TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Machine learning of the global consumer market

УДК 004.853:339.9:330.567.2

Студент

Группа	ФИО	Подпись	Дата
8ПМ9И	Солиев Искандар Бегалиевич		21.06.2021 г.

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф-м.н.		21.06.2021 г.

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Гончарова Н. А.	к.э.н.		22.02.2021 г.

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.б.н.		19.02.2021 г.

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А.О.	к.т.н.		21.06.2021 г.

Томск – 2021

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП

по направлению 09.04.04 «Программная инженерия»

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований
ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности
ОПК(У)-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию вариантов архитектуры программного средства
ПК(У)-2	Способен разрабатывать и администрировать системы управления

	базам данных
ПК(У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК(У)-4	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ПК(У)-5	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Савельев А.О.
(подпись) (дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	**ФИО**
8ПМ9И	Солиев Искандар Бегалиевич

Тема работы:

Machine learning of the global consumer market	
Утверждена приказом директора (дата, номер)	№ 40-5/с от 09.02.2021 г.

Срок сдачи студентом выполненной работы:	15.06.2021 г.

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Предметом исследования является алгоритм машинного обучения и исследовательский анализ данных для определение наиболее выгодного способа продажи товаров и проводить процедур прогнозирования для общей прибыли глобальный потребительский рынка.

Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1. Обзор литературных источников (прогнозирование, статистический анализ, очистка и подготовка данных, машинное обучение). 2. Интерпретация исходных данных. 3. Исследовательский анализ данных мирового рынка для определения выгодного способа продажи продукции. 4. Реализация алгоритма машинного обучения для прогнозирования исторических данных глобальный рынка. 5. Результаты проведенного исследования. 6. Финансовый менеджмент, устойчивость ресурсоэффективности и ресурсосбережение. 7. Социальная ответственность.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. UML-диаграммы. 2. Скриншоты веб-форм.

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Основная часть	Доцент ОИТ ИШИТР, к.ф-м.н., доцент Губин Е. И.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Доцент ОСГН ШБИП, к.э.н., доцент Гончарова Н. А.
Социальная ответственность	Доцент ООД ШБИП, к.б.н., доцент Антоневич О. А.

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	1.03.2021 г.

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф-м.н.		1.03.2021 г.

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ9И	Солиев Искандар Бегалиевич		1.03.2021 г.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2020 /2021 учебного года

Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	15.06.2021

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
01.06.2021	Теоретическая часть	20
01.06.2021	Использованные технологии, их эффективность и возможности на глобального рынке	20
01.06.2021	Подготовка данных и алгоритм машинного обучения для анализа глобального рынка	20
01.06.2021	Результаты проведенных исследований	20
01.06.2021	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
01.06.2021	Социальная ответственность	10

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф-м.н.		

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А. О.	к.т.н.		

TASK FOR SECTION
«FINANCIAL MANAGEMENT, RESOURCE EFFICIENCY AND RESOURCE SAVING»

To the student:

Group	Full name
8PM9I	Soliev Iskandar Begalievich

School	Information Technology and Robotics	**Division**	Information Technology
Degree	Master	**Educational Program**	09.04.04 Software Engineering

Input data to the section «Financial management, resource efficiency and resource saving»:	
1. Resource cost of scientific and technical research (STR): material and technical, energetic, financial and human	– Salary costs – 198000; – STR budget – 13945,2;
2. Expenditure rates and expenditure standards for resources	– Electricity costs – RUB 5,8 per kWh;
3. Current tax system, tax rates, charges rates, discounting rates and interest rates	– Labor tax – 27,1%; – Overhead costs – 30%;
The list of subjects to study, design and develop:	
1. Assessment of commercial and innovative potential of STR	– comparative analysis with other researches in this field;
2. Development of charter for scientific-research project	– SWOT-analysis;
3. Scheduling of STR management process: structure and timeline, budget, risk management	– calculation of working hours for project; – creation of the time schedule of the project; – calculation of scientific and technical research budget;
4. Resource efficiency	– integral indicator of resource efficiency for the developed project.
A list of graphic material *(with list of mandatory blueprints):*	
1. Competitiveness analysis *2. SWOT- analysis* *3. Gantt chart and budget of scientific research* *4. Assessment of resource, financial and economic efficiency of STR* *5. Potential risks*	

Date of issue of the task for the section according to the schedule	

Task issued by adviser:

Position	Full name	Scientific degree, rank	Signature	Date
Associate professor	N.A. Goncharova	PhD		22.02.2021

The task was accepted by the student:

Group	Full name	Signature	Date
8PM9I	Soliev Iskandar Begalievich		22.02.2021

TASK FOR SECTION
«SOCIAL RESPONSIBILITY»

To student:

Group	Full name		
8PM9I	Soliev Iskandar Begalievich		
School	Information Technology and Robotics	**Department**	Information Technology
Degree	Master programmer	**Specialization**	09.04.04 Software Engineering

Title of graduation thesis:

Marketing analysis of the global market using machine learning	
Initial data for section «Social Responsibility»:	
1. Information about object of investigation (matter, material, device, algorithm, procedure, workplace) and areaof its application	– Base exploratory data analysis to choosing the most profitable way to sell products – Machine learning algorithm to performing forecasting total profit of the global market – Working area: desktop and personal computer
List of items to be investigated and to be developed:	
1. Legal and organizational issues to provide safety: – Special (specific for operation of objects of investigation, designed workplace) legal rules of labor legislation; – Organizational activities for layout of workplace.	– GOST 12.2.032-78 SSBT. Workplace when performing work while sitting. General ergonomic requirements. – SP 2.4.3648-20. Sanitary and Epidemiological Requirements for Organizations of Education and Training, Recreation and Recreation of Children and Youth
2. Work Safety: 2.1. Analysis of identified harmful and dangerous factors 2.2. Justification of measures to reduce probability of harmful and dangerous factors	Dangerous and harmful factors: – Increased levels of electromagnetic radiation; – Insufficient illumination of workplace – Excessive noise – Increased / decreased air humidity in the workplace; – Electric shock – Ionizing radiation
3. Ecological safety:	– Insufficiency of atmosphere and hydrosphere – Lithosphere: when disposing of fluorescent lamps and office equipment
4. Safety in emergency situations:	– Fire safety
Assignment date for section according to schedule	

The task was issued by consultant:

Position	Full name	Scientific degree, rank	Signature	date
docent professor	Antonevich O.A	PhD		19.02.2021

The task was accepted by student

Group	Full name	Signature	date
8PM9I	Soliev Iskandar Begalievich		19.02.2021

ABSTRACT

The final qualifying work contains 87 pages, 29 figures, 5 tables, and 28 sources.

The goal of work is to determine a profitable way to sell products and performing forecasting total profit of global market using technologies of machine learning algorithm.

As a result of the work, autoregressive additive models were applied to predict the total profit of a marketing field with a significant historical data. Forecasting was carried out for the entire field using an autoregressive additive model, as well as for individual groups of wells, using machine learning algorithms. The forecast results were compared with the traditional method - the analysis of flow rate decline curves. It was determined that statistical methods are capable of significantly better predicting field production over a short period of time than traditional ones.

The area of application of the technique is global market that fields at various stages of development to sell order and profit at product.

The economic efficiency / significance of the work consists in a more accurate profit calculation associated with well-executed process of forecasting.

Definitions, Designations, Abbreviations

Machine Learning (ML). Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Artificial Intelligence (AI). Artificial intelligence is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals, which involves consciousness and emotionality. The distinction between the former and the latter categories is often revealed by the acronym chosen.

Exploratory Data Analysis (EDA). In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods

Forecasting. Forecasting is the process of making predictions based on past and present data and most commonly by analysis of trends. A commonplace example might be estimation of some variable of interest at some specified future date. Prediction is a similar, but more general term.

Time Series. In mathematics, a time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data.

Global Market. Global marketing is defined as marketing on a worldwide scale reconciling or taking global operational differences, similarities and opportunities in order to reach global objectives.

Prophet. The Prophet library is an open-source library designed for making forecasts for univariate time series datasets. It is easy to use and designed to automatically find a good set of hyperparameters for the model in an effort to make skillful forecasts for data with trends and seasonal structure by default.

Table of Contents

Introduction

Relevance of work. Exploratory data analysis and forecasting is an important part of the decision-making process. This task infuses: data cleaning and preparation, statistical analysis, data visualization, forecasting. In this work, an autoregressive additive model was created to forecast the historical data of the global market.

Purpose of work. The objectives of this work are: determine a profitable way to sell products and performing forecasting total profit of global market using technologies of machine learning algorithm.

Problem of statement. Forecasting is the process of making some kind of judgment about the future, using available information, including historical data, as well as considering possible future events that may affect the forecast event. In many areas, forecasting plays a significant role, being an important part of the decision-making process. In marketing, forecasting the total profit of the global market is one of the most important tasks, it allows you to evaluate the way products are sold in global marketing. Forecasting in the marketing sphere is not an easy task for a number of reasons: a huge amount of scattered information comes in every day, which requires a significant amount of time to prepare and interpret it; almost all measurements are highly dependent on the world stock exchange. Therefore, it is important to evaluate the quality of the incoming information and properly prepare it before using it for forecasting. Forecasts can be made for long and short time intervals. Long-term forecasts are usually carried out using numerical models that are time-consuming and expensive. Such forecasts are very important throughout the "life" of the field, they allow making strategic decisions regarding development, reserves assessment, and surface development.

Subject of study. The subject of the research is carry out exploratory data analysis and methods machine learning used in forecasting the total profit of the global market.

Chapter 1. Theoretical part

1.1 Global Marketing and impacts nowadays

In nowadays-global marketplace, corporations and their leaders must master certain areas before entering and competing in new (foreign) markets. Martin and Chney point out those organizations (including their leaders and management) need to build environmental competencies in order to fulfill the needs and desires of global marketing, creating analytical competence to be able to assess global marketing prospects, strategic competence. To be able to develop global marketing policies and approaches.

Countries around the world are connected not only geographically, but also through various multidimensional networks that are made possible by new technological developments, for example, economic relations, social, cultural and political ties. Therefore, before entering any new market or entering the global market, one must understand the global marketing environment, which is the place to enter and offer products and services.

Even if it seems to be an advantage for companies entering new markets and through this movement, creating new connections, penetration and participation in those markets will become more important and complex. Based on new ideas and knowledge pertaining to new markets that companies obtain, they will find themselves richer (in the knowledge of customer needs) but more vulnerable to foreign competition. As they (companies entering new markets) face those threats by competition and working environment; hence, becoming more vulnerable, will move the concerns related to global trade and finance into the political arena.

According to Bearden et al., organizations and their leaders should comprehend the nature of the marketing environment and why it is important to marketers [13], especially when entering new markets, or in other words, when becoming global. Companies and organizations should be able to describe the major components of the environment within the market planning to enter, how trends in the new

environment affect marketing and how they are related to the organization's products. Not only selling means that organizations will succeed in new markets, but also foremost one should see how the political and/or legal environment within those new markets can be turned into opportunities and what possible threats are.

There is no exact procedure when entering new markets. It is more than a guide. The right solution is that organizations should study how to enter a specific market as each market has its own characteristics and differs from other markets, in particular comparing with firms domestic one. Market sizes, buyer behavior and marketing practices all very important and curtail knowing them, meaning that companies entering new global marketers must carefully evaluate all market segments in which they expect to compete.

Trying to adapt to the new reality, companies have become more open to new ideas and businesses. The main drivers of standardization in marketing have been the globalization of business and the appearance of multinational companies. There were many drivers, which have shown that globalization is possible and through which it was made easy for companies, such as advances in technology, transportation and communication. Hence, those companies, which have adopted the standardization of products, harvested the benefits of those markets on which they have expanded their presence, by being able to produce cheaper and with same or higher quality.

This is nothing new and it is not from today. The debate about effective global marketing management is present since the technology has advanced to that extent, where big companies have snuffed new investment possibilities, with better income possibilities. There were different arguments from 70s until the end of the last century, standardization or adaptation, globalization or localization, and at the end, the argument was global integration or local awareness. The essential question was whether the global homogenization of the buyer's demands allowed global standardization of the marketing mix. Different companies are responding

differently to this issue. Some of them are adopting standardization strategy, whereas others choose for localization strategy in global markets.

As the companies are getting bigger and more capable to extend their production capability, many of them see the chance to grow by entering into global markets. Without any doubt, global market research ha becomes an important part of strategic planning. Depending on the product variety and depending on market conditions and requirements, companies which intend to expand to those companies, all strategies needs to be considered carefully. There are many indicators, which can affect the success of the company within those markets. One may think that it is nothing special when expending to global markets. Is it? Companies are prepared for any market. Are they? It is not so easy, and many companies are yet not ready to make that move. Global marketing research is a very costly process and the results are not, in all cases, 100% reliable.

When preparing to enter an global market, companies must take into consideration many issues, which play a crucial role in those markets, such as: the political situation, the cultural pattern in that society and how do those affect the consumption, distribution networks, business culture, Internal business regulations and policies, political stability, and many others. There are many countries, which intend to protect their own production by eliminating foreign competition through different measures, but the better way would be to offer the opportunity to those companies to continuously improve their products in their home markets, and at the same time to expand into other global markets.

In today's economy, the global company's position in the global market will affect also their position within their own local markets. Global companies are active and operate in different world countries. Their entire strategies are based on global research and not only on their domestic market. R&D, production, marketing, logistic, and financial issues are not available only to their domestic competitors. Those companies plan, coordinate their activities and operate on a global basis. Ford

is one of the most knows car manufacturer in the USA. On the other site, if we consider where those cars (trucks) are produced and assembled, then one may say that Ford is a global brand. For instance, Ford's world truck has an EU made cab, a North American chassis, and all that will be assembled in Latin America (Brazil) and then will be sold in the USA.

As first, companies must decide whether to go abroad and to enter new markets. It is the very first and most important issue to be solved. Entering other markets, means also learning other languages, cultures, laws, and many other details related to that market (or country). Many companies, yet, prefer to remain on their domestic markets. According to a study, 70% of 8000 polled by the British Chambers of Commerce (BCC) offers their products on the UK market. Mainly, all companies said that their products and services were unsuitable for consumption overseas and they see as a risky movement to try to improve their products and services to adjust to requirements from other countries. On the other side, a fifth of those companies interviewed, were quite happy living off their existing sales and they see every additional change as a risky move, which they think it is not necessary to do.

If the potential is there for companies to move forward and go abroad, the next step would be to decide which market to enter. Most of the companies seek for those markets (countries) which are similar, or the differences are not so big comparing to their domestic market. Companies need to define their strategies and policies, and then seeking for the best opportunity to enter foreign markets. Many companies start small and then they try to grow gradually, as they see their chance for improvement. In meanwhile, the company must prepare the strategy to enter those markets. To be more precise in strategy preparation, they (company management) must decide on how many markets they are able to enter. They must prepare the plan where all details about those markets (countries) are listed and what the company must improve to be able to enter and to remain there.

The way the companies enter a new global market is crucial for surviving in those markets. When deciding to enter solely or through a local partner company, then the company must see how high are the chances to succeed when deciding the way they are interested to do. The risk is there, solo or through a local partner. However, determinate to select one of these options is that which of these methods offers more security on returning the investment. There are companies that cannot enter as solely new global markets, therefore they select a local partner, who is already in the market, has his own logistic and network, and is familiar with consumer habits and needs, culture, law and policies, etc. There are many institutions, which helps companies to export/import their products and services to global markets.

Based on the facts gathered from investigation and researches of global markets, based on the indicators above, the company can decide if, how, when, how many, markets to enter.

1.2 Machine Learning approach to marketing

Why should ML be applied to marketing? There are many possible answers to this question rooted both in academic and applied practices of the discipline. For practitioners, for example, ML is disrupting many industries with new business models, products, and services. In academia, the impact appears to equally substantial. For example, the lack of generalization of scientific discoveries is at the center of the so-called “replication crisis,” which has affected many of the life and social sciences, including the fields of management and marketing. This crisis has occurred because researchers have found that many of the most important scientific studies are difficult or impossible to replicate. As this monograph will discuss, the fundamental goal of machine learning is to generalize beyond the examples provided by training data, looking for generalizability. Thus, one of the potential contributions of ML to marketing (and to management in general) lies in its robustness for the generation, testing, and generalization of scientific discoveries. With these different academic and practical perspectives in mind, the goal is to provide marketing with an overview of ML and to analyze required learning, and future developments involved in applying ML to marketing.

Machine learning (ML) refers to the study of methods or algorithms designed to learn the underlying patterns in the data and make predictions based on these patterns [14]. ML tools were initially developed in the computer science literature and have recently made significant headway into business applications. A key characteristic of ML techniques is their ability to produce accurate out-of-sample predictions[14].

Academic research in marketing has traditionally focused on causal inference. The focus on causation stems from the need to make counterfactual predictions. For example, will increasing advertising expenditure increase demand? Answering this question requires an unbiased estimate of advertising impact on demand. However, the need to make accurate predictions is also important to marketing practices. For example, which consumers to target, which product configuration a consumer is

most likely to choose, which version of a banner advertisement will generate more clicks, and what the market shares and actions of competitors are likely to be. All of these are prediction problems. These problems do not require causation; rather, they require models with high out-of-sample predictive accuracy. ML tools can address these types of problems.

ML methods differ from econometric methods in both their focus and the properties they provide. First, ML methods are focused on obtaining the best out-of-sample predictions, whereas causal econometric methods aim to derive the best-unbiased estimators. Therefore, tools that are optimized for causal inference often do not perform well when making out-of-sample predictions. As we will show below, the best-unbiased estimator does not always provide the best out-of-sample prediction, and in some instances, a biased estimator performs better for out-of-sample data.

Second, ML tools are designed to work in situations in which we do not have an a priori theory about the process through which outcomes observed in the data were generated. This aspect of ML contrasts with econometric methods that are designed for testing a specific causal theory. Third, unlike many empirical methods used in marketing, ML techniques can accommodate an extremely large number of variables and uncover which variables should be retained and which should be dropped. Finally, scalability is a key consideration in ML methods, and techniques such as feature selection and efficient optimization help achieve scale and efficiency. Scalability is increasingly important for marketers because many of these algorithms need to run in real-time.

To illustrate these points, consider the problem of predicting whether a user will click on an ad. We do not have a comprehensive theory of users' clicking behavior. We can, of course, come up with a parametric specification for the user's utility of an ad, but such a model is unlikely to accurately capture all the factors that influence the user's decision to click on a certain ad. The underlying decision process

may be extremely complex and potentially affected by a large number of factors, such as all the text and images in the ad, and the user's entire previous browsing history. ML methods can automatically learn which of these factors affect user behavior and how they interact with each other, potentially in a highly non-linear fashion, to derive the best functional form that explains user behavior virtually in real-time. ML methods typically assume a model or structure to learn, but they use a general class of models that can be very rich.

Broadly speaking, ML models can be divided into two groups: supervised learning and unsupervised learning. Supervised learning requires to be input data that has both predictor (independent) variables and a target (dependent) variable whose value is to be estimated. By various means, the process learns how to predict the value of the target variable based on the predictor variables. Decision trees, regression analysis, and neural networks are examples of supervised learning. If the goal of an analysis is to predict the value of some variable, then supervised learning is used. Unsupervised learning does not identify a target (dependent) variable but rather treats all of the variables equally. In this case, the goal is not to predict the value of a variable, but rather to look for patterns, groupings, or other ways to characterize the data that may lead to an understanding of the way the data interrelate. Cluster analysis, factor analysis (principal components analysis), EM algorithms, and topic modeling (text analysis) are examples of unsupervised learning [15].

We focus on supervised learning because marketing researchers are already familiar with many of the unsupervised learning techniques. In this section, we presented by a detail of the major approach of ML to global market [16].

Chapter 2. Used technologies and their potency and opportunities in the global market

Technology is superior and it has become one of the key points now. Particularly, technologies approach to market also has become the majority sing. Present scientific work also has been performed with the help of high-performance technologies like Oracle DB, Power BI, and Python. These technologies have played a pivotal role in creating more efficient and potential visual graphs, accurate calculations, and more.

2.1 Data processing, and cleansing with Oracle Database

Data Processing and Advanced Analytics is the foundation to producing good intelligence. However, analytics means many things to many people. Advanced analytics utilizes data of different types, from different sources and applies precise algorithmic processing. Valued intelligence results from the timely correlations and insights amongst this data, the algorithm results, and the interrelationships that exist from different data sources.

The challenge to producing mission results in data processing and analytics comes from numerous areas. Frequently, different data types are managed in different data stores. This causes information fragmentation and hinders analysis. Additionally, there are inefficiencies in the rigorous ETL (Extract, Transform, and Load) and integration required to do analysis. Many believe 80% of the time and cost spent is just preparing the data for analysis. Likewise, the pugh and pull of the developer community between NoSQL, SQL, Schema, Schema-less, Hadoop, No-Hadoop has created a pile of failed programs with limited analytical results. All these issues result in significant delays in mission execution that is unnecessary given the advancements with today’s technology.

To overcome what we see our enterprise customer face on a daily basis, Oracle and our thousands of developers have created an efficient and robust standards-based

platform that addresses the practical needs of the analytic enterprise. Oracle has addressed major issues like multi-data type support in the same data engine, easier data wrangling, automation, machine learning and unification of Hadoop, Relational and In Memory eco-systems. Our technology supports rapid, low cost, iterative and adaptive analytics. By empowering analysts with a self-service platform, users can perform rapid tests and evaluations on the data with current analytical methods. Oracle's technology provides a solution that is a fully complete analytic environment that supports full-spectrum data ingest, wrangling, data exploration & discovery through advanced and predictive analytics. It represents the combination of software, cloud computing and/or supporting hardware that has been professionally engineered, optimized, developed and deployed to support the analytic challenges faced today.

A key differentiated objective is to empower analysts to explore, test, and evaluate in a self-service fashion, thus reducing the need for costly programmers and data scientists.

Unfortunately, the overloaded use of the word "analytics" creates not only ambiguity in conversation but also incomplete and inefficient solution architectures. The reason has to do with the fact that the analytical tools used to perform the above processes vary as widely as the definition of analytics itself [13]. The reality is that the different tools and different ways we store and manage data for analytics creates impedance in doing higher level, result-oriented advanced analytics.

Oracle has created a holistic, standards-based and unified approach to provide integrated analysis for all data types, analytic methods and user classes.

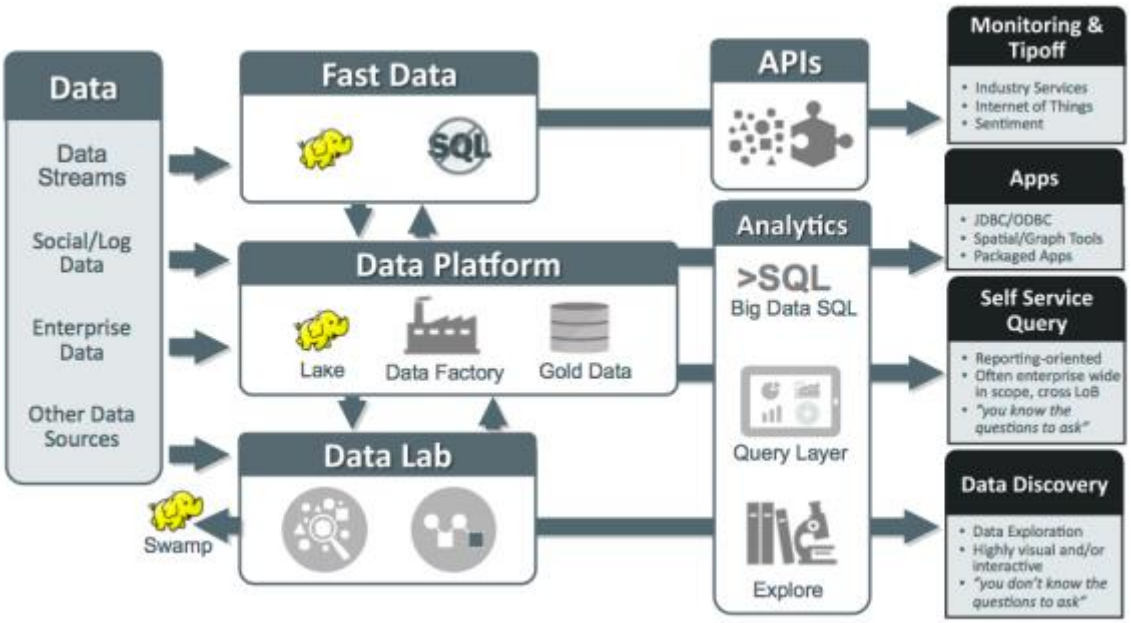

Fig 1. Pipeline data processing on the Oracle analytic platform

This figure is an excellent roadmap to understand the landscape of data processing and analytics. Each area is explained in more detail.

The Oracle NoSQL Database provides network-accessible multi-terabyte distributed key/value pair storage with predictable latency. Data is stored in a very flexible key-value format, where the key consists of the combination of a major and minor key (represented as a string) and an associated value (represented as a JSON data format or opaque set of bytes). It offers full Create, Read, Update and Delete (CRUD) operations, with adjustable durability and consistency guarantees [18]. It also provides a powerful and flexible transactional model (with ACID) that eases application development.

The Oracle NoSQL Database is designed to be a highly available and extreme scalable system, with predictable levels of throughput and latency, while requiring minimal administrative interaction.

It is also network topology and latency aware. The database driver working in conjunction with highly scalable, fault tolerant, high throughput storage engine enables a more granular distribution of resources and processing, which reduces the

incidence of hot spots and provides greater performance on commodity-based hardware.

Oracle provides several enabling technologies for data analytics, statistical analysis, time-series analysis, modeling, and machine learning requirements. These technologies can actually be used in both the product data platform and the data lab. Traditional advanced analytics are inherently weak at information technology management such as:

- data extracts and data movement
- data duplication resulting in no single-source of truth
- data security exposures
- multiple analytical tools (commercial and open source) and languages (SAS, R, SQL, Python, SPSS, etc.)

Problems become particularly egregious during a deployment phase when the worlds of data analysis and information management collide [19].

Oracle delivers an analytics platform and visualization that eliminates the traditional extract, move, load, analyze, export, paradigm when wanting to apply an advance algorithm on data [19]. The Oracle Advanced Analytics Option in Oracle 12c and the Advanced Analytics for Big Data options perform analytic functions on data where it resides. Data remains in the database or HDFS cluster, which reduces data movement performance impacts, promotes data security control mechanisms, and provides greater performance and scalability.

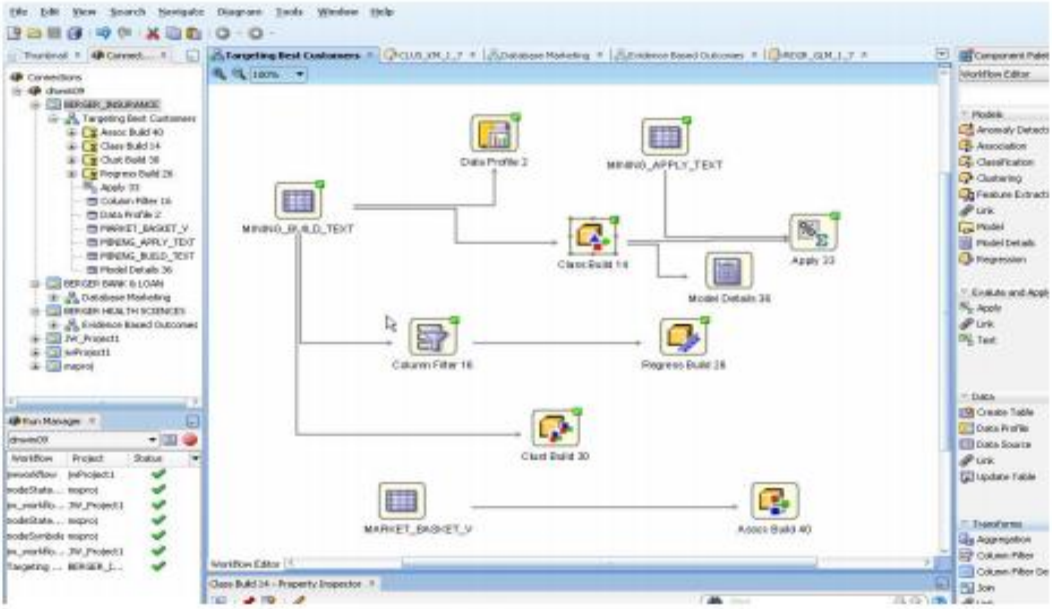

Fig 2. The schema of Oracle DB analytic platform workflows

Oracle Advanced Analytics Option extends the database into a comprehensive advanced analytics platform for data mining and data analytics. It delivers scalable, parallelized, in-database implementations of 20+ analytics algorithms (e.g., clustering, regression, prediction, associations, text mining, associations analysis, anomaly detection, etc.) as SQL functions within the Oracle Database 12c. Oracle Advanced Analytics exposes these predictive algorithms as SQL functions accessible via Oracle Data Miner, the Oracle Data Miner "drag and drop" workflow GUI, an extension to Oracle SQL Developer, and through tight integration with open source R (Oracle R Enterprise) [18].

Oracle Advanced Analytics provides a mechanism to not only study the data through statistical means, and put those statistical and machine leaning modules into production. The ease of migration from the data study/analysis phase, into an enterprise data production environment using these models is a key differentiator our technology has over other competitors. This allows our customers to quickly modify or employ multiple machine learning techniques based on business or mission need.

Machine Learning is the ability to automatically sift through large amounts of data to create models that find previously hidden patterns, discover valuable insights, and make predictions through the Big Data, Fast Data, and All Data use of analytics models. These capabilities are important to multiple commercial industries including; banking, financial, retails, and the DOD. The ability to predict fraud on-line is a key element for these sectors, and these models are based on past behavior and methods to allow the model to predict and the system to take alternative measures with the transaction.

Oracle Advanced Analytics and Enterprise R technology allow the data scientist to review the data via built in statistical algorithms and extend R algorithms to determine key data patterns that define the behavior or characteristic under investigation. These models can then be tested and validated, and placed into production to run against real-time data coming into the system and provide NRT prediction on the system behavior and/or characteristic under investigation.

Oracle Data Integrator delivers high-performance data movement and transformation across enterprise platforms with its open and integrated E-LT (Extract, Load, Transform) architecture and extended support for Big Data. Oracle Data Integrator is critical to leveraging data integration initiatives on premise or in the cloud, such as Big Data management, Service Oriented Architecture and Business Intelligence. An easy-to-use user interface combined with a rich extensibility framework helps Oracle Data Integrator improve 12 Big Data, Fast Data, All Data productivity, reduce development costs and lower total cost of ownership among data-centric architectures.

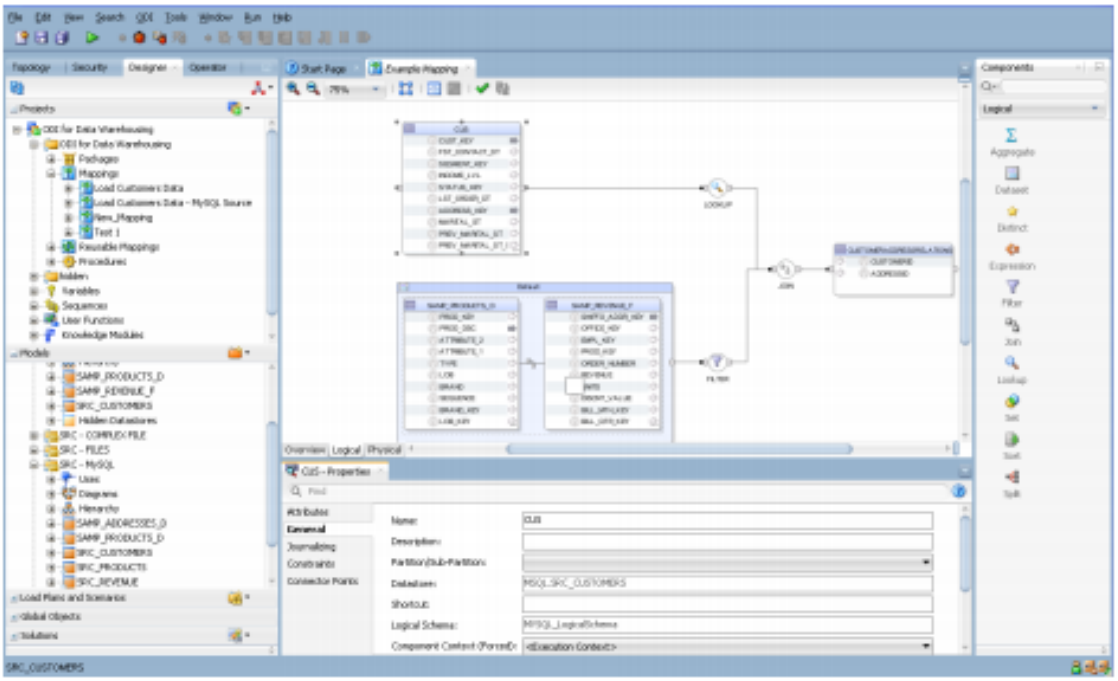

Fig 3. The Oracle Data Integrator

Oracle Data Integrator's extract, load, transform (E-LT) architecture advantages disparate relational database management systems (RDBMS) or Big Data engines to process and transform the data. This approach optimizes performance and scalability and lowers overall solution costs [21]. Instead of relying on a separate, conventional ETL transformation server, the architecture generates native code for disparate RDBMS or big data engines (SQL, HiveQL, PySpark, Pig Latin or bulk loader scripts, for example). The E-LT architecture extracts data from the disparate sources, loads it into a target, and executes transformations using the power of the database or Hadoop.

Oracle Data Integrator uses a declarative flow-based user interface for enhanced user experience and productivity. The interface combines the simplicity and ease-of-use of the declarative approach with the flexibility and extensibility of configurable flows. This blend simplifies common data integration design and deployment use cases, shortening implementation times. Data integration designers describe source and target data formats and data integration processes. The business user or the developer can focus on describing what to do, not how to do it. Oracle

Data Integrator generates, deploys and manages the code required to implement those processes across the various source and target systems [21].

Data is very rarely available in a completely neat and ordered fashion. Typical problems include:

- Constructed fields, where a customer ID may be made up of a location code, a customer reference, and an account manager code
- Misfielded data, such as names, comments, or telephone numbers in address blocks
- Poorly structured data such as addresses, where data can flow from one field to the next.
- Notes fields that store information that the data structure doesn't support, but that contain useful semi-structured data that normally cannot be analyzed or extracted

All of these problems can be solved using Oracle Enterprise Data Quality. Using a data-driven approach to rapidly tag or describe data, it can manipulate a single record by parsing it into multiple structured elements (and, if required, records) and standardize results according to predefined rules. 13 Big Data, Fast Data, All Data Innovative parsing and phrase analysis technology uniquely allows you to find hidden knowledge within any text field and create rules to standardize it into structured data.

Oracle Enterprise Data Quality provides a rich palette of functions to transform and standardize data using easily managed reference data and a simple graphical configuration. In addition to functions for basic numeric, string, and date fields, functions for contextual data such as names addresses and phone numbers are provided. Users can also quickly configure, package, share and deploy new functions that encapsulate rules specific to their data and industry without any coding [22].

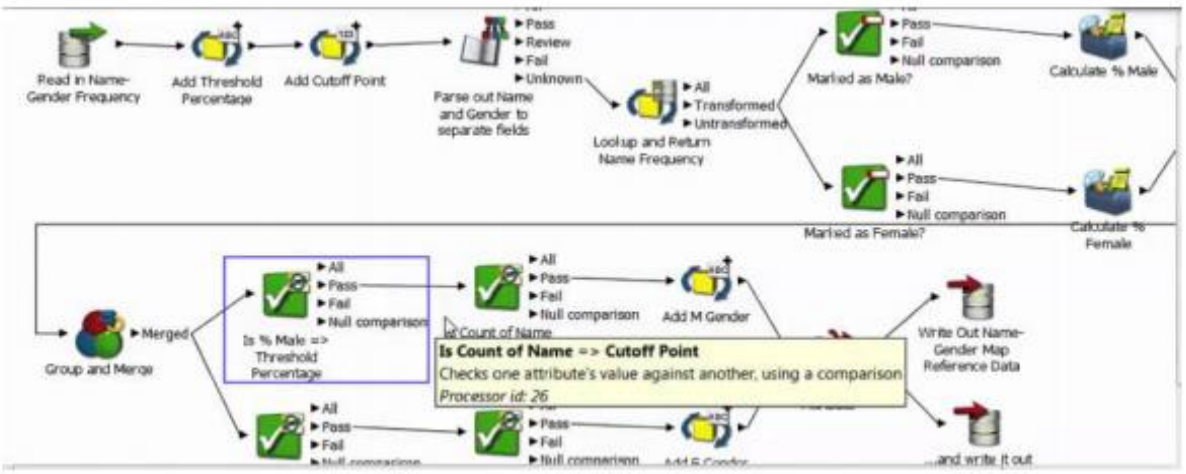

Fig 4. The main workflow of the Oracle Enterprise Data Quality

Using a web browser, workers and managers can monitor and review ongoing data quality against defined metrics. Data quality dashboards allow problems to be quickly identified and dealt with before they start to cause significant business impact. Graphical views show data quality trends over time, helping your organization protect its investment in data quality, by giving visibility to the right people [22].

2.2 Microsoft Power BI and building essential graphs

Microsoft introduced the idea of self-service business intelligence (BI) back in 2009, announcing Power Pivot for Microsoft Excel 2010. Ironically, at the time, it did not make big announcements, hold conferences, or run a big marketing campaign for it. Everything started slowly, some users enthusiastically embraced the new technology, but the vast majority of people were not even aware of its existence. As members of the business intelligence community, we were very surprised by this approach. At the time, we could see clearly benefits for users to start using Power Pivot as a tool for gathering insights from data, so no marketing at all was somewhat disappointing.

So over the years, as a community kept asking Microsoft what they were looking for, what the delay is in promoting self-service BI to a larger audience of data analysts, data scientists, decision makers, and business analyst enthusiasts

across the board planet. We asked for the ability to share reports with the team and the answer was to use SharePoint, on-premises or online, with the first release of Power BI - an experience that has not been fully implemented yet satisfactorily. While we waited for Microsoft to fix issues with previous versions and start advertising current products, it did something different, which, in hindsight, seemed to be the perfect choice. Microsoft collected user feedback, took a close look at what was missing in the end-user business intelligence world, and then made the version of Power BI available for nowadays.

Through Power BI, technologies will perform essential graphs marketing as Animated Bar Charts. An animated bar chart is basically a fascinating animated trend chart, with bars that race to the top based on ranks. There are usually 3 variables involved in making an animated bar chart, one of which is, more often than not, the time variable. The remaining two variables are

- The variable representing categories such as Country names, Company names and so on
- The variable representing the corresponding value of the each category.

The following are some of the examples of the animated bar charts (images from current scientific project):

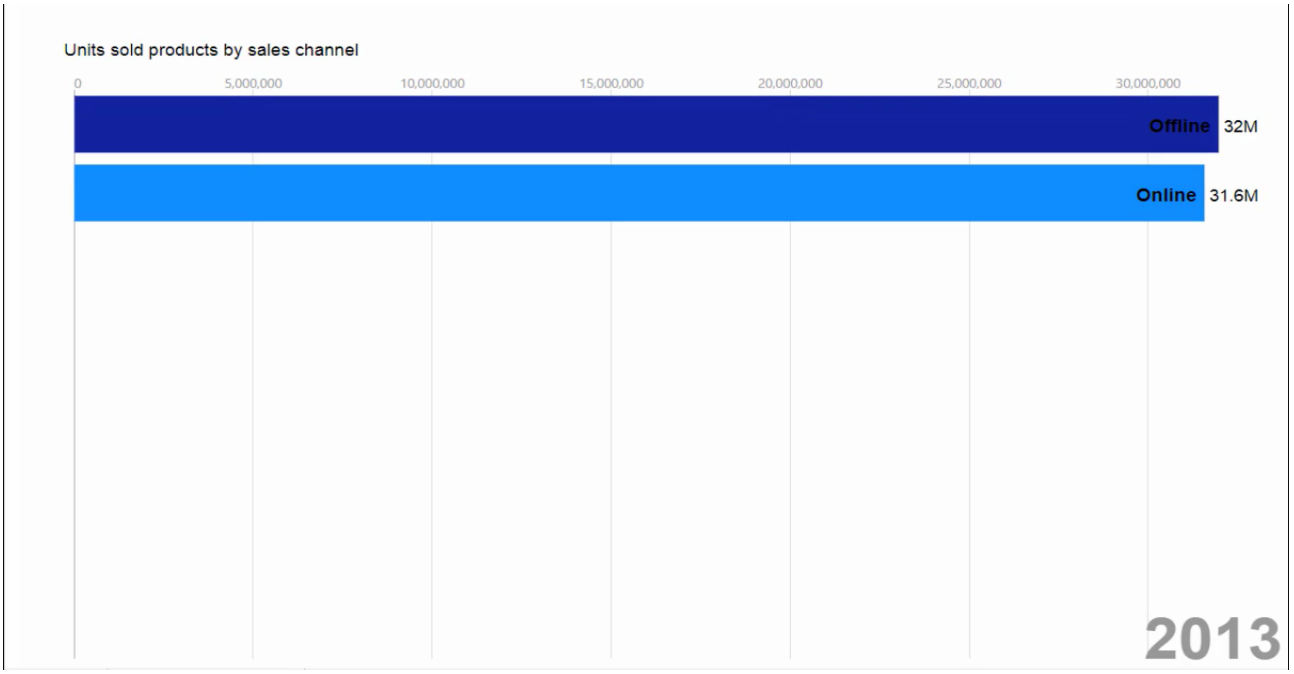

Fig 5. Animated bar chart race units sold product by sales channel

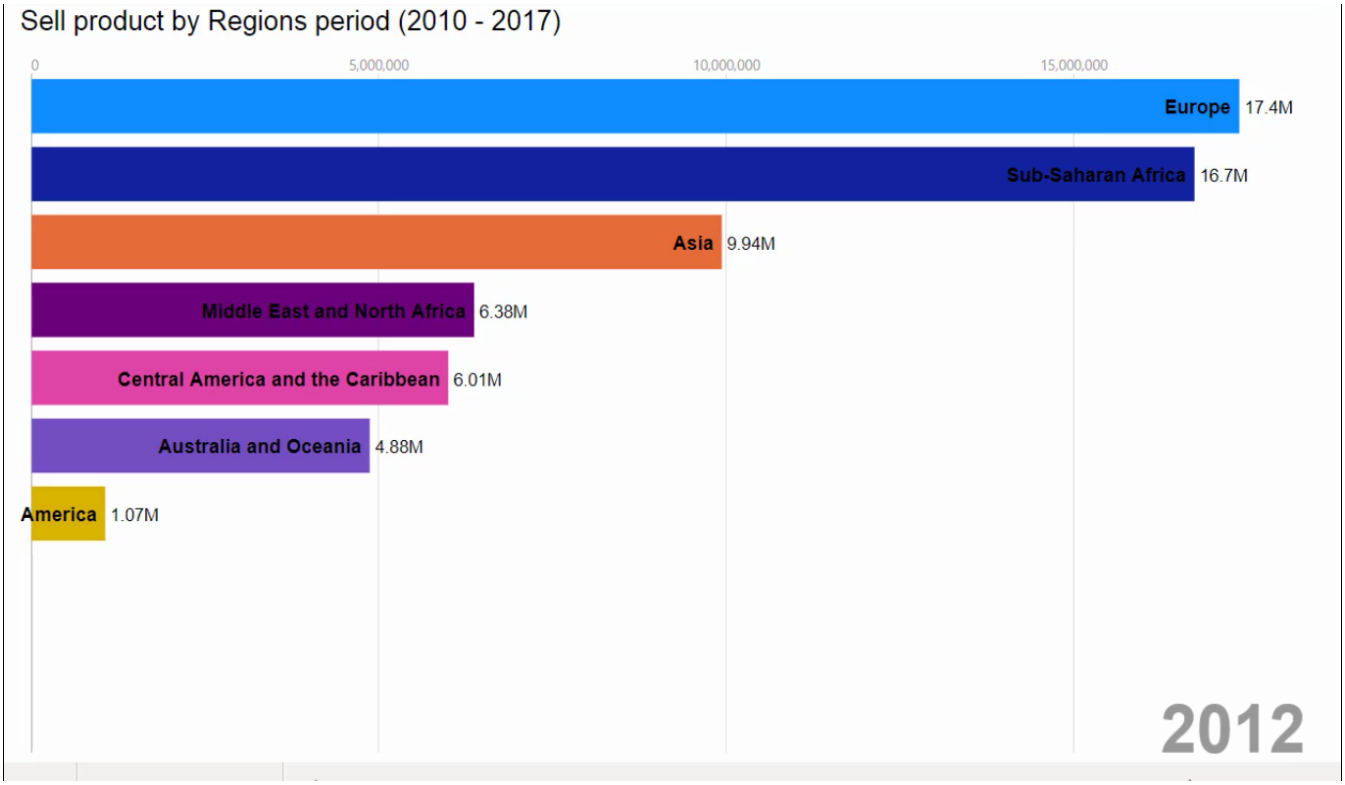

Fig 6. Animated bar chart race units sold product by regions

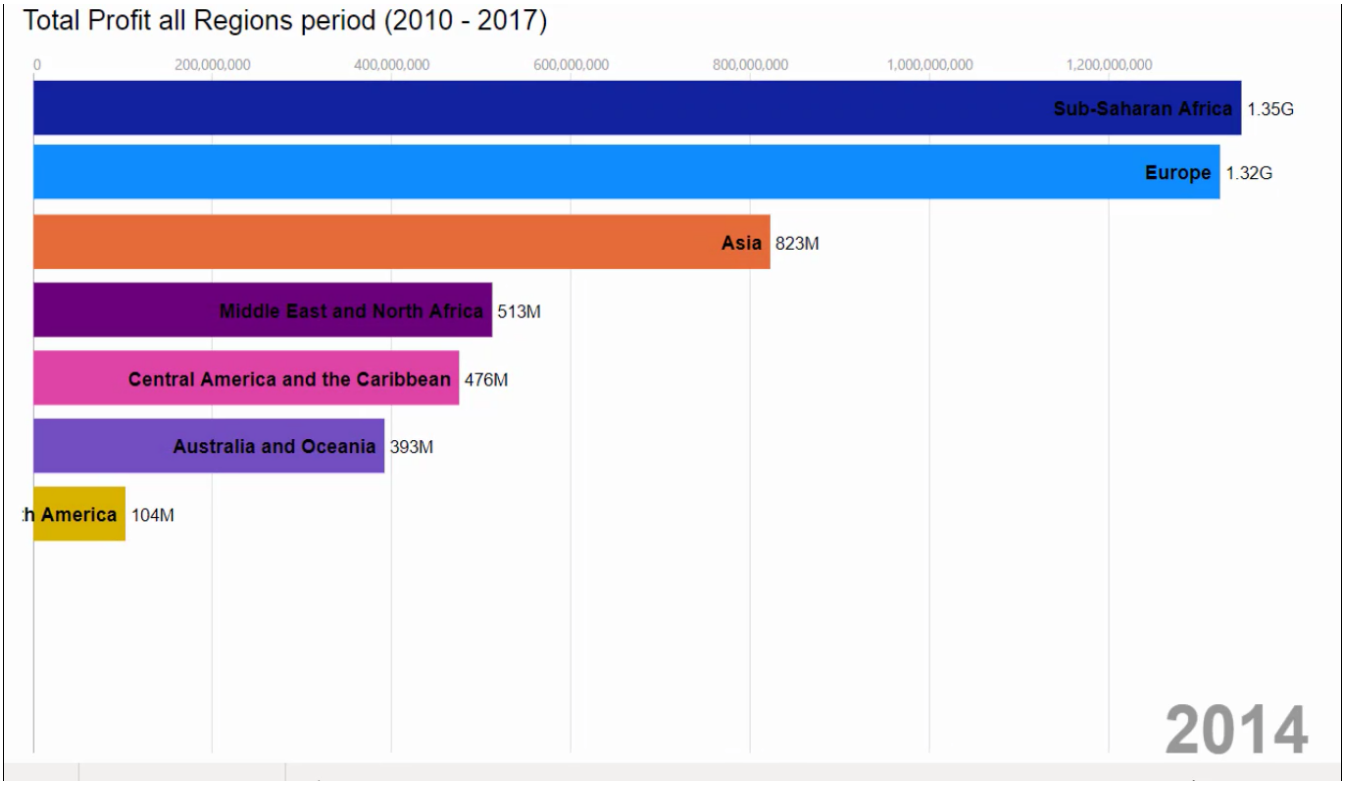

Fig 7. Animated bar chart race total profit by regions

The animated bar chart helps us visualize the change in trends over time, these type of charts are very popular, especially on social media as they provide a holistic

data story/insight in a concise and easy-to-understand chart.

2.3 Python and performing common features of data analysis

Python is a popular multi-purpose programming language widely used for its flexibility, as well as its extensive collection of libraries, which are valuable for analytics and complex calculations.

Python's extensibility means that it has thousands of libraries dedicated to analytics, including the widely used Python Data Analysis Library (also known as Pandas).For the most part, data analytics libraries in Python are at least somewhat derived from the NumPy library, which includes hundreds of mathematical calculations, operations, and functions.

Python analytics tools have become popular due to the computer language's widespread adoption and its versatility when it comes to developing multifaceted solutions. The fact that it is a truly general-purpose language means that it can also add deeper functionality to data analytics software than domain-specific languages that have a narrow scope and functionality. Additionally, Python's performance capability is much higher than other popular languages used in data analytics, and its compatibility with a greater array of other languages means that it is simply more convenient in most cases.

Python's relatively light usage of memory and other processing resources means that it can quickly outstrip languages like MatLab or R, which are built specifically for statistical analysis.

There are several ways you can integrate python data analytics into your existing business intelligence and analytics tools. One of the most common uses for Python is in its ability to create and manage data structures quickly — Pandas, for instance, offers a plethora of tools to manipulate, analyze, and even represent data structures and complex datasets. This includes time series and more complex data structures such as merging, pivoting, and slicing tables to create new views and perspectives on existing sets.

Elsewhere, tools like Scikit-learn (also known as Sklearn) provides advanced

analytics tools combined with complex machine learning capabilities. This allows you to build more sophisticated models, performing more complex and multivariate regressions, as well as data preprocessing. Combined with libraries such as IPython and NumPy itself, these tools can form the foundation of a powerful data analytics suite.

Additionally, data analytics can use Python to write him own data analysis algorithms that can be directly integrated into their business intelligence tools via API.

Chapter 3. Preparation data and Machine Learning algorithm for global market analysis

Data preparation or data preprocessing, one of the initial stage of data analysis that will be providing warranty, efficient and accurate of data. Data preparation process of transforming raw data so that data scientists and analysts can run it through machine learning algorithms to uncover insights or make predictions.

Most Machine Learning algorithms require data to be formatted in a very specific way, so datasets generally require some amount of preparation before they can yield useful insights. Some datasets have values that are missing, invalid, or otherwise difficult for an algorithm to process. If data is missing, the algorithm cannot use it. If data is invalid, the algorithm produces less accurate or even misleading outcomes. Some datasets are relatively clean but need to be shaped (e.g., aggregated or pivoted) and many datasets are just lacking useful business context (e.g., poorly defined ID values), hence the need for feature enrichment. Good data preparation produces clean and well-curated data, which leads to more practical, accurate model outcomes.

There is a pipeline explaining the workflow of a science project. According to the pipeline, each stage has a definition and functionality. The pipeline workflow of a scientific project was presented below:

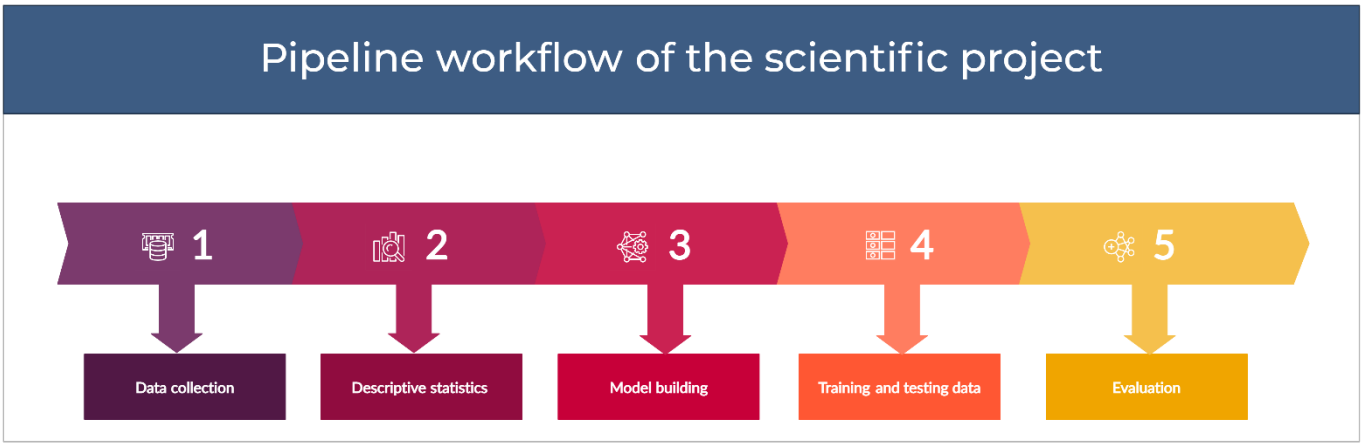

Fig 8. The pipeline workflow of a scientific project

3.1 Data collection

Data collection is defined as the procedure for collecting, measuring, and analyzing accurate information for research using standard proven methods. The researcher can evaluate his hypothesis based on the collected data. In most cases, data collection is the first and most important stage of research, regardless of the field of study. Data collection approaches differ for different fields of study, depending on the information required.

The basic problem data collection was directed to getting correspond dataset that contains essential features of the global market. Conformity cleaned and transformed obtain of the global market dataset by help Oracle Database results was presented below:

	REGION	COUNTRY	ITEM_TYPE	SALES_CHANNEL	QUALITY_PRODUCT	ORDER_DATE	ORDER_ID	SHIP_DATE	UNITS_SOLD	UNIT_PRICE	UNIT_COST	TOTAL_COST
1	Central America and the Caribbean	Trinidad and Tobago	Fruits	Offline	C	13.04.14	810956513	15.04.14	8908	9.33	6.92	61643.36
2	Middle East and North Africa	Israel	Furniture	Offline	L	16.03.11	694069278	05.05.11	9378	668.27	502.54	4712820.12
3	Australia and Oceania	Papua New Guinea	Fruits	Offline	M	26.09.16	895744359	24.10.16	9516	9.33	6.92	65850.72
4	Middle East and North Africa	Iraq	Cosmetics	Offline	H	01.04.14	597280585	01.04.14	7453	437.20	263.33	1962598.49
5	Sub-Saharan Africa	Sudan	Baby Food	Online	M	19.11.10	400289285	31.12.10	3350	255.28	159.42	534057.00
6	Europe	Vatican City	Furniture	Offline	C	25.01.11	976694285	07.02.11	4398	668.27	502.54	2210170.92
7	Europe	Norway	Furniture	Offline	M	19.09.12	818731915	01.11.12	9823	668.27	502.54	4936450.42
8	Europe	Kosovo	Furniture	Online	H	14.01.11	288951003	28.02.11	5878	668.27	502.54	2953930.12
9	Australia and Oceania	Papua New Guinea	Cereal	Offline	L	04.06.17	270212876	21.07.17	850	205.70	117.11	99543.50
10	Middle East and North Africa	Kuwait	Meat	Online	C	13.04.10	815691363	27.04.10	56	421.89	364.69	20422.64
11	Australia and Oceania	Tuvalu	Baby Food	Offline	C	02.07.14	698872792	29.07.14	845	255.28	159.42	134709.90
12	Central America and the Caribbean	Nicaragua	Cereal	Online	H	24.05.14	563555991	27.05.14	7596	205.70	117.11	889567.56
13	Europe	Croatia	Clothes	Online	M	25.06.10	673852288	19.07.10	6721	109.28	35.84	240880.64
14	Central America and the Caribbean	Haiti	Fruits	Online	L	16.09.15	464185512	24.09.15	1663	9.33	6.92	11507.96
15	Australia and Oceania	Tuvalu	Meat	Offline	M	18.10.13	789291989	26.11.13	7894	421.89	364.69	2878862.86
16	Europe	Croatia	Meat	Offline	H	13.01.16	774220550	18.02.16	7920	421.89	364.69	2888344.80
17	Europe	Czech Republic	Cereal	Offline	H	21.01.12	964332520	24.02.12	6277	205.70	117.11	735099.47
18	Middle East and North Africa	Egypt	Cosmetics	Offline	H	30.06.11	175301730	05.07.11	4532	437.20	263.33	1193411.56
19	Sub-Saharan Africa	Sierra Leone	Vegetables	Offline	H	23.06.17	443114507	29.07.17	6284	154.06	90.93	571404.12
20	Sub-Saharan Africa	Rwanda	Fruits	Online	L	03.04.15	274069678	22.05.15	2163	9.33	6.92	14967.96
21	Europe	Macedonia	Meat	Offline	M	14.11.13	494778573	17.12.13	4830	421.89	364.69	1761452.70
22	Europe	Andorra	Meat	Online	L	28.01.12	202165019	08.02.12	3977	421.89	364.69	1450372.13
23	Asia	China	Clothes	Online	C	01.04.17	813465344	26.04.17	1327	109.28	35.84	47559.68
24	Central America and the Caribbean	Panama	Meat	Offline	M	04.05.11	216228377	04.05.11	5637	421.89	364.69	2055757.53
25	Australia and Oceania	Tonga	Baby Food	Online	L	20.03.15	418315446	17.04.15	7691	255.28	159.42	1226099.22

Fig 9. Imagination basic features of the global market dataset

The global market dataset includes more than 95 000 data and 13 columns that collected by the Yahoo statistics of financial marketing. To understand the structure dataset need to display description columns according to the essential features of the dataset. Below putted tables presents the description of essential columns of the global market dataset.

Table 1. Description of essential columns of the global market dataset

Assignment	**Amount**
Region	7
Country	177
Item Type	12
Sales Channel	2
Quality Product	4
Order Date	10/10/2010 - 9/9/2017
Ship Date	10/10/2010 - 9/30/2017

Table 2. Description of column region

Region
North America
Central America and the Caribbean
Sub-Saharan Africa
Europe
Australia and Oceania
Asia
Middle East and North Africa

Table 3. Description of column item type

Item Type
Medical Products
Meat
Vegetables
Office Supplies
Appliances
Cosmetics
Clothes
Fruits
Cereal
Furniture
Optical Instruments
Baby Food

Table 4. Description of column Sales Channel

Sales Channel
Offline
Online

Table 5. Description of column Quality Product

Quality Product
Critical
Low
Medium
High

3.2 Descriptive statistics

Descriptive statistics are measures that summarize important features of data, often with a single number. Producing descriptive statistics is a common first step to take after cleaning and preparing a data set for analysis. We have already seen several examples of deceptive statistics in earlier lessons, such as means and medians. In this lesson, we will review some of these functions and explore several new ones.

In Python, descriptive statistics measures include the following methods: describe, mean, dtypes, nunique, correlation, etc.

Below are the descriptive statistics results obtained by help Python.

	ORDER_ID	UNITS_SOLD	UNIT_PRICE	UNIT_COST	TOTAL_COST	TOTAL_PROFIT
count	2.004000e+03	2004.000000	2004.000000	2004.000000	2.004000e+03	2.004000e+03
mean	5.577227e+08	4916.576347	262.769137	185.782769	9.142408e+05	3.947314e+05
std	2.620999e+08	2915.670258	217.795228	176.679449	1.144149e+06	4.975489e+05
min	1.001106e+08	3.000000	9.330000	6.920000	1.075200e+02	1.566000e+02
25%	3.311899e+08	2339.500000	81.730000	56.670000	1.500635e+05	9.124191e+04
50%	5.593735e+08	4866.500000	154.060000	97.440000	4.400451e+05	2.830084e+05
75%	7.982649e+08	7444.500000	421.890000	364.690000	1.130287e+06	5.501549e+05
max	9.986635e+08	10997.000000	668.270000	524.960000	5.241201e+06	9.297511e+06

Fig 10. Describing a column from a data by accessing an attribute

REGION	object
COUNTRY	object
ITEM_TYPE	object
SALES_CHANNEL	object
QUALITY_PRODUCT	object
ORDER_DATE	datetime64[ns]
ORDER_ID	int64
SHIP_DATE	datetime64[ns]
UNITS_SOLD	int64
UNIT_PRICE	float64
UNIT_COST	float64
TOTAL_COST	float64
TOTAL_PROFIT	float64

Fig 11. Data type of each column

	ORDER_ID	UNITS_SOLD	UNIT_PRICE	UNIT_COST	TOTAL_COST	TOTAL_PROFIT
ORDER_ID	1.000000	0.017688	0.015246	0.023825	0.024448	-0.015308
UNITS_SOLD	0.017688	1.000000	0.018739	0.018997	0.458660	0.471753
UNIT_PRICE	0.015246	0.018739	1.000000	0.986840	0.752776	0.454607
UNIT_COST	0.023825	0.018997	0.986840	1.000000	0.765077	0.398248
TOTAL_COST	0.024448	0.458660	0.752776	0.765077	1.000000	0.586518
TOTAL_PROFIT	-0.015308	0.471753	0.454607	0.398248	0.586518	1.000000

Fig 12. Pairwise correlation of dataset columns

REGION	7
COUNTRY	176
ITEM_TYPE	12
SALES_CHANNEL	2
QUALITY_PRODUCT	4
ORDER_DATE	1430
ORDER_ID	2004
SHIP_DATE	1417
UNITS_SOLD	1811
UNIT_PRICE	12
UNIT_COST	12
TOTAL_COST	1990
TOTAL_PROFIT	1990

Fig 13. Counts of unique elements in each position

3.3 Model building – selecting the right Machine Learning algorithm

Various samples data are can used for machine learning procedures.

However, how do we will realize that relationships samples data is strong or not. There are more approaches, which can figure out that problem such as:

- Correlation - can indicate strong relationships
- Box plot - can identify outliers
- Density plot - show the spread of data
- Scatter plot - can describe bivariate relationships

The above approaches can help with data sampling. Successfully fetched the outcomes with Python given below:

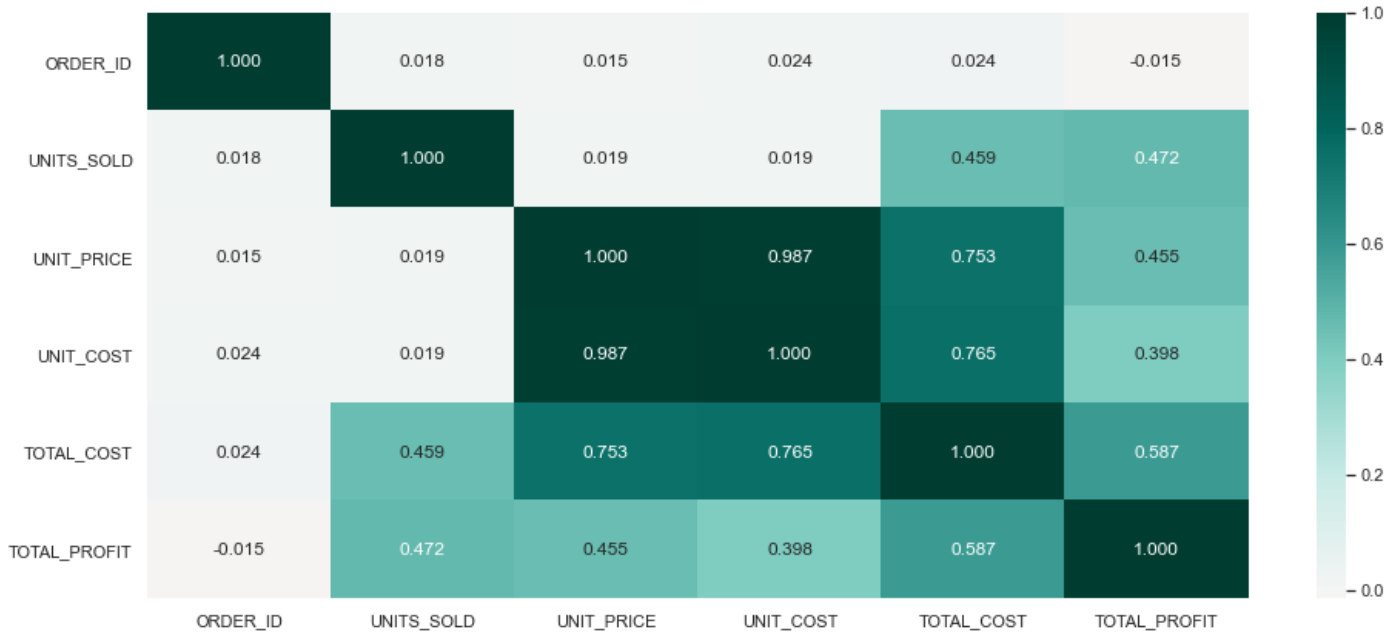

Fig 14. Correlation between dependent and independent variables

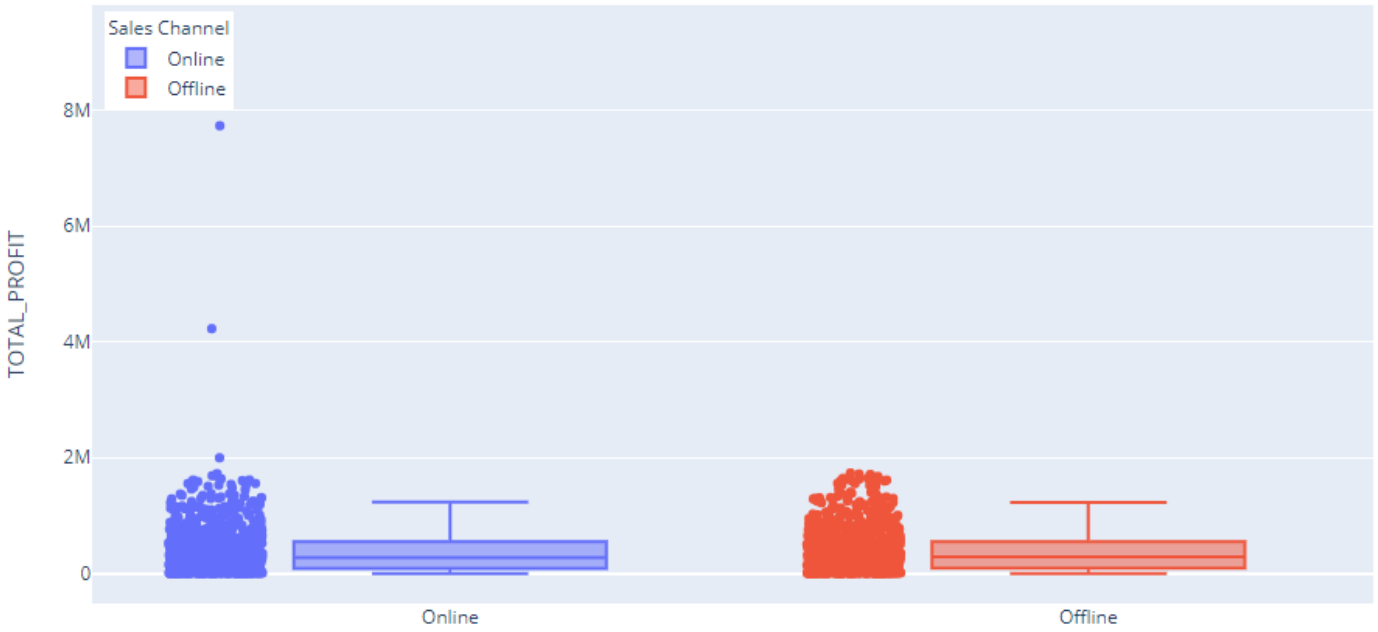

Fig 15. Box plot outliers of total profit by sales channel

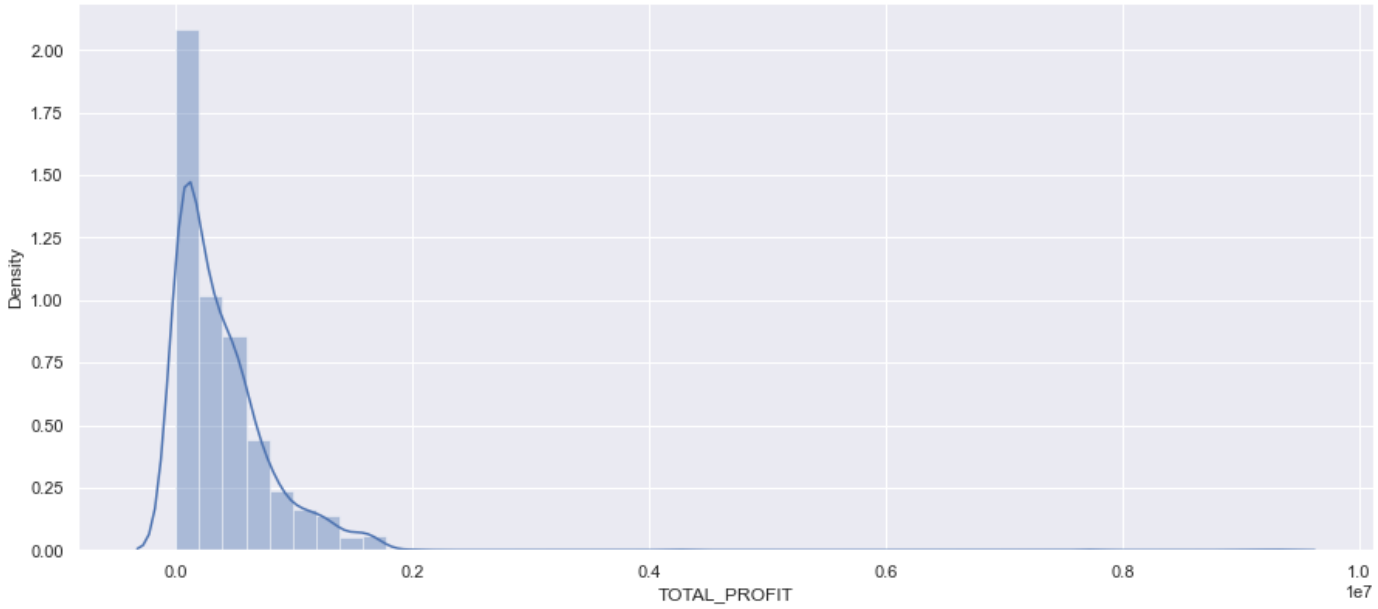

Fig 16. Density plot spread of total profit

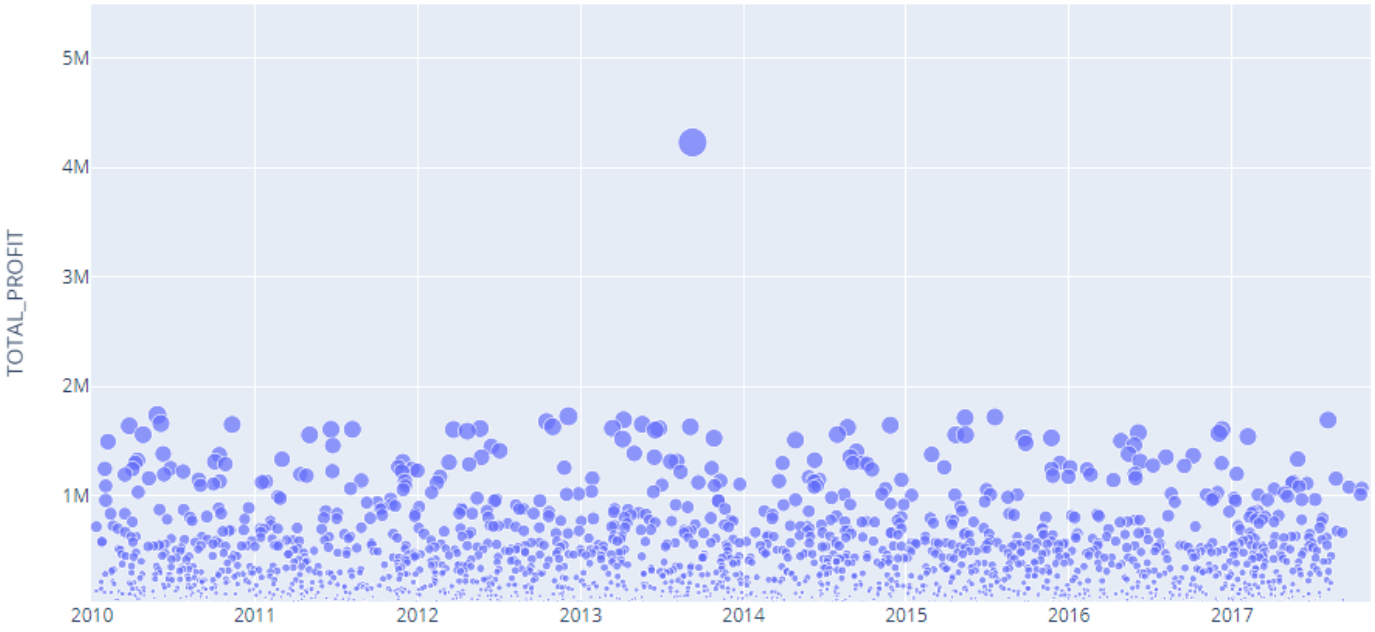

Fig 17. Scatter plot bivariate relationships between total profit by date

The basic building block of the proposed framework for sales forecasting and product portfolio classification is a tool/method for generating high-quality time-series forecasts. Despite the fact that there are numerous tools/methods that can be applied, it was decided to use Facebook's Prophet tool for this research since it is capable of generating forecasts of a reasonable quality at scale. Nevertheless, the main workflows were focused on selecting an appropriate machine learning algorithm that could predict the future based on historical data. Prophet is an open source package for predicting time series data based on an additive model in which nonlinear trends correspond to annual, weekly and daily seasonality plus efficiency. This works best with time series that have strong seasonal effects and multiple seasons of historical data [23]. The Prophet is resilient to missing data and trend changes and usually handles outliers well. Prophet is an open source software released by the core Data Science team at Facebook. Moreover, the main approach to forecasting will be implemented by the time series method. The time series method predicts the simple assumption that the future is a function of the past. In other words, they look at what happened over a period of time and use a series of past data to predict [23].

The Prophet machine-learning algorithm is most appropriate algorithm that can figure it out the business forecasting challenges, typically have any of the following characteristics:

- Hourly, daily or weekly observations with a history of at least several months (preferably a year).
- Strong multiple "human" seasonality: day of the week and season.
- Reasonable number of missing observations or large outliers.
- Historical trend changes, for example due to product launches or logging of changes.
- Trends that are non-linear growth curves where the trend reaches a natural limit or saturates

The Prophet's procedure is an additive regression model with four main components:

- Piece wise linear or logistic trend of the growth curve.
- Prophet automatically detects changes in trends by selecting points of change from the data.
- Annual seasonal component modeled using Fourier series.
- Weekly seasonal component using dummy variables.

The basic equation of the Prophet's algorithm shown in formula (1):

$$y(t) = g(t) + s(t) + h(t) + \epsilon(t) \quad (1)$$

Here $g(t)$ is the trend function which models non-periodic changes in the value of the time series, $s(t)$ represents periodic changes (e.g., weekly and yearly seasonality), and $h(t)$ represents the effects of holidays which occur on potentially irregular schedules over one or more days. The error term $\epsilon(t)$ represents any idiosyncratic changes, which not accommodated by the model.

Estimation of the parameters of the fitted Prophet's algorithm is performed

using the principles of Bayesian statistics.

A module that would be able to forecast any business case with a reasonably high accuracy, augmented by the module for highly-reliable classification of the product portfolio according to the expected level of forecastability, would be of great use for any company operating in the retail industry [24]. The Prophet methodology model for successful forecasting based on real-world data is shown in Fig 18.

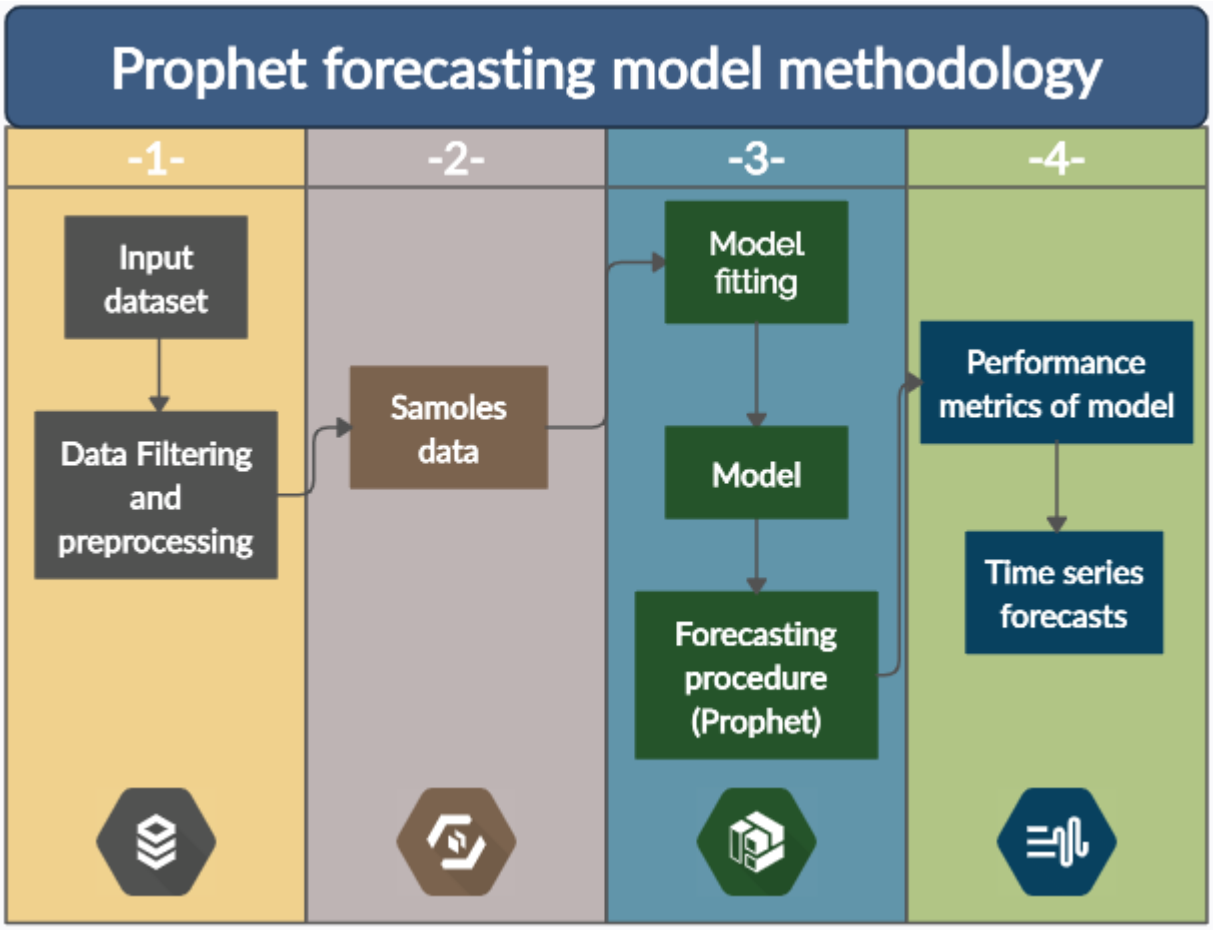

Fig 18. Prophet forecasting procedure methodology

Each phase individually affects the workflow of a machine learning model. Moreover, these phases are represented by a sub-process, and it will be carried out in stages. The quality process of a machine learning model is based on cleaned, trained and transformed data that the model can easily process on its own. Therefore, data preparation will play a major role in analytics - this is one of the important parts of the machine learning process.

3.4 Training and testing data

Training data and testing data are two important concepts in machine learning. The observations in the training set form the experience that the algorithm uses to learn. In supervised learning problems, each observation consists of an observed output variable and one or more observed input variables. The testing set is a set of observations used to evaluate the performance of the model using some performance metric [25]. It is important that no observations from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it.

In practice, data usually will be split randomly 70-30 or 80-20 into train and test datasets respectively in statistical modeling, in which training data utilized for building the model and its effectiveness will be checked on test data:

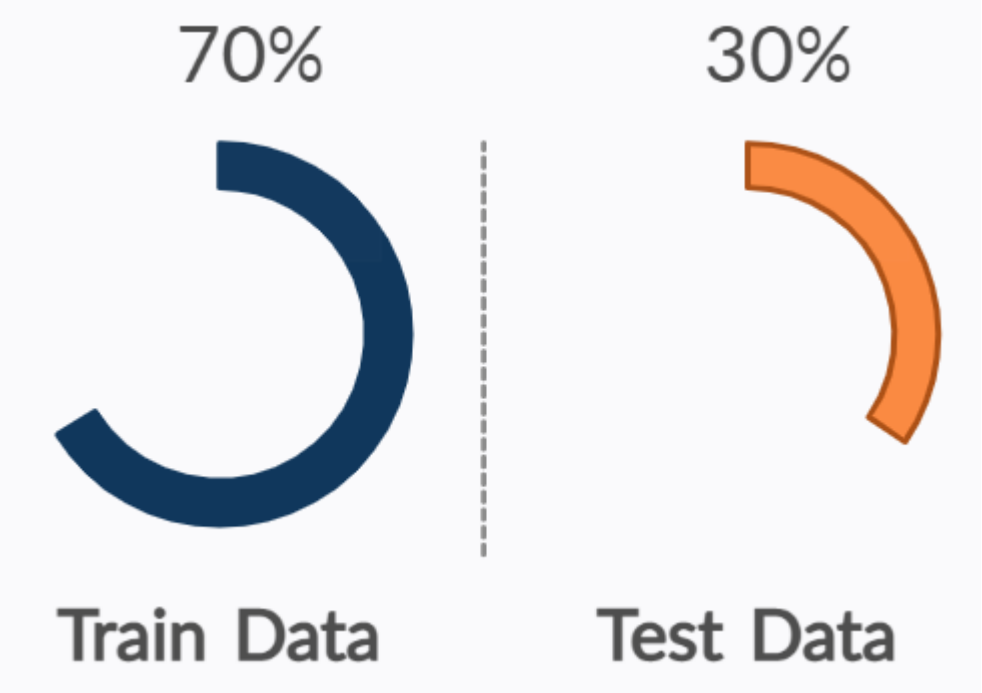

Fig 19. Statistical modeling methodology

We split the data into training and test data by 70-30 percent. However, before that, we must determine what features will be included in the major and target

features. There is already separated data for the major and target features. In the below was represented all features.

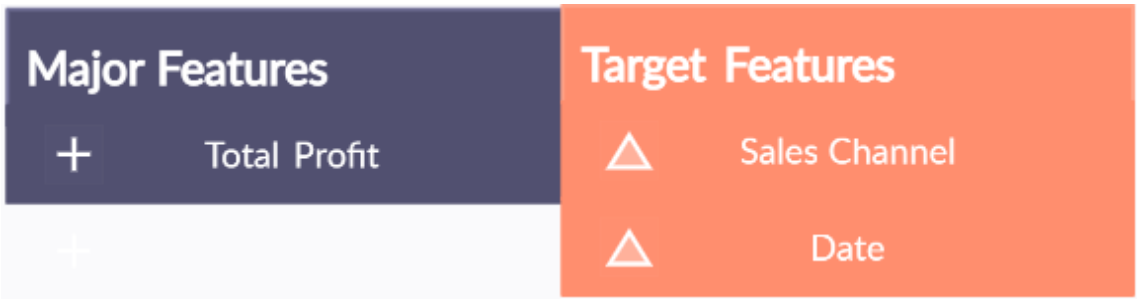

Fig 20. Major and target features

Nevertheless, in machine learning, a programmer usually inputs the data and the desired behavior, and the logic is elaborated by the machine. This is especially true for deep learning. Therefore, the purpose of machine learning training and testing is, first of all, to ensure that this learned logic will remain consistent, no matter how many times we call the program. Moreover, every ML model needs not only to be trained or tested but also evaluated. Our model should generalize well. This is not what we usually understand by training and testing, but evaluation is needed to make sure that the performance is satisfactory.

3.5 Evaluation

Before models can be deployed for use within an organization, it is important that they are fully evaluated and proved to be fit for the purpose. This phase of machine learning workflow covers all the evaluation tasks required to show that a prediction model will be able to make accurate predictions after being deployed and that it does not suffer from overfitting or underfitting.

Evaluating the performance of the model using different metrics is integral to every data science project. Here is what you have to keep an eye on [28]:

– **Mean Squared Error (MSE)**

The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value

[27]. Since it is differentiable and has a convex shape, it is easier to optimize.

The mathematical formula of the Mean Squared Error method is given below.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (2)$$

– **Mean Absolute Error (MAE)**

This is simply the average of the absolute difference between the target value and the value predicted by the model [28]. Not preferred in cases where outliers are prominent.

The mathematical formula of the Mean Absolute Error method is given below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (3)$$

– **R-squared or Coefficient of Determination**

This method helps us to calculate the relative error. This technique helps us to judge, which algorithm is better based on their mean squared errors.

The mathematical formula of the R – Squared method is given below.

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \quad (4)$$

– **Root Mean Squared Error (RMSE)**

This is the square root of the average of the squared difference of the predicted and actual value R-squared error is better than RMSE [28]. This is because R-squared is a relative measure while RMSE is an absolute measure of fit (highly dependent on the variables — not a normalized measure). Basically, RMSE is just the root of the average of squared residuals. We know that residuals are a measure of how distant the points are from the regression line. Thus, RMSE measures the scatter of these residuals.

The mathematical formula of the Root Mean Squared Error method is given below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (5)$$

Conformity of the performance metrics of model, we can assess efficiency and potentiality of machine learning model. During machine learning workflow, we are received next performance metrics of model and below represented given result.

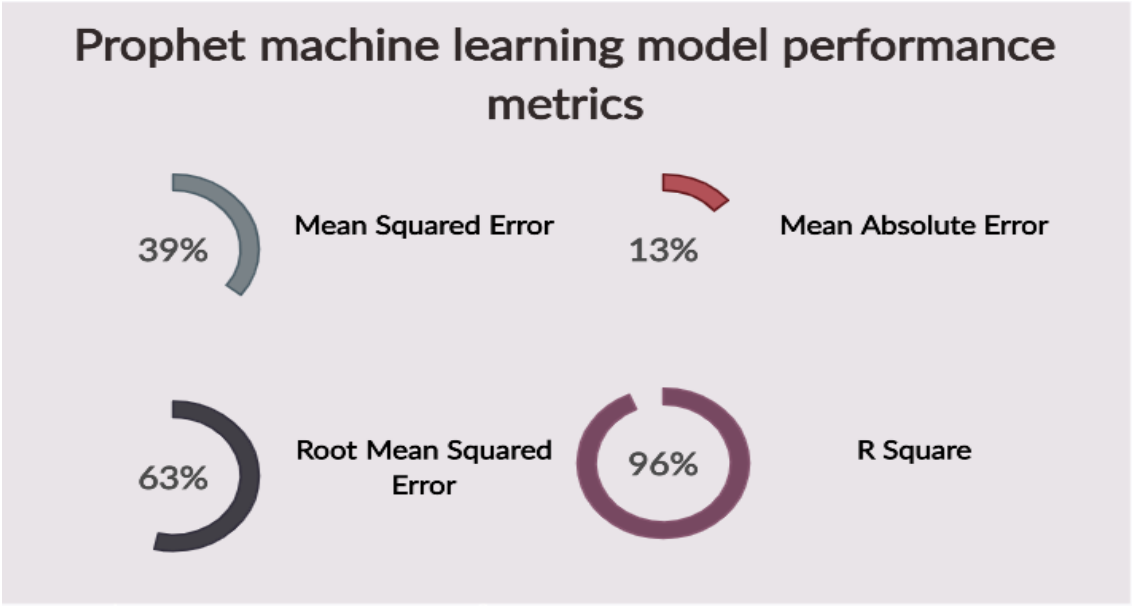

Fig 21. Prophet machine learning model performance metrics by percentage indicates

	Model	Mean Squared Error	Mean Absolute Error	Root Mean Squared Error	R Square
0	Prophet	0.39244	0.132538	0.626451	0.958263

Fig 22. Prophet machine learning model performance metrics by original indicates

Performing ML performance metrics is necessary if we care about the quality of the model. ML model has a couple of peculiarities: it demands that our test the quality of data, not just the model, and go through a couple of iterations adjusting the hyperparameters to get the best results. However, if we perform all the necessary procedures, we can be sure of its performance metrics of model.

Chapter 4. Results of the performed studies

This chapter summarizes the main findings of the study. The results were successfully obtained using Python. In the following sections, the resulting calculations and graphs will be represented.

4.1 Results of the performed Exploratory Data Analysis to determine the most profitable way to sell products on the global market

To determine most profitable way sell products have been used essential columns of global market dataset: units sold, total profit, sales channel and region. These columns played a main role in the working process to obtain decision problem. Figures 23 and 24 presents the quantity of the units sold products by regions and years.

REGION	UNITS_SOLD
Europe	131866843
Sub-Saharan Africa	125310746
Asia	75626139
Middle East and North Africa	52056553
Central America and the Caribbean	45464509
Australia and Oceania	37890346
North America	7929433

Fig 23. Quantity units sold products by years

year	UNITS_SOLD
2013	63590951
2014	63100285
2016	62894265
2015	62823893
2012	62497188
2011	62046529
2010	58272104
2017	40919354

Fig 24. Quantity units sold products by years

To obtain a profitable way of selling products have to aggregate columns units sold and total profit by the year, region and sales channel. These results will clarify which way of selling products more comfortable, thrifty, and profitable to the global market. Figures 24 and 25 present the graph quantity of the units sold products by sales channel and years, and graph obtained total profit on the units sold products by

sales channel and years.

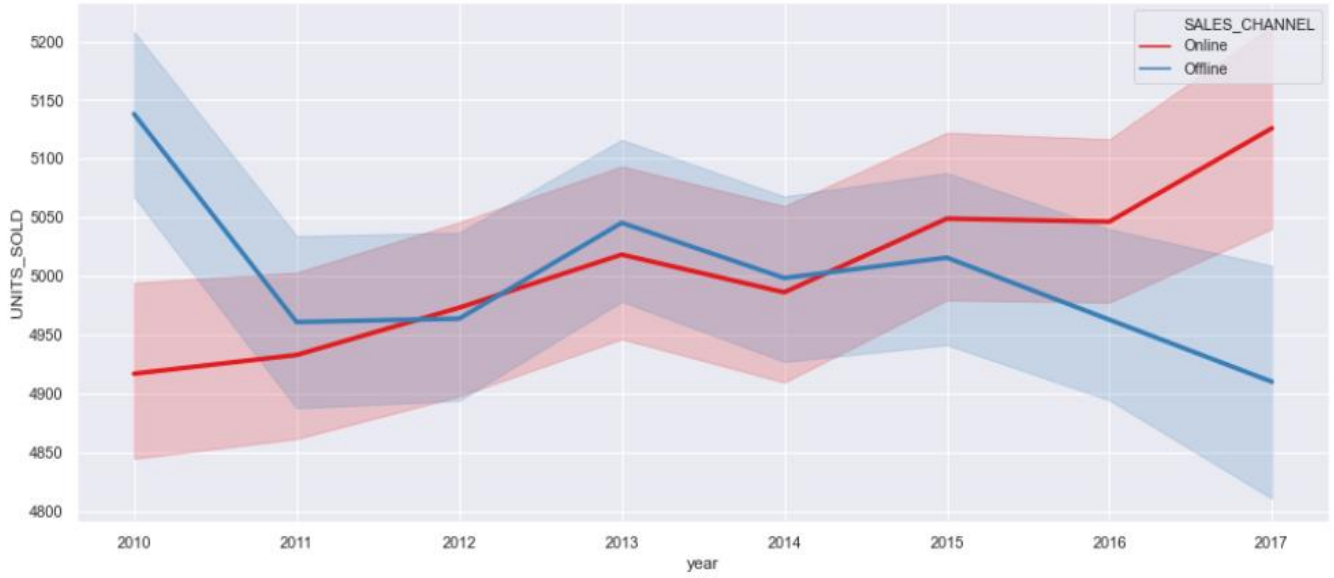

Fig 24. Graph units sold products by sales channel and years

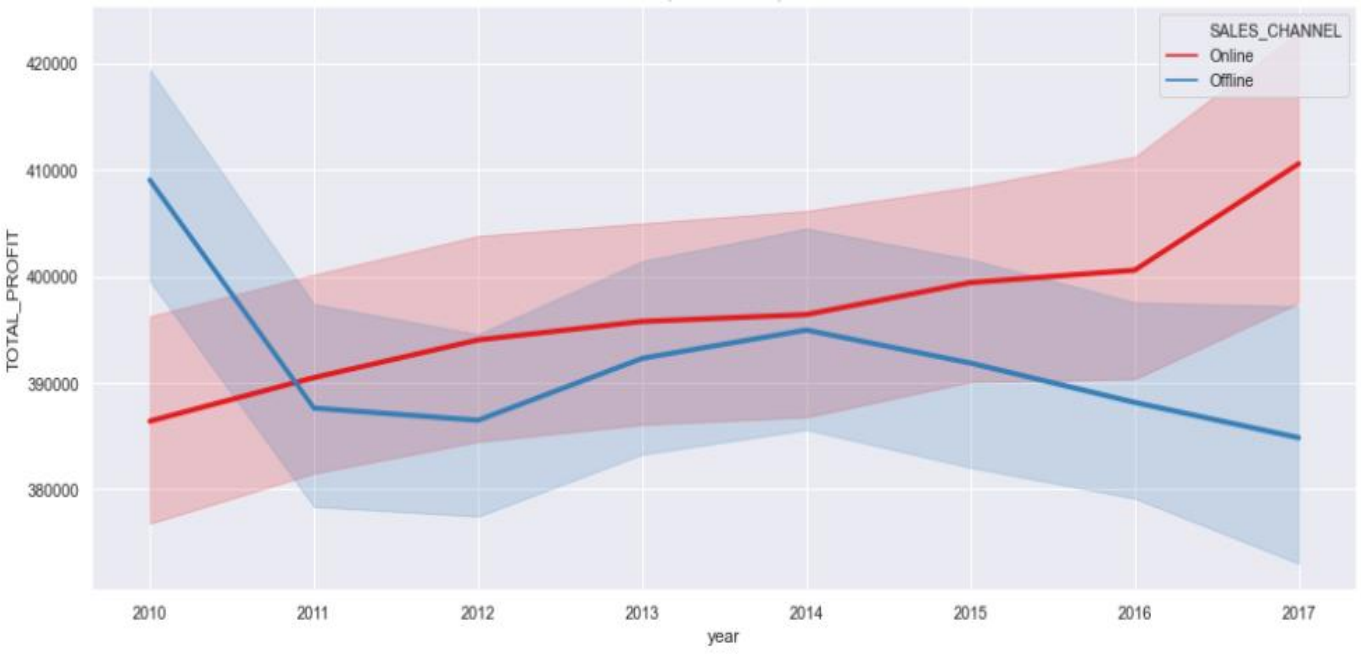

Fig 25. Graph total profit from units sold products by sales channel and years

The basic significance of the units sold and total profit on the products located in aggregate column region by sales channel and year. This outcome is more comprehensible to the final selection most profitable way. Figures 26 and 27 presents the graph quantity of the units sold products by region, sales channel and

years, and graph total profit on the units sold products by region, sales channel and years.

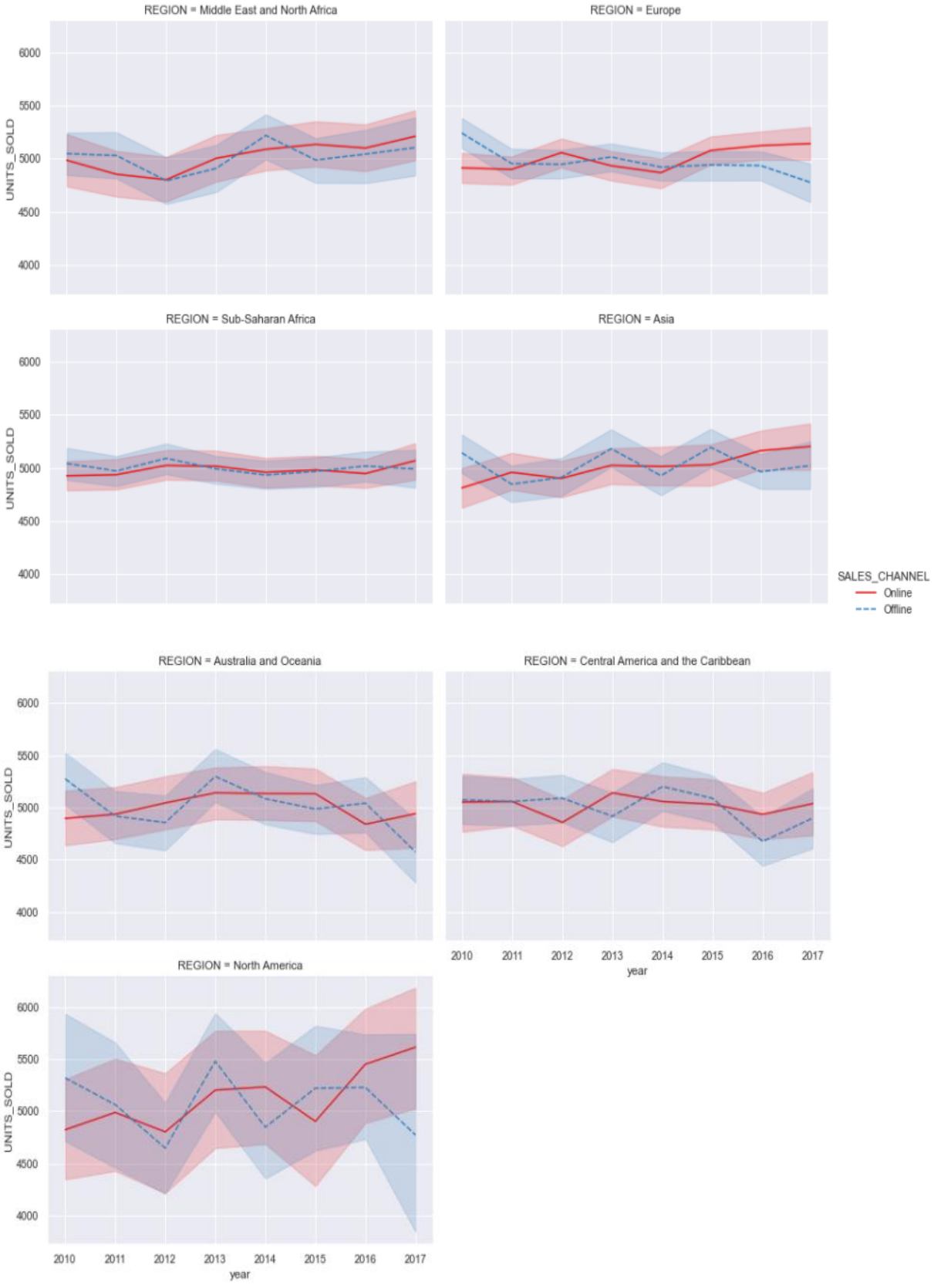

Fig 26. Graph units sold products by region, sales channel and years

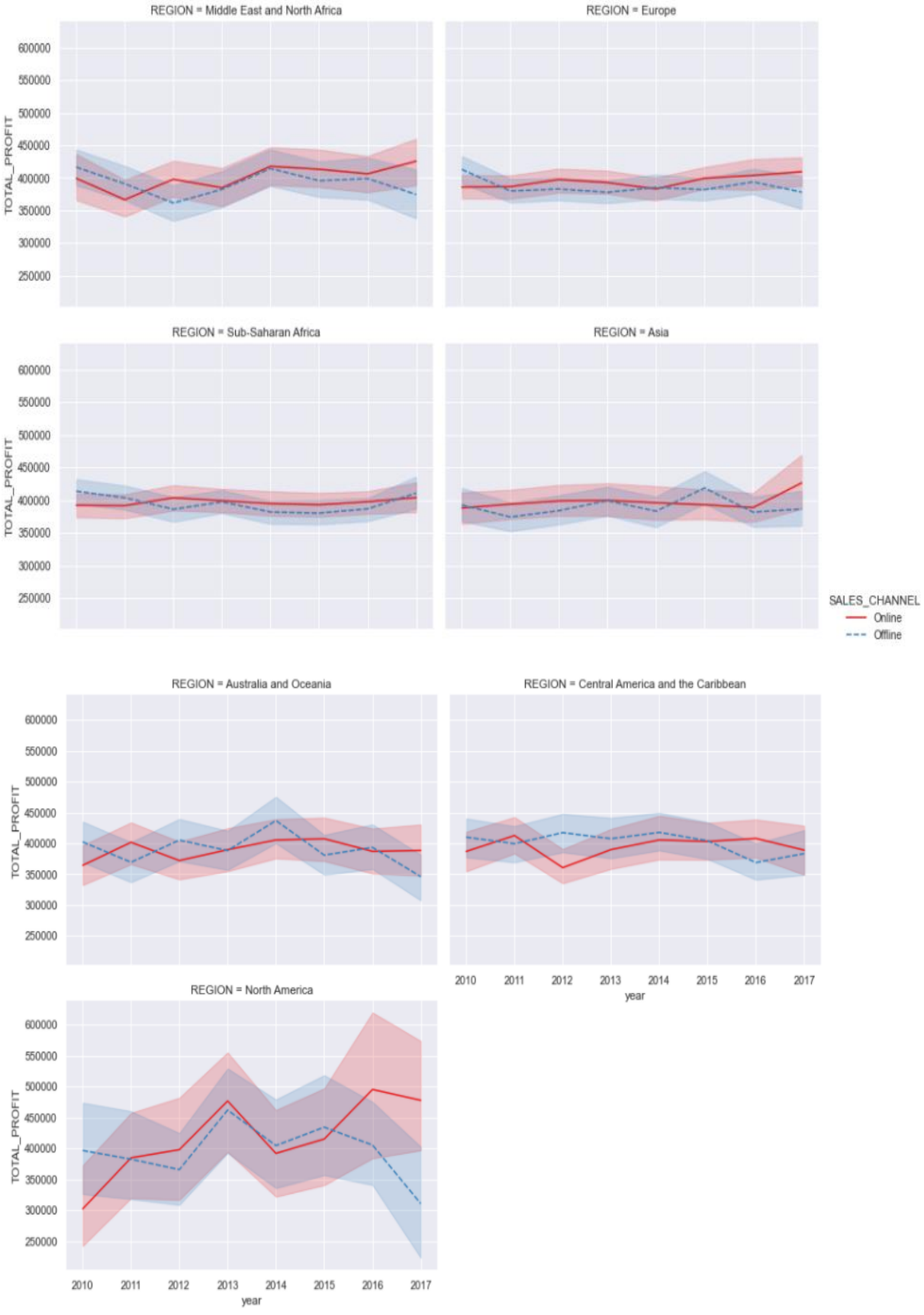

Fig 27. Graph total profit on the units sold products by region, sales channel and years

After getting multiple results, we are able to offer marketers or customers a more convenient and safer way to sell the product. In addition, we have determined that the best way to sell a product is through the online sales channel.

There are several reasons why we have chosen this path for product sales:

- At the moment, civilization of the world has deep ties to innovative technology.
- The last decade in the status of a business market has changed the ways of selling and ordering a product, the use of technology.
- In addition, in the future, innovative technologies will play a major role in business marketing, which to a greater extent ensures quality, safety, comfort and profitability for future marketers and customers.

4.2 Results of the performed forecasting procedure based on the Machine Learning algorithm

Before performing the forecasting procedure, it is necessary to review the hypothesis on historical data. Let us plot the hypotheses between the total profit by sales channel. Figure 28 present the plot the hypotheses between the total profit by sales channel:

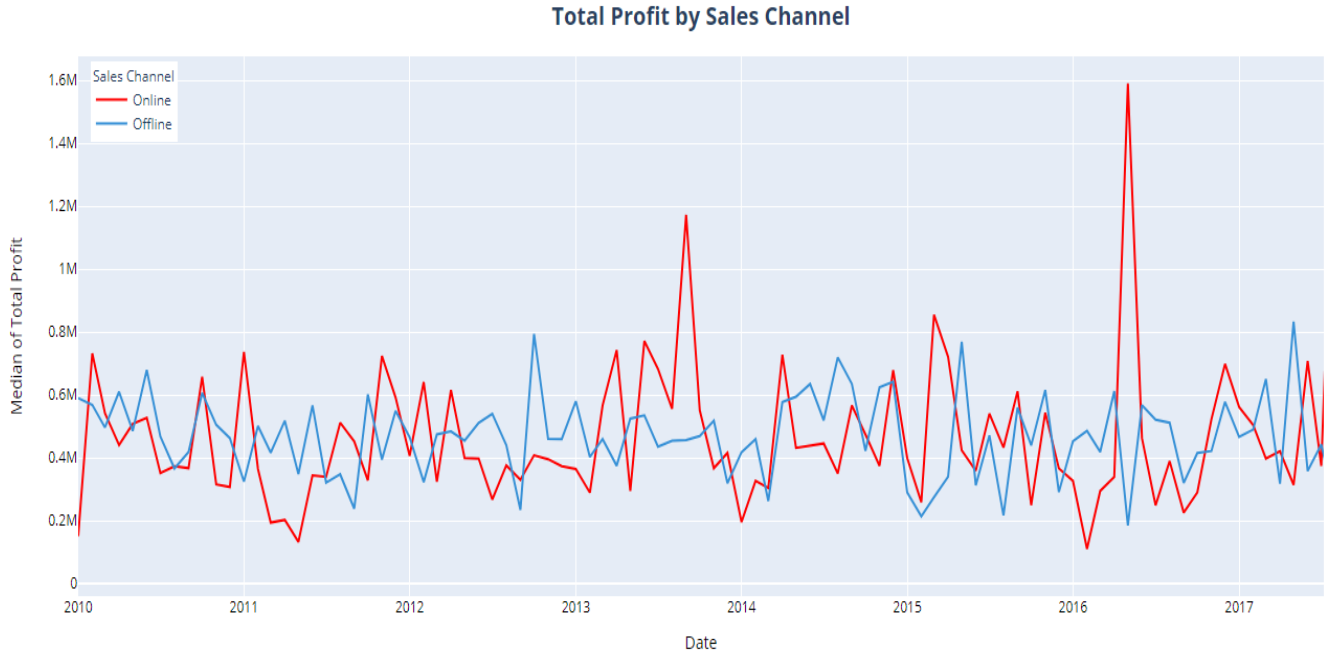

Fig 28. Graph total profit on the units sold products by region, sales channel and years

Understanding workflow of Prophet machine learning algorithm was important as realize definition, mathematical aspects, and working process of algorithm because that's playing a principal role to understand advantages and disadvantages of prophet's algorithm. The following stage is forecasting procedure apply of machine learning algorithm. Suppose we need to make a forecast of the total profits by sales channel for the next 3 years that worth define total profit by sales channel. Matching with the prophet's algorithm will sort out that problem a much better way. Figure 29 present forecasting the outcome:

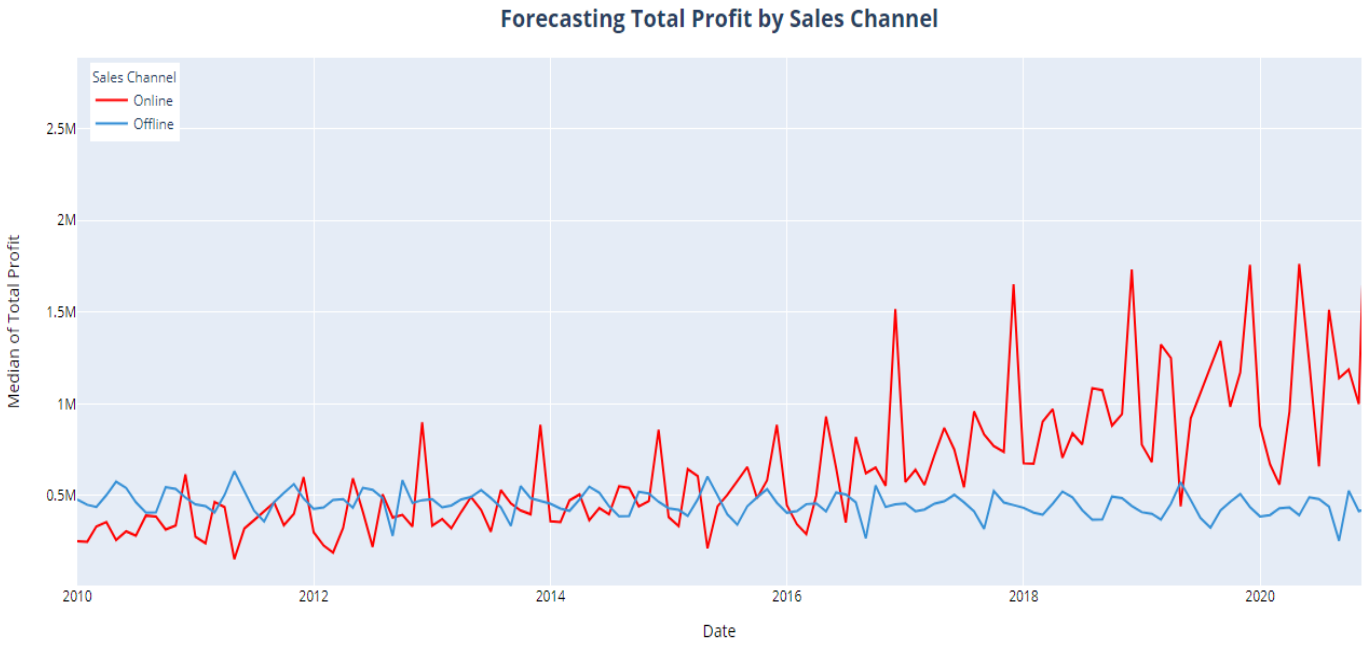

Fig 29. Forecasting the outcome of the Prophet machine learning algorithm

Conformity of forecasting procedure outcomes the plot of the hypothesis total profit by sales channel entirely changed in the next 3 years. The supposable hypothesis of total profit by online sales channel will grow in the next 3 years and the hypothesis of total profit by offline sales channel will reduce or will remain regular. Nevertheless, the major and essential issues of the current work were performing forecasting procedure the total profit through the sales channels, and the successful results obtained are consistent with the current statistics of the global market.

Conclusion

The global market data analysis is based on business marketing. However, some aspects of the global market remain problematic. The ways of selling and ordering products through innovative technology not be reducing or worsened, conversely will improves from every side. The dataset introduces a tidy representation of global market data that explicitly conveys the relation between market and technology. Furthermore, carried out exploratory data analysis shown confidence outcomes. Supposable until 2014 according to the indicators global market dataset online selling products has been uncomfortable, unsafely and was not popular. However, it only began improvement technology and start certainly impact to the global market. We used techniques of machine learning to the operation of a forecast total profit of global market. In our simplified forecasting procedure, we subdivided the major feature global market dataset the total profit into two categories of sales channels: Online, offline. The major objective of the work was performing the corresponding action to identify which hypothesis total profit through a sales channel going to grow or reduce in the next 3 years. Moreover, outcomes present the hypothesis total profit through the online sales channel will grow, however, the hypothesis total profit through the offline sales channel will remain as it was before or reduce.

Chapter 5. Financial management, resource efficiency and resource saving

Research master's thesis is a scientific work related to scientific research, conducting research in order to obtain scientific generalizations, finding principles, and ways to create (modernize) products. Currently, the prospect of scientific research is not determined to scale of the discovery that in the early stages of the life cycle of high-tech and the resource-efficient product is hard enough to reach. Therefore, commercial value is as important as development research. Evaluation of the commercial value (potential) of the development is a necessary condition when searching for sources of funding for scientific research and commercialization of its results. The commercial value is essential for developers who need to understand the state and prospects of ongoing research.

The purpose of this section is to discuss the issues of competitiveness, resource efficiency and resource saving, as well as financial costs regarding the object of the study of the Master thesis. The competitiveness analysis is carried out for this purpose. The SWOT analysis helps to identify strengths, weaknesses, opportunities and threats associated with the project, and decide how to deal with them in each particular case. The development of the project requires funds that go to thesalaries of project participants and the necessary equipment (the list is given in the respective section). The calculation of the resource efficiency indicator helps to make a final assessment of the technical decision on individual criteria and in general.

5.1 Pre-project analysis

Nowadays, the perspective of scientific research is determined not so much by the scale of discovery, which is difficult to estimate at the first stages of the life cycle of a high-tech and resource-efficient product, but by the commercial value of the development. Assessment of the commercial value of the development is a necessary condition when searching for sources of financing for scientific research and commercialization of its results. It is important for developers, who should

represent the state and prospects of ongoing scientific research.

It is necessary to understand that the commercial attractiveness of scientific research is determined not only by the excess of technical parameters over previous developments but also by how quickly the developer will be able to find answers to such questions - whether the product will be in demand in the market, what will be its price, what is the budget of the scientific project, how long it will take to enter the market, etc.

The achievement of the goal is ensured by solving the tasks:

- evaluation of the commercial potential and prospects of scientific research;
- identifying possible alternatives to scientific research that meets current resource efficiency and resource conservation requirements;
- research planning;
- resource (resource-saving), financial, budgetary, social, and economic efficiency of research.

5.2 Competitiveness analysis of technical solutions

In order to find sources of financing for the project, it is necessary, first, to determine the commercial value of the work. The analysis of competitive technical solutions in terms of resource efficiency and resource saving allows us to evaluate the comparative effectiveness of the scientific development. This analysis is advisable to carry out using an evaluation card.

This analysis was carry out using an evaluation card (see Table 1). Two competitive developments were selected for this. The criteria for comparing and evaluating resource efficiency and resource conservation, shown in Table 1, were

selected based on the selected objects of comparison, taking into account their technical and economic features of development, creation and operation.

First, it is necessary to analyze possible technical solutions and choose the best one based on the considered technical and economic criteria.

The evaluation map analysis is presented in Table 1. The position of your research and competitors is evaluated for each indicator by you on a five-point scale, where 1 is the weakest position and 5 is the strongest. The weights of the indicators determined by you in the amount should be 1. The analysis of competitive technical solutions is determined by the formula:

$$C = \sum W_i \cdot P_i$$

C - the competitiveness of research or a competitor;Wi– criterion weight;

Pi – point of i-th criteria.

Table 1. Evaluation card for comparison of competitive technical solutions

Evaluation criteria ***example***	**Criterion weight**	**Points**			**Competitiveness** Taking into account weight coefficients		
		P_f	P_{i1}	P_{i2}	C_f	C_{i1}	C_{i2}
1	2	3	4	5	6	7	8
Technical criteria for evaluating resource efficiency							
1. Energy efficiency	0,18	4	2	2	0,72	0,36	0,36
2. Reliability	0,13	5	2	5	0,65	0,26	0,65
3. Safety	0,2	4	4	2	0,4	0,4	0,25
4. Functional capacity	0,14	3	5	5	0,42	0,7	0,7
Economic criteria for performance evaluation							
1. Development cost	0,2	4	4	4	0,4	0,4	0,4
2. Market penetration rate	0,08	5	4	4	0,5	0,4	0,4

3. Expected lifecycle	0,07	3	5	2	0,18	0,3	0,15
Total	1	28	26	24	3,27	2,82	2,91

This analysis suggests that the study is effective because it provides acceptable quality results. Further investment in this development can be considered reasonable.

5.3 SWOT analysis

The complex analysis solution with the greatest competitiveness is carried out with the method of the SWOT analysis: Strengths, Weaknesses, Opportunities and Threats. The analysis has several stages. The first stage describes the strengths and weaknesses of the project, identifies opportunities and threats to the project that have emerged or may appear in its external environment. The second stage identifies the compatibility of the strengths and weaknesses of the project with the external environmental conditions. This compatibility or incompatibility should help to identify what strategic changes are needed.

The SWOT analysis of this research project is presented in Table 2.

Table 2. Matrix of SWOT-analysis

	Strengths: S1. Forecasting total profit products for the global market. S2. Representing statistics and analysis global marketing. S3. Revealing patterns of development global marketing.	**Weaknesses:** W1. With a large amount of information, the requirement powerful computing technology. W2. Insufficient data on regions in the global market
Opportunities: O1. Using machine learning technology for basic analysis of global marketing. O2. Development of the global market.	1. Compilation of a more efficient machine learning algorithm for the development of global market. 2. Analysis and correction of the program development for the development of global market.	1. Obtaining licensed software from potential partners 2. The ability to adapt new software for less powerful computers
Threats: T1. Lack of demand for innovative technologies.	1. Threats to the project are related to temporary difficulties. In this case, in the long term the future	1. To implement this development financial and labor personnel costs, which is a deterrent on the way

T2. Lack of the possibility of implementation. T3. Slow program performance.	development of the project based on its economic efficiency and technological benefits are integral development of a complex of emergency automation. Competent developer support will reduce the likelihood of system slow operation.	implementation of development.

Based on the results of the analysis of this matrix, it can be concluded that the difficulties and challenges that this research project may face in one way or another can be addressed by the existing strengths of the research.

5.4 Project Initiation

The initiation process group consists of processes that are performed to define a new project or a new phase of an existing one. In the initiation processes, the initial purpose and content are determined and the initial financial resources are fixed. The internal and external stakeholders of the project who will interact and influence the overall result of the research project are determined.

5.4.1 Project objectives and results

This section describes the project stakeholders, the hierarchy of project objectives and the criteria for achieving the objectives.

Project stakeholders refer to individuals or organizations that are actively involved in the project or whose interests may be affected positively or negatively during project implementation or completion. Information about project stakeholders provides in Table 3.

Table 3. Stakeholders of the project

Project stakeholders	Stakeholder expectations
Exploratory Data Analysis	Selecting the most profitable method of sale of products on the global market
Machine Learning	Baseline forecast of the total profit of the global market using machine learning

Table 4 shows information on the hierarchy of project objectives and criteria for achieving the objectives.

Table 4. Purpose and results of the project

Purpose of project:	Machine learning of the global market
Expected results of the project:	The machine learning algorithm for performing the operation of forecasting the total profit of the global market corresponds to historical data
Criteria for acceptance of the project result:	The Facebook Prophet machine learning algorithm with high precision to performing forecasting operation according to the algorithm was obtained high precision forecasting a total profit of the global market
Requirements for the project result:	The research carried out under this project should be completed by May 25, 2021
	The results obtained must meet the criteria for accepting the project result
	If unsatisfactory results are obtained, additional studies should be conducted using received results

5.5 Organizational structure of the project

It is necessary to solve some questions: who will be part of the working group of this project, determine the role of each participant in this project, and prescribe the functions of the participants and their number of labor hours in the project.

Table 5. Participant of the project

№	Participant	Role in the project	Functions	Labor time, hours.
1	Gubin E. I. PhD, Associate Professor TPU	Supervisor	Consultations. Drafting and approval terms of technique task. Review master's dissertation.	1010 hours
2	Iskandar B. S. Master student, TPU	Executor	Scheduling of works on the topic. Selection and study of materials on the topic. Analysis of initial data. Choosing a method of performing work. Writing a program. Analysis of the obtained results of	3450 hours

			work Drawing up a report on the work.	
Total:				4460hours

5.6 Project limitations

Project limitations are all factors that can be as a restriction on the degreeof freedom of the project team members. Table 6 shows the main limitations of this work.

Table 6. Project limitations

Factors	**Limitations / Assumptions**
3.1. Project's budget	198000 rub
3.1.1. Source of financing	TPU
3.2. Project timeline:	22/07/2020 to 14/05/2021
3.2.1. Date of approval of plan of project	20/03/2021
3.2.2. Completion date	25/05/2021

5.7 Project Schedule

As part of planning a science project, you need to build a project timeline and a Gantt Chart.

Table 7. Project Schedule

Job title	Duration, working days	Start date	Date of completion	Participants
Preparation of technical specifications and choice of research direction	**41 days**	November 1, 2019.	December 28, 2019.	Supervisor
Selection and study of materials on the topic	**198 days**	December 27, 2019.	January 9, 2020.	Student
Calendar planning of work on the topic	**345 days**	November 1, 2019.	June 1, 2021.	Head of Division

Description and analysis subject area	**60 days**	March 1, 2020.	May 31, 2020.	Supervisor/ Student
Obtaining the necessary analysis data and checking the outcomes	**205 days**	June 1, 2020	March 27,2021.	Supervisor/ Student
Processing the received data	**140 days**	September 2, 2020.	March 27, 2020.	Student
Writing a master's thesis	**96 days**	December 2, 2021.	May 29, 2021.	Student
Master's thesis defense	**14 days**	June 15, 2021.	June 28, 2021.	Student

A Gantt chart, or harmonogram, is a type of bar chart that illustrates a project schedule. This chart lists the tasks to be performed on the vertical axis, and time intervals on the horizontal axis. The width of the horizontal bars in the graph shows the duration of each activity.

On the basis of Table 8, a time schedule is constructed (see Table 9). The schedule is constructed for the maximum duration of the research project by months and decades (10 days) for the period of writing the diploma. At the same time, the works on the schedule should be distinguished by different shading depending on the executors responsible for this or that work (– Supervisor, – student).

Table 8. Gantt chart of Project Schedule

№	Activities	Participants	T_c, days	Duration of the project								
				March			April			May		
				1	3	27	21	1	23	5	4	10
1	Drafting and approval of the graduation work assignment	Supervisor/ Student	1									
2	Scheduling of works on the topic	Student	3									

3	Selection and study of materials on the topic	Student	27										
4	Analysis of initial data	Student	21										
5	Choosing a method of performing work	Supervisor/ Student	1										
6	Writing a program	Student	23										
7	Testing the program	Student	5										
8	Analysis of work results	Supervisor/ Student	4										
9	Drawing up a report on the work	Student	10										

5.8 Scientific and technical research budget

The amount of costs associated with the implementation of this work is the basis for the formation of the project budget. This budget will be presented as the lower limit of project costs when forming a contract with the customer.

To form the final cost value, all calculated costs for individual items relatedto the manager and the student are summed.

In the process of budgeting, the following grouping of costs by items is used:

- material costs of scientific and technical research;
- costs of special equipment for scientific work (Depreciation ofequipment used for design);
- basic salary;
- additional salary;
- labor tax;
- overhead.

5.8.1 Calculation of material costs

The calculation of material costs is carried out according to the formula:

$$C_m = (1 + k_T) \cdot \sum_{i=1}^{m} P_i \cdot N_{consi}$$

where m – the number of types of material resources consumed in theperformance of scientific research;

N_{consi} – the amount of material resources of the i-th species planned to be used when performing scientific research (units, kg, m, m^2, etc.);

P_i – the acquisition price of a unit of the i-th type of material resources consumed (rub./units, rub./kg, rub./m, rub./m^2, etc.);

k_T– coefficient taking into account transportation costs.

Prices for material resources can be set according to data posted on relevant websites on the Internet by manufacturers (or supplier organizations).

Energy costs are calculated by the formula:

$$C = P_{el} \cdot P \cdot F_{eq}$$

where P_{el} − power rates (5.8 rubles per 1 kWh for Tomsk);

P − power of equipment, kW;

F_{eq} − equipment usage time, hours.

Table 9. Material costs

Name	Unit	Amount	Price per unit, rub.	Material costs, rub.
Papers		300	1	300
Electricity of computer	kWh	150	5.8	870
Printing on A4 sheet		400	4	1,900

Pen		10	40	400
Internet	Month	10	356	3,560
Total				7,030

5.8.2 Basic salary

This point includes the basic salary of participants directly involved in the implementation of the work on this research. The value of salary costs is determined based on the labor intensity of the work performed and the current salary system

The basic salary (S_b) is calculated according to the following formula:

$$S_b = S_a \cdot T_w$$

where S_b – basic salary per participant;

T_w– the duration of the work performed by the scientific and technical worker, working days;

S_a - the average daily salary of an participant, rub.

The average daily salary is calculated by the formula:

$$S_d = \frac{S_m \cdot M}{F_v}$$

where

S_m – monthly salary of an participant, rub.;

M – the number of months of work without leave during the year:

at holiday in 48 days, M = 10.4 months, 6 day per week;

at holiday in 24 days, M = 11.2 months, 5 day per week;

F_v – valid annual fund of working time of scientific and technical staff.

Table 10. The valid annual fund of working time

Working time indicators	**Working group of the project**
Calendar number of days	365
The number of non-working days - weekend - holidays	104 14
Loss of working time - vacation - sick absence	52
The valid annual fund of working time	195

Monthly salary is calculated by formula:

$$S_{month} = S_{base} \cdot (k_{premium} + k_{bonus}) \cdot k_{reg}$$

where S_{base} - base salary, rubles;

$k_{premium}$ - premium rate;

k_{bonus} - bonus rate;

k_{reg} - regional rate.

Table 11. Calculation of the base salaries

Performers	*Sbase*, rubles	*kpremium*	*kbonus*	*kreg*	*Smonth*, rub.	W_d, rub.	Tp, work days	*Wbase*, rub.
Supervisor	33300	0,3	1,28	1,3	56277	2360	5	11800
Student	14874	0,3	1,28	1,3	25137	1054	77	81168,01
Total:								92968,01

5.8.3 Additional salary

This point includes the amount of payments stipulated by the legislation on labor, for example, payment of regular and additional holidays; payment of time associated with state and public duties; payment for work experience, etc.

Additional salaries are calculated on the basis of 10-15% of the base salary of workers:

$$W_{add} = k_{extra} \cdot W_{base}$$

where W_{add} – additional salary, rubles; k_{extra} – additional salary coefficient;W_{base} – base salary, rubles.

Additional salary of the supervisor:

$$W_{add} = 0{,}15 \cdot 11800 = 1770\ RUB$$

Additional salary of the student:

$$W_{add} = 0{,}15 \cdot 81168{,}01 = 12175{,}2\ RUB$$

Total additional salary is equal 13945,2 rub.

5.8.4 Labor tax

Tax to extra-budgetary funds are compulsory according to the norms established by the legislation of the Russian Federation to the state social insurance (SIF), pension fund (PF) and medical insurance (FCMIF) from the costs of workers.

Payment to extra-budgetary funds is determined of the formula:

$$P_{social} = k_b \cdot (W_{base} + W_{add})$$

where k_b – coefficient of deductions for labor tax.

In accordance with the Federal law of July 24, 2009 No. 212-FL, the amount of insurance contributions is set at 30%. Institutions conductingeducational and scientific activities have rate - 27.1%.

Table 12. Labor tax

	Project leader	**Engineer**
Coefficient of deductions	27,1%	
Salary, rubles	11800	81168,01
Labor tax, rubles	3400,92	10890
Total:	107258,93	

5.8.5 Overhead costs

Overhead costs include other management and maintenance costs that can be allocated directly to the project. In addition, this includes expenses for the maintenance, operation and repair of equipment, production tools and equipment, buildings, structures, etc.

Overhead costs account from 30% to 90% of the amount of base and additional salary of employees.

Overhead is calculated according to the formula:

$$C_{ov} = k_{ov} \cdot (W_{base} + W_{add})$$

where k_{ov} – overhead rate.

Table 13. Overhead

	Project leader	**Engineer**
Overhead rate	30%	

Salary, rubles	11800	81168,01
Overhead, rubles	22069,02	88036,11
Total:	203073,14	

5.9 Formation of budget costs

The calculated amount of research costs is the basis for the formation of a project cost budget, which, when forming a contract with a customer, is protected by a scientific organization as the lower limit of costs for the development of scientific and technical products. Determining the budget for the scientific research is given in the table 14.

Table 14. Budget for scientific and technical research.

Name	**Cost, rubles**
1. Material costs	7030
2. Basic salary	92968,01
3. Additional salary	13945,2
4. Labor tax	107258,93
5. Overhead	203073,14
Total planned cost	424275.28

Conclusion

Thus, in this section we developed stages for the design and creation of the competitive development that meets the requirements of the field of resource efficiency and resource saving.

These stages include:

- development of the economic project idea, formation of the project concept;
- organization of the work on the research project;
- identification of possible research alternatives;
- research planning;
- assessing the commercial potential and prospects of scientific research from the standpoint of resource efficiency and resource saving;
- determination of resource (resource saving), financial, budget, social and economic efficiency of the project.

In the course of performing the economic part of the qualification masterwork, calculations were made of the planned cost of research and the time spent.

Chapter 6. Social responsibility

Introduction

The developed project is intended for baseline forecast total profit of global market using machine learning. The development of the program was carried out only with the help of computer.

In this section, harmful and dangerous factors affecting the work of personnel will be considered, the impact of the developed program on the environment, legal and organizational issues, measures in emergency situations will be considered.

The work was carried out in the hall of residence of TPU (16th floor). Room 1607A was a research execution place.

The layout of the room is shown in Figure 1

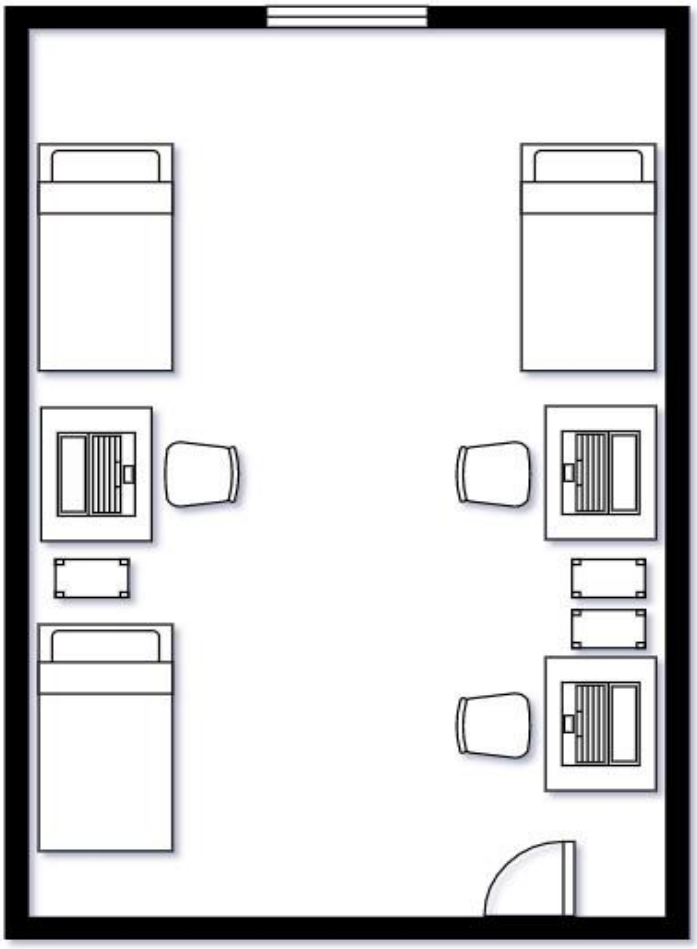

Figure 1. Room layout 1607A

6.1 Legal and organizational issues in providing safety

Nowadays one of the main way to radical improvement of all prophylactic work referred to reduce Total Incidents Rate and occupational morbidity is the widespread implementation of an integrated Occupational Safety and Health management system. That means combining isolated activities into a single systemof targeted actions at all levels and stages of the production process.

Occupational safety is a system of legislative, socio-economic,organizational, technological, hygienic and therapeutic and prophylactic measures and tools that ensure the safety, preservation of health and human performance in the work process.

According to the GOST 12.2.032-78 SSBT [1], every employee has the right:

- To have a workplace that meets Occupational safety requirements;
- To have a compulsory social insurance against accidents at manufacturing and occupational diseases;
- To receive reliable information from the employer, relevant government bodies and public organizations on conditions and Occupational safety at the workplace, about the existing risk of damage to health, as well as measures to protect against harmful and (or) hazardous factors;
- To refuse carrying out work in case of danger to his life and health due to violation of Occupational safety requirements;
- Be provided with personal and collective protective equipment in compliance with Occupational safety requirements at the expense of the employer;
- For training in safe work methods and techniques at the expense of the employer;
- For personal participation or participation through their representatives in consideration of issues related to ensuring safe working conditions in his workplace, and in the investigation of the accident with him at work or occupational disease;
- For extraordinary medical examination in accordance with medical

recommendations with preservation of his place of work (position) and secondary earnings during the passage of the specified medical examination;

- For warranties and compensation established in accordance with this Code, collective agreement, agreement, local regulatory an act, an employment contract, if he is engaged in work with harmful and (or) hazardous working conditions.

The labor code of the Russian Federation states that normal working hours may not exceed 40 hours per week, The employer must keep track of the time worked by each employee.

Rules for labor protection and safety measures are introduced in order to prevent accidents, ensure safe working conditions for workers and are mandatory for workers, managers, engineers and technicians.

6.2 Basic ergonomic requirements for the correct location and arrangement of researcher's workplace

The workplace when working with a PC should be at least 6 square meters. The legroom should correspond to the following parameters: the legroom height is at least 600 mm, the seat distance to the lower edge of the working surface is at least 150 mm, and the seat height is 420 mm. It is worth noting that the height of the table should depend on the growth of the operator.

The following requirements are also provided for the organization of the workplace of the PC user: The design of the working chair should ensure the maintenance of a rational working posture while working on the PC and allow the posture to be changed in order to reduce the static tension of the neck and shoulder muscles and back to prevent the development of fatigue.

The type of working chair should be selected taking into account the growth of the user, the nature and duration of work with the PC. The working chair should be lifting and swivel, adjustable in height and angle of inclination of the seat and back,

as well as the distance of the back from the front edge of the seat, while the adjustment of each parameter should be independent, easy to carry out and have a secure fit [2].

6.3 Occupational safety

Workplace safety is the responsibility of everyone in the organization.

Occupational hygiene is a system of ensuring the health of workers in the process of labor activity, including legal, socio-economic, organizational and technical, sanitary and hygienic, treatment and prophylactic, rehabilitation and other measures.

Working conditions - a set of factors of the working environment and the labor process that affect human health and performance.

Harmful production factor is a factor of the environment and the work process that can cause occupational pathology, temporary or permanent decrease in working capacity, increase the frequency of somatic and infectious diseases, and lead to impaired health of the offspring.

Hazardous production factor is a factor of the environment and the labor process that can cause injury, acute illness or sudden sharp deterioration in health, death.

In this subsection it is necessary to analyze harmful and hazardous factors that can occur during research in the laboratory, when development or operation of the designed solution (on a workplace).

GOST 12.0.003-2015 "*Hazardous and harmful production factors. Classification"* must be used to identify potential factors, that can effect on a worker (employee).

Table 1 - Potential hazardous and harmful production factors

Factors (**GOST 12.0.003-2015)**	Stages of work			Legislation documents
	developing	manufacturing	operation	
1. Excessive levels of noise, vibration	+	+		**GOST 12.1.003-2014** Occupational safety standards system. Noise. General safety requirements
2. Insufficient illumination	+			**SanPiN 2.2.1/2.1.1.1278-03** Hygienic requirements for natural, artificial and mixed lighting of residential and public buildings
3. Electromagnetic fields	+	+	+	**SanPiN 2.2.4.1329-03** Requirements for protection of personnel from the impact of impulse electromagnetic fields
4. Abnormally high voltage value in the circuit, the closure which may occur through the human body		+	+	**Sanitary rules GOST 12.1.038-82 SSBT.** Electrical safety. Maximum permissible levels of touch voltages and currents.

6.3.1 Excessive levels of noise, vibration

Noise and vibration worsen working conditions; have a harmful effect on the human body, namely, the organs of hearing and the whole body through the central nervous system. It result in weakened attention, deteriorated memory, decreased response, and increased number of errors in work.

Noise can be generated by operating equipment, air conditioning units, daylight illuminating devices, as well as spread from the outside.

When working on a PC, the noise level in the workplace should not exceed 50 dB [3].

6.3.2 Insufficient illumination

Light sources can be both natural and artificial. The natural source of the light in the room is the sun, artificial light are lamps. With long work in low illumination conditions and in violation of other parameters of the illumination, visual perception decreases, myopia, eye disease develops, and headaches appear [4].

According to the SanPiN 2.2.2 / 2.4.1340-03 [4] standard, the illumination on the table surface in the area ofthe working document should be 300-500 lux. Lighting should not create glare on the surface of the monitor. Illumination of the monitor surface should not be more than 300 lux.

The brightness of the lamps of common light in the area with radiation angles from 50 to 90° should be no more than 200 cd/m, the protective angle of the lamps should be at least 40°. The ripple coefficient should not exceed 5%.

6.3.3 Electromagnetic fields

In this case, the sources of increased intensity of the electromagnetic field are a personal computer. 8- is considered acceptable. An hour's working day for an employee at his workplace, with the maximum permissible level of tension, should be no more than 8 kA / m, and the level of magnetic induction should be 10 mT. Compliance with these standards makes it possible to avoid the negative effects of electromagnetic radiation.

To reduce the level of the electromagnetic field from personal it is recommended to connect no more than two computers to one outlet, make a protective grounding, connect the computer to the outlet through an electric field neutralizer.

Personal protective equipment when working on a computer includes spectral computer glasses to improve image quality and Protection against excessive energy flows of visible light and for Prof. Glasses reduce eye fatigue by 25-30%.

They are recommended to be used by all operators when working more than 2 hours a day, and in case of visual impairment by 2 diopters or more - regardless of

the duration of work.

Sources of electromagnetic radiation in the workplace are system units and monitors of switched on computers. To bring down exposure to such types of radiation, it is recommended to use such monitors, the radiation level is reduced, as well as to install protective screens and observe work and rest regimes.

According to the intensity of the electromagnetic field at a distance of 50 cm around the screen along the electrical component should be no more than [5]:

- in the frequency range 5 Hz - 2 kHz - 25 V / m;
- in the frequency range 2 kHz - 400 kHz - 2.5 V / m.

The magnetic flux density should be no more than:

- in the frequency range 5 Hz - 2 kHz - 250 nT;
- in the frequency range 2 kHz - 400 kHz - 25 nT.

There are the following ways to protect against EMF:

- increase the distance from the source (the screen should be at least 50 cm from the user);
- the use of pre-screen filters, special screens and other personal protective equipment.

When working with a computer, the ionizing radiation source is a display. Under the influence of ionizing radiation in the body, there may be a violation of normal blood coagulability, an increase in the fragility of blood vessels, a decrease in immunity, etc. The dose of irradiation at a distance of 20 cm to the display is 50 μrem/hr. According to the norms [8], the design of the computer should provide the power of the exposure dose of x–rays at any point at a distance of 0,05 m from the screen no more than 100 μR/h. Fatigue of the organs of vision can be associated with both insufficient illumination and excessive illumination, as well as with the wrong direction of light.

6.3.4 Abnormally high voltage value in the circuit

The mechanical action of current on the body is the cause of electrical injuries. Typical types of electric injuries are burns, electric signs, skin metallization, tissue tears, dislocations of joints and bone fractures.

The following protective equipment can be used as measures to ensure the safety of working with electrical equipment:

- disconnection of voltage from live parts, on which or near to which work will be carried out, and taking measures to ensure the impossibility of applying voltage to the workplace;
- posting of posters indicating the place of work;
- electrical grounding of the housings of all installations through a neutral wire;
- coating of metal surfaces of tools with reliable insulation;
- inaccessibility of current-carrying parts of equipment (the conclusion in the case of electroporation elements, the conclusion in the body of current carrying parts) [6].

6.4 Ecological safety

Presently section discusses the environmental impacts of the project development activities, as well as the product itself as a result of its implementation in production. The software product itself, developed during the implementation of the master's thesis, does not harm the environment either at the stages of its development or at the stages of operation. However, the funds required to develop and operate it can harm the environment.

There is no production in the laboratory. The waste produced in the premises, first of all, can be attributed to paper waste - waste paper, plastic waste, defective parts of personal computers and other types of computers. Waste paper is recommended accumulate and transfer them to waste paper collection points for further processing. Place plastic bottles in specially designed containers.

Modern PCs are produced practically without the use of harmful substances

hazardous to humans and the environment. Exceptions are batteries for computers and mobile devices. Batteries contain heavy metals, acids and alkalis that can harm the environment by entering the hydrosphere and lithosphere if not properly disposed of. For battery disposal it is necessary to contact special organizations specialized in the reception, disposal and recycling of batteries [8].

Fluorescent lamps used for artificial illumination of workplaces also require special disposal, because they contain from 10 to 70 mg of mercury, which is an extremely dangerous chemical substance and can cause poisoning of living beings, and pollution of the atmosphere, hydrosphere and lithosphere. The service life of such lamps is about 5 years, after which they must be handed over for recycling at special reception points. Legal entities are required to hand over lamps for recycling and maintain a passport for this type of waste. An additional method to reduce waste is to increase the share of electronic document management [8].

6.5 Safety in emergency

In the working environment of the PC operator, the following manufactured emergencies may occur [9]:

- Fires and explosions in buildings and communications;
- Collapse of buildings.

Possible natural disasters include meteorological (hurricanes, showers, frosts), hydrological (floods, floods, flooding), and natural fires.

Emergencies of a biological and social nature include epidemics, epizootics, and epiphytotic. Environmental emergencies can be caused by changes in the state, lithosphere, hydrosphere, atmosphere and biosphere as a result of human activities.

The most typical for the object where the working rooms are located, equipped with a personal computer, the emergency is a fire. Premises for work of PC operators according to the classification system of categories premises for explosion and fire hazard belongs to category D (out of 5 categories A, B, B1-B4, D, D), because applies to premises with non-combustible substances and materials in a cold

state[12].

All employees of the organization must be familiar with the fire safety instructions, undergo safety instructions and strictly observe it. It is forbidden to use electrical appliances in conditions that do not meet the requirements of the manufacturer's instructions, or have various kinds of malfunctions that, in accordance with the instructions for use, may lead to a fire, as well as use electrical wires and cables with damaged or lost protective properties of insulation.

Before leaving the office, it is required to inspect it, close the windows, and make sure that there are no sources of possible ignition in the room, all electrical appliances are turned off and the lighting is turned off.

With a frequency of at least once every three years, it is necessary to measure the insulation resistance of current-carrying parts of power and lighting equipment. The increase in sustainability is achieved through the implementation of appropriate organizational and technical measures, training of personnel to work in emergencies[11].

Upon detecting a fire or signs of combustion (smoke, burning smell, temperature increase, etc.), an employee must:

- It is required to stop work, call the fire department by phone "01";
- If possible, take measures to evacuate people and material values;
- Disconnect electrical equipment from the mains;
- Start extinguishing the fire with the available fire extinguishing means;
- Inform the immediate or superior supervisor and notify the surrounding employees;
- In case of a general signal of danger, leave the building in accordance with the "Plan for the evacuation of people in case of fire and other emergencies."

To extinguish a fire, use manual carbon dioxide fire extinguishers (type OU-2, OU-5) located in the office premises, and a fire hydrant internal fire-fighting water supply. They are designed to extinguish the initial fires of various substances and materials, with the exception of substances that burn without air access. Fire

extinguishers must be kept in good working order at all times and ready for action. It is strictly forbidden to extinguish fires in office premises using chemical foam fire extinguishers (type OHP-10) [11].

Conclusion

Each employee must carry out professional activities with taking into account social, legal, environmental and cultural aspects, issues health and safety, be socially responsible for the solutions, be aware of the need for sustainable development.

In presently section covered the main issues of observance of rights employee to work, compliance with the rules for labor safety, industrial safety, ecology and resource conservation.

It was found that the researcher's workplace satisfies safety and health requirements during project implementation, and the harmful impact of the research object on the environment is not exceeds the norm.

List of publications and speeches

1. Soliev Iskandar Begalievich (Томский политехнический университет) “Choosing the most profitable way to sell products on the global market” // Молодежь и современные информационные технологии: сборник трудов XVIII Международной научно практической конференции студентов, аспирантов и молодых ученых, Томск, 22-26 марта 2021 г.

2. Soliev Iskandar Begalievich (Томский политехнический университет) “Marketing analysis of the global market using machine learning” // Молодежь и современные информационные технологии: сборник трудов XVIII Международной научно практической конференции студентов, аспирантов и молодых ученых, Томск, 22-26 марта 2021 г.

List of references

1. GOST 12.2.032-78 SSBT. Workplace when performing work while sitting. General ergonomic requirements.
2. SanPiN 2.2.2 / 2.4.1340-03. Sanitary-epidemiological rules and standards "Hygienic requirements for PC and work organization".
3. GOST 12.1.003-2014 SSBT. Noise. General safety requirements.
4. SanPiN 2.2.1 / 2.1.1.1278-03. Hygienic requirements for natural, artificial and combined lighting of residential and public buildings.
5. SanPiN 2.2.2 / 2.4.1340-03 "Hygienic requirements for personal computers and work organization ".
6. GOST 12.1.038-82 Occupational safety standards system. Electrical safety.
7. Federal Law "On the Fundamentals of Labor Protection in the Russian Federation" of 17.07.99 № 181 – FZ.
8. GOST R ISO 1410-2010. Environmental management. Assessment of life Cycle. Principles and structure.
9. GOST R12.1.004-85 Occupational safety standards system. Fire safety.
10. GOST 12.2.003-91 Occupational safety standards system. Industrial equipment. General safety requirements.
11. GOST Industrial equipment. General safety requirements to working places.
12. GOST 12.2.003-91 Occupational safety standards system. Industrial equipment. General safety requirements.
13. Alan Talevi, Juan Francisco Morales, Gregory Hather, Jagdeep T Podichetty, Sarah Kim, Peter C Bloomingdale, Samuel Kim, Jackson Burton, Joshua D Brown, Almut G Winterstein, Stephan Schmidt, Jensen Kael White, Daniela J Conrado “Machine Learning in Drug Discovery and Development Part 1: A Primer” Mar 11 2020.

14. Handbook of Marketing Analytics: Methods and Applications in Marketing Management, Public Policy, and Litigation Support | Natalie Mizik, Dominique M. 06 Jun 2021.
15. DTREG. Predictive Modeling Software. Phillip H. Sherrod 20 Mar 2012.
16. Comparative Analysis of Univariate Forecasting Techniques for Industrial Natural Gas Consumption 20 Mar 2018.
17. Dave Chappelle “Big Data & Analytics Reference Architecture”, September 2013.
18. Charlie Berger “Big Data Analytics with Oracle Advanced Analytics”, January 2015.
19. Omaha “Enterprise Application Integration Services” 2019.
20. Siebel Information Development Team “Customer Hub (UCM) Master Data Management Reference”, January 2020.
21. Antonio Fernandez “Forecasting in a Spare Parts Business”, May 14, 2020.
22. Jana Fabiano, Peter Kamari, Miroslav Milner, and Peter Michelin “Using a Software Tool in Forecasting: a Case Study of Sales Forecasting Taking into Account Data Uncertainty” July 24, 2016.
23. Rakesh Kumar Pandey, Anil Kumar Dahiya, Ajay Mandal “Identifying Applications of Machine Learning and Data Analytics Based Approaches for Optimization of Upstream Petroleum Operations”, 23 October 2020.
24. Yannis J. Trakadis Sameer Sardaar Anthony Chen Vanessa Fulginiti Ankur Krishnan “Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes”, April 2018.
25. Eniafe Festus Ayetiran, Adesesan. B. Adeyemo “A Data Mining-Based Response Model for Target Selection in Direct Marketing”, February 2012.
26. Cora Nelson “GLOBAL COMPANY PROFILE Walt Disney Parks & Resorts: Forecasting Provides a Competitive Advantage for Disney”, 2016.
27. Saratendu Sethi “Machine Learning with SAS”, March 2018.

28. NVS Yashwath “Evaluation metrics & Model Selection in Linear Regression”, October 7 2020.