conversions 25
Chapter 5: User Analytics, Algorithms and Event Prediction 26
5.1 User Analytics 26
5.1.1 User Features 27
5.1.2 Segmentation 27
5.1.2.1 k-Means 27
5.1.3 Classification 29
5.1.3.1 Logistic Regression 30
5.1.3.2 Support Vector Machines (SVM) 31
5.1.3.3 Random Forest 32
5.1.3.4 Light Gradient Boosting Machine (LightGBM) 33
5.1.3.5 XGBoost (eXtreme Gradient Boosting) 34
5.2 Google Analytics Dataset 35
5.2.1 Dataset Definition 35
5.2.2 Explanatory Data Analysis 36
5.2.3 Methods 40
5.3 Event Prediction & Algorithms and Application 41
5.3.1 Segmentation Data Preparation & Application 41
5.3.2 Prediction Data Preparation & Application 43
5.4 Model Results 45

5.4.1 Segmentation Results 45
5.4.2 Prediction Results 46
Chapter 6: Conclusion 48
REFERENCES 49

LIST OF TABLES

TABLES

Table 1 Web analytics data 9
Table 2 Dimensions of Google Analytics 15
Table 3 Metrics of Google Analytics 17
Table 4 Raw dataset of the study 36
Table 5 Top 10 browser types and counts 37
Table 6 Segmentation result 45
Table 7 Cross validation results 46
Table 8 Classification results 47
Table 9 Confusion matrix 47
Table 10 Confusion matrix without segmentation 47

LIST OF FIGURES

FIGURES

Figure 1 Estimated amount of data on the internet in one minute in 2020..... 2

Figure 2 Data collection process of Google Analytics (Cutroni, J. Google Analytics: Understanding Visitor Behavior, 2010) 22

Figure 3 Google Analytics reports..... 23

Figure 4 kMeans steps (Arai, K., Barakbah, A., 2007) 28

Figure 5 Hyperplanes and support vectors 31

Figure 6 Random forest process 32

Figure 7 Leaf based process of LightGBM 33

Figure 8 Leaf based process of XGBoost..... 34

Figure 9 Statistical data information 36

Figure 10 Most used browser, device, source, channel types..... 38

Figure 11 Monthly visit count 38

Figure 12 Monthly order revenue..... 39

Figure 13 Outlier boxplot for source and browser 40

Figure 14 Elbow method 42

Figure 15 Correlations between variables 44

Figure 16 Feature importances 46

LIST OF ABBREVIATIONS

CRM	Customer Relationship Management
ERP	Enterprise Resource Planning
SQL	Structured Query Language
RDBMS	Relational Database Management System
AWS	Amazon Web Services
GA	Google Analytics
GATC	Google Analytics Tracking Code
OS	Operating Service
URL	Uniform Resource Locator
LightGBM	Light Gradient Boosting Machine
ANOVA	Analysis of variance
GBDT	Gradient Boosting Decision Tree

Chapter 1

Introduction

With the developing technologies, many institutions try to reach consumers over the internet. Many users browse the internet while watching television and can log into the application with a notification on their phone while doing a job. As a result of that, many transactions are made online over the internet. In online applications, there are e-commerce sites that users can fulfill their needs, some service applications provide some benefits to make users' life easier, game applications make users having more fun. These applications are becoming more popular in recent years and they are storing a lot of data through users' behaviors. Using these data make the organizations reach to their users in an easy and efficient way. Organizations are forced to use multiple channels to explain themselves to consumers with an enhancing social media and Internet environment. They try to produce advertisements, campaigns, products, and content that can attract the attention of users and deliver them to users through different channels. This is a topic of interest both marketing and data science since it is tried to be handled in the smartest way today. This topic requires the collection of a lot of data and observing the actions of users involves majority of analytics work. As seen in Figure 1, huge data is collected every single minute in the internet environment. For example, Netflix, a TV series and movie application, displays approximately 404,444 videos per minute, on Instagram, a photo and video sharing application, users share 347,222 stories a minute, and Zoom, a video chat application, organizes meetings for 208,333 participants every minute, on Amazon.com, an e-commerce site, 6,659 packages are sent in every minute.

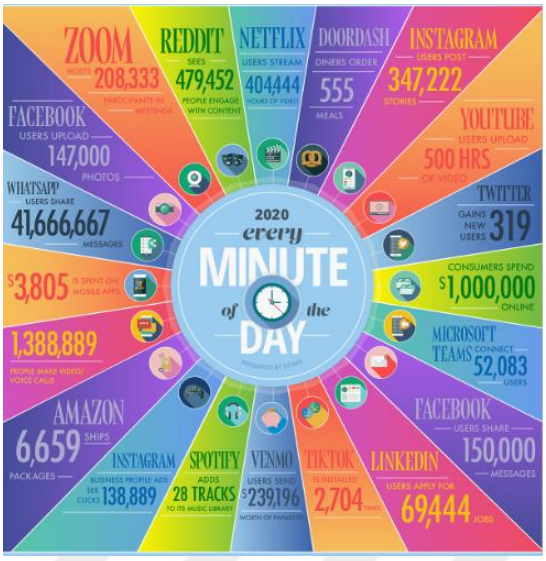

Figure 1. Estimated amount of data on the Internet in one minute in 2020 (Venture Capitalist, 2020)

Collecting and processing this data is one of the important issues of our day, this issue called "big data" is on the agenda of many people. Most of the users' behaviors are now tracking by websites & applications. User data has many contents such as website visits, transactions made, monetary values of transactions and so on.

1.1 Theoretical Framework

In this study, in first part, an information is given about user data, how it can be collected, some specific topics and problems using user data, and application of these problems' solutions in different departments and sectors. For instance, in order to observe user data, Google has launched an application which is called "Google Analytics" that web & mobile application owners can track the users' behavior in a smart way. The data has been collected, processed, configured and many reports can be prepared by using this data. Users' behavior can be observed through their visits, purchases, bounces concerning the product, channel, source, and many other relations. In the study, there is a literature review part mentioning about event prediction, Google Analytics and customer analytics topics, moreover, in the other sections there are also literature examples about related topic. In the third part of the study, Web Analytics and User & Customer Behavior topics are evaluated in details.

In the fourth part of this study, it is mentioned about Google Analytics, its concepts and detailed information. To analyze the data in user based, there is a need to use Google BigQuery tool in order to query the specific information. In this study, Google Analytics sample dataset has been used and the data subjected to manipulation process by using Python and BigQuery connections. Google Analytics has nested data structure, so there is need to unnest the features as hits, which show the interactions of users and totals which shows the aggregated form of bounces, pageviews etc. In the fifth part of the study, for the prediction, some features have been selected and Explanatory Data Analysis process has been applied to understand the data. Also, clustering results have been added as a feature. After these transactions, the data is divided into train, test and validation parts considering the timeline then some prediction models have launched, and the best performed model, XGBoost has selected for the main prediction algorithm. In the end of the study, there is a conclusion part mentioning about the results of the project.

1.2 Statement of the Problem

Prediction is a popular topic, and, in many sectors, it can be applied to solve different problems. It is mentioned about algorithms that can be used in prediction, also the subject of this study may consist of time series problems and algorithms. The results and the comments will be involved in the results part of this study.

1.3 Purpose of the Study

When it comes to this studies' applications in real world, predicting next event time of a user is crucial for websites and applications that they can prepare for this users' action in advance. For instance, there is an e-commerce website being mentioned about, if they predict if a user visits in a specific period, they can assign coupons or offer some deals to user. Or may be in social media platforms, if it is known that a user opens the application in a specific time period, they may arrange the content related with to this users' interests and hobbies. This framework can be applied to many different tools and applications.

1.4 Significance of the Study

This study gets a new perspective to predicting customer behavior as it is fed from customer segmentation. Segmentation layer is used as an input of prediction.

Chapter 2

Literature Review

There are many studies that clarifies clickstream data and purchase prediction that inspire to research about online data and prediction subjects. One of these studies is about creating a model for prediction customer satisfaction in terms of e-commerce data, and the findings are insightful that states that in the case of the study the features as arrival time, customer location determine the satisfaction level of customers (Maria, K. 2021). In addition, in the study of Hendriksen, M., Kuiper, E., Nauts, P., Schelter, S., Rijke, M. (2020) analyzed the difference of anonymous and identified customers' behavior and their intention of the purchase and found that the day of week and page features are important for prediction of anonymous users and for the identified users behavior prediction, number of orders, days since last order features is important. Moreover, Mohammadifard, N. (2013) had an approach by using Hidden Markov Models and Logistic Regression in order to model e-commerce users towards advertisements. Also, Kulik, R. (2020) took basket sessions to find intention of purchase by creating a hybrid approach from session-based features and click sequences.

Also, there can be a need of using analyzes and prediction models actualize in businesses such as combining them with sending notifications according to these models and analyzes. Jeng, J., Drissi, Y. (2000) gets a framework that creating a predictive notification system can be implemented into CRM systems.

In the next sections, other literature research will be given based on the topics they are related to.

Chapter 3

Web Analytics and User & Customer Behavior

Web analytics is a process starting from web data collection, reporting, and analyzing the users' behaviors while users spending time in web channel. Today, many users browse websites to have fun, to meet their needs, to learn, to search that they are wondering. These activities create data and events while users having an interaction with the websites. Understanding these data and events is provided from reports and analyzing them creates web analytics environment. Jarvinen and Karjaluoto (2015) has research for application Web Analytics and their effects to Digital Marketing, and they've found that 2 of the 3 companies that participated the study are not satisfied with their usage of Web Analytics. Web Analytics has a lot of details and technical issues that needs to be understood and implemented to the action of the organizations.

3.1 Understanding User & Customer Behavior

Jansen (2009) explained behavior term and behavior in web analytics in a detailed way. Behavior is referred as a thing that can be detected and they are reactive responses. User behavior is the response in the Web and collecting and analyzing it is an important subject. In the study, user behavior divided into parts as: view results, selection, view document, execute, navigation, browser, relevance action, view/implement assistance.

- View results: Users' viewing or scrolling behavior

- Selection: Users' selection behavior from the results

- View document: Users' viewing a document behavior in the results.

- Execute: Users' action behavior

- Navigation: Users' coming back behavior to Home or Back.

- Browser: Users' opening browser behavior

- Relevance action: Users' save, print actions

- View/Implement assistance: User's viewing assistance behavior

There are also some metrics to analyze the behavior data and getting some insights from it. Burby, Brown and Web Analytics Association Standards Committee (2007) determined the basic metrics such as page, page views, visits/sessions, unique visitors. Knowing these terms make the web analytics environment clearer. Phippen, Sheppard & Furnell (2004) explained about traditional web metrics such as user definition, visit, pageview definitions and they offer some brand-new approaches to web metrics, such as implementing segmentation, customer lifecycle analysis also these new approaches can give a shape to website design, campaign scenarios and other related actions. Chaffey and Patron (2012) developed a Key Performance Indicator (KPI) framework as RACE (Reach, Act, Convert, Engage) from the perspective of tracking metrics, performance drivers, customer centric KPIs, business value KPIs.

3.1.1 Collecting user & customer data.

User and customer data in web can be collected in many ways. Kaushik (2007) explained these elements as:

- From web log files, web beacons, JavaScript tags, packet sniffing
- Qualitative data such as surveys, CRM or ERP system data, etc.
- Competitive data from other companies

These methods can be used separately, or they can be merged to get meaningful insights. In order to describe these elements, starting from web data collection part will be crucial. Web logs provides data collection as the main source. Web beacons are used for tracking the visitors to a website. JavaScript tags are popular nowadays, it is a data collection method and preferred for reducing the dependence of departments to IT. Packet sniffing is also a data collection method but not popular as the other data collection methods. Other types of data are the data which are gathered from CRM of ERP systems. Briefly, CRM (Customer Relationship Management) is a division that manage the customer data, perform works to gain new customers and retain existing customers. Organizations gather and process the customer data via CRM systems, if they have not an advanced CRM system, they might collect the data in ERP systems. In addition, surveys

are also an insight channel that organizations launch to understand the customers' opinion and needs with asking some questions. There are also third-party data sources that are coming from tools involving sector players data, also organizations give importance to these data to see their positions in the sector.

With using these techniques, Table 1 indicates the main example of elements that generates Web Analytics data. First, the features of the users can be mentioned.

- Demographics: Users' age, gender, location information

- Preferences: Users' channel & browser preferences while reaching Website

- Visits: Users' visit counts, unique visitor metrics as an important information

- New/ Existing Users: Users' information about their first visit or they are also a member of the Website

- Users' Monetary Values: Users purchased or viewed products monetary values

- Segments: User's segmentation according to their behaviors, monetary values, etc.

Behavior part is the actions that users make on the Web.

- Bounces: Users' exiting behavior of the website

- Page Views Counts/ Durations: Users' page view behavior and count and durations of it

- Favorite Categories/ Products: Users' frequently visiting/ purchasing products

- Actions to Communications: There are some communication tools as email, push notifications and the users' reaction to them.

Traffic part is the triggers that makes the users come to website.

- Sources: Users' landing behavior to website

- Browsers: User's browser preferences that forwarding to website

- Campaigns: Campaign and promotions that make the users come to website

Table 1

Web Analytics Data

Visitors/ Users	**Behavior**	**Traffic**
Demographics	Bounces	Sources
Preferences - Channel /Browser	Page Views Counts / Durations	Browsers
Visit Count, Unique Visitors	Landing / Existing Pages	Campaigns
New/ Existing Users	Favorite Categories / Products	
Monetary Values of Users	Actions to Communications	
Segments		

3.1.2 Processing user & customer data.

Gathering and collecting user and customer data is of the essence for organizations. After collecting them, there is a need to process and getting insights from these data. For those purposes, databases and data processing tools helps to process the data.

3.1.2.1 Databases – traditional & cloud systems.

Databases are a source of storing data. It consists of three main items as tables, columns, and rows. Tables demonstrates the information collected. Columns shows the elements of information in the tables and rows are the records that are collecting as real data. (Petersen J., 2002) Some of the examples of database management systems are:

- Microsoft SQL Server
- Oracle RDBMS
- MySQL

Using these systems can provide analysis by writing queries with database connections. Also, there are systems such as Hadoop that can scale big data in a faster way. With the growing Web ecosystem, the need of those systems will increase. Cloud

ecosystem is also includes actively used systems. Google Bigquery, Amazon AWS are the most popular ecosystems.

3.1.2.2 Google Analytics – Google Bigquery & Data Studio.

It is a digital marketing tool that we can learn about the people who come to the Google Analytics website and their behavior. The data here can be processed with SQL-NoSQL queries via Bigquery, the Big Data tool. It will be explained in the next part in detail.

3.1.2.3 Open-source programming languages.

Open-source programming languages are popular nowadays, since they are free and able to process data faster, at the same time they have a user-friendly language. R and Python are most common used tools.

R is a statistical software, queries or data conversions can be performed on the data. At the same time, machine learning models can be made in a short time. Being open source allows you to download it to your computer and use it immediately.

Python is a programming language that has become increasingly popular lately. Like R, Python is used as open source. While processing big data quickly, it also enables the application of concepts such as machine learning and deep learning.

3.1.2.4 Software.

Software as SAS, SPSS and other software are also used for processing data. These two programs are used frequently by companies. It provides connection with databases and is used while processing data and provides an easy use in terms of modeling. These programs are purchased under license, and then support can be obtained from consultancy firms for system integration.

3.2 Data in Marketing

Marketing is a field that building relationships with customers in a profitable way. This can be achieved by creative communication offerings to the customers. In a traditional perspective, marketing activities may appear as advertisements heard on

television and radio, on outdoor banners or on magazines. Online social networks or websites are the environments of marketing activities of late years (Armstrong, Adam, Denize, Kotler, 2015).

In order to keep up and cover all activities Kumar, Chattaraman, Neghina, Skieria, Aksoy, Buoye & Henseler (2013) divide the data types into three: traditional data, neurophysiological and digital data. Neurophysiological data covers the data that comes directly from human's brain with using elecctromyography or other scientific methods. Traditional data covers focus groups, surveys, transactions, product & sales data. Also, digital data includes surveys, transactions and product reviews but also covers the clickstream, social media, search data. In this study, mostly digital data is considered. These data are collected and by merging them, some insights maintained by considering these data.

3.2.1 Structured data

Structured data is the data stored in relational databases includes with stable rows and columns. In marketing, there are some information that needed mostly such as transactions, customers data, and there are some other data as products can be applicable in marketing concepts.

3.2.1.1 Transactions.

Transactions or orders data is important to track the customers' behaviors. There are some information in the purchase of a product or service. There can be fields as users' id, number of the order, product/ service id that purchased, quantity of the purchased product or service, monetary values of the products or services, the date of the purchase event.

3.2.1.2 Customer data.

Customer data involves the customers' id, customers' mail, address, phone, email, age, gender, customers' communication permissons. These information are collected according to General Data Protection Regulation (GDPR) rules.

3.2.1.3 Other dtructured data examples.

Product data is a structured data that shows the products information like size of the product, raw material of a product, gender of the product, etc. It can be used to know the customers' preferences better.

3.2.2 Unstructured data.

Unstructrued data is data type that cannot be collected by traditional databases mostly, it includes data such as text, images, videos, social media posts, emails, survey responses and many other data. In marketing there are some data that can be processed and for getting insights from them.

3.2.2.1 Surveys.

Open- ended survey answers are a great example of unstructured data. By looking at marketing perspective, asking to customers whether they like the product, adverstisement, or anything that they are wondering is a great way to understand customer needs. By asking open-ended text questions, they are allowing users make their comments. These data are collected and for instance with using text mining, a lot of information can be gathered.

3.2.2.2 Images, videos, audios.

Data such as images, videos and audios can be gathered in different ways and can be suitable for table format. For instance, it can be used for many areas as text to speech, or sending notifications with image or video.

3.2.2.3 Social media.

Users' comments for an organization, their service, or product is important for the organizations. They can be stored and processed via machine learning or Artificial Intelligence techniques to get meaningful insights. Organizations can easily take actions from these social media events of users.

3.2.2.4 Website.

By following user behavior on organizations' website, that is a way to collect information about users. There are some heatmap visuals that organizations can get ideas of the features that takes customers' attention.

Chapter 4

Google Analytics

Google Analytics is a tool that let the organizations track the web or application behavior of users and provides statistics, insights, reporting for web analytics. It is launched in November 2005. According to the research of Hotjar, Google Analytics is the most favorite tool for web analytics in 2020 (https://www.hotjar.com/web-analytics/tools). The application is free, but in the high volumes of data, Google can apply sampling for not showing all the data.

Nevertheless, they also offer Google 360 that provides data without sampling with real numbers, in addition to this, they give an access to advanced tools in Google Analytics.

There are also other web analytics tools such as Adobe Analytics, Mixpanel, Matomo, Yandex Metrica, but in this study sample Google Analytics data is used and processed.

4.1 Previous Studies with Google Analytics

There are many studies on Google Analytics which have many applications in different areas. One of them is the study of Plaza, B. (2011) that have been made a time series application for tourism field with using Google Analytics data.

Moreover, Hasan, L., Morris, A., & Probets, S. (2009) made a general analysis and found some metrics that can be easily followed by e-commerce sites and make these processes better by looking these metrics.

Online education is getting popular nowadays. Filva, D. A., Guerrero, M. J. C., & Forment, M. A. (2014) had a study that track students behavior on a learning enviroment and prepared an approach that can be applied all other online learning platforms.

Pakkala, H., Presser, K., & Christensen, T. (2012) cited about Google Analytics usage in food composition websites, also Google Analytics can be a great tool for comparison of websites by looking their metrics and improving website design.

Tupikovskaja-Omovie, Z. & Tyler, D. (2020) compared Google Analytics and eye-tracking tool performances whether gathering consumers behavior properly. Google

Analytics can collect half of the eye-tracking tool results, so there will be a need of getting more data on Google Analytics.

Vecchione, A., Brown, D. , Allen, E. & Baschnagel A. (2016) made a study about decreasing bounce rate, improving Academic Library Website using Google Analytics.

4.2 Google Analtyics Dimensions and Metrics

As Google's own definiton of dimensions are attributions of the data (https://support.google.com/analytics/answer/1033861?hl=en). Dimensions are listed in Table 2.

Table 2

Dimensions Of Google Analytics

Dimension	**Explanation**	**Example**
User	Features of users	User Type
Session	Features of sessions	Session Duration, Session Count
Traffic Sources	Features of source of visit	Source/Medium, Campaign
Ecommerce	Features of order and products	Product, Product Category, Transaction ID
Audience	Features of audience	Age, Gender
Channel	Channel Preference	Default Channel Grouping
Time	Timing	Date, Minute, Hour, Day, Month, Year
Platform	Features of device	Browser, Device Category, Operating System
Adwords	Adwords related features	Google Ads Campaign ID, Google Ads Ad Group ID

Table 2. (cont.d)

Goal Conversions	Features of goals	Goal Completion Location
Geo Network	Location of sessions	Country, Region
System	Features of system	Java Support, Flash Version
Page Tracking	Features of pages	Page, Landing Page, Exit Page
Internal Search	Features of audiences' searches	Search Term, Site Search Status
App Tracking	Features of app	App Name, Screen Name
Event Tracking	Features of events	Event Label, Event Category
Social Interactions	Features of social media interactions	Social Type, Social Action
User Timings	Measuring periods of time	Timing Label, Timing Category
Exceptions	Features of crashes and exceptions	Exception Description
Content Experiments	Features of comparing tests	Experiment ID
Custom Variables or Columns	Features of additional segments	Custom Dimensions, Custom Variables
Double Click Campaign Manager (Campaign Manager)	Campaign features	CM Ad, CM Advertiser
Lifetime Value and Cohorts	Measuring montetary values of users in period basis	Acquisition Channel, Acquisition Source
DoubleClick Bid Manager (Display & Video360)	Efficiency of campaign features	DV360 Advertiser
DoubleClick Search (Search Ads 360)	Efficiency of campaign features	SA360 Campaign, SA360 Keyword

Metrics are listed in Table 3. Double Click, Ad Exchange, Adsense, Adwords, Site Speed , Goal Conversions, Exceptions metrics are not included.

Table 3

Metrics Of Google Analytics

Metric Type	Metric	Explanation
User	Users	Number of users
User	New Users	Number of users that shows the first session of them
User	% New Sessions	Sessions percentage of coming to the website at the first time
User	1 Day Active Users	1 day users for each day in a given date period
User	7 Day Active Users	7 day users for each day in a given date period
User	14 Day Active Users	14 day users for each day in a given date period
User	28 Day Active Users	28 day users for each day in a given date period
User	30 Day Active Users	30 day users for each day in a given date period
User	Number of Sessions per User	Total transactions divided by total users
Session	Sessions	Total number of sessions
Session	Bounces	Total single page sessions
Session	Bounce Rate	Percentage of single page sessions in whole sessions
Session	Session Duration	Users' session duration
Session	Avg. Session Duration	Average session duration of users
Session	Hits	Number of hits for the view
Session	Organic Searches	Number of organic searches in a session
Traffic Sources	Page Value	The average value of pages
Page Tracking	Entrances	The number of entrances to the property measured as the first pageview in a session, typically used with landingPagePath.
Page Tracking	Entrances / Pageviews	The percentage of pageviews of the entrance page
Page Tracking	Pageviews	Number of pageviews
Page Tracking	Pages / Session	The average number of pageview event in a session
Page Tracking	Unique Views	The number of unique content group views
Page Tracking	Unique Pageviews	Sessions that a page viewed unique

Table 3 (cont.d)

Page Tracking	Avg. Time on Page	The average time users spent in pages
Page Tracking	Exits	Number of exits from the website
Page Tracking	% Exit	The percentage of exits from the website that in all pageviews
Page Tracking	Results Pageviews	The number of times a search result page was viewed
Internal Search	Total Unique Searches	Unique keywords from internal searches in a session
Internal Search	Results Pageviews / Search	Average number of times people viewed a page as a result of a search
Internal Search	Sessions with Search	The total number of sessions that included an internal search
Internal Search	% Sessions with Search	The percentage of sessions with search.
Internal Search	Search Depth	Number of subsequent page views made after searching keywords
Internal Search	Avg. Search Depth	The average number of pages users viewed after a search
Internal Search	Search Refinements	The total number of times a transition occurs between internal keywords search in sessions
Internal Search	% Search Refinements	The percentage of the number of times a transition occurs between internal keywords search within a session.
Internal Search	Time after Search	The average time of users spent on the website after searching
Internal Search	Search Exits	The number of exits that happened after search results
Internal Search	% Search Exits	The percentage of searches that resulted in an immediate exit from the website
Internal Search	Site Search Goal XX Conversion Rate	The percentage of search sessions which resulted in a conversion to the goal
Internal Search	Site Search Goal Conversion Rate	The percentage of search sessions which resulted in a conversion to goals

Table 3 (cont.d)

Internal Search	Per Search Goal Value	The average goal value of a search
Internal Search	Screen Views	Number of screenviews
App Tracking	Unique Screen Views	Unique screen views
App Tracking	Screens / Session	Average number of screenviews per session
App Tracking	Time on Screen	The time spent viewing in the screen
App Tracking	Avg. Time on Screen	Average time of users spent on a screen
App Tracking	Total Events	The total number of events for the users
Event Tracking	Unique Dimension Combinations	Unique dimension-values for each dimension in a report
Event Tracking	Unique Events	Unique events count
Event Tracking	Event Value	Total value of events
Event Tracking	Avg. Value	The average value of an event
Event Tracking	Sessions with Event	Sessions count with events
Event Tracking	Events / Session with Event	The average number of events
Event Tracking	Transactions	The total number of transactions
Ecommerce	Ecommerce Conversion Rate	The average number of transactions in a session
Ecommerce	Revenue	Total revenue of the transaction
Ecommerce	Avg. Order Value	The average revenue of a transaction
Ecommerce	Per Session Value	Average transaction revenue value of a session
Ecommerce	Shipping	The total cost of shipping
Ecommerce	Tax	Total tax for the transaction
Ecommerce	Total Value	Total values for the property
Ecommerce	Quantity	Total number of items in a transaction
Ecommerce	Unique Purchases	The number of unique products purchased in a transaction
Ecommerce	Avg. Price	The average revenue per item
Ecommerce	Product Revenue	The total revenue from purchased products in a transaction
Ecommerce	Avg. QTY	The average quantity of this item in a transaction

Table 3 (cont.d)

Ecommerce	Local Shipping	Transaction shipping cost in local currency
Ecommerce	Local Tax	Transaction tax in local currency
Ecommerce	Local Product Revenue	Product revenue in local currency
Ecommerce	Buy-to-Detail Rate	Percentage of users purchased after viewing the products
Ecommerce	Cart-to-Detail Rate	Percentage of users add to basket after viewing the products
Ecommerce	Internal Promotion CTR	The rate at which users clicked through to view the promotion
Ecommerce	Internal Promotion Clicks	The number of clicks on an internal promotion
Ecommerce	Internal Promotion Views	The number of views of an internal promotion
Ecommerce	Local Product Refund Amount	Refund amount in local currency for a given product
Ecommerce	Local Refund Amount	Total refund amount in local currency for the transaction
Ecommerce	Product Adds To Cart	Count of times the product was added to the shopping cart
Ecommerce	Product Checkouts	Count of times the product was included in the check-out process
Ecommerce	Product Detail Views	Count of times users viewed the product-detail page
Ecommerce	Product List CTR	The rate at which users clicked through on the product in a product list
Ecommerce	Product List Clicks	Count of times users clicked the product in the product list
Ecommerce	Product List Views	Count of times the product appeared in a product list
Ecommerce	Product Refund Amount	Total refund amount associated with the product
Ecommerce	Product Refunds	Count of times a refund was issued for the product
Ecommerce	Product Removes From Cart	Count of times the product was removed from the shopping cart
Ecommerce	Product Revenue per Purchase	Average product revenue per purchase
Ecommerce	Quantity Added To Cart	Product count of added to the shopping cart

Table 3 (cont.d)

Ecommerce	Quantity Checked Out	Product count included in check out
Ecommerce	Quantity Refunded	Product count of refunded
Ecommerce	Quantity Removed From Cart	Product count of removed from a shopping cart
Ecommerce	Refund Amount	Currency amount refunded for a transaction
Ecommerce	Revenue per User	The total sale revenue of the transaction divided by users count
Ecommerce	Number of Sessions per User	The total number of sessions divided by users count
Ecommerce	Refunds	Refund count that have been issued
Ecommerce	Transactions per User	Transaction count divided by users count
Social Interactions	Social Actions	Social interactions count
Social Interactions	Unique Social Actions	Sessions count of during which the specified social actions
Social Interactions	Actions Per Social Session	Social interactions count per session.
User Timings	User Timing (ms)	Total number of milliseconds for user timing.
User Timings	User Timing Sample	The number of hits sent for a particular userTimingCategory, userTimingLabel, or userTimingVariable.
User Timings	Avg. User Timing (sec)	The average elapsed time.
User Timings	Exceptions	The number of exceptions sent to Google Analytics.
Custom Variables		A variable that assigned to users using the JavaScript Tracking API

Source: https://ga-dev-tools.appspot.com/dimensions-metrics-explorer/

Source2:

https://www.digishuffle.com/blogs/list-of-dimensions-metrics-google-analytics/

4.3 Google Analytics Data Collection & Data Processing

Google Analytics collects the data via page tags mechanism, that a small JavaScript code added into website pages and tracking starts under favor of these small piece of codes named Google Analytics Tracking Code (GATC). Data collection process of Google Analytics is shown in Figure 2. The process starts with a visitor visits the website. Sending a request a page forwarded to web servers, server responds back and show the page. The browser interacted with GATC, the code starts to process the data about the visitor and her behavior. After that, there are cookies set and collects information from the visitor's machine, the tracking code works and waits for sending these data to Google Analytics. The browser starts downloading a ga.js file from GA servers and the data as pageview, e-commerce data or other data is sent to Google. The pageview transmitted to GA servers via a gif file and GA servers receive the file, stores it in a temporary data storage. This pageview stored in logfile the process is finished. This data collection part is completed, then processing part started, in every 3 hours GA processed the data in the logfile, but before 24 hours it doesn't completed. In processing part, logfile splitted into each line as attributes of a pageview, also GA creates data parts as a field, that fields generate dimensions. Then, GA uses these dimensions for reporting. (Cutroni, J. 2010, Google Analytics: Understanding Visitor Behavior)

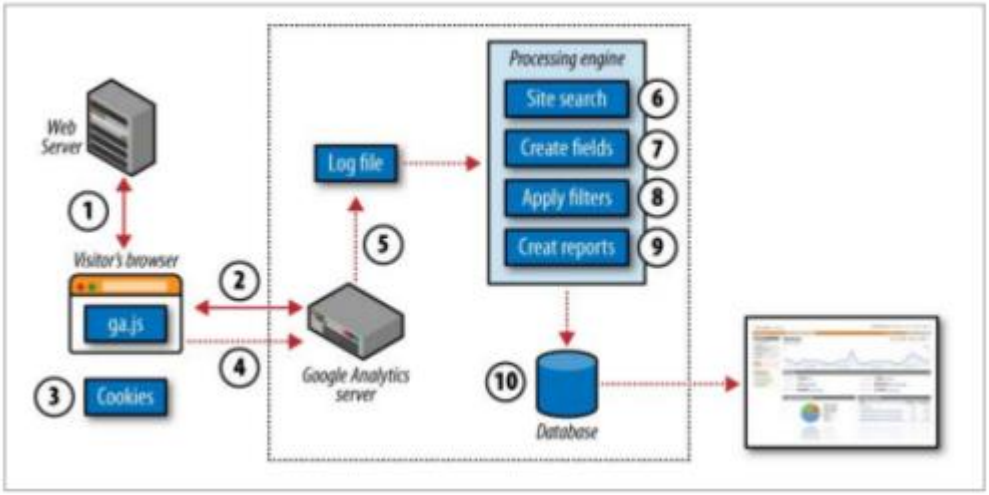

Figure 2. Data collection process of Google Analytics (Cutroni, J. Google Analytics: Understanding Visitor Behavior, 2010)

4.4 Google Analytics Reporting

In Google Analytics application there is reporting part consists of 5 parts as real-time, audience, acquisition, behavior and conversions as shown in Figure 3. It provides time-based analysis to compare metrics in custom dates that users select in Google Analytics panel.

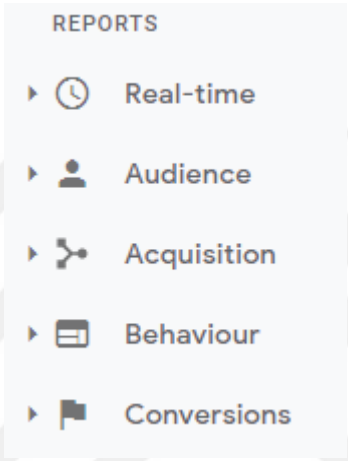

Figure 3. Google Analytics reports

4.4.1 Real-time.

Real-time reporting part shows what is happening in the website at this moment as giving insights real-time.

It has tabs as Overview, Locations, Traffic Sources, Content, Events, Conversions.

Overview part gives insights about active users, pageviews, top referrals, top active pages, top social traffic at this moment.

Locations part gives information about active users by country base at this moment.

Traffic sources gives insgihts about users coming from source and medium based at this moment.

Content part gives information about users viewing active pages and page titles at this moment.

Events part gives insights about defined events and users performing these events now.

Conversions part gives information about conversions as they happen now.

4.4.2 Audience.

The Audience part provides a detailed analysis of the users visiting the website. The Overview tab gives an a complete status of the visitors and their activities. There are tabs in the Audience part as:

- Demographic tab provides information about the age and gender of the users.
- Interests tab gives information about the interests of the users such as lifestyle, technology, sports, media, photography, and more. Also, in-market behavior part, it gives information about purchase intent by categories such as financial services, real estate, electronics, education and more.
- Geo tab provides information about the language and location of the users.
- Behavior tab gives insights about the behavior of the users as their session count, recency distributions, if they are new or returning visitors, and more.
- Technology tab shows the browser, OS, and network of users.
- User Flow tab shows the flow of users starting their journeys and following steps on the website.

4.4.3 Acquisition.

The Acquisition part provides an overview of where the traffic is creates sessions from such as direct, organic search, referral, social, or campaign communications. Traffic sources can be analyzed deeply by viewing them by channel, source/medium, and more in a detailed way.

In Google Adwords tab, Adwords campaigns can be provided detailly.

The Search Engine Optimization gives information about the keywords that users searched for to come to the website.

The Social tab gives insights about the traffic provided via social networks as Facebook, Twitter, Instagram, and others.

4.4.4 Behavior.

The Behavior part provides information about the visitors' behavior on the website considering the pages they visit, the events they do in the website. The Overview gives insights about pageviews, bounce rates, exits, average time on pages and pages that users visit. There are tabs in the Behavior part as:

- Behavior Flow shows the information of journeys from the first landing page to the exit page.
- Site Content shows the information such as the best content, landing and exit pages.
- Site Speed shows the information of pages with slow performance and other problematic issues about the website.
- Site Search shows the information of visitors that searched on the website.
- Events shows the information about the users and their actions on the website as viewing products, adding the products to the basket.
- Publisher shows the information about the AdSense publisher data right within Analytics.
- Experiments shows the information of A/B testing to compare landing page variations creates best of the conversion goals.
- In-page analytics shows the information of page statistics to the front end of the website.

4.4.5 Conversions.

- Goals shows the performances of the goals which are previously set.
- E-commerce shows the revenues, conversions, performances of campaigns & products.
- Multi-Channel Funnels shows the paths to conversion as assisted channels, top conversion paths.

Chapter 5

User Analytics, Algorithms and Event Prediction

Customer/ user analytics stands out as a very important concept today. The reason for this is the need for businesses and organizations to do their marketing activities in a more personalized way, taking into account the costs. However, many analytical studies, machine learning, artificial intelligence studies are carried out. In this study, it is tried to find a solution to the problem of predicting the next visit.

5.1 User Analytics

User analytics is analytical studies of the users in digital platforms as web or applications. Differently from customer analytics, it is able to search for the behaviors of even they do not sign in. These data can be used for understanding user behavior, the bottlenecks of preventing a user to continue their behavior, predicting their next steps, these activities provides business growth and user retention. There are many studies in the literature with a subject of user analytics. Collecting the event data of users are crucial for the digital platforms, Bernaschina C., Brambilla M., Mauri A., Umuhoza E. (2017) developed a big data analysis framework for collections and streaming of data. Also, finding patterns of user features and behavior is important and Yahia, S. (2017) had a study related with it. In addition, user analytics can be applied in many area, such as for Zhang, J., Tjhi, W., Lee,B., Lee, K., Vassileva, J., Looi, C. (2010) developed a framework for course management in their study. However, Bahad, P., Saxena, P. Kamal, R. (2020) explored and predicted user preferences with Lego toys dataset. Zhu, J., Zhou, Y., Guan, L., Hou, L., Shen, A., Lu, H. (2019) made their studies about user analytics and effects of social media. In addition to these, Akutota, T., Choudhury,S. (2017) pointed of security challenges of big data and resulting from privacy risks of users. Adding to these, recommendation for interests of users are a great system that increase the interactions, Koutrika, G. (2017) had a study about user analytics in recommendation systems.

5.1.1 User features.

Understanding the users' features plays an important role for segmenting the users or predicting their behaviors. The features can be divided as demographic features, behavioral features, geographic features, channel features, value based features, and so on. Demographic features as age, gender, maritial status can be collected via forms in platforms. Geographic features as city, country and channel features as a user prefers to use app or web can be collected, behavior based can be collected by events that users' behaviors while using the app. Also value based features can be get from monetary values of the events that users'.

5.1.2 Segmentation.

After collecting user data, in order to understand and determine the patterns of users' behavior, there is a need to segmentation. Segmenting the users starts from determining the features of users and purpose of segmentation. Machine learning techniques especially unsupervised learning methods are applied in segmentation and they gives great results and diminish the human effect. Segmentation is widely used in social media platforms. Risius, M. And Aydingul, O. (2018) analyzed and classified Facebook profiles into different segments and identified the most important features to segment the users in their study. An, J., Kwak, H., Jung, S., Salminen, J., Jansen, B. (2018) segmented online customer data for products gathered from online social media platforms. Lee Y., Park I., Cho S., Choi J. (2018) applied a segmentation for app usage of smartphone users. Fu, X., Chen, X., Shi, Y., Bose, I., Cai, S. (2017) segmented users in online social games in their study. Moon, S., Jalali N., Erevelles, S. (2021) had a study of segmenting reviewers and businesses on social media from the survey and transaction data, also the reviews.

5.1.2.1 k-Means.

k-Means is a method to group similar items in the same group, and unsimilar items to different groups. It is based on distances between data points and clustering the items with different distance metrics and divides the data in k groups. The goal is minimizing the group distances and maximizing the distances between different groups. K-Means is developed in 1967 by Mac Queen and its popularity comes from creating quick and fast

solutions even in huge data and the steps are figured in Figure 4 (Arai, K., Barakbah, A., 2007).

Algorithm steps are defined as in the study of Shobha, G., & Rangaswamy, S. (2018):

Step 1: Randomly selecting the data points (k) , setting the first cluster centers.

Step 2: Matching each data point to the closest center

Step 3: Creating new clusters by calculating the centers and assigning the data points to them

Step 4: If the cluster logic doesn't work, repeating from step 2

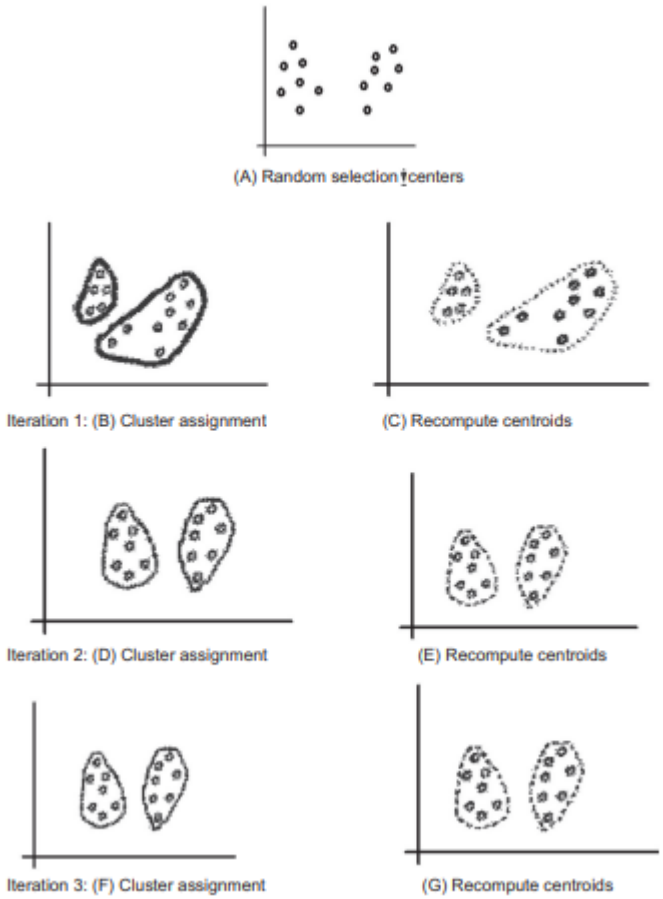

Figure 4. kMeans steps (Arai, K., Barakbah, A., 2007)

K-Means clustering can be applied in many areas as customer segmentation, image and pattern recognition, document clustering, and so on. Beginning from recommender systems, Zahra, S. Ghazanfar, M., Khalid, A., Azam, M., Naeem, U. (2015) applied k-

Means in recommendation systems and by considering of centroid selection how it effects the performance of recommendation. Jing, L., Ng, M., Xu, J., Huang, J. (2005) studied on large text data to cluster them and based on their calculation of feature weights and use them to discover the clusters. Liu, H. (2010) take advantage of k-Means for analyzing and proposing a model for internet public hotspot detection. Zhong, W., Altun, G., Harrison, R., Tai, P., Pan, Y. (2005) took advantage of k-Means for finding local protein patterns. Perez, J., Mexicano, A., Santaolaya, R., Hidalgo, M., Moreno, A. (2012) applied k-Means algorithm for building bee honeycomb structure with spending less time.

k-Means can be applied via different systems as open source programmes, SQL, Cloud systems, etc. Ordonez, C. (2006) studied on applying k-Means clustering with Relational Database Management Systems using SQL. Gopalani, S., Arora, R. (2015) compared Apache Spark and Map Reduce using k-Means and founded that Spark is a powerful framework in many ways.

5.1.3 Classification.

Classification is a machine learning technique that provides to categorize the data groups into different classes. Classifying buyer/ user features is so crucial for the organizations to understand the behaviors of users. According to these features, with appling the machine learning algorithms predicting user behavior and taking actions considering these results. By doing this, organizations can be more profitable since they have known about what will be happening in advance.

The classification model gets the training data, it has features as input and target as the value that needs to be predicted. Then, a classification algorithm run and predict the new values, with comparing these values to the test set, the accuracy of the model can be found.

Classification is used in many areas and many studies. In medical base, Zhang, J., Xie, Y., Wu, Q., Xia, Y. (2019) has studied in medical area that they classified medical images with a synergic deep learning model by using deep convolutional neural networks. Lashari,S., Ibrahim, R., Senan, N., Taujuddin, N. (2018) has studied on medical data and applied data mining models to classify them. There are many studies in city planning area that Kranjcic, N., Medak, D., Zupan, R., Rezo, M. (2019) applied classification models as

support vector machine, random forest, artificial neural network, and the naïve Bayes classifier for detecting green urban areas from satellite images. Liu, S., Su, H., Cao, G., Wang, S., & Guan, Q. (2019) also applied classification methods for findind annual land covers in urban areas. When it comes to e-commerce area, there are huge number of studies applied in the literature. For instance, in the e-commerce sites, there are many products and many reviews about them, analyzing these reviews is a challenging process, Xu, F., Pan, Z., Xia, R. (2020) studied about this topic and applied a Naive Bayes model to classify product reviews. Wazarkar, S., Keshavamurthy, B. (2018) applied a classificaion model for classifying the fashion images using linear convolution and appending points using local features.

5.1.3.1 Logistic regression.

Logistic Regression is a classification algorithm that provides to make predictive analysis. It is a supervised learning process that the target labels are static and known before the modelling, and the prediction values is in these target values.

It is based on logistic function that takes the value between 0 and 1. It is called as Sigmoid function and the equation is shown in Equation 1.

y = 1/ (1+ e*-x) [1]

There are inputs as features and they are combined with weights and the intercept or bias to predict the value. It represented in Equation 2.

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x)) [2]

In this equation, x are the inputs, b1 is the coefficient for input, b0 is the bias/intercept value.

For instance, this algorithm is be applied in biological classifications in gene data by Sartor, M., Leikauf, G., Medvedovic, M. (2008) for identifying biological groups. Lee, H., Tu, Z., Deng, M., Sun, F., & Chen, T. (2006) developed using logistic regression a kernel logistic regression to enlarge interactions of closest proteins. These are the applications in biological area. As an example, one of the areas that logistic regression applied is e-commerce. Yanfang, Q., & Chen, L. (2017) applied logistic regression model for churn prediction in e-commerce area and created a new model based on logistic regression and its predictive power is high. In the study of Mohammadifard, N. (2013)

logistic regression is applied to predict user behavior on e-commerce data, also Hidden Markov Models is added to make a hybrid approach.

5.1.3.2 Support vector machines.

Support Vector Machine is supervised machine learning algorithm applied for classification. Support Vector Machine tries to develop the best limits that divides the space into different groups. The limit line is the hyperplane and the extreme points are chosen as Support Vector Machine to create the hyperplane. There is a term called margin as the closest distance between the nearest points, and the best margin value is the maximum. In the process of SVM, hyperplanes are generated at first and as it can be seen in the equation 3, samples(X) has weights and after a new sample joined the equation becomes as in equation 4.

$$W.X + W_0 \text{ [3]}$$

$$W.X_p + W_0 \text{ [4]}$$

Selecting the true hyperplane that seperate the data from the nearest data points. In the case of the hyperplane is not linear, SVM uses kernels. It transforms the input space ino a high dimensional space. It adds more dimensions to make the data seperable. And the figure 5 shows the hyperplane selection progress.

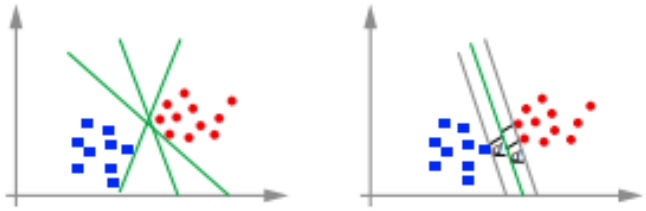

Figure 5. Hyperplanes and support vectors

Kernels are divided as:

- Linear kernels
- Polynomial kernels
- Radial basis function kernels

In the biological applications Support Vector Machines are being widely used as Noble, W. (2006) go around in their study that in the gene dataset, Support Vector

Machines is the best performed algorithm. However, Support Vector Machine is can be applied in Customer Analytics area that Guo-en, X., Wei-dong, J. (2008) performed SVM (Support Vector Machine) for customer chrun prediction and in the specific dataset, SVM performed best when comparing with other prediction algorithms.

5.1.3.3 Random forest.

Random Forest is a supervised machine learning algorithm and can be used both prediction and classification. It is an ensemble method that applies same algorithm and creates a powerful algorithm fed from their errors and learnings. Random forests are made by the combination of multiple decision trees. By doing this, biases are decreased with the strength of multiple combinations, the model stability is increased. Both numerical and categorical variables work well in random forest algorithms. It sometimes spend more hours than other classification algorithms. Its working principle starts with choosing N random values then building a decision tree based on these values. According the required tree number, these steps repeated. Each tree predicts the value of the new value. At the end of the process, the value of the majority defines the new value. The process is stated in the study of Misra, S., & Li, H. (2020) and is shown in Figure 6.

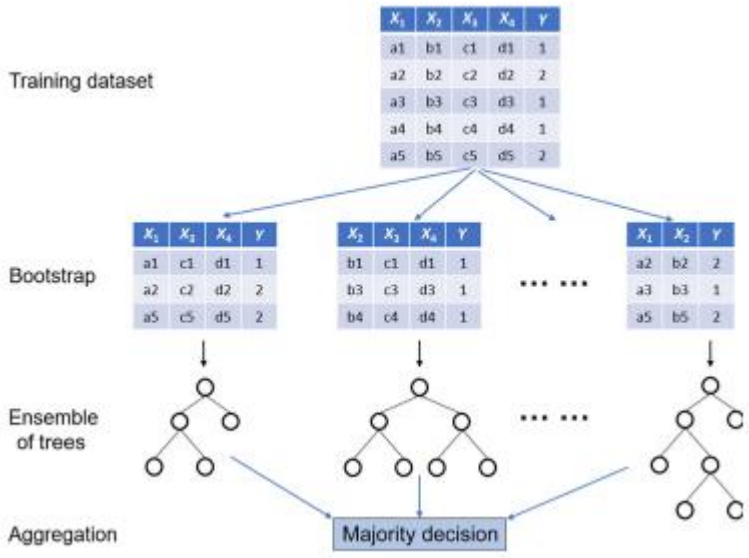

Figure 6. Random forest process

Random forests are algorithms that widely applied, for instance Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018) used random forests to detect the fraud transactions with credit card in a e-commerce dataset from China. In another case, we can observe that random forests are used for skin color classification and found the robust skin types in the study of Khan, R., Hanbury, A., & Stoettinger, J. (2010). Also, in the study of Xu, B., Guo, X., Ye, Y., Cheng, J. (2012), it can be observed that random forests are good classifiers for text categorization, it is successful in high dimensional data.

5.1.3.4 Light gradient boosting machine (LightGBM).

Light Gradient Boosting Machine is algorithm that based on the framework of gradient boosting trees. It is good at memory usage, based on histograms and replace the continuous observations into discrete bins. It is fast in training process comparing with Gradient Boosting Decision Tree (GBDT) with the same accuracy as for Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. (2017). The steps of LightGBM algorithm is citated below as in the study of Zhang, D., & Gong, Y. (2020).

Step 1: Defining a loss function

Step 2: Sampling with larger gradients

Step 3: Histogram for identifying the optimum segmentation point

Step 4: Feature dimension (Exclusive Feature Bundling)

Step 5: For increasing accuracy, appliyng leaf-wise algorithm with depth limitation

Step 6: The leaf nodes include the samples are united for fitting the residuals

Step 7: Splitting the nodes with tree structure scores

Step 8: End of the process with a final decision tree

The leaf based process is stated in the study of Al Daoud, E. (2019) and is shown in Figure 7.

Figure 7. Leaf based process of LightGBM

As it is a newly developed algorithm, it is applied in many areas. For instance, in the study of Xiaolei, S., Mingxi, L., Zeqian, S. (2018) LightGBM performed well for forecasting cyrptocurrency prices in the specific dataset. Also, as for the study of Daoud, E. (2019) in the home credit dataset LightGBM gives better results than Catboost and XGBoost as it is faster and more accurate. Also with combining both XGBoost and LightGBM, better accuracy results can be get in the traffic flow prediction as for the study of Mei, Z., Xiang, F., Zhen-hu, L. (2018).

5.1.3.5 XGBoost (eXtreme gradient boosting).

Extreme Gradient Boosting is a common supervised algorithm, can be used for classification and regression. It creates ensemble decisions from weak predictions for aiming optimization of loss function.

The steps of eXtreme Gradient Boosting algorithm is citated below as in the study of Zhang, D., & Gong, Y. (2020):

Step 1: Defining a loss function

Step 2: Transforming objective function

Step 3: Create initial tree for prediction of samples

Step 4: Extending loss function by applying the second order derivative Taylor rule

Step 5: For simplifying the the objective function ignoring the constant term

Step 6: Adding the residuals of the leaf nodes

Step 7: Scoring the tree structure and splitting the nodes of the tree

Step 8: End of the process with a final decision tree

The leaf based process is stated in the study of Al Daoud, E. (2019) and is shown in Figure 8.

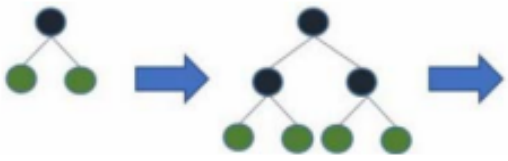

Figure 8. Leaf based process of XGBoost

XGBoost algorithms are widely used in many areas as biology, pricing and sociology topics. In the study of Ogunleye, A., & Wang, Q. G. (2019), chronic kidney disease is researched and found a perspective developed with using XGBoost for diagnosing chronic kidney disease and with like these studies, it would be possible to save lives. Also, Li, M., Fu, X., & Li, D. (2020) used XGBoost for predicting diabetes and it provides preventing the diabetes.

As a different area, Shi, X., Li, Q., Qi, Y., Huang, T., & Li, J. (2017) applied XGBoost for predicting urban fire accidents XGBoost is used for predicting crude oil prices as for the study of Gumus, M., & Kiran, M. S. (2017).

5.2 Google Analytics Dataset

In this study, Google Analytics sample dataset which was launched by Google has applied. It can be processed by using Google BigQuery. The dataset belongs to Google Analytics 360, gathered from Google Merchandise Store which is selling Google-brand .

- Traffic source data: This part includes where the visitors come from as organic traffic, paid search traffic, display traffic, etc.
- Content data: This data includes the patterns of the visitors in the website as page URLs, content information, etc.
- Transactional data: This data includes order data gathered from visitors orders.

Source:

(https://support.google.com/analytics/answer/7586738?hl=en#zippy=%2Cin-this-article)

5.2.1 Dataset definition.

In this study, there are some metrics and dimensions gathered for the raw data represented in the Table 4.

Table 4

Raw Data Of The Study

Feature	Data Type	Explanation
fullVisitorId	object	Id of a user
uniqueId	object	A key merged fullvisitorId and uniqueId, represents a unique session
visitStartTime	object	Exact time of a visit starts
Date	object	Date format of visitStartTime
visitNumber	int64	Rank of a visit
visitId	int64	Id of a visit
browser	object	Browser type of a visit
deviceCategory	object	Device type of a visit
source	object	Source of visit
medium	object	Medium of visit
channelGrouping	object	Channel of visit
hitNumber	int64	Hit number of a visit
total_visits	int64	Total visits in that session
transactions	object	Transaction of a visit
transactionrevenue	object	Revenue of a visit
pageviews	object	Pageviews of a visit
totaltimeOnSite	object	Total time spend on a visit

5.2.2 Explanatory data analysis.

The dataset includes 903,653 rows and 17 dimensions. In the Figure 9, some statistical information about the numeric values has given.

	visitNumber	visitId	hitNumber	total_visits	transactions	transactionrevenue	pageviews	totaltimeOnSite
count	903653.000000	9.036530e+05	903653.000000	903653.0	11552.000000	11515.000000	903553.000000	451894.000000
mean	2.264897	1.485007e+09	4.613173	1.0	1.048736	154.593941	3.849764	262.612141
std	9.283735	9.022124e+06	9.681080	0.0	0.455401	720.482634	7.025274	485.509238
min	1.000000	1.470035e+09	1.000000	1.0	1.000000	1.200000	1.000000	1.000000
25%	1.000000	1.477561e+09	1.000000	1.0	1.000000	29.990000	1.000000	32.000000
50%	1.000000	1.483949e+09	2.000000	1.0	1.000000	55.610000	1.000000	83.000000
75%	1.000000	1.492759e+09	4.000000	1.0	1.000000	116.615000	4.000000	259.000000
max	395.000000	1.501657e+09	500.000000	1.0	25.000000	47082.060000	469.000000	19017.000000

Figure 9. Statistical data information

When the browser type discussed, for instance, the data includes a huge number of types and there are some outliers that these data will be excluded. In the Table 5, top 10 browser types are shown.

Table 5

Top 10 Browser Types And Counts

Browser	Count
Chrome	620,364
Safari	182,245
Firefox	37,069
Internet Explorer	19,375
Edge	10,205
Android Webview	7,865
Safari (in-app)	6,85
Opera Mini	6,139
Opera	5,643
UC Browser	2,427

Briefly, finding the all values of browser, device, source and channel, top 10 most values shown in Figure 10. It can be seen that,

- Chrome, Safari, Firefox and Internet Explorer in browsers,
- Desktop in devices,
- Direct, Google, youtube.com in sources,
- Organic Search, Social, Direct, Referral, Paid Search in channel are the most used features.

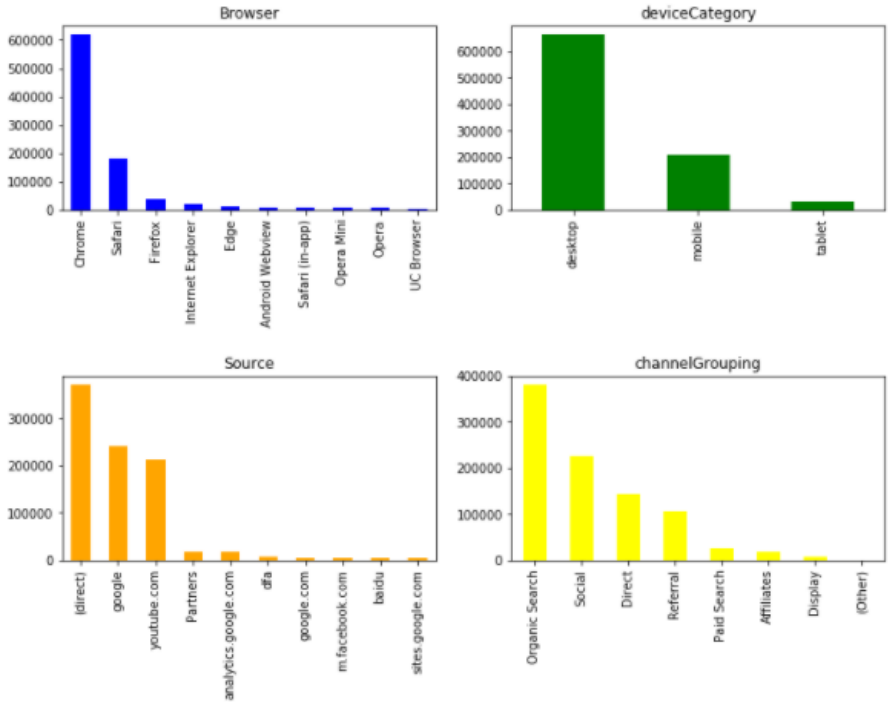

Figure 10. Most used browser, device, source, channel types

The data starts from 2016-08 and ends in 2017-08 as it is a 1 year data. In Figure 11, it can be observed the monthly visit counts. There are peak on 2016-11 which started to increase on 2016-10.

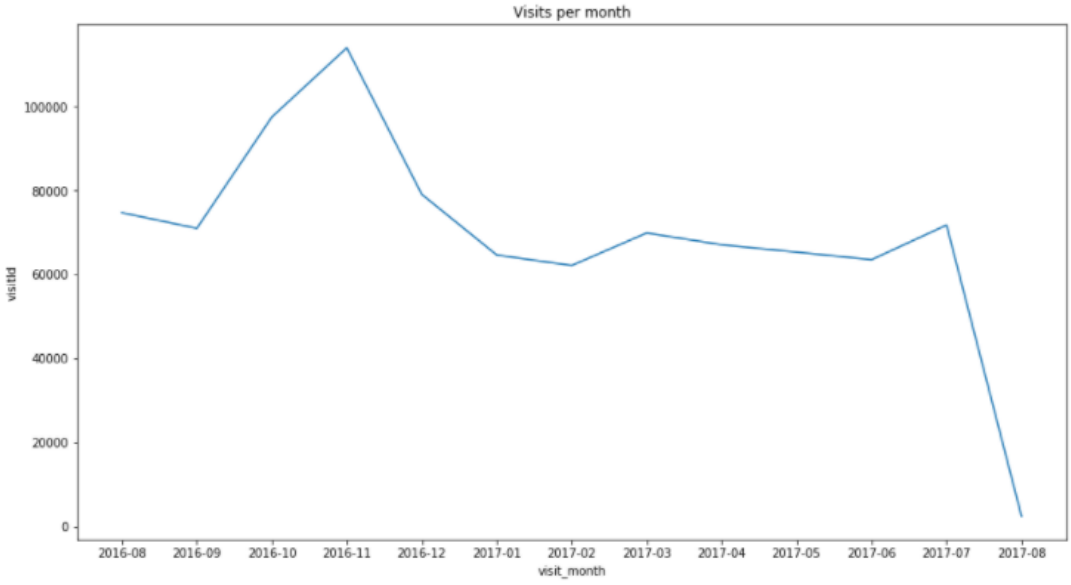

Figure 11. Monthly visit count

In Figure 12 the distribution of total order revenues can be observed in monthly basis. There are peaks on 2016-12, 2017-04 which is the highest number and 2017-07.

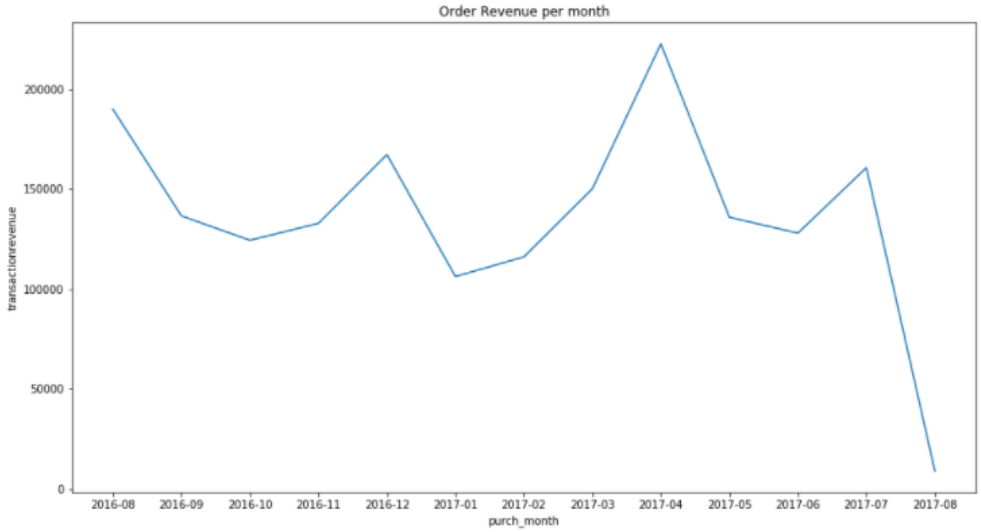

Figure 12. Monthly order revenue

There are outliers in browser, source, medium, device and channel based, if the count of the visit is less than 50 it is eliminated as a business perspective. As a outlier elimination technique, IQR approach has applied in this study. In the study of Yang, J., Rahardja, S, , Franti, P. (2019) stated that IQR is one of the most common approach handling with outliers. According to IQR values of source and browser counts, the data has between lower and upper limits of IQR has been stayed in the raw data and outliers are excluded. Outliers can be observed in Figure 13. Final data count has become 902,392 from 903,653.

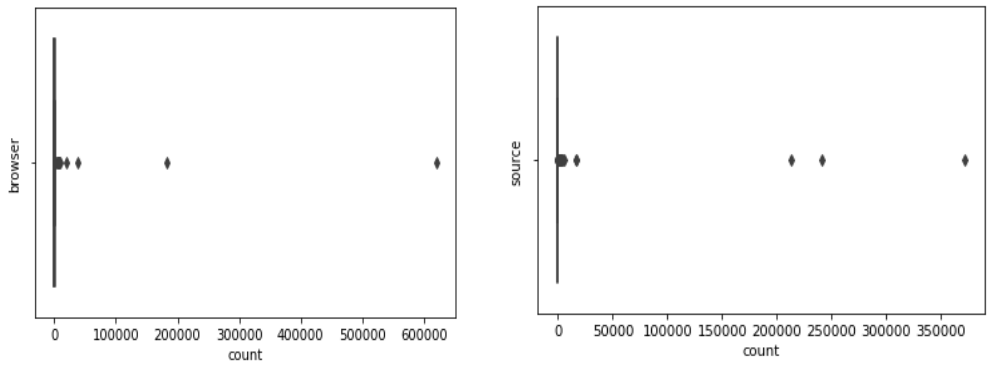

Figure 13. Outlier boxplot for source and browser

5.2.3 Methods.

In this part of the study, the methods used for data preparation & feature selection is explained briefly.

Boxplots are graph types in order to show the distrbution of data with showing the quartiles of values.

One-Hot Encoding is a method to create binary features from categorical variables. It helps the machine learning algorithms to seperate values in categorical columns. Seger, C. (2018) shows when One-Hot Encoding improves prediction performance in the study. Even, in the study of Yu, L., Zhou, R., Chen, R. (2020) One-Hot Encoding can be applied in missing data preproccesing.

Under Sampling is a method for oversampling and helps to solve imbalance in the data. It replicates the high volume classes and decreasing randomly to provide the balance. In the study of Qazi, N., & Raza, K. (2012), in the process of Network Intrusion Detection, founded that undersampling method is more efficient than SMOTE technique. Also, while detecting breast cancer, undersampling method has applied for better performance of algorithms (Qazi, N., & Raza, K. (2012).

Qazi, N., & Raza, K. (2012, March). Effect of feature selection, SMOTE and under sampling on class imbalance classification. In *2012 UKSim 14th International Conference on Computer Modelling and Simulation* (pp. 145-150). IEEE.

Hsu, J. L., Hung, P. C., Lin, H. Y., & Hsieh, C. H. (2015). Applying under-sampling techniques and cost-sensitive learning methods on risk assessment of breast cancer. *Journal of medical systems*, *39*(4), 1-13.

Boruta is a recent algorithm for feature selection on datasets. It evaluates the importance scores for features. High number of variables could not be efficient for model performance, and Boruta helps to select important features. Kursa, M., Jankowski, A., Rudnicki, W. (2010) founded this algorithm for finding all of the important variables in a dataset. The method works by duplicating the dataset and shuffling each feature, after this process, a classifier works on the dataset and compares the real features and predicted features, which have higher importances. After lots of iterations, if a feature performs well, it is labeled as important.

ANOVA is a statistical approach that determines the means of dimenions are significantly different from each other. The effect impact of the features controlled by comparing the means of different samples in ANOVA method.

Elbow method is applied for finding the optimal number of clusters into the data for unsupervised algorithms. Distortion metrics is evaluated by calculating the average of the squared distances from the cluster centers of the clusters considering the Euclidean distance.

The feature of the data may have different values as age, hours, meters and it causes difference in the scaling and a difficulties for the model. MinMaxScaler helps to normalize the data set that the features are transformed in the range [0, 1].

5.3 Event Prediction & Algorithms and Application

In this study, the goal is to figure our that a user visits in specific time periods by concerning their past visits.

5.3.1 Segmentation data preparation & application.

After Explanatory Data Analysis, there is a segmentation layer, users are aggregated by their mostly used source, browser, channelGrouping, deviceCategory, medium values. After this process, the data consists of source, browser, channel grouping, device category, medium, preferences and total transactions, transaction revenue, pageviews, total time on site of the visitors.

After completing these processes, categorical columns are processed to encoding since they are categorical columns and this is not suitable for segmentation. It created new

binary variables that are constituted from categorical variables. In this data, these values are browser, source, channel grouping, device category and medium.

There is a Recency, Frequency, Monetary and Tenure layer calculation and another frequency term named as AvgFrequency calculated by the difference of Tenure and Recency of the visitors divided by visit count. To make the segmentation more efficient, buyers who have more than one visit included to the study as a rule based logic. Unique visitor count became 91,836.

All of the manipulated data combined and prepared for the kmeans algorithm. There is a need to scale the data 1 by using MinMaxScaler, since it involves different values since kMeans algorithm is a distance based approach and to maket he points between 0 and 1. To sum up, the data consists, visitorId, their source, medium, browser, channel and device preferences, transaction behavior and visit recency, count of visit, monetary, tenure and frequency of visits of each visitor.

For finding ideal cluster numbers, Elbow method has applied. Segmentation is applied according to elbow method values as k=5 with 5 segments as it it shown in Figure 14.

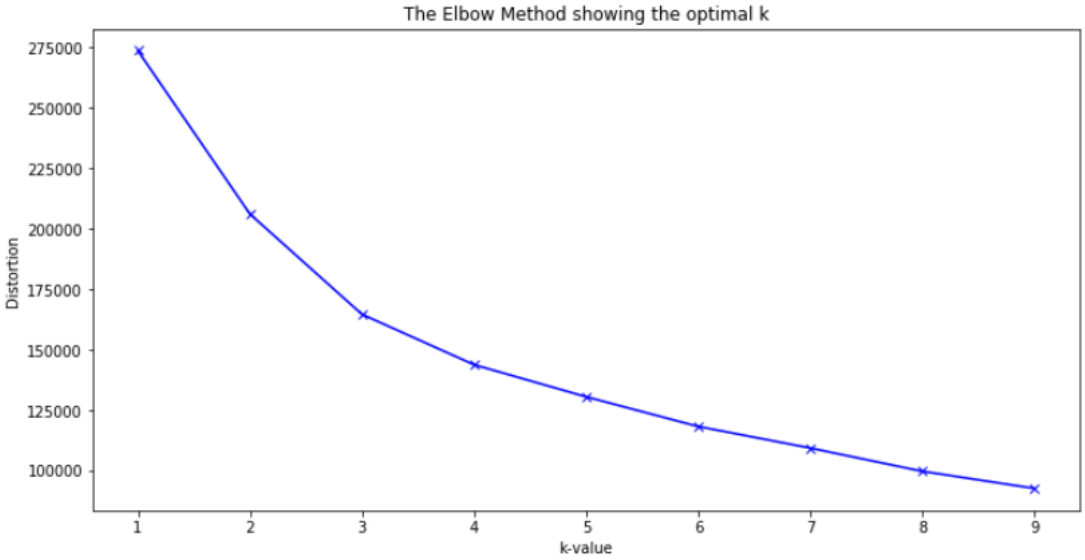

Figure 14. Elbow method

Segmentation results will be represented in the ‘Model Results’ part. The cluster labels will be added as a feature of prediction data.

5.3.2 Prediction data preparation & application.

For prediction, first the data is sorted considering the importance of dates and the historical events as it is a time based data.

Prediction data is prepared based on user and visit day. On daily basis, source, browser, channel, device, medium, visit count, transaction count, transaction revenue, total page views, total time on site values are calculated. After this step, the data as browser, transactions, source and so on are collected for the date before the last visit for predicting the next event.

The previous day before the visits of each user from the dataset and the differences between them calculated and the mean of these differences added as a feature. Also, clustering results have been added as a prediction dimension. According the discretizing the data for DayDiff feature, the data has been divided into 3 bins. The bins will be used as a target value.

Since adding the date feature to the model is not efficient, 'date' feature is divided into different dimensions as year, month, day, dayofyear, weekofyear, weekday, quarter, if the date is the beginning of the month, if the date is the end of the month.

After this process, since the data involves categorical variables, there is a need to encode the source, browser, channelGrouping, deviceCategory and medium features. These features also encoded seperately for modeling.

Finding the correlations between features is so crucial in order to prevent the dependencies between variables. The Pearson correlation between variables is represented in Figure 15. Dimensions as date, Day Difference and Day Difference Mean dimensions are eliminated for not to dominate the other dimensions. Also, there are dimensions with high correlations between each other as quarter, weekofyear, dayofyear, year, Tenure, pageviews, is_month_start, is_month_end are eliminated from the data since there are still some dimensions exist by representing these dimensions. Since ANOVA can be used for finding correlations between categorical dimensions and target values, applying ANOVA test, weekday, transaction revenue and recency columns are eliminated.

Also, categorical values as source, channelGrouping, browser, medium, deviceCategory and cluster dimensions are encoded for better prediction.

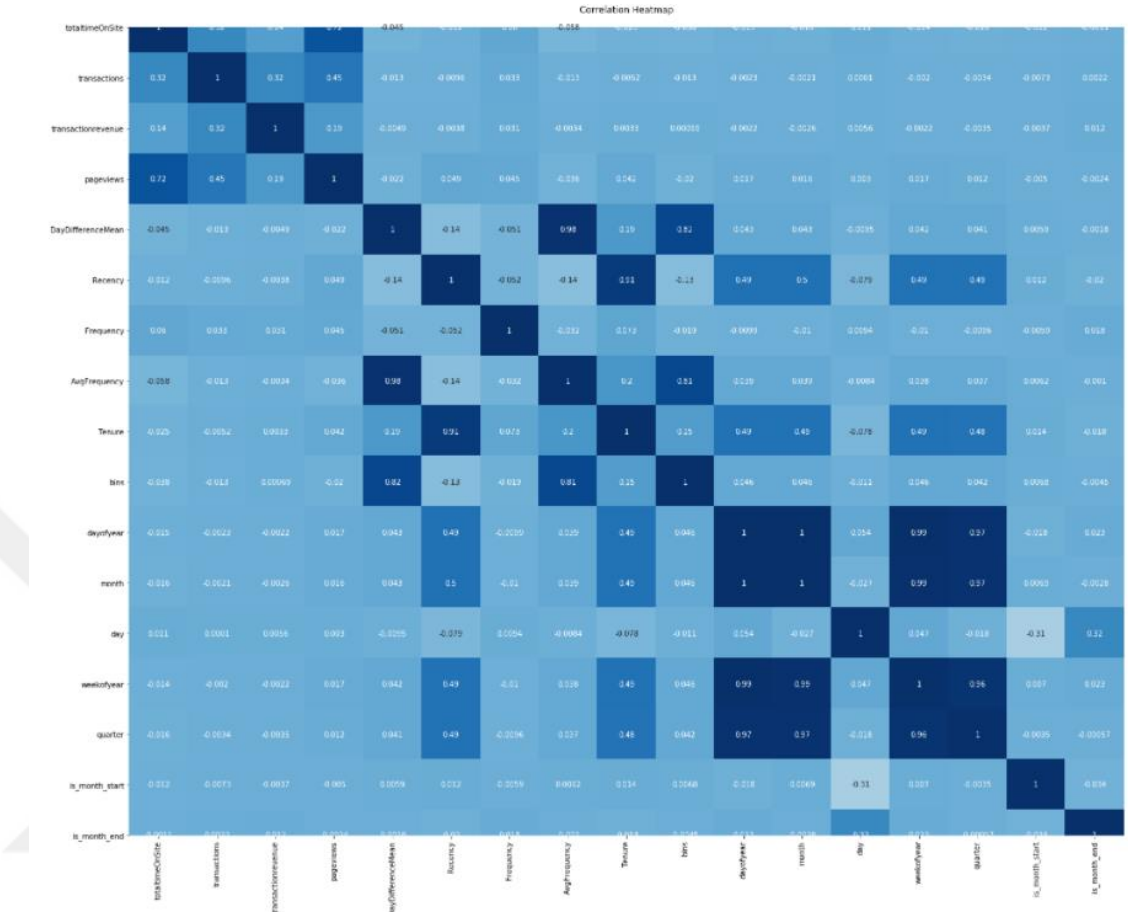

Figure 15. Correlations between variables

With completing these processes, bins label, gathered from next visit day, is seperated for the prediction. Since there is different sizes of data values, there is a need occurs to scale the data by MinMax method, and the data is divided into train and test as %80 of the data creates the train data and %20 of the data creates the test data.

However, the data consists high number of ‘0’ label and other labels have low volume, in order to prevent this inequality, Under Sampling method has applied and it balanced the data. In addition, finding the best effective features is crucial so Boruta algorithm has applied and best ranked features has involved in the model. These features are totaltimeOnSite, Frequency, AvgFrequency, month, source_direct, cluster, browser_Safari , browser_Chrome, medium_(none).

In the prediction step, Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest and LightGBM, k Nearest Neighbors, XGBoost, Gaussian

Naive Bayes methods has applied with cross validation. With comparing the accuracy, F1 score metrics the best method was found as XGBoost. It applied to the data and confusion matrix will be shown in 'Model Results' part.

5.4 Model Results

Model results are stated in two parts as the segmentation layer and prediction layer. In segmentation results parts segments' specialities has demonstrated. In the prediction part, accuracy, F1 score values can be observed.

5.4.1 Segmentation results.

After the segmentation process, according the mobile, recency, tenure, frequency, behaviors, clusters can be explained as in Table 6.

Table 6

Segmentation Result

cluster	browser_Chrome_ mean	browser_ Safari_ mean	Avg Frequency_ mean	pageviews_ mean	Device Category_ mobile_ mean	Tenure _ mean	Recency _ mean	deviceCategory _desktop_ mean
0	0,6	0,4	3,7	12,2	0,9	198,1	186,3	0
1	0,9	0	6,7	23,2	0	205,8	183,1	1
2	0,5	0,5	5	12,1	0,9	156,1	142,4	0
3	0,8	0,1	6	17,2	0	167,4	150,2	1
4	0,8	0,1	4,7	9,3	0,1	213,7	200,6	0,8

Cluster features can be described as:

-Cluster 0: The users in Cluster 0 tend to use mobile and Chrome & Safari mostly with the lowest frequency.

-Cluster 1: The users in Cluster 1 tend to use Desktop and Chrome mostly, they are old visitors and less active visitors, also they are frequent.

-Cluster 2: The users in Cluster 2 tend to use more mobile, newly started to use this site and more recent.

-Cluster 3: The users in Cluster 3 tend to use Desktop and Chrome, visit many pages, frequently visit the site.

-Cluster 4: The users in Cluster 4 tend to use Desktop and Chrome mostly, and they have less pageviews, they are not recent but they have highest tenure.

5.4.2 Prediction results.

Feature importances graph is shown in Figure 16. Most important features are AvgFrequency, totaltimeOnSite, month, Frequency, cluster, browser_Safari, source_direct , browser_Chrome, medium_(none) with order.

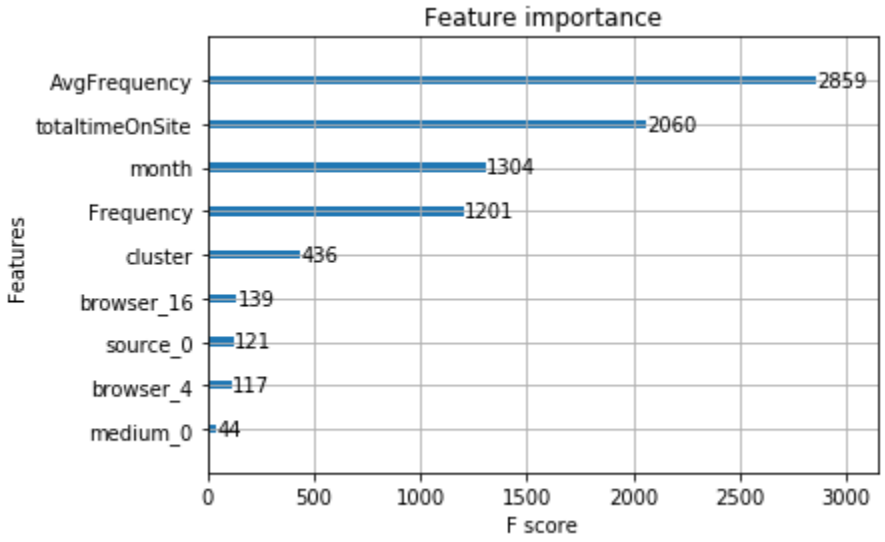

Figure 16. Feature importances

A few methods as Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest and LightGBM, k Nearest Neighbors, Gaussian Naive Bayes XGBoost methods has applied and cross validation results is shown in Table7. As it can be observed, best results belong to XGBoost and for the prediction XGBoost has applied.

Table 7

Cross Validation Results

	LightGBM	**Logistic Regression**	**Support Vector Classifier**	**Decision Tree**	**Random Forest**	**Gaussian Naive Bayes**	**XGBoost**	**k Nearest Neighbors**	**Best Model**
Accuracy	0.923218	0.694144	0.744810	0.898319	0.638654	0.814049	0.924145	0.796016	**XGBoost**
F1 Score	0.955951	0.758472	0.811346	0.940723	0.743091	0.886542	0.956336	0.875163	**XGBoost**

Accuracy would be 0.966 and F1 score would be 0.966. The detailed classification report is shown in Table 8.

Table 8

Classification Results

	precision	recall	f1-score	support
Class 1	0.98	0.99	0.98	11061
Class 2	0.85	0.78	0.81	1149
Class 3	0.90	0.90	0.90	257
accuracy			0.97	12467
macro avg	0.91	0.89	0.90	12467
weighted avg	0.97	0.97	0.97	12467

Confusion matrix is shown in Table 9. In addition to that ROC-AUC score is 0.90.

Table 9

Confusion Matrix

	0	**1**	**2**
0	10927	136	3
1	228	898	23
2	4	21	232

When the cluster dimension eliminated, as there is no segmentation layer, the features are changed. Accuracy would be 0.962 and F1 score would be 0.962. Confusion matrix is shown in Table 10.

Table 10

Confusion Matrix Without Segmentation

	precision	recall	f1-score	support
Class 1	0.98	0.98	0.98	11061
Class 2	0.77	0.80	0.79	1149
Class 3	0.87	0.75	0.81	257
accuracy			0.96	12467
macro avg	0.87	0.84	0.86	12467
weighted avg	0.96	0.96	0.96	12467

Since Average Frequency metric is highly effective to predict to result, when this metric is excluded, prediction ability is decrased and accuracy metric would be 0.886 and F1 score metric would be 0.835.

Chapter 6

Conclusion

With this study, which combines segmentation and prediction, a flow that can be used in real e-commerce sites is provided by estimating the next purchase time of the customer. The outputs here can be used by growth-oriented departments such as marketing or sales, and special communications can be made to people who are approaching the next purchase day, or if the person has more than the estimated purchase time, a discount can be defined for that person and actions can be taken to shorten the purchase time.

There is a system that is created in this study, that analyzing users' behavior and segmenting by considering their lifetime behavior. These segments are used as a dimension for prediction step. The target value is made from the difference between users's last event date and the event before the last date. There are lots of methods applied as boxplot, under sampling, bin discretizer, scaling, encoding, Boruta algorithms. In addition to these, there are many algorithm compared to find the best performed algorithm, such as kmeans algorithm which is a supervised learning algorithm, and many classifers as Support Vector Classifier, Logistic Regression, Decision Tree which are supervised learning algorithms. Also, there are ensemble algorithms applied in this study such as XGBoost, LightGBM algorithms.

For better model performance, features can be added or removed, and the dataset can be expanded. Ensemble methods can be tested by combining different models. A model can be built based on transaction prediction.

REFERENCES

Akutota, T., & Choudhury, S. (2017). Big data security challenges: An overview and application of user behavior analytics. *Int. Res. J. Eng. Technol*, *4*, 1544-1548.

Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. International Journal of Computer and Information Engineering, 13(1), 6-10.

Amer-Yahia, S. (2017, July). Toward interactive user data analytics. In *British International Conference on Databases* (pp. 3-6). Springer, Cham.

An, J., Kwak, H., Jung, S. G., Salminen, J., & Jansen, B. J. (2018). Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, *8*(1), 1-19.

Armstrong, G., Adam, S., Denize, S., & Kotler, P. (2014). *Principles of marketing*. Pearson Australia.

Bahad, P., Saxena, P., & Kamal, R. (2021, March). Exploratory and Predictive Analytics of User Preferences from Kaggle LEGO-Toys Datasets Using Spark ML. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1099, No. 1, p. 012019). IOP Publishing.

Bernaschina, C., Brambilla, M., Mauri, A., & Umuhoza, E. (2017, June). A big data analysis framework for model-based web user behavior analytics. In *International Conference on Web Engineering* (pp. 98-114). Springer, Cham.

Burby, J., Brown, A., & WAA Standards Committee. (2007). Web analytics definitions. *Washington DC: Web Analytics Association*.

Chaffey, D., & Patron, M. (2012). From web analytics to digital marketing optimization: Increasing the commercial value of digital analytics. *Journal of Direct, Data and Digital Marketing Practice*, *14*(1), 30-45.

Filvà, D. A., Guerrero, M. J. C., & Forment, M. A. (2014, June). Google analytics for time behavior measurement in Moodle. In *2014 9th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE.

Fu, X., Chen, X., Shi, Y. T., Bose, I., & Cai, S. (2017). User segmentation for retention management in online social games. *Decision Support Systems*, *101*, 51-68.

Gumus, M., & Kiran, M. S. (2017, October). Crude oil price forecasting using XGBoost. In *2017 International conference on computer science and engineering (UBMK)* (pp. 1100-1103). IEEE.

Hasan, L., Morris, A., & Probets, S. (2009, July). Using Google Analytics to evaluate the usability of e-commerce sites. In *International Conference on Human Centered Design* (pp. 697-706). Springer, Berlin, Heidelberg.

Hendriksen, M., Kuiper, E., Nauts, P., Schelter, S., & de Rijke, M. (2020). Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers. *arXiv preprint arXiv:2012.08777*.

Jansen, B. J. (2009). Understanding user-web interactions via web analytics. *Synthesis lectures on information concepts, retrieval, and services*, *1*(1), 1-102.

Järvinen, J., & Karjaluoto, H. (2015). The use of Web analytics for digital marketing performance measurement. *Industrial Marketing Management*, *50*, 117-127.

Jeng, J. J., & Drissi, Y. (2000, October). Pens: a predictive event notification system for e-commerce environment. In *Proceedings 24th Annual International Computer Software and Applications Conference. COMPSAC2000* (pp. 93-98). IEEE.

Kaushik, A. (2007). *Web analytics: an hour a day*. John Wiley & Sons.

Kotsokechagia, M. (2021). Predictive model for customer satisfaction in e-commerce.

Kulik, R. (2020). *Basket Sessions-Purchase Intent Prediction and Analysis* (Doctoral dissertation, Ben-Gurion University of the Negev).

Kumar, V., Chattaraman, V., Neghina, C., Skiera, B., Aksoy, L., Buoye, A., & Henseler, J. (2013). Data-driven services marketing in a connected world. *Journal of Service Management*, *24*(3), 330-352.

Lee, Y., Park, I., Cho, S., & Choi, J. (2018). Smartphone user segmentation based on app usage sequence with neural networks. *Telematics and Informatics*, *35*(2), 329-339.

Li, M., Fu, X., & Li, D. (2020, March). Diabetes prediction based on xgboost algorithm. In *IOP Conference Series: Materials Science and Engineering* (Vol. 768, No. 7, p. 072093). IOP Publishing.

Misra, S., Li, H., & He, J. (2020). Noninvasive fracture characterization based on the classification of sonic wave travel times. In *Machine Learning for Subsurface* Characterization (pp. 243-287). Gulf Professional Publishing.

Mohammadifard, N. (2013). Modeling user behavior from e-commerce data with hidden Markov models and logistic regression.

Moon, S., Jalali, N., & Erevelles, S. (2021). Segmentation of both reviewers and businesses on social media. *Journal of Retailing and Consumer Services*, *61*, 102524.

Ogunleye, A., & Wang, Q. G. (2019). XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics*, *17*(6), 2131-2140.

Pakkala, H., Presser, K., & Christensen, T. (2012). Using Google Analytics to measure visitor statistics: The case of food composition websites. *International Journal of Information Management*, *32*(6), 504-512.

Phippen, A., Sheppard, L., & Furnell, S. (2004). A practical evaluation of Web analytics. *Internet Research*.

Plaza, B. (2011). Google Analytics for measuring website performance. *Tourism Management*, *32*(3), 477-481.

Risius, M., & Aydingul, O. (2018). Facebook user segmentation to enable targeted social advertisement.

Shi, X., Li, Q., Qi, Y., Huang, T., & Li, J. (2017, November). An accident prediction approach based on XGBoost. In *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)* (pp. 1-7). IEEE.

Shobha, G., & Rangaswamy, S. (2018). Machine learning. In *Handbook of statistics* (Vol. 38, pp. 197-228). Elsevier.

Tupikovskaja-Omovie, Z., & Tyler, D. (2021). Eye tracking technology to audit google analytics: Analysing digital consumer shopping journey in fashion m-retail. *International Journal of Information Management*, *59*, 102294.

Vecchione, A., Brown, D., Allen, E., & Baschnagel, A. (2016). Tracking user behavior with Google Analytics events on an academic library web site. *Journal of Web Librarianship*, *10*(3), 161-175.

Yang, J., Rahardja, S., & Fränti, P. (2019, December). Outlier detection: how to threshold outlier scores?. In *Proceedings of the international conference on artificial intelligence, information processing and cloud computing* (pp. 1-6).

Zhang, D., & Gong, Y. (2020). The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure. IEEE Access, 8, 220990-221003.

Zhang, J., Tjhi, W. C., Lee, B. S., Lee, K. K., Vassileva, J., & Looi, C. K. (2010, November). A framework of user-driven data analytics in the cloud for course management. In *Proceedings of the 18th International Conference on Computers in Education* (Vol. 29, pp. 698-702). Asia-Pacific Society for Computers in Education.

Zhu, J. J., Zhou, Y., Guan, L., Hou, L., Shen, A., & Lu, H. (2019). Applying user analytics to uses and effects of social media in China. *Asian Journal of Communication*, *29*(3), 291-306.