MARKETPLACE OPTIMIZATION BASED ON CUSTOMER ANALYTICS

DOĞUKAN AHUÇALAR

JANUARY 2022

MARKETPLACE OPTIMIZATION BASED ON CUSTOMER ANALYTICS

THESIS SUBMITTED TO THE

GRADUATE SCHOOL

OF

BAHCESEHIR UNIVERSITY

BY

DOĞUKAN AHUÇALAR

IN PARTIAL FULLFILMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF MASTER OF COMPUTER ENGINEER

IN THE DEPARTMENT OF NATURAL AND APPLIED SCIENCES

JANUARY 2022

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL

...../...../.....

MASTER THESIS APPROVAL FORM

Program Name:	**COMPUTER ENGINEERING**
Student's Name and Surname:	**Doğukan AHUÇALAR**
Name Of The Thesis:	**MARKETPLACE OPTIMIZATION BASED ON CUSTOMER ANALYTICS**
Thesis Defense Date:	**26 / 01 /2022**

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Prof. Dr. Ahmet ÖNCÜ
Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	Signature
Thesis Advisor's	**Assoc. Prof. Serkan TÜRKELİ**	
Member's	**Assist. Prof. Betül ERDOĞDU**	
Member's	**Assist. Prof. Tarkan AYDIN**	

I hereby declare that all information in this document has been obtained and presentedin 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 :

ABSTRACT

MARKETPLACE OPTIMIZATION BASED ON CUSTOMER ANALYTICS

AHUÇALAR DOĞUKAN
Computer Engineering Master's Program
Thesis Supervisor: Assoc. Prof. Serkan TÜRKELİ

January 2022, 54 Pages

Measuring the credibility of a merchants and enforcing them to provide good product and services to the customers via scoring the merchants based on their performance, brings reliability to the marketplace platforms / e-commerce companies.

When it's considered in terms of customer viewpoint, customers rely on other customers' feedback first under any circumstances. After that merchant scores and tags that approved by marketplace platform brings reliability to the merchant. According to that, merchant sales are going to increase considerably. For these reasons, it is required for the marketplace platforms calculate merchant scores to protects and give information to the customers which will buy products.

In the competitive environment that exist in the marketplace platforms, marketplace platforms should get the responsibility to provide reliable scores to the merchants those using their platforms and gave right information to the customers about the merchant. When trying to give reliable information to the customer, marketplace platforms should provide analyze tools and reporting tools that provide information about the giving score to encourage them to increase their score. By design, scoring merchants will extend their work ethics unintentionally.

In this study, a qualitative observation has been made by analyzing real data sets which contains information about the complete sales process which includes pricing, shipment, pre-support, after support and customer feedback etc.

Keywords: E-Commerce, Merchant analysis, Knowledge Management, Credibility

ÖZET

MÜŞTERİ ANALİZLERİ İLE PLATFORM PAZAR YERİ OPTİMİZASYONU

AHUÇALAR DOĞUKAN
Bilgisayar Mühendisliği Tez Programı
Tez Danışmanı: Doç. Dr. Serkan TÜRKELİ

Ocak 2022, 54 Sayfa

Bir üye işyerinin güvenilirliğini ölçmek ve üye işyerlerini performanslarına göre puanlayarak müşterilere iyi ürün veya hizmet sunmaya zorlamak, pazaryeri platformlarına / e-ticaret şirketlerine güvenilirlik getirir.

Müşteri bakış açısı açısından bakıldığında, müşteriler her koşulda önce diğer müşterilerin geri bildirimlerine güvenirler. Ardından üye işyeri puanları ve pazaryeri platformu tarafından onaylanan etiketler üye işyerine güvenilirlik kazandırmaktadır. Buna göre, tüccar satışları önemli ölçüde artacaktır. Bu nedenlerle, pazaryeri platformlarının ürün alacak olan müşterileri korumak ve bilgi vermek için üye işyeri puanlarını hesaplaması gerekmektedir. Pazaryeri platformlarında var olan rekabet ortamında, pazaryeri platformları, platformlarını kullanan tüccarlara güvenilir puanlar verme ve müşterilere tüccar hakkında doğru bilgileri verme sorumluluğunu almalıdır. Müşteriye güvenilir bilgi vermeye çalışırken, pazaryeri platformları, puanlarını artırmaya teşvik etmek için verilen puan hakkında bilgi veren analiz araçları ve raporlama araçları sağlamalıdır. Tasarım gereği, puanlama, tüccarların çalışma etiğini istemeden genişletecektir.

Bu çalışmada, satış sürecinin tamamı hakkında bilgi içeren gerçek veri setleri analiz edilerek nitel bir gözlem yapılmakla birlikte; fiyatlandırma, sevkiyat, destek öncesi, destek sonrası ve müşteri geri bildirimleri vs. ile ilgili konular ele alınmıştır.

Anahtar Kelimeler: E-Ticaret, Satıcı Analizleri, Bilgi Yönetimi, Güvenilirlik

TABLE OF CONTENTS

ETHICAL CONDUCT iii
ABSTRACT iv
ÖZET vi
TABLE OF CONTENTS viii
LIST OF TABLES x
LIST OF FIGURES xi
LIST OF ABBREVIATIONS xiiiii
Chapter 1: Introduction 1
Chapter 2: Literature Review 3
Chapter 3: Methodology 7
3.1 Data Collection and Process 7
3.1.1 Data Understanding 8
3.1.2 Data Preparation 9
3.1.3 Business Understanding 9
3.1.4 Modeling 11
3.1.5 Evaluation 12
3.1.6 Deployment 12
3.1.7 Confusion Matrix 14
3.2 Knowledge Management 14
3.3 Qualitative Methods and Structural Observation 15
3.3.1 Interview 15
3.3.2 Focus Group 16
3.3.3 Participatory Action Research 16
Chapter 4: Architecture 18
Chapter 5: Findings 24
5.1 Product and Reports 24
5.1.1 Merchant Rating Report & Analysis 24
5.1.2 Sales Operations Reports 25
5.1.3 High Potential Sales 27
5.1.4 Customers Price Alerts 28
5.1.5 Stock Offer 30
5.1.6 Live Performance 31
5.1.7 Dashboard 32

5.2 Analysis and Observation 33
Chapter 6: Discussions And Conclusions 37
REFERENCES 38

LIST OF TABLES

TABLES

Table 1 Rating Overall Table 25
Table 2 Sales Operations Report Table 26
Table 3 High Potential Sales Table 28
Table 4 Customers Price Alert Table 29
Table 5 Stock Offer Table 31
Table 6 Confusion matrix 34

LIST OF FIGURES

FIGURES

Figure 1 DIKW Hierarchy (Rowley, 2007) 5
Figure 2 CRISP-DM Process Diagram (IBM, 2021) 7
Figure 3 Agile Lifecyle 11
Figure 4 Deployment 13
Figure 5 Champion-Competitor Approach 14
Figure 6 Look, Think, Act Mechanism, adapted from (Türkeli & Erçek, 2010). 16
Figure 7 Gathering Data Workflow 20
Figure 8 Daily Operation Cycle Model 21
Figure 9 Rating Overall Model 22
Figure 10 General System Structure 22
Figure 11 Merchant Rating Analysis Interface 24
Figure 12 Sales Operations Report Interface 26
Figure 13 High Potential Sales Interface 27
Figure 14 Customers Price Alert Interface 28
Figure 15 Stock Offer Interface 30
Figure 16 Live Performance Interface 32
Figure 17 Complete Rota Interfaces 32

LIST OF ABBREVIATIONS

DIKW	Data – Information -Knowledge- Wisdom
SQL	Structured Query Language
ETL	Extraction - Transformation – Loading
CRISP-DM	Cross Industry Standard Process for Data Mining
WOE	Weight of Evidence
DB	Database
KM	Knowledge Management
HDFS	Hadoop Distributed File System
SME	Small or Medium Sized Enterprises
API	Application Programming Interface

CHAPTER 1

INTRODUCTION

It is not possible for marketplace platforms that do not approach every merchant, which uses their system, in a same way and that do not provide services by considering the risk and the value of the merchant could not continue to their activities and existence in a competitive environment. Marketplace platforms continuously provide services to merchants that is the main source of income due to competitive environment (Kleindl, B, 2000).

In the absence of systems to enable this approach, and in an environment of full competition, the institution will have no value in the long run. Marketplace platforms which do not provide easy to use environment, analyze tools for sales reports, campaigns etc. and transparency during scoring to merchants will lose their trust in stock market. Because the main source of income of the marketplace platforms are the commission get by every sale from any of the merchant. To illustrate with a few examples (Mehrotra & Carterette, 2019).

i. In case of a merchant who is not expert of complicated procedures and newcomer to the platform must guide properly during creation of the market in marketplace platform. Easy to use user interface with a few steps will increase merchant count.

ii. Merchants should be able to see their negative sides and easily report them to their supervisors to increase their sales by increasing their score. At the same time, after the recovery process of their negative sides, merchant should be easily aware of its potential and report it to its supervisor. Transparency at the scores will improve merchant trust to marketplace.

iii. Merchants should have the ability to migrate their own tools, external tools which they are used to use and the marketplace tools to work with both.

In the calculation of the scoring process of merchants, a single information will not be enough. In order to make the assessment real data sets which contains information about the complete sales process which includes pricing, shipment, pre-support, after support and customer feedback etc. Flexible and dynamic structures are

required that can be evaluated real time or daily calculations should be made using data (Agarwal, Dahleh & Sarkar, 2019).

Loss due to the absence of such a system; complex user interface, complex procedures, migration issues, limitations of essential tools etc. will cause the merchants choose another platform to use and that cause lack of commission income and customer portfolio of the merchant which will be most likely to use marketplace platform (Morrisey, Rivlin & Nathan, 2017).

In this study, an institution in the e-commerce sector will be analyzed and systems that solves mentioned above will be created. During the qualitative observation of the process, encountered problems, improvements, system designs and the third-party tools will be mentioned.

In the introductory part, which describes the general framework, information about the input data subject of the study and how the data plays a role in our acquisition of information is given. Knowledge management principles that used to find out chosen data is the source of truth. Known problems from the past experience will be mentioned.

In the next chapters, modern solutions created by the team managed via agile methodology. Refactoring of legacy systems, used technologies, and built system will be mentioned.

In the conclusion part, solutions and systems has succeeded and failed, emerged and unsolved problems, pondering questions about the methodological management systems will be mentioned, during one year of work.

CHAPTER 2

LITERATURE REVIEW

Information that has been converted into a form that is efficient for movement or executing is called data. Data is unprocessed and meaningless state of information. On the other hand, information is the transaction of data for a purpose. A wisdom hierarchy which is named DIKW (Data – Information -Knowledge- Wisdom) explains the confusion of information and data concepts, which generally mixed with each other. By the road of the DIKW hierarchy, structural observation model developed within the scope of this study is assumed to be wisdom, which makes easier to understand the requirement of data. DIKW model declares a range of models for representing meaningful structural and functional, scalable relationships between knowledge existence, wisdom and every grain of information (Rowley, 2007).

In this study, just as explained in the DIKW hierarchy description, customer data acquired from different sources like separated databases was transformed into information with the help of SQL tool and then stored in the data warehouse as information for the modeling stage.

To sequentially dive into the wisdom hierarchy as a summary.

Data: Data is explained in the following figures by the light of the wisdom hierarchy (Allen, 2004):

i. As data does not include any connection and subjective sentences and does not contain data relations, it does not mean any sense and value with just itself.
ii. The data do not contain any specific information as they are individual objective cases and observations that have not been organized and processed.
iii. Data items are simple and basically recorded definition of events, actions, existences, activities.
iv. Working on it and getting experiences

Knowledge: While open source programmed and recorded knowledge is virtually ready for sharing, shared knowledge is only available to a select few. It is a tough to define and comprehend idea that is founded on collected information (Aven, 2013).

i. Over time, a synthesis of many sources of information
ii. Expert opinion, skill set, and belief system
iii. Study and experience
iv. The process of accumulating rules by mixing what has been taught and experienced with them
v. Understand and absorb the concepts

can be interpreted in different ways as above.

Information: Organized and structured data means the explanation of information term. The configuration processes make the data relevant for a specific purpose and therefore can make the data relevant, valuable, meaningful and useful (Ahsan, 2006).

Wisdom: Wisdom is the ability to approach any new situation or problem in a critical and practical way, based on moral judgements. Though wisdom is an accumulated experience, gathered event by event, it is actually a concept that is hard to understand and directly related to human thinking, intuition, and interpretation rather than systems (Ahsan, 2006).

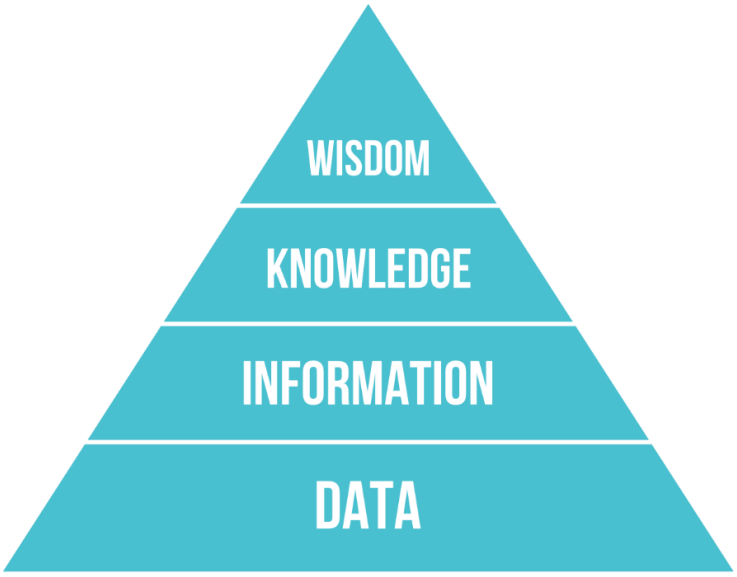

Figure 1 DIKW Hierarchy (Rowley, 2007)

In a competitive environment, analyzing and executing data has been vital fact for companies and has become an inevitable necessity nowadays. This situation has caused demand for storage systems to improve and increase day by day. It has become easier and cheaper to own in storage areas than in the past with improving and advancing this kind a technology. Companies that have the ability to invest more in data storage and interpret data have begun to make more flexible decisions (Haapasalo, Harkonen, Silvola & Vehkapera, 2009).

Every event that humanity make in their daily lives is an instance of data source for companies; the most important need of companies is to access the right data in a timely manner under this bunch of data bombardment. In order to use data effectively and fast responsibly, it is necessary to use ETL called (Extraction - Transformation - Loading) and data mining information technology tools. Therefore, it would be correct to consider technology as a strategic part of collecting information (Vassiliadis, Simitsis & Skiadopoulos, 2002). Unvaluable, incomplete, empty, and idle data has no meaning other than taking up space. Data is valuable only if it can reply decision making, director questions and provide guidance in situations of uncertainty. For this case, the compulsory features of the data can be ordered as:

Should be reached on time, easily, completely, accurately. Data that cannot be accessed in a timely manner and cannot be easily managed can cause decisions to be made late, to spend more time than necessary, or not to be made at all. These, in turn,

can cause the company to miss an opportunity or lose income. Inaccurate data can cause the company to make wrong decisions (Wang, 1998). Data must be of high quality and completely accurate along access to information through data analysis, if it is considered on contrary, the reasons for the poor quality of the data;

i. Inconsistent data entries
ii. Data retries
iii. Invalid, wrong type modeled data formats
iv. Uncorrelated, false data
v. Data entries that are out of date

can be listed.

The results may be meaningless and may cause the work to be in vain optimistically, to take a long time to work while all these situations are out of the topic. Consequently, paying attention to data quality and integrity is not considered a waste of time, but rather play a vital role that will increase time and efficiency.

CHAPTER 3

METHODOLOGY

3.1 Data Collection and Process

The value of using, analyzing, and interpreting techniques that can process the increased amount of data with basic methods has gained important role for those working in decision-maker authorities with the increase in data storage power.

Data mining can be defined as follows, with large amounts of data to determine the current situation and predict the future; They are analyzing that transform data into meaningful information. The most strategic feature that data mining techniques prepare to companies is that they reveal the resemblances, trends, and behaviors among the related, associated data groups. Decisions that are not based on documents cannot go beyond intuition, therefore they risk making wrong decisions.

Data mining techniques help decision support system for companies. It is possible to do data mining almost in every sector where there is data sheltered. Instances of these sectors can be given as medicine, active sports, banking, industrial fields. The standard process between industries in the world which is called CRISP-DM (Cross Industry Standard Process for Data Mining) cycle, summarizes the phases that need to be studied exactingly in a data mining study (Schäfer, Zeiselmair, Otten & Becker, 2018).

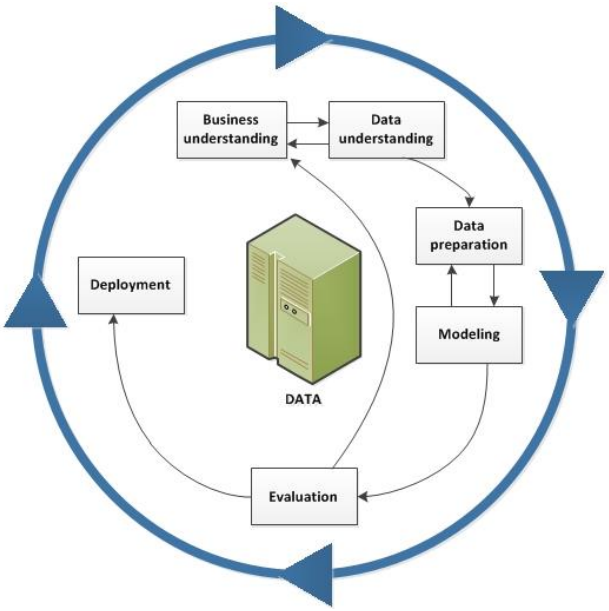

Figure 2 CRISP-DM Process Diagram (IBM, 2021)

The arrows above represent the most essential and common dependencies between the phases in this six-phase process cycle. In most projects, there is no specific order to the phases, and they can go back and forth as needed. The CRISP-DM cycle is versatile and adaptable. It enables the creation of models for unique requirements.

Because CRISP-DM is a high-level methodology, the processes indicated in the model can be done in a variety of ways, sequences, and technologies to meet business objectives. This model has a total of six steps will be explained.

3.1.1 Data Understanding

In order to minimize potential problems in the next phase of data preparation, it is critical to examine the data attentively and make notes on what you observe. For better understanding of data and data quality, graphs and tables are helpful (Wirth & Hipp, 2000).

i. Which variables can you use to get the information you need?
ii. What data is not necessary?
iii. Where can these variables be found in the table?
iv. Is there enough data to use the chosen modeling method and generate a prediction?
v. Empty fields, that is, unobserved values in the data source need to be filled?

At this point, looking for answers to questions that indicated above is beneficial.

Furthermore, the following critical qualities for characterizing the data and determining whether there are any situations that positively or negatively affect the study should be considered (Huber & Wiemer and Schneider, 2019).

i. Table record numbers
ii. Encoding types of values in tables
iii. Availability of data on established hypotheses
iv. Documentation of data needed for hypotheses

The most vital step in modeling is to recognize the logical relationship between data and business requirements. As a result, it is critical that the analyses of the task needs specified in the previous step be completed at this stage.

3.1.2 Data Preparation

The data preparation phase of the modeling process is the most important and time-consuming, accounting for approximately 50-70 percent of the total project time (Huber, Wiemer & Schneider, 2019). If enough time is spent on the phases of understanding the business need and learning the facts, this duration can be halved. The data preparation steps, according on the study's purpose as well as the company's data sources, essentially involve the following

i. Association and relation between data sets
ii. Taking totals according to the main distinguishing features (according to segments, features in the definition of need, etc.)
iii. When it is judged required, acquiring additional factors that can help in defining the business requirements
iv. Removing or combining slings and extreme values when it is necessary to examine
v. Unobserved, empty values to be removed or filled when it is necessary to examine
vi. Separation of continuous variables into significant and ordered groups from smallest to largest or largest to smallest
vii. Scanning the variables that are most relevant or unrelated to the model and that are predicted to be included in the model with variable selection algorithms

3.1.3 Business Understanding

It should be communicated clearly before beginning data mining work what the expectations are for the data mining activity. To put it another way, enough time should be allocated to figure out what the business requirement is. Because of this, essential persons should be involved in determining the job requirement, and talks should be documented. Despite the fact that there are no any meetings or conversations scheduled at this time, it is vital to guarantee that every person has a common

denominator when it comes to a certain business necessity. The most significant factors for forming hypotheses about the model outcome and testing the accuracy of the prediction model created by data mining are people with some industry expertise. Also, all types of collection, law, notice and other similar procedures that are used in the company and have modeling points (Martinez, Contreras & Ferri, 2019).

Agile Methodology: In general perspective, agile methodologies are common methods of software development which focus on quick actions of any occasions. These methods based on iterative and gradual development. Agile methodology focus on planning and understanding the problems before any action taken.

All agile techniques have four key characteristics: adaptable planning, iterative and evolutionary development, quick and flexible reaction to change, and communication oriented (Begel & Nagappan, 2007).

Its major emphasis is on adhering to the "Light but Sufficient" ideals and being people-oriented and communication-centered. Because it is referred to as a lightweight method, it is better suited for the creation of small projects (Maher, 2009). Business owners understands the need then take the project apart. After the separation of the project by the business owners, development team involves the process and make predictions about the restrictions and first requirements. Details of the requirements or restrictions, roadblocks could be affecting the project in the lifecycle of the agile methodology.

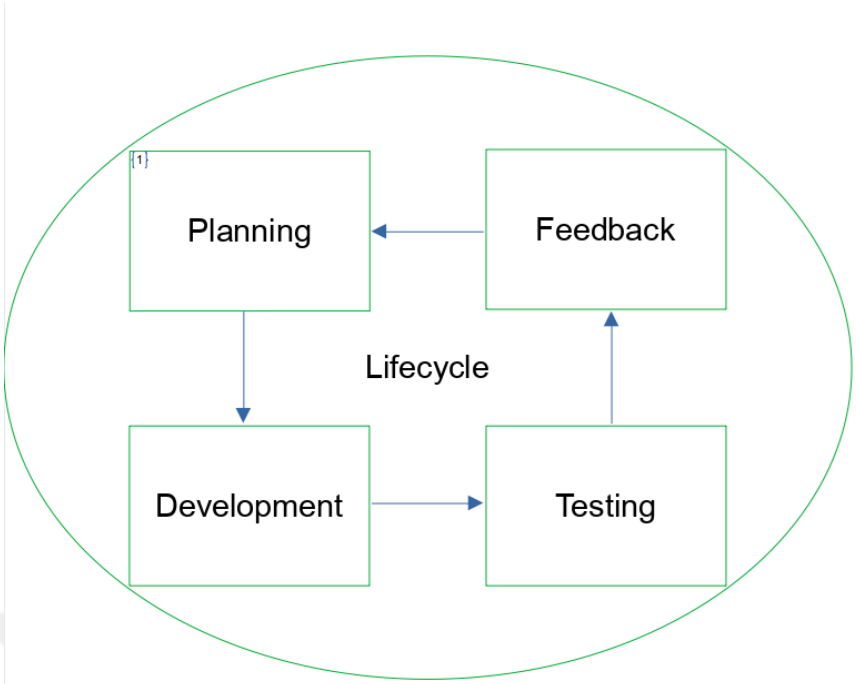

Figure 3 Agile Lifecyle

After the planning phase ends which means business owner and development team agreed on minimum viable product. Software is designed to meet fewer criteria at first but is created in such a way that additional requirements may be easily incorporated, resulting in more adaptability. Because of the decreased degree of functionality, this strategy makes the development time is reduced, and the program may be improved more quickly and for a longer length of time (Davis & Bersoff, 1988). Testing is incremental as the development process. That's because development is changeable and flexible, testing importance is increasing day by day. After the most viable product has released, feedback of the business owners and customers change and add feature to the projects continuously. Then the lifecycle of the agile methodologies repeats.

3.1.4 Modeling

Hundreds of modeling techniques or different combinations of the same model can be created for a single business need. As a result, the modeling process can be repeated until the model that is considered to be the best is found. While preparing the data, it is vital to have a foresight about the modeling technique that will be employed.

The structural observation method will be applied in this work. The model's validity must be checked before it is finished (Zuo, Yang & Chen, 2010).

3.1.5 Evaluation

This is the step in which the model is evaluated against the previously specified success criteria. Apart from technical information, the variables representing the model must have a logical explanation including business information. The values generated by the model, such as profitability, can be confused with analyses for the institution in this comparison. However, the entire process is examined to identify what steps are skipped and what steps need to be repeated, and the next step can be passed if required.

3.1.6 Deployment

Within the organization of the deployment process, the common point of various strategies is the use of the same model output in entire stage. Processing the scores generated at the model output into the database might also be one of the strategies to be improved. In summary, the information obtained in the model output is acquired as a new printing angle:

i. Actual data is combined to create presentations.

ii. It is determined when the established model needs to be updated because changing business rules and strategies can be factors that disrupt the continuity of the model.

iii. After summarizing the findings, it's decided how the model will be integrated into the database, whether systemic development is necessary, and what compelling results will be presented to the top echelon.

iv. The model's score points can be run with SQL or any other source code and then successfully combined into the decision-making system.

v. Sometimes the scores are possible to manually integrated into existing customer management systems with the help of a special interface (IBM,2021).

Figure 4 Deployment

Champion – Competitor: The Champion – Competitor technique allows for the use of score points in addition to existing risk assessment guidelines. At the beginning of the deployment of the newly developed model, it should be determined whether the new model data record risk assessment or modeling methods.

For this purpose, while most of the score calculations are made using existing methods (which called champion), the remaining score calculations are operated with the newly developed model (called competitor). The effects of the various strategies are compared. If the newly constructed model does not cause some many problems with the computed points, it is utilized instead of the old technique, and a new champion is chosen. While it is a competitor, the new model can be accepted in the risk evaluation if it is required. The development process can go on forever until the situation improves (Shao, Zoldi & Cameron, 2007).

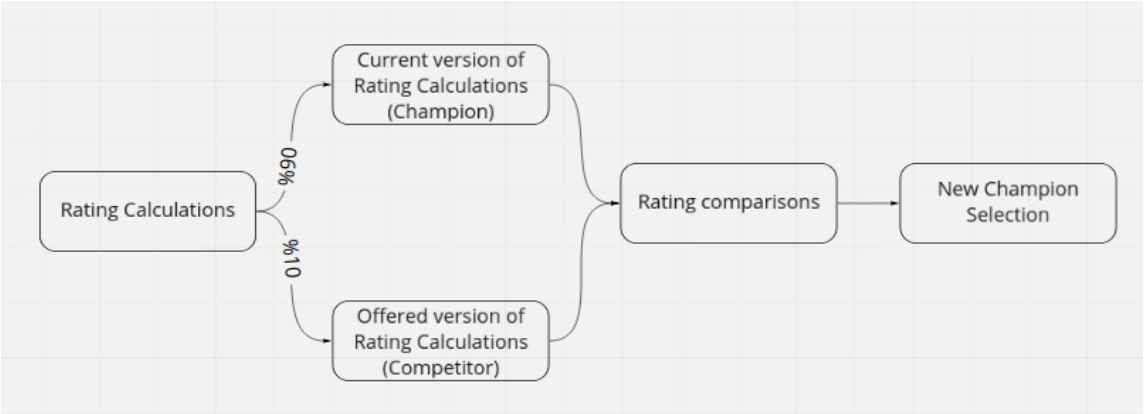

Figure 5 Champion-Competitor Approach

3.1.7 Confusion Matrix

A confusion matrix is a tool used in predictive analytics. Particularly it is a table that presents and compares actual values to the model's projected values. A confusion matrix is a statistic used in machine learning to evaluate how well a machine learning classifier performed on a dataset. The precision, accuracy, specificity, and recall measures are visualized using a confusion matrix (Visa, Ramsay & Ralescu, 2011).

3.2 Knowledge Management

The concept of knowledge management has been used for over three decades. These techniques tend to increase data-driven organizational performance, particularly where systems need to filter data to provide understandable data.

The majority of well-known Knowledge Management (KM) studies have tended to focus on large organizations. For Small or Medium Sized Enterprises (SME), getting the data / information is more important than redirecting or filtering it for purposeful use. SME's could benefit from new inventions in terms of knowledge utilization strategies or research. For huge corporations, data collection is no longer a challenge. Companies tend to focus on reliability on the data and uses the knowledge ability to filter the data (Durst & Edvardsson, 2012).

Knowledge can be constituted inside with collaboration of internal sources data or gathered completely from external source. Committed resources are those that an organization creates within the institution for a specific purpose by dedicating select staff members or an entire department. Individual pieces of data could generate fresh pieces of information to predict decision making in the future (Anand & Singh, 2011).

KM implementation enables an organization to learn from its corporate memory, share knowledge and identify competencies in order to become a forward thinking and learning organization" (Patil & Kant, 2014).

Data knowledge has a direct impact on scoring, and whether or not to accept the data is totally trustful on one's grasp of the necessary information. Data scoring motivates us to think forward and create predictions and observations that will put our team ahead of the competition. Furthermore, the system should be self-contained in order to relieve the developer and technical teams of their responsibilities.

3.3 Qualitative Methods and Structural Observation

There is no straight forwarded, beautiful, and widely accepted definition of qualitative research in qualitative methodologies. On the other hand, qualitative researchers know what it means and what its qualities are, and this knowledge guides them through each stage of their research. Qualitative research aims to understand people from their own point of view (Walczak, 2008).

Structured observation is a data collection strategy in which researchers and participants collaborate on a project (Fidel, 1993).

3.3.1 Interview

Interviewing participants is another way of data collecting that is often utilized in qualitative research. The goal of interviews is to learn about professionals' perspectives, experiences, and attitudes on certain issues.

There are three basic interview methods: organized, semi-structured, and unstructured. In structured method, tool name Hotjar used to analyze customer behavior on the pages. Hotjar allows to make surveys on the pages, users can criticize the pages with left comments on them.

3.3.2 Focus Group

The aim of the focus groups approach is to obtain data from individuals who have been grouped into a group or groups where they may talk at the same time. It's also a type of group interview that people talk about or share their experiences about given topic (Stewart, Patterson, Baldwin & Araujo, 2010).

3.3.3 Participatory Action Research

This approach entails a group of individuals working together to solve an issue, designing a solution, then attempting to monitor the effects, and ultimately adjusting behavior in response to input (O'Brien, 1998). As indicated in Figure 6, participatory behavioral research (PBR) can be characterized as a three-step procedure of looking, thinking, and acting. Researchers are said to be directly active in the project environment, according to the method (Türkeli & Erçek, 2010).

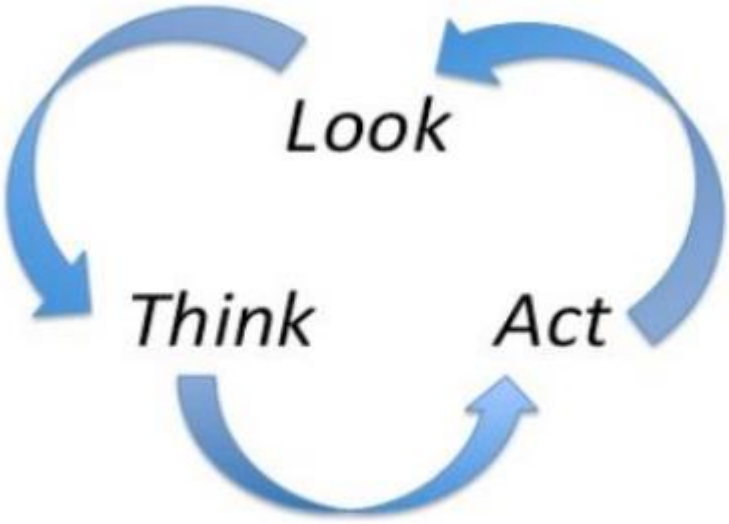

Figure 6 Look, Think, Act Mechanism, adapted from (Türkeli & Erçek, 2010).

Steps of this circulation can be ordered as:

Look: This is the stage which user gather information from the project environment about user's goals and requirements.

Think: After collecting the data, this process involves evaluating prospective actions, hypotheses, and assumptions.

Act: This is the point at which you take steps to improve your ideas and assumptions, make radical decisions. If an issue or new concept occurs after this phase, the cycle continues with the initial step of documenting the changes and events that have happened

CHAPTER 4

ARCHITECTURE

In this section of this work, brief introduction of the technologies which implemented during the whole phase of designing algorithms will be given.

i. Hadoop & HDFS
ii. Docker
iii. Kubernetes
iv. MongoDB
v. Apache Products (Apache Kafka, Apache Druid)
vi. Hotjar

Hadoop and HDFS: Hadoop is a scalable, open source, fault-tolerant Virtual Grid operating system architecture for data storage and processing. It runs on commodity hardware, it uses HDFS (Hadoop Distributed File System) which is fault-tolerant high-bandwidth clustered storage architecture (Shvacko, Kuang, Radia & Chansler, 2010). It's used for process or querying big data and brings expected benefits with it, its distributed architecture its more invulnerable to hardware failures (Borthakur, 2008).

Docker: Docker provides tools to make creating and working with containers as easy as possible. Container kind of lightweight equivalent of a virtual machine. Container level virtualization has done on operating system level. It's used for run different software environments easily and manage software dependencies (Medel, Rana, Bañares & Arronategui, 2016). Docker increases resource usage, allows developer to ship code faster, standardize application operations, seamlessly move code, make projects more scalable.

Kubernetes: In Kubernetes, the fundamental working unit is a pod, which is an abstraction of a group of containers that are tightly tied with certain common resources. Kubernetes provides persistence to the deployment of single containers via this concept. A pod is scheduled to operate on a single computer, with all containers within the pod deployed on the same machine a pod has a local IP address inside the cluster

network; and all containers within the pod share the same port space (Merkel, 2014). It is used for orchestrate containerized applications like network between all containerized applications. Load balancing between containers, local storage access, health management etc.

MongoDB: It's a NoSQL which means document-oriented/schema-less management system. NoSQL is a type of database management system that differs from standard relational databases in that data is not stored using set table structures. Its primary use is to act as a database system for large web-scale applications, where it outperforms standard relational databases (Estrada, 2018).

Apache Kafka: Kafka provides easy to implement message queue mechanism to catch almost real time data by streaming when the data created or manipulated. Publishers of the data from source of the raw data produce/publish the data via topic name created by the origin. Subscribes on the other hand, subscribe the topics to consume data streaming (Boicea, Radulescu & Agapin, 2012).

Apache Druid: "Apache Druid is a real-time analytics database designed for fast slice-and-dice analytics on large data sets. Most often, Druid powers use cases where real-time ingestion, fast query performance, and high uptime are important." (Druid Apache, 2019) It provides performance, real time and batch ingestions, approximation to provide more performance, indexes for quick filtering and time-based storage which almost full supports SQL via Apache client. Main difference between Mongo is Druid keeps records time-based. It ensures bring data faster for queries which filtered by time boundaries.

Hotjar: Hotjar logs and saves the actions and events that users click on and presents them to the developer, which is used to determine the areas that consumers are interested in and frequently follow. By the light of this information set, the developer develops the optimization issues on top of it.

After the business understanding meet ups with the responsible team member. Business understanding of CRISP-DM flow matched with the data understanding part of the system. Collection of raw data of the entire ecosystem of the company lies on the Apache Hadoop already.

To begin with, we should determine which data is raw. If we try to be more precise and equitable in our scoring system. System should not be use affected by business logic of external teams or sources. It could affect our fairness in system. And the first flow of reliable data flow affects our other products during the service. So it's the most crucial part of our system, the base of our data collection part of our system.

i. Determine the raw data reachability
ii. Chosen tables are business free?
iii. Were the tables gives the final truth of the wanted information, or need to be joined some other tables?
iv. Were the data updated daily?

So because of the main part of the system rely on data analysis of Hadoop environment. Question like above were answered and then system built on top of it.

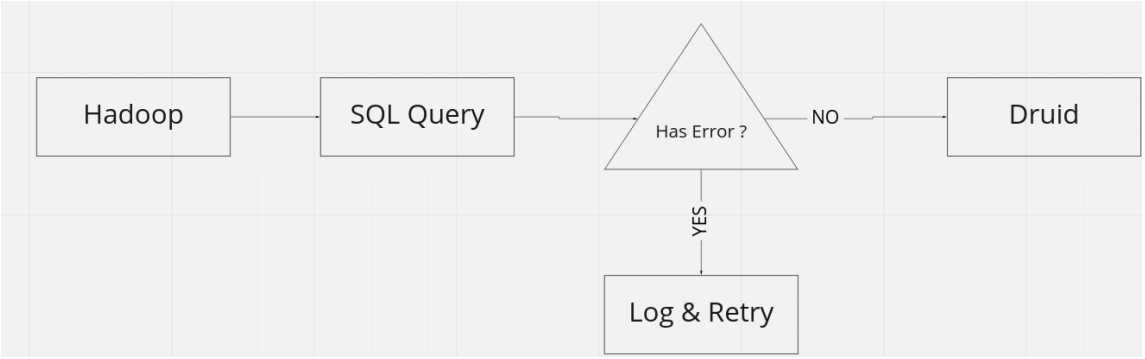

Figure 7 Gathering Data Workflow

Calculation of rating seems to be done in certain intervals. That lead us to use historical databases (Apache Druid) which uses time segments and approximations.

Druid keep the data root value as date (time) type. And use approximations to serve it fast as possible. Some of the merge operations already done in this stage of this project but some of them still need to be merged in runtime due to business necessity. At overall our system work schema should look like Figure 9.

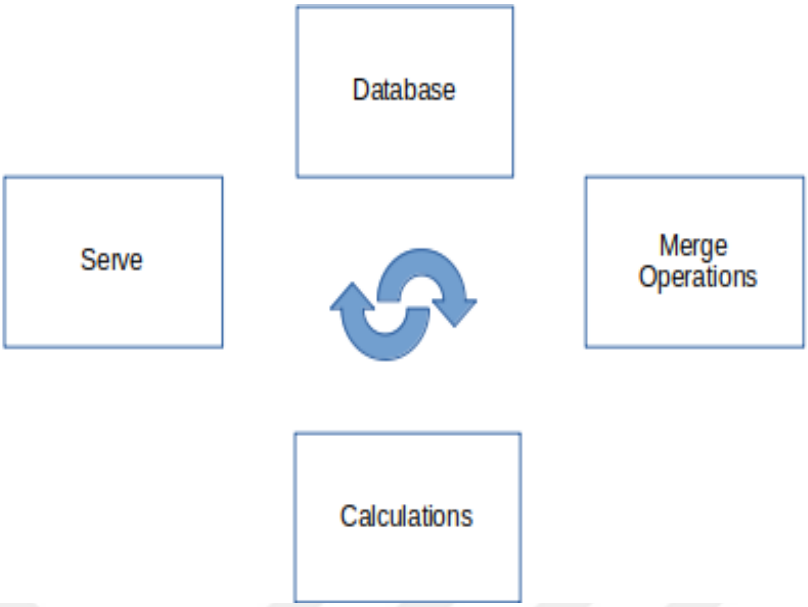

Figure 8 Daily Operation Cycle Model

For the given cycle of figure above database merge operations done. For the runtime expected data subscribing Apache Kafka topics should be done and, in the meantime, if the consumed data needs for later. Consumers has written, to read messages from Kafka message queues then without any business logic written to Mongo Databases. Application of business logic to consumed data, corrupts the reliability of the data.

For the calculation of the rating part, gathered metrics and the constant coefficients already found by the maximization approach for the positive merchant ratings.

To serve the data gathered and present merchants to analytic and ratings, micro service architecture is on followed.

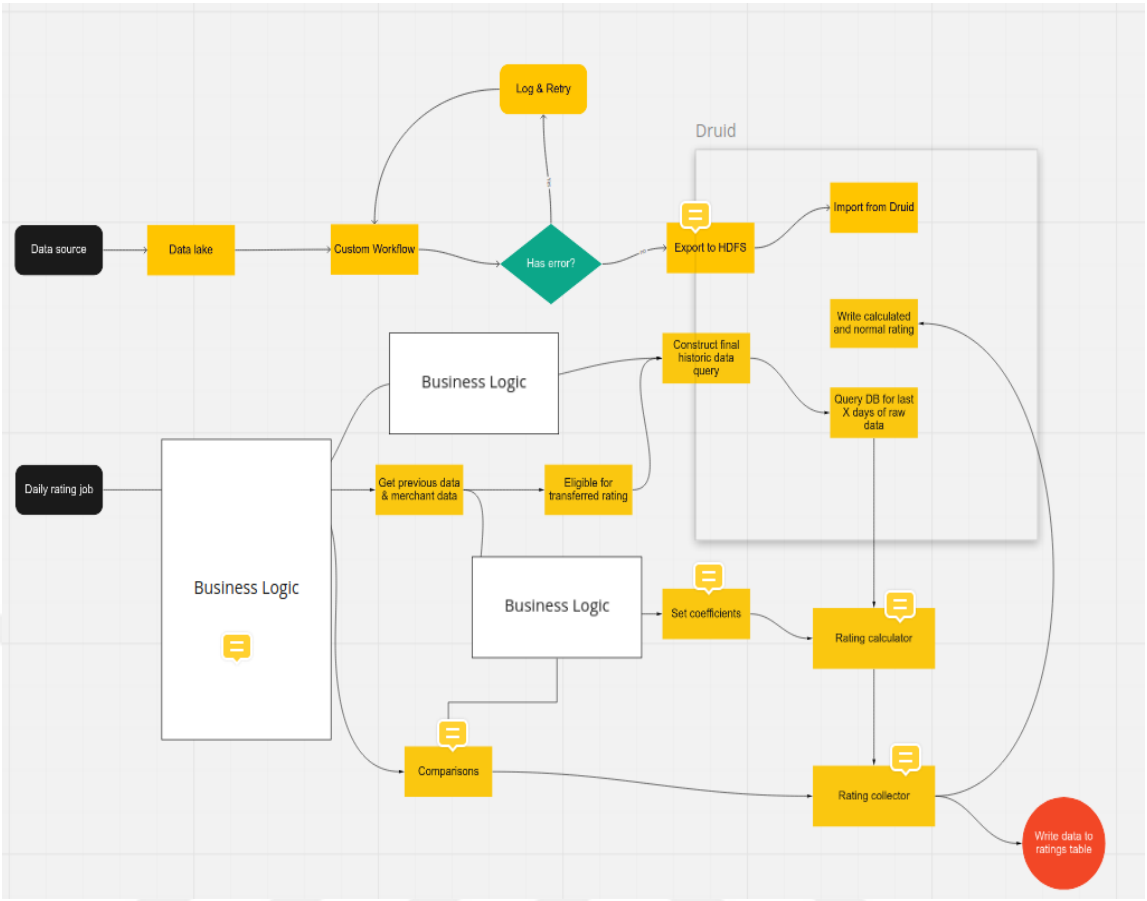

Figure 9 Rating Overall Model

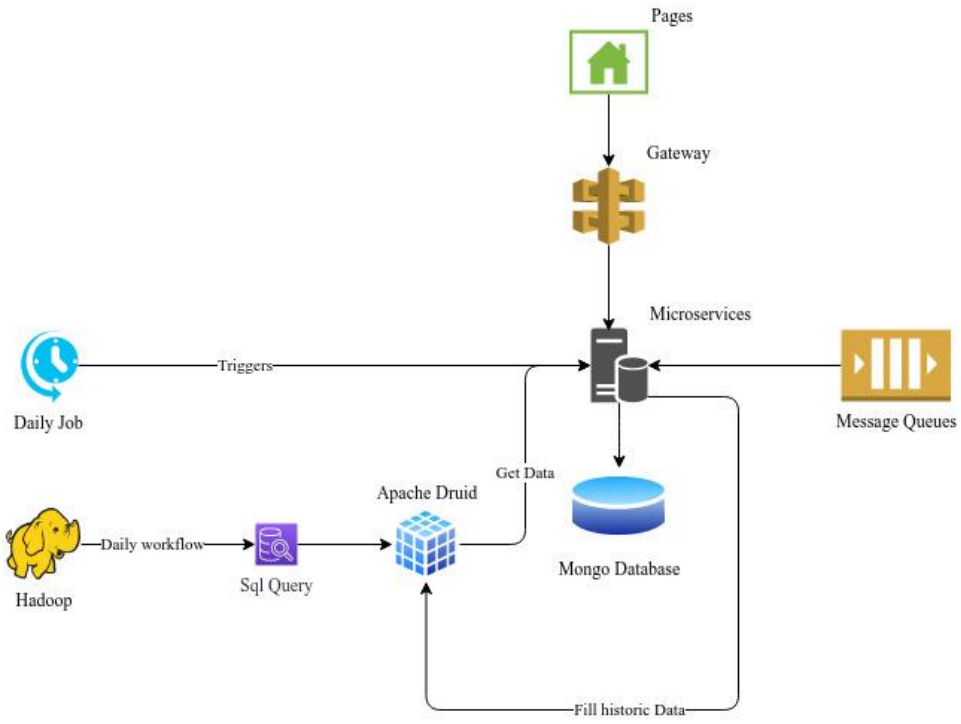

Figure 10 General System Structure

Data gathering is essential for data processing. Challenges of gathering data in cloud computing systems has presented in this study. Binding data from external sources even for just one crucial data could be tiring. But worst thing is constant need of changing business expectations. Some of the business logic cannot be shared at this moment but the overall rating calculation workflow seems like below.

On the Rating score analysis page, user can observe its historical rating changes and some of the other metrics to provoke the merchant work ethics.

Gathered data could be used for giving more information, after the druid database filled and the real time consumers written on Mongo database. API of the database tables which serves the data's combined. Information pages to provide detailed reports to the merchants could be served.

Merchants who observe and improve their quality of work based on their score needs to see their performance on those metrics on time-based periods. Because of the system of the base design provide us time-based data, graphical presentation could be made easily. Merchant could check time-based data which used in rating score and optimize it to get more rating score which will improve their sales (final goal of the business).

Base of the data process successfully completed after this point only the servant of different pages should be created on API level and bind to back-end for front-end layer of the product. Beside the scoring merchant, giving merchants detailed reports based on their sales to provide them complete analysis of their canceled, returned or successful sales reports under different tabs means different grouping under the same tables already gathered in druid and servant layer of Mongo databases.

In the meantime, decided parameters like packaging, delay rate, cancel rate etc. and its coefficients based on maximization of the ratings to not make protestation merchants. The difference between calculated score beforehand (which was not calculated a while) and calculated score now should not be colossal.

CHAPTER 5

FINDINGS

5.1 Product and Reports

From all these methods, technologies, information systems and explained data collection and flow, there are couple of findings and results that sailing many improvements compared to absence of development.

5.1.1 Merchant Rating Report & Analysis

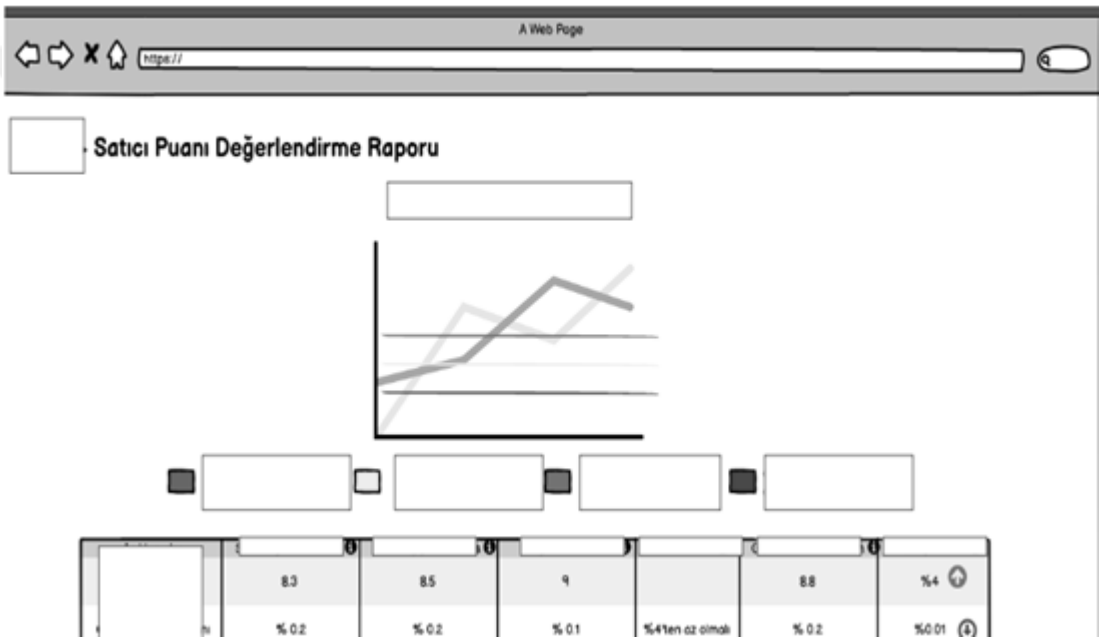

Figure 11 Merchant Rating Analysis Interface

Merchants could not see their ratings and their rating related parameters changes by time. Goal of this page was to show their rating related parameters like feedbacks, delays, cancels and their changes in time. Showing them the rating of a merchant has huge impact of their sales. Rating is one of the important parameters to list of merchants who sells the same item. And increase of rating parameter will increase their place at the marketplace.

Table 1

Rating Overall Table

Time	Rating	Cancel	Delay	Packaging	Feedback
April	100.00%	85.31%	58.37%	99.24%	26.85%
May	97.64%	84.98%	73.11%	100.00%	44.21%
June	97.02%	91.25%	93.05%	88.10%	59.06%
July	96.87%	97.58%	95.43%	88.74%	69.05%
August	96.80%	100.00%	100.00%	90.70%	75.90%
September	97.64%	98.74%	69.95%	93.52%	82.85%
October	98.55%	82.35%	52.04%	94.29%	98.24%
November	98.14%	78.92%	68.21%	94.66%	100.00%

This table expresses the percentage according to max value of every parameter.

Through April to November cancels of merchant's side decrease at overall. It means if merchants who analyze their data tried to decrease their cancels to increase their scores. Delay rate and cancel rates max value month of August, ratings are overall lowest. But merchants cannot affect delays directly because of third party dependencies of cargo companies. Packaging in the meantime didn't change significantly. But through April to November customers comprehends the importance of feedback parameters with their observation. Customers give much more interest to give feedback to merchants through the time and comprehend the importance of merchant ratings. As a result, overall score didn't change that much.

5.1.2 Sales Operations Reports

In this section, provided metrics of the filtered or ordered sales operations to merchants in different tabs. Complete sales report, cancels report, returned items report, campaign reports and so on shown to a merchant who use this platform. Giving them an overall data to observe their sales, returns, cancels etc. Merchant and filter its data by ID, name, brand, or category and able to sort its value. In case of merchant expand the product, variant of the product will be listed below. Sales report will show the merchant its total sales, platform average price and total price columns. In detailed view parameters changed by the report type. In detail of sales report livestock and live price parameters will be shown. For cancel and return pages which show total count

of the canceled or returned product and total loss of these products. In detail of these tabs, merchants could be able to see the reasons of cancels, returns. Merchant could be able to comprehend their position in the marketplace platform. Provide filtered and sorted data could be downloaded by the merchant to keep the data for further analysis on their cold storages.

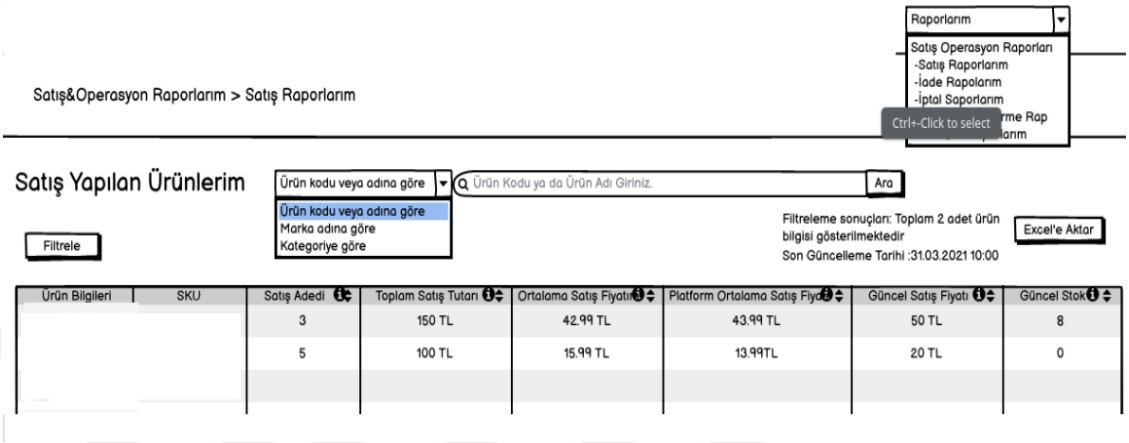

Figure 12 Sales Operations Report Interface

Table 2

Sales Operations Report Table

Time	Sales	Returns	Cancels
July	32.03%	29.00%	29.64%
August	52.13%	48.53%	49.58%
September	88.56%	80.35%	88.18%
October	100.00%	100.00%	100.00%
November	95.61%	89.44%	88.37%

This table expresses the percentage according to max value of every parameter.

Result: To analyze the report pages, given table provide merchant activity on these report pages separately. As we can define from the table 5-2. All of the merchant who are using the platform daily basis interest the report pages significantly. Through benchmark value of July to September. Merchants understands the working logic of the pages and after September there are non-changes on the entrance of report pages. Which means all the active users who use this platform to run their business learned to how to use this page to make their analysis. At the previous analysis of Table 5-1.

Cancels intended to decrease at overall even in the Black Friday era in November. In October all the report pages at their max value, Black Friday era at November, merchants try to analysis their all data to prepare the Black Friday sales. It seems like almost 88% of the merchants use the report pages regularly.

5.1.3 High Potential Sales

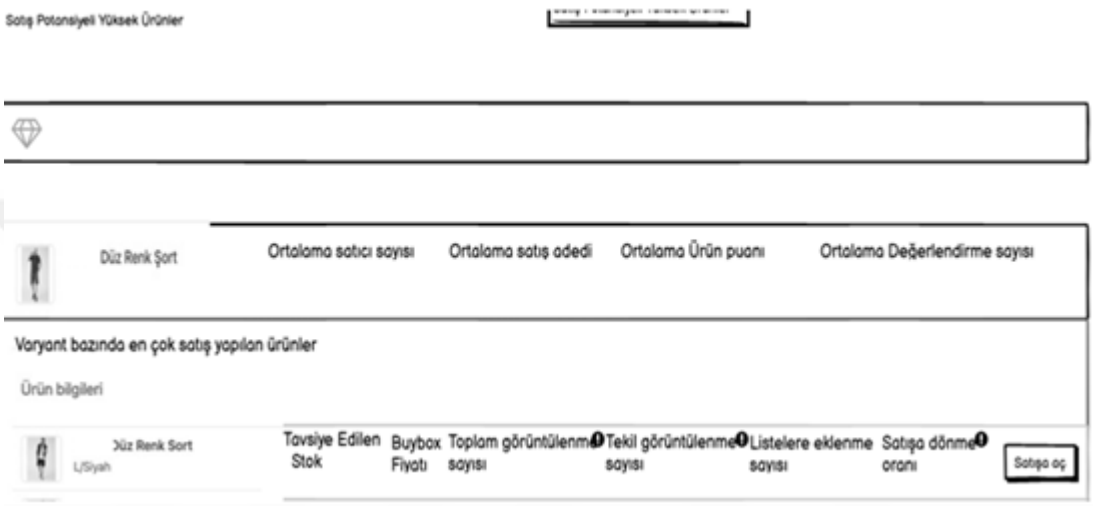

Figure 13 High Potential Sales Interface

Aim of this page is to show merchants a new products or variants of the products like its own products. To do that we observe the categories and brands which sold by merchants. Cross referenced with most shown products via customers. At this page we show the high potential sale product with merchant number whom sales this product as well. Average number of sales, average of product score, and average evaluation number has given. In product based shown, only parameters mentioned above are acquirable. But if level of cross-referencing done via variant. Additional parameters can be shown like proposed stock information, buybox winner price, total number of shown rate, add list count and rate of return sale rate. After the observation made by merchant could redirect itself to the page which can open a product for sale.

Table 3

High Potential Sales Table

Time	High Sales Potential	Open Sales	Sold
September	100.00%	32.00%	9.88%
October	19.88%	4.00%	1.88%
November	59.25%	6.38%	2.50%

This table expresses the percentage according to max value of high sales potential product parameter.

According to table, At September is the most merchant affected by high sales potential feature. Because of the sellable products couldn't change a lot, highest sales potential product shown to a merchant in September. Most open for sales after proposing happen in September again. Merchants who open the product on sale relatively 30% of were sold. In October same domain logic couldn't detect as much as September that's why accepted offer return sale percentage decreased 25%. At November product filters and changes on the business logic increased the chosen product of high sales potential.

5.1.4 Customers Price Alerts

Figure 14 Customers Price Alert Interface

Customers could open notification for certain products. When the product reached its demanded feature, mostly the price. Customers notified about the product

they have been interested. With this information, products could be analyzed by their shown rate and the number of added goods for notification. At this page, customers who demands the notification, remaining stock of the product, live price and the recommended price information for the product shown. There is redirect link for the further analysis to sales report page with the product-based filter.

Table 4

Customers Price Alert Table

Time	Page Enter Count	Page Enter Get Proposed	Propose Accepted	Turn Into Sale
September	90.19%	86.25%	32.13%	16.21%
October	93.20%	74.83%	18.09%	16.13%
November	100.00%	96.88%	36.47%	20.56%

This table expresses the percentage according to max value of high sales potential product parameter.

From the first day of the Price Alert page, merchants interest are quite high. In September 95% of the merchants get proposed, related price alert from the customers. 37% propose acceptation rate from the merchants and related to accepted proposals. 50% of the product sold in September propositions. In October 80% of the merchants get proposed. 24% propose acceptation rate from the merchants and related to accepted proposals. 89% of the product sold in October propositions. In November 96.88% of the merchants get proposed. 37% propose acceptation rate from the merchants and related to accepted proposals. 56.37% of the product sold in November propositions.

5.1.5 Stock Offer

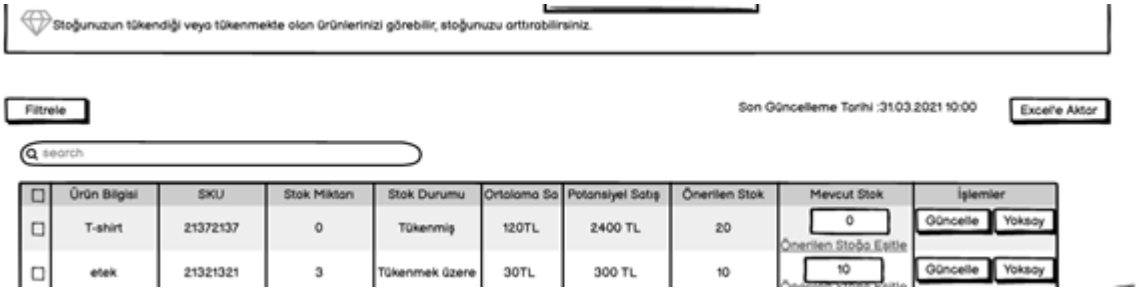

Figure 15 Stock Offer Interface

Aim of this page is giving merchants propose to update their stock on the marketplace based on some analysis. If the product stock out, making a proposal has not much value. But before the limit for the stock out alert could be preventing the potential loss of sales. First determined the critic stock on our behalf, calculated with total sales count parameter. Customers sometimes add the product on the cart but cannot buy them because merchant is out of stock and this data is valuable when offering stock either. After the merge of the data under the critic stock and loss sales data. Total sales count, total day which is indicate the sales rate (sales date difference), average price of the platform, total merchant count and live price are the parameters of stock offer. All the above the calculated by per product. But algorithm behind the parameters is internal information and must remain implicit.

By observing rate of decreasing stock, this system manages to guess the sales potential in a few days and notify the merchants about the stocks. For instance, product ‘X’ has 10 stocks, but it has no sales at last 2 months. Prediction of the sales is 0, so system would not suggest updating the stock of this product. In the meantime, product ‘Y’ has 10000 stock, and last month product has 9300 sales. So, prediction would be for product ‘Y’ for this month need update at the stock by checking the rates of previous sales. That’s why these suggestions calculated live.

Table 5

Stock Offer Table

Time	Page Enter	Page Enter Get Proposed
September	100.00%	29.25%
October	83.28%	33.09%
November	75.73%	33.90%

In September merchants figure out the page as we can get from the page entrance data and most. After the monthly gathering data reached its climax points at September. Page Enter Propose which means shown propositions count who enter the stock recommendation is 30%. Entrance rate is decreasing day by day. In October and November 33% stabilized.

5.1.6 Live Performance

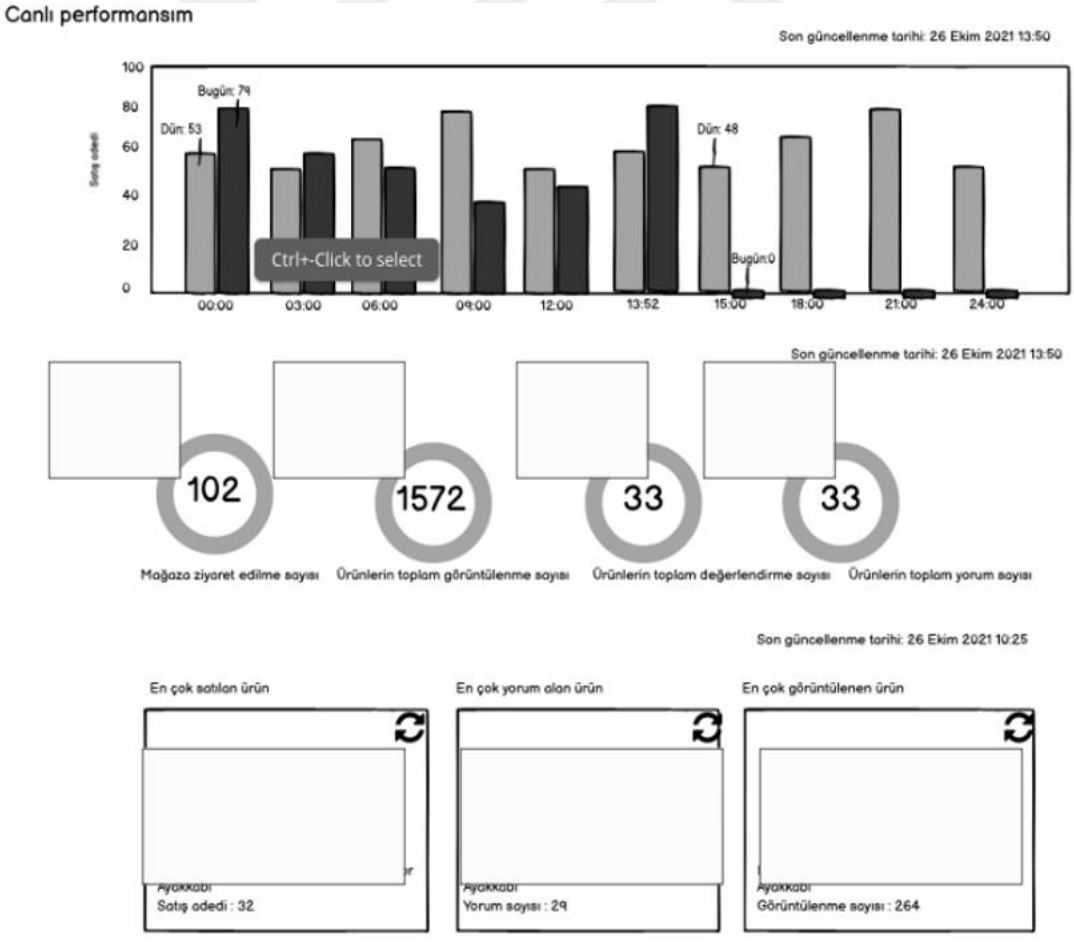

Figure 16 Live Performance Interface

At this page, merchants could observe its live performance at hourly time based graphics. Especially at the campaign period live sales performance has much more importance to analyze if the campaign works or not. In this page edge cases of product like, most commented product, most shown product, most sold product shown to the merchants. Merchants should gather information daily on graphs, pie graphs and so on. Merchant could be able to screen shot the whole screen for presentation purposes.

This is the newest product that's because there is no data on this page.

5.1.7 Dashboard

Figure 17 Complete Rota Interfaces

On this summary page it shows the merchants' summary of all report pages could be explicitly reached. It is not detailed as specific report pages, yet it is possible to give the merchants overall perspective. Merchants may wish to see their sales, delays, and its cancels on weekly basis. Merchants could redirect themselves to specific pages which has the information in detail shown in charts. All the pages mentioned above are gathered in one place a holistic way.

5.2 Analysis and Observation

From all these products and reports it is possible to make some inferences by comparing expected and actual results. This process can be made iteratively with following.

First, to calculate actual and predicted data results sample size should be defined. To define it Sample Calculation Formula (Naing, Winn & Rusli, 2006).

$$n = \frac{z^2 pq}{e^2}$$

n: $sample\ size, z$: $statistic\ for\ a\ level\ of\ confidence, p$: $expected\ prevalance\ or\ proportion$

Sample size is identified for:

$$error\ rate, e: \%5$$

$$expected\ proportion, p: 0.5$$

$$level\ of\ confidence, z: 1.96$$

$$n = \frac{1.96^2(0.5)(1-0.5)}{0.05^2} \cong 384$$

To acquire accuracy, precision, miss rate and some other informative and beneficial metrics, should be prepared a confusion matrix.

Confusion Matrix: From dataset which obtained by created reports, confusion matrix can be built.

Sample size obtained with minimum value as 384 for %0.05 error rate, yet it is picked a greater value which is 600.

Determined definition of the confusion matrix is the merchants who are accepted, or High Sales Potential products offer, or Price alert offers (Stock offers could affect the sales but it contains lots of layers or assumptions to accept the offer as success).

Table 6

Confusion matrix

$n = 600$	$actual_{positive}$	$actual_{negative}$	Total
$predicted_{positive}$	TP = 101	FP = 85	186
$predicted_{negative}$	FN = 194	TN = 220	414
Total	295	305	

$TN: True\ Negative, FP: False\ Positive, FN: False\ Negative, TP: True\ Positive$

i. True Positive: If merchant accept any of the offer mentioned above then increase it sales after the accept date means our prediction for our product is true and the merchant sales increased.
ii. False Positive: Merchant accept any of the offer mentioned above, with the prediction of increasing sales. But eventually merchant sales rate didn't change or decreased.
iii. False Negative: If merchant didn't accept our offers' expected result is their sales rate remains the same or decreased but these merchants sales rate increased even if they didn't accept the offers.
iv. True Negative: These merchants didn't accept our offers and their sales rate decreased or remains the same.

Output: Scope is to see benefits of prepared system upon merchants, if merchant uses system, it should increase numbers for sales. On the contrary, if merchant does not use system, it should see that there is no influence, or any impact upon sales.

Observations: In the confusion matrix step, which will be used for the metric calculation of the system, 186 of 600 sellers use the system, while 414 do not. As can be seen, the reason for the high number of non-using merchants is that the systems are relatively new, data range is small, and the reputation and reliability are not fully established for produced systems. The 414 merchants who did not use the system were not expected to see an increase in sales It was observed that while 220 of the sales did not accelerate their sales as expected and observed that the sales of 194 sellers increased. One of the main reasons for this was that the analysis process coincided with the Black Friday campaign in November, when the system was just starting to settle in.

On the one hand, the fact that people thought that the exchange rates would increase in the future, caused them to buy the product which they would probably buy later.

On the other hand, it was discovered that 54.3 percent of the merchants who agreed and began utilizing the system received the benefits of the system.

The resulting metrics are as follows:

Accuracy (ACC): Accuracy is perfect indicator that represents of well a model performs overall. To experiment accuracy rate for all the cases (positive and negative):

$$accuracy: \frac{(TP + TN)}{(TP + TN + FP + FN)} = 0.535 = \%53.5$$

Precision (PPV): Precision measures the percentage of cases that are positive.

$$precision: \frac{TP}{(TP + FP)} = 0.543 = \%54.3$$

F1 Score (F1): The F1 Score was intended to have a balanced metric between recall and accuracy, taking into account the opposing features of precision and recall. The average may not be a reliable representation of model performance if the

distribution of positive and negative cases is too uneven. In this case, F1 score is a great picture to see how the model performs.

$$F1\ Score = 2 * \frac{TP}{(2TP + FP + FN)} = 0.42$$

Sensitivity (TPR): The number of correct positive predictions divided by the total number of positives is how sensitivity (SN) is computed. It is also known as the recall rate (REC) or the true positive rate (TPR).

$$sensitivity: \frac{TP}{(TP + FN)} = 0.3424 = \%34.2$$

34.2% sensitivity means our modal is not good enough, our prediction and actual true positives which indicates the merchants who accept our offer, sales will be increased. But in our case problem is not the merchants who accept the offer and increased their sales, problem that faced here is the economical inconsistency and worldwide campaign data (known as black Friday) increased the False negative which stands for merchant who doesn't use the system increased their sales.

Specifity (SPC): The number of valid negative predictions divided by the total number of negatives is used to compute specificity (SP).

$$specifity: \frac{TN}{(TN + FP)} = 0.7213 = \%72.13$$

72% specificity means our modal is quite good at prediction of true negatives which indicates our merchants, who doesn't accept the offer, sales are remains the same.

CHAPTER 6

DISCUSSIONS AND CONCLUSIONS

Nowadays, many various optimization methods have developed for the benefit of merchants and many products, analysis and reporting systems are emerging thanks to these methods. The reason of developing optimization algorithms can be explain with this way. From buyer perspective of view, buyers always rely first on other customers' feedback and discourses. After, merchant scores and tags approved by the marketplace platform bring credibility to the merchant. By the light of this way, merchant sales will increase significantly. With this road, marketplace platforms calculate merchant points to protect, support and inform customers who will buy products to supply demands which are demonstrated.

A qualitative observation and inferences are made by extracting, analyzing actual data collections which contain information about the whole sales processes. Also pricing, shipping, pre and post support customer feedbacks issues researched, studied and discussed.

In this thesis, the production and compilation of data and its accuracy and reliability; the arrival process of the acquired accurate data to the developer; the working discipline of developer group in the optimization algorithm which will be designed process; the details of the tools and technologies which will be used throughout the merchant rating optimization derivation; the overall system architecture which generated rating algorithm; and other products that produced by represented whole system model is explained. After all these analytics, prediction - actual data balance it is acquired sensitivity, accuracy and other values for data result will be TPR=0.342, SPC=0.721, PPV=0.543, F1= 0.42, ACC= 0.535.

In the future, developers, researchers who will be set sail to various methods parallel to the algorithms explained in detail in the thesis may contribute their development by the shadow of the system model and architecture, and by making use of the tools, packages, libraries, and disciplines which are specified during the study.

REFERENCES

Agarwal, A., Dahleh, M., & Sarkar, T. (2019, June). A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation* (pp. 701-726).

Ahsan, S., & Shah, A. (2006). Data, information, knowledge, wisdom: A doubly linked chain. In *the proceedings of the 2006 international conference on information knowledge engineering* (pp. 270-278).

Allen, G. D. (2004). Hierarchy of Knowledge: From Data to Wisdom. *International Journal of Current Research in Multidisciplinary (IJCRM)*, *2*(1), 15-23.

Andrew Begel, Nachiappan Nagappan, "Usage and Perceptions of Agile Software Development in an Industrial Context:An Exploratory Study", First International symposium on empirical software engineering and measurement, pp. 255-264, 2007.

Anfan Zuo, Jing Yang, Xiaowen Chen, "Research of Agile Software Development Based on Formal Methods", International Conference on Multimedia Information Networking and Security, pp. 762-766, 2010.

Apurva Anand, M.D.Singh, (2011), "Understanding Knowledge Management: literature review "

Aven, T. (2013). A conceptual framework for linking risk and the elements of the data–information–knowledge–wisdom (DIKW) hierarchy. *Reliability Engineering & System Safety*, *111*, 30-36.

Boicea, A., Radulescu, F., & Agapin, L. I. (2012, September). MongoDB vs Oracle--database comparison. In *2012 third international conference on emerging intelligent data and web technologies* (pp. 330-335). IEEE.

Borthakur, D. (2008). HDFS architecture guide. *Hadoop apache project*, *53*(1-13), 2.

Davis, A. M., Bersoff, E. H., & Comer, E. R. (1988). A strategy for comparing alternative software development life cycle models. *IEEE Transactions on software Engineering*, *14*(10), 1453-1461.

Estrada, R. (2018). *Apache Kafka Quick Start Guide: Leverage Apache Kafka 2.0 to simplify real-time data processing for distributed applications*. Packt Publishing Ltd.

Fidel, R. (1993). Qualitative methods in information retrieval research. *Library and information science research, 15*, 219-219.

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. *Procedia Cirp, 79*, 403-408.

IBM, 2021. IBM SPSS Modeler CRISP-DM Guide s.l.:IBM Corporation.

Kleindl, B. (2000). Competitive dynamics and new business models for SMEs in the virtual marketplace. *Journal of Developmental Entrepreneurship, 5*(1), 73.

Kropsu-Vehkapera, H., Haapasalo, H., Harkonen, J., & Silvola, R. (2009). Product data management practices in high-tech companies. *Industrial Management & Data Systems*.

literature review", Journal of Knowledge Management, Vol. 16 Issue: 6, pp.879-903

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., ... & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*.

Medel, V., Rana, O., Bañares, J. Á., & Arronategui, U. (2016, December). Modelling performance & resource management in kubernetes. In *Proceedings of the 9th International Conference on Utility and Cloud Computing* (pp. 257-262).

Mehrotra, R., & Carterette, B. (2019, September). Recommendations in a marketplace. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 580-581).

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux journal, 2014*(239), 2.

Morrisey, M. A., Rivlin, A. M., Nathan, R. P., & Hall, M. A. (2017). Five-state study of ACA marketplace competition: A summary report. *Risk Management and Insurance Review*, *20*(2), 153-172.

Naing, L., Winn, T. B. N. R., & Rusli, B. N. (2006). Practical issues in calculating the sample size for prevalence studies. *Archives of orofacial Sciences, 1*, 9-14.

O'Brien, R. (1998). An overview of the methodological approach of action research.

Peter Maher, "Weaving Agile Software Development Techniques into a Traditional Computer Science Curriculum", Proc. of 6th IEEE International Conference on Information Technology: New Generation, pp. 1687-1688, 2009.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. Journal of information science, 33(2), 163-180.

Sachin K. Patil, R. Kant, (2014), "Methodological literature review of knowledge management research"

Schäfer, F., Zeiselmair, C., Becker, J., & Otten, H. (2018, November). Synthesizing CRISP-DM and quality management: a data mining approach for production processes. In *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)* (pp. 190-195). IEEE.

Shao, M., Zoldi, S., Cameron, G., Martin, R., Drossu, R., Zhang, J. G., & Shoham, D. (2007). *U.S. Patent No. 7,191,150*. Washington, DC: U.S. Patent and Trademark Office.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010.

Stewart, M. A., Patterson, L., Baldwin, S., Araujo, J., & Shearer, R. (2010). Look, think, act: Using critical action research to sustain reform in complex teaching/learning ecologies.

Susanne Durst, Ingi Runar Edvardsson, (2012), "Knowledge management in SMEs:

Türkeli, S., & Erçek, M. (2010). Designing a capability-focused strategic management model for a turkish public hospital: learning from failure. *Systemic Practice and Action Research*, *23*(5), 353-370.

Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002, November). Conceptual modeling for ETL processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP* (pp. 14-21).

Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *MAICS*, *710*, 120-127.

Walczak, S. (2008). Knowledge management and organizational learning: An international research perspective. *The Learning Organization*.

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), 58-65.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1). London, UK: Springer-Verlag.