Data Science Based Prediction System for a Company

Project Report (CS892)

Bachelor of Technology in Computer Science & Engineering

B. P. Poddar Institute of Management & Technology

Under

Maulana Abul Kalam Azad University of Technology

Under the supervision of

Subhasis Mallick, Assistant Professor, B.P.P.I.M.T.

In fulfillment for the award of the degree

Submitted By

Name	University Roll Number	University Registration Number
Sandipan Chowdhury	**11500116049**	**161150110091**
Nishant Kumar	**11500116073**	**161150110067**
Olivia Mukherjee	**11500117011**	**171150120013**
Rahul Singh	**11500116061**	**161150110079**

Academic Year: 2019 - 2020

Department of Computer Science & Engineering
B.P. Poddar Institute of Management & Technology
137, V.I.P. Road, Poddar Vihar, Kolkata – 700052

CERTIFICATE

This is to certify that the project work, entitled "Data Science Based Prediction System for a Company" submitted by Group No. B-18, comprising of Sandipan Chowdhury, Nishant Kumar, Olivia Mukherjee, Rahul Singh; has been prepared according to the regulation of the degree B.Tech. in Computer Science & Engineering of the Maulana Abul Kalam Azad University of Technology, West Bengal. The candidates have fulfilled the requirements for the submission of the project work.

(Signature of HOD)
Dept. of Computer Science & Engg.

(Signature of the Supervisor)
Dept. of Computer Science & Engg.

ACKNOWLEDGEMENT

It is a great pleasure for us to express our earnest and great appreciation to Mr. Subhasis Mallick, our project guide. We are very much grateful to him for his kind guidance, encouragement, valuable suggestions, innovative ideas, and supervision throughout this project work, without which the completion of the project work would have been difficult one.

We would like to express our thanks to the Head of the Department, Professor Ananya Kanjilal for her active support.

We also express our sincere thanks to all the teachers of the department for their precious help, encouragement, kind cooperation and suggestions throughout the development of the project work.

We would like to express our gratitude to the library staff and laboratory staff for providing us with a congenial working environment.

Table of Content

DEPARTMENTAL MISSION

Enrich students with sound knowledge in fundamentals and cutting edge technologies of Computer Science and Engineering to excel globally in challenging roles in industries and academics.

Emphasize quality teaching, learning and research to encourage creative thoughts through application of professional knowledge and skill.

Inspire leadership and entrepreneurship skills in evolving areas of Computer Science and Engineering with social and environmental awareness.

Instill moral and ethical values to attain the highest level of accomplishment and personal growth.

DEPARTMENTAL VISION

Developing competent professionals in Computer Science and Engineering, who can adapt to constantly evolving technologies for addressing industrial and social needs through continuous learning.

PROGRAM EDUCATIONAL OBJECTIVES (PEO)

Graduates of Computer Science and Engineering program will have good knowledge in the core concepts of systems, software and tools for analyzing problems and designing solutions addressing the dynamic requirements of the industry and society, while employed in industries or work as entrepreneurs.

Graduates of Computer Science and Engineering program will opt for higher education and research in emerging fields of Computer Science & Engineering towards building a sustainable world.

Graduates of Computer Science and Engineering will have leadership skills, communication skills, ethical and moral values, team spirit and professionalism.

PROGRAM SPECIFIC OUTCOMES (PSO)

Students will have proficiency in fundamental engineering and computing techniques and knowledge on contemporary topics like artificial intelligence, data science and distributed computing towards development of optimized algorithmic solutions.

Students will have capabilities to participate in the development of software and embedded systems through synergized teams to cater to the dynamic needs of the industry and society.

PROGRAM OUTCOMES (PO)

PO1: Engineering Knowledge: Apply the knowledge of Mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO2: Problem Analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

PO3: Design / Development of Solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO4: Conduct Investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5: Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

PO6: Engineer and Society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO7: Environment and Sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO8: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO9: Individual and Team Work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11: Project Management and Finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO12: Life-long Learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Data Science Based Prediction System for a Company

PO AND PSO MAPPING OF PROJECT

1. Slight (low) 2. Moderate (Medium) 3. Substantial (High)

PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	P012	PSO1	PSO2
3	3	3	3	3	3	3	3	3	3	3	3	2	2

JUSTIFICATIONS OF MAPPING

The project aims at solving a major challenge, providing a solution thereby benefiting the society, thus justifying PO2, PO3, PO4, PO6, and PO7. The project can achieve its objective only when each and every member works as an individual completing his/her assigned work and also by acting like a team, consulting each other, thus justifying PO9, PO10, and PO11. The project works on modern techniques such as Machine Learning and Artificial Intelligence, thus justifying PO5. Knowledge of Statistics is required, thus justifying PO1. Business Ethics is related with this project, thus satisfying PO8. The project can be extended and modified accordingly with requirements and time, and also the completion of this project will contribute to the team members a life-long experience and learning, thus justifying PO12.

ABSTRACT

In Section Activity Chart, an activity chart is given to represent when and how we completed this project. It describes when we started this project, when we completed each part of project and when we finished the project.

In Section Introduction, we discussed about exactly what we are doing in this project.

In Section Theory, detailed discussions are made about some topics. These topics are Data Science, Logistic Regression, Confusion Matrix, Linear Regression, Predictive Models and E-Commerce. Each of these has significance in this project.

In Section Used Hardware and Software, we told about the computer where we did our project and the software, programming language which was used to do the project.

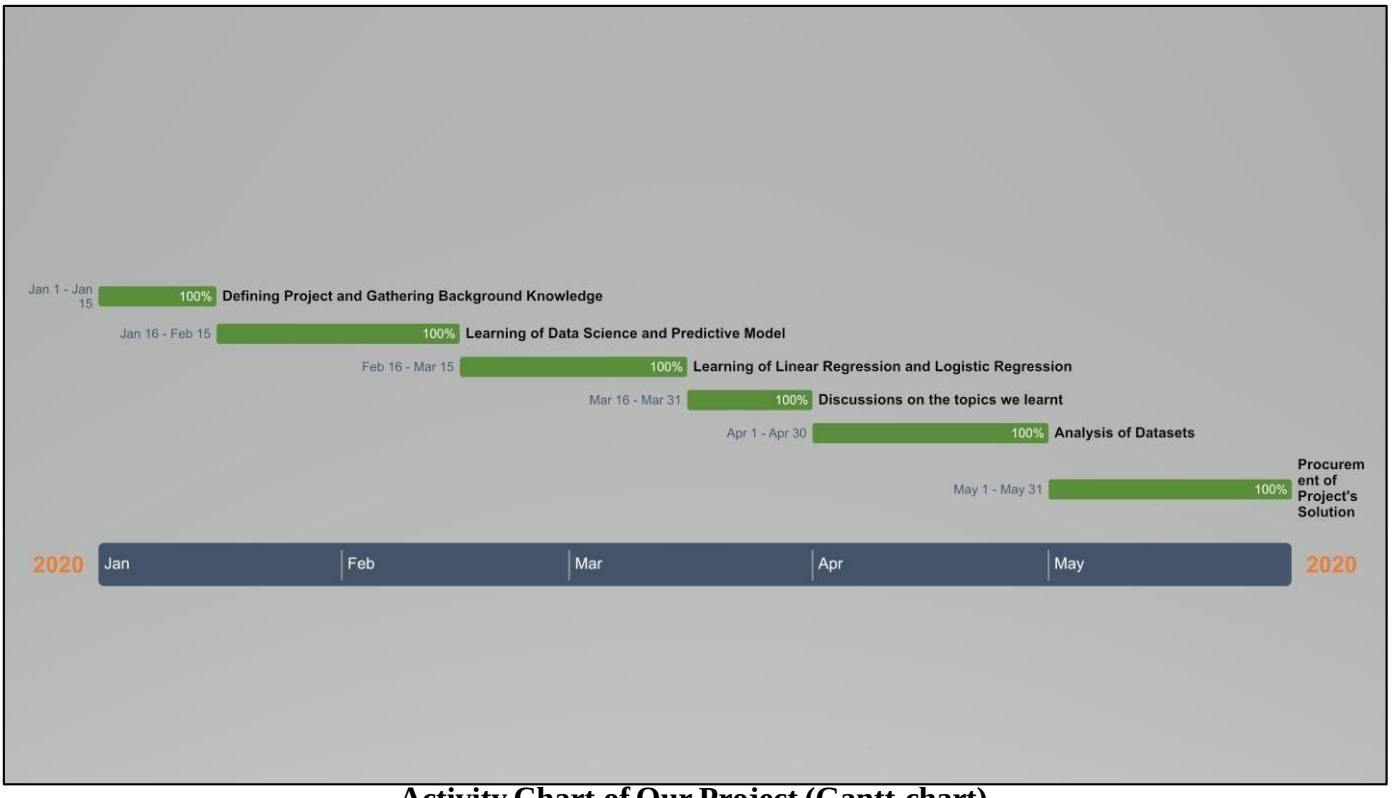
In Section Mathematical Formulation, we discussed about mathematical formula of regressions.

In Section Results and Discussions, we discussed about the result of project and how we executed the main part of project.

In Section Future Plans, we discussed about what we could have done to make this project better than now.

In Section References, we have written the internet links and books from where we learnt many things and we had help.

ACTIVITY CHART



Activity Chart of Our Project (Gantt chart)

We started our project from 1st January 2020. In January month, we defined the project and gathered background knowledge. In January and February months, we learnt about data science and predictive models. In February and March months, we learnt about logistic regression and linear regression. In March month, we discussed about what we learnt. In April month, we analyzed our dataset. In May month, we procured a solution for our project.

INTRODUCTION

In this 21st century of technology, we are progressing towards smarter concepts and tools day by day. Nowadays it is possible to shop while sitting at home, by e-commerce. Just a mobile number or an email id and a bank account are needed to do online shopping. Various companies (like Amazon, Flipkart, and Snapdeal etc.) provide us the facility of good online shopping.

These companies also seek help from various tools and concepts to get an idea about how they can increase their profit. Such a concept is Data Science. A Data Scientist can analyze the sales data of an e-commerce company, and build a recommendation system to give an optimized business strategy to that company.

In this project, we have 1 dataset of Amazon Company. We will analyze this dataset and calculate the optimized business strategy for Amazon Company for next year.

THEORY

1) Data Science

Introduction:-
Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems".

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science. Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge. In 2015, the American Statistical Association identified database management, statistics and machine learning, and distributed and parallel systems as the three emerging foundational professional communities.

In 2012, when Harvard Business Review called it "The Sexiest Job of the 21st Century", the term "data science" became a buzzword. It is now often used interchangeably with earlier concepts like business analytics, business intelligence, predictive modeling, and statistics. Even the suggestion that data science is sexy was paraphrasing Hans Rosling, featured in a 2011 BBC documentary with the quote; "Statistics is now the sexiest subject around." Nate Silver referred to data science as a sexed up term for statistics. In many cases, earlier approaches and solutions are now simply rebranded as "data science" to be more attractive, which can cause the term to become "dilute beyond usefulness". While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents. To the discredit of the discipline, however, many data-science and big-data projects fail to deliver useful results, often as a result of poor management and utilization of resources.

History:-
The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage, it was used as a substitute for computer science by Peter Naur in 1960. Naur later introduced the term "datalogy". In 1974, Naur published Concise Survey of Computer Methods, which freely used the term data science in its survey of the contemporary data processing methods that are used in a wide range of applications.

The modern definition of "data science" was first sketched during the second Japanese-French statistics symposium organized at the University of Montpellier II (France) in 1992. The attendees acknowledged the emergence of a new discipline with a specific focus on data from various origins, dimensions, types and structures. They shaped the contour of this new science based on established concepts and principles of statistics and data analysis with the extensive use of the increasing power of computer tools.

In 1996, members of the International Federation of Classification Societies (IFCS) met in Kobe for their biennial conference. Here, for the first time, the term data science is included in the title of the conference ("Data Science, classification, and related methods"), after the term was introduced in a roundtable discussion by Chikio Hayashi.

In November 1997, C.F. Jeff Wu gave the inaugural lecture entitled "Statistics = Data Science?" for his appointment to the H. C. Carver Professorship at the University of Michigan. In this lecture, he characterized statistical work as a trilogy of data collection, data modeling and analysis, and decision

making. In his conclusion, he initiated the modern, non-computer science, usage of the term "data science" and advocated that statistics be renamed data science and statisticians data scientists. Later, he presented his lecture entitled "Statistics = Data Science?" as the first of his 1998 P.C. Mahalanobis Memorial Lectures. These lectures honor Prasanta Chandra Mahalanobis, an Indian scientist and statistician and founder of the Indian Statistical Institute.

In 2001, William S. Cleveland introduced data science as an independent discipline, extending the field of statistics to incorporate "advances in computing with data" in his article "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics", which was published in Volume 69, No. 1, of the April 2001 edition of the International Statistical Review / Revue Internationale de Statistique. In his report, Cleveland establishes six technical areas which he believed to encompass the field of data science: multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation, and theory.

In April 2002, the International Council for Science (ICSU): Committee on Data for Science and Technology (CODATA) started the Data Science Journal, a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues. Shortly thereafter, in January 2003, Columbia University began publishing The Journal of Data Science, which provided a platform for all data workers to present their views and exchange ideas. The journal was largely devoted to the application of statistical methods and quantitative research. In 2005, The National Science Board published "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century" defining data scientists as "the information and computer scientists, database and software and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection" whose primary activity is to "conduct creative inquiry and analysis."

Around 2007, Turing award winner Jim Gray envisioned "data-driven science" as a "fourth paradigm" of science that uses the computational analysis of large data as primary scientific method and "to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other."

In the 2012 Harvard Business Review article "Data Scientist: The Sexiest Job of the 21st Century", DJ Patil claims to have coined this term in 2008 with Jeff Hammerbacher to define their jobs at LinkedIn and Facebook, respectively. He asserts that a data scientist is "a new breed", and that a "shortage of data scientists is becoming a serious constraint in some sectors", but describes a much more business-oriented role.

In 2013, the IEEE Task Force on Data Science and Advanced Analytics was launched. In 2013, the first "European Conference on Data Analysis (ECDA)" was organized in Luxembourg, establishing the European Association for Data Science (EuADS). The first international conference: IEEE International Conference on Data Science and Advanced Analytics was launched in 2014. In 2014, General Assembly launched student-paid boot camp and The Data Incubator launched a competitive free data science fellowship. In 2014, the American Statistical Association section on Statistical Learning and Data Mining renamed its journal to "Statistical Analysis and Data Mining: The ASA Data Science Journal" and in 2016 changed its section name to "Statistical Learning and Data Science". In 2015, the International Journal on Data Science and Analytics was launched by Springer to publish original work on data science and big data analytics. In September 2015 the Gesellschaft für Klassifikation (GfKl) added to the name of the Society "Data Science Society" at the third ECDA conference at the University of Essex, Colchester, UK.

Relationship of Data Science with Statistics:-

"Data science" has recently become a popular term among business executives. However, many critical academics and journalists see no distinction between data science and statistics, whereas others consider it largely a popular term for "data mining" and "big data". Writing in Forbes, Gil Press argues that data science is a buzzword without a clear definition and has simply replaced "business analytics" in contexts such as graduate degree programs. In the question-and-answer section of his keynote address at the Joint Statistical Meetings of American Statistical Association, noted applied statistician Nate Silver said, "I think data-scientist is a sexed up term for a statistician ... Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn't berate the term statistician."

On the other hand, responses to criticism are as numerous. In a 2014 Wall Street Journal article, Irving Wladawsky-Berger compares the data science enthusiasm with the dawn of computer science. He argues data science, like any other interdisciplinary field, employs methodologies and practices from across the academia and industry, but then it will morph them into a new discipline. He brings to attention the sharp criticisms of computer science, now a well-respected academic discipline, had to once face. Likewise, NYU Stern's Vasant Dhar, as do many other academic proponents of data science, argues more specifically in December 2013 that data science is different from the existing practice of data analysis across all disciplines, which focuses only on explaining data sets. Data science seeks actionable and consistent pattern for predictive uses. This practical engineering goal takes data science beyond traditional analytics.
Now the data in those disciplines and applied fields that lacked solid theories, like health science and social science, could be sought and utilized to generate powerful predictive models.

In an effort similar to Dhar's, Stanford professor David Donoho, in September 2015, takes the proposition further by rejecting three simplistic and misleading definitions of data science in lieu of criticisms. First, for Donoho, data science does not equate to big data, in that the size of the dataset is not a criterion to distinguish data science and statistics. Second, data science is not defined by the computing skills of sorting big datasets, in that these skills are already generally used for analyses across all disciplines. Third, data science is a heavily applied field where academic programs right now do not sufficiently prepare data scientists for the jobs, in that many graduate programs misleadingly advertise their analytics and statistics training as the essence of a data science program. As a statistician, Donoho, following many in his field, champions the broadening of learning scope in the form of data science, like John Chambers who urges statisticians to adopt an inclusive concept of learning from data, or like William Cleveland who urges to prioritize extracting from data applicable predictive tools over explanatory theories. Together, these statisticians envision an increasingly inclusive applied field that grows out of traditional statistics and beyond.

For the future of data science, Donoho projects an ever-growing environment for open science where datasets used for academic publications are accessible to all researchers. US National Institute of Health has already announced plans to enhance reproducibility and transparency of research data. Other big journals are likewise following suit. This way, the future of data science not only exceeds the boundary of statistical theories in scale and methodology, but data science will revolutionize current academia and research paradigms. As Donoho concludes, "the scope and impact of data science will continue to expand enormously in coming decades as scientific data and data about science itself become ubiquitously available."

Applications:-
Using data science, companies have become intelligent enough to push & sell products as per customers' purchasing power & interest. Here's how they are ruling our hearts and minds:

Internet Search
There are many search engines like Yahoo, Bing, Ask, AOL, Duckduckgo etc. All these search engines (including Google) make use of data science algorithms to deliver the best result for our searched query in fraction of seconds. Considering the fact that, Google processes more than 20 petabytes of data every day. Had there been no data science, Google wouldn't have been the Google we know today.

Digital Advertisements (Targeted Advertising and re-targeting)
If you thought Search would have been the biggest application of data science and machine learning, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various websites to the digital bill boards at the airports – almost all of them are decided by using data science algorithms.
This is the reason why digital ads have been able to get a lot higher CTR than traditional advertisements. They can be targeted based on users' past behavior. This is the reason why I see ads of analytics trainings while my friend sees ad of apparels in the same place at the same time.

Recommender Systems
Who can forget the suggestions about similar products on Amazon? They not only help you find relevant products from billions of products available with them, but also add a lot to the user experience.
A lot of companies have fervidly used this engine / system to promote their products / suggestions in accordance with user's interest and relevance of information. Internet giants like Amazon, Twitter, Google Play, Netflix, LinkedIn, IMDB and many more uses this system to improve user experience. The recommendations are made based on previous search results for a user.

Image Recognition
You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. Similarly, while using Whatsapp web, you scan a barcode in your web browser using your mobile phone. In addition, Google provides you the option to search for images by uploading them. It uses image recognition and provides related search results.

Speech Recognition
Some of the best examples of speech recognition products are Google Voice, Siri, and Cortana etc. Using speech recognition feature, even if you aren't in a position to type a message, your life wouldn't stop. Simply speak out the message and it will be converted to text. However, at times, you would realize, speech recognition doesn't perform accurately.

Gaming
EA Sports, Zynga, Sony, Nintendo, Activision-Blizzard have led gaming experience to the next level using data science. Games are now designed using machine learning algorithms which improve / upgrade themselves as the player moves up to a higher level. In motion gaming also, your opponent (computer) analyzes your previous moves and accordingly shapes up its game.

Price Comparison Websites
At a basic level, these websites are being driven by lots and lots of data which is fetched using APIs and RSS Feeds. If you have ever used these websites, you would know the convenience of comparing the price of a product from multiple vendors at one place. PriceGrabber, PriceRunner, Junglee, Shopzilla, DealTime

are some examples of price comparison websites. Nowadays, price comparison website can be found in almost every domain such as technology, hospitality, automobiles, durables, apparels etc.

Airline Route Planning

Airline Industry across the world is known to bear heavy losses. Except a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air fuel prices and need to offer heavy discounts to customers has further made the situation worse. It wasn't for long when airlines companies started using data science to identify the strategic areas of improvements. Now using data science, the airline companies can:

a) Predict flight delay
b) Decide which class of airplanes to buy
c) Whether to directly land at the destination, or take a halt in between (For example: A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.)
d) Effectively drive customer loyalty programs
e) Southwest Airlines, Alaska Airlines are among the top companies who have embraced data science to bring changes in their way of working.

Fraud and Risk Detection

One of the first applications of data science originated from Finance discipline. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paper work while sanctioning loans. They decided to bring in data science practices in order to rescue them out of losses. Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

Delivery logistics

Who says data science has limited applications? Logistic companies like DHL, FedEx, UPS, Kuhne+Nagel have used data science to improve their operational efficiency. Using data science, these companies have discovered the best routes to ship, the best suited time to deliver, the best mode of transport to choose thus leading to cost efficiency, and many more to mention. Furthermore, the data that these companies generate using the GPS installed provides them a lot of possibilities to explore using data science.

2) Linear Regression

Introduction:-

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regressions. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quintile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed dataset of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.
Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

History:-

Least squares linear regression, as a means of finding a good rough linear fit to a set of points was performed by Legendre (1805) and Gauss (1809) for the prediction of planetary movement. Quetelet was responsible for making the procedure well-known and for using it extensively in the social sciences.

Applications:-

Linear regression is widely used in biological, behavioral and social sciences to describe possible relationships between variables. It ranks as one of the most important tools used in these disciplines.

Trend line

A trend line represents a trend; the long-term movement in time series data after other components has been accounted for. It tells whether a particular dataset (say GDP, oil prices or stock prices) have increased or decreased over the period of time. A trend line could simply be drawn by eye through a set of data points, but more properly their position and slope is calculated using statistical techniques like linear regression. Trend lines typically are straight lines, although some variations use higher degree polynomials depending on the degree of curvature desired in the line.

Trend lines are sometimes used in business analytics to show changes in data over time. This has the advantage of being simple. Trend lines are often used to argue that a particular action or event (such as training, or an advertising campaign) caused observed changes at a point in time. This is a simple technique, and does not require a control group, experimental design, or a sophisticated analysis technique. However, it suffers from a lack of scientific validity in cases where other potential changes can affect the data.

Epidemiology

Early evidence relating tobacco smoking to mortality and morbidity came from observational studies employing regression analysis. In order to reduce spurious correlations when analyzing observational data, researchers usually include several variables in their regression models in addition to the variable of primary interest. For example, in a regression model in which cigarette smoking is the independent variable of primary interest and the dependent variable is lifespan measured in years, researchers might include education and income as additional independent variables, to ensure that any observed effect of smoking on lifespan is not due to those other socio-economic factors. However, it is never possible to include all possible confounding variables in an empirical analysis. For example, a hypothetical gene might increase mortality and also cause people to smoke more. For this reason, randomized controlled trials are often able to generate more compelling evidence of causal relationships than can be obtained using regression analyses of observational data. When controlled experiments are not feasible, variants of regression analysis such as instrumental variables regression may be used to attempt to estimate causal relationships from observational data.

Finance

The capital asset pricing model uses linear regression as well as the concept of beta for analyzing and quantifying the systematic risk of an investment. This comes directly from the beta coefficient of the linear regression model that relates the return on the investment to the return on all risky assets.

Economics

Linear regression is the predominant empirical tool in economics. For example, it is used to predict consumption spending, fixed investment spending, inventory investment, and purchases of a country's exports, spending on imports, the demand to hold liquid assets, labor demand, and labor supply.

Environmental science

Linear regression finds application in a wide range of environmental science applications. In Canada, the Environmental Effects Monitoring Program uses statistical analyses on fish and benthic surveys to measure the effects of pulp mill or metal mine effluent on the aquatic ecosystem.

Machine learning

Linear regression plays an important role in the field of artificial intelligence such as machine learning. The linear regression algorithm is one of the fundamental supervised machine-learning algorithms due to its relative simplicity and well-known properties.

3) Predictive Models

Introduction:-

Predictive modeling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modeling can be applied to any type of unknown event, regardless of when it occurred. For example, predictive models are often used to detect crimes and identify suspects, after the crime has taken place.

In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data, for example given an email determining how likely that it is spam.

Models can use one or more classifiers in trying to determine the probability of a set of data belonging to another set. For example, a model might be used to determine whether an email is spam or "ham" (non-spam).

Depending on definitional boundaries, predictive modeling is synonymous with, or largely overlapping with, the field of machine learning, as it is more commonly referred to in academic or research and development contexts. When deployed commercially, predictive modeling is often referred to as predictive analytics.

Predictive modeling is often contrasted with causal modeling/analysis. In the former, one may be entirely satisfied to make use of indicators of, or proxies for, the outcome of interest. In the latter, one seeks to determine true cause-and-effect relationships. This distinction has given rise to a burgeoning literature in the fields of research methods and statistics and to the common statement that "correlation does not imply causation".

Applications:-

Uplift modeling

Uplift modeling is a technique for modeling the change in probability caused by an action. Typically this is a marketing action such as an offer to buy a product, to use a product more or to re-sign a contract. For example, in a retention campaign you wish to predict the change in probability that a customer will remain a customer if they are contacted. A model of the change in probability allows the retention campaign to be targeted at those customers on whom the change in probability will be beneficial. This allows the retention program to avoid triggering unnecessary churn or customer attrition without wasting money contacting people who would act anyway.

Archaeology

Predictive modeling in archaeology gets its foundations from Gordon Willey's mid-fifties work in the Virú Valley of Peru. Complete, intensive surveys were performed then covariability between cultural remains and natural features such as slope, and vegetation were determined. Development of quantitative methods and a greater availability of applicable data led to growth of the discipline in the 1960s and by the late 1980s, substantial progress had been made by major land managers worldwide.

Generally, predictive modeling in archaeology is establishing statistically valid causal or covariable relationships between natural proxies such as soil types, elevation, slope, vegetation, proximity to water, geology, geomorphology, etc., and the presence of archaeological features. Through analysis of these

quantifiable attributes from land that has undergone archaeological survey, sometimes the "archaeological sensitivity" of unsurveyed areas can be anticipated based on the natural proxies in those areas. Large land managers in the United States, such as the Bureau of Land Management (BLM), the Department of Defense (DOD), and numerous highway and parks agencies, have successfully employed this strategy. By using predictive modeling in their cultural resource management plans, they are capable of making more informed decisions when planning for activities that have the potential to require ground disturbance and subsequently affect archaeological sites.

Customer relationship management

Predictive modeling is used extensively in analytical customer relationship management and data mining to produce customer-level models that describe the likelihood that a customer will take a particular action. The actions are usually sales, marketing and customer retention related.

For example, a large consumer organization such as a mobile telecommunications operator will have a set of predictive models for product cross-sell, product deep-sell (or upselling) and churn. It is also now more common for such an organization to have a model of severability using an uplift model. This predicts the likelihood that a customer can be saved at the end of a contract period (the change in churn probability) as opposed to the standard churn prediction model.

Auto insurance

Predictive modeling is utilized in vehicle insurance to assign risk of incidents to policy holders from information obtained from policy holders. This is extensively employed in usage-based insurance solutions where predictive models utilize telemetry-based data to build a model of predictive risk for claim likelihood. Black-box auto insurance predictive models utilize GPS or accelerometer sensor input only. Some models include a wide range of predictive input beyond basic telemetry including advanced driving behavior, independent crash records, road history, and user profiles to provide improved risk models.

Health care

In 2009 Parkland Health & Hospital System began analyzing electronic medical records in order to use predictive modeling to help identify patients at high risk of readmission. Initially the hospital focused on patients with congestive heart failure, but the program has expanded to include patients with diabetes, acute myocardial infarction, and pneumonia.

In 2018, Banerjee et al. proposed a deep learning model—Probabilistic Prognostic Estimates of Survival in Metastatic Cancer Patients (PPES-Met)—for estimating short-term life expectancy (>3 months) of the patients by analyzing free-text clinical notes in the electronic medical record, while maintaining the temporal visit sequence. The model was trained on a large dataset (10,293 patients) and validated on a separated dataset (1818 patients). It achieved an area under the ROC (Receiver Operating Characteristic) curve of 0.89. To provide explain-ability, they developed an interactive graphical tool that may improve physician understanding of the basis for the model's predictions. The high accuracy and explain-ability of the PPES-Met model may enable the model to be used as a decision support tool to personalize metastatic cancer treatment and provide valuable assistance to the physicians.

Algorithmic trading

Predictive modeling in trading is a modeling process wherein the probability of an outcome is predicted using a set of predictor variables. Predictive models can be built for different assets like stocks, futures, currencies, commodities etc. Predictive modeling is still extensively used by trading firms to devise strategies and trade. It utilizes mathematically advanced software to evaluate indicators on price, volume, open interest and other historical data, to discover repeatable patterns.

4) Logistic Regression

Introduction:-

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

In a binary logistic regression model, the dependent variable has two levels (categorical). Outputs with more than two values are modeled by multinomial logistic regression and if the multiple categories are ordered, by ordinal logistic regression (for example the proportional odds ordinal logistic model). The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier. The coefficients are generally not computed by a closed-form expression, unlike linear least squares. The logistic regression as a general statistical model was originally developed and popularized primarily by Joseph Berkson, beginning in Berkson (1944), where he coined "logit".

History:-

A detailed history of the logistic regression is given in Cramer (2002). The logistic function was developed as a model of population growth and named "logistic" by Pierre François Verhulst in the 1830s and 1840s, under the guidance of Adolphe Quetelet. In his earliest paper (1838), Verhulst did not specify how he fit the curves to the data. In his more detailed paper (1845), Verhulst determined the three parameters of the model by making the curve pass through three observed points, which yielded poor predictions.

The logistic function was independently developed in chemistry as a model of autocatalysis (Wilhelm Ostwald, 1883). An autocatalytic reaction is one in which one of the products is itself a catalyst for the same reaction, while the supply of one of the reactants is fixed. This naturally gives rise to the logistic equation for the same reason as population growth: the reaction is self-reinforcing but constrained.

The logistic function was independently rediscovered as a model of population growth in 1920 by Raymond Pearl and Lowell Reed, published as Pearl & Reed (1920), which led to its use in modern statistics. They were initially unaware of Verhulst's work and presumably learned about it from L. Gustavedu Pasquier, but they gave him little credit and did not adopt his terminology. Verhulst's priority was acknowledged and the term "logistic" revived by Udny Yule in 1925 and has been followed since. Pearl and Reed first applied the model to the population of the United States, and also initially fitted the curve by making it pass through three points; as with Verhulst, this again yielded poor results.

In the 1930s, the probit model was developed and systematized by Chester Ittner Bliss, who coined the term "probit" in Bliss (1934), and by John Gaddum in Gaddum (1933), and the model fit by maximum likelihood estimation by Ronald A. Fisher in Fisher (1935), as an addendum to Bliss's work. The probit model was principally used in bioassay, and had been preceded by earlier work dating to 1860. The probit model influenced the subsequent development of the logit model and these models competed with each other.

The logistic model was likely first used as an alternative to the probit model in bioassay by Edwin Bidwell Wilson and his student Jane Worcester in Wilson & Worcester (1943). However, the development of the logistic model as a general alternative to the probit model was principally due to the work of Joseph Berkson over many decades, beginning in Berkson (1944), where he coined "logit", by analogy with "probit", and continuing through Berkson (1951) and following years. The logit model was initially dismissed as inferior to the probit model, but "gradually achieved an equal footing with the logit", particularly between 1960 and 1970. By 1970, the logit model achieved parity with the probit model in use in statistics journals and thereafter surpassed it. This relative popularity was due to the adoption of the logit outside of bioassay, rather than displacing the probit within bioassay, and its informal use in practice; the logit's popularity is credited to the logit model's computational simplicity, mathematical properties, and generality, allowing its use in varied fields.

Various refinements occurred during that time, notably by David Cox, as in Cox (1958).

The multinomial logit model was introduced independently in Cox (1966) and Thiel (1969), which greatly increased the scope of application and the popularity of the logit model. In 1973 Daniel McFadden linked the multinomial logit to the theory of discrete choice, specifically Luce's choice axiom, showing that the multinomial logit followed from the assumption of independence of irrelevant alternatives and interpreting odds of alternatives as relative preferences; this gave a theoretical foundation for the logistic regression.

Applications:-

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression. Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.). Another example might be to predict whether a Nepalese voter will vote Nepali Congress or Communist Party of Nepal or Any Other Party, based on age, income, sex, race, state of residence, votes in previous elections, etc. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription, etc. In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

5) Confusion Matrix

Introduction:-

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

6) E-Commerce

It is the activity of electronically buying or selling of products on online services or over the Internet. Electronic commerce draws on technologies such as mobile commerce, electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. E-commerce is in turn driven by the technological advances of the semiconductor industry, and is the largest sector of the electronics industry.

Modern electronic commerce typically uses the World Wide Web for at least one part of the transaction's life cycle although it may also use other technologies such as e-mail. Typical e-commerce transactions include the purchase of online books (such as Amazon) and music purchases (music download in the form of digital distribution such as iTunes Store), and to a less extent, customized/personalized online liquor store inventory services. There are three areas of e-commerce: online retailing, electronic markets, and online auctions. E-commerce is supported by electronic business.

USED SYSTEM / SOFTWARE

Used Hardware

Machine → HP 658TX (Laptop)
Ram Memory → 8 GB
Hard Disk Memory → 1024 GB

Used Software

Operating System → Windows 10 Enterprise (64 bit)
Application Software → Jupyter Notebook
Programming Language → Python 3.7.3

MATHEMATICAL FORMULATION

Logistic Regression

Formula of logistic regression is

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

In above formula ---

- x = input value
- y = predicted output
- β_0 = intercept
- β_1= coefficient for input x

Linear Regression:-

Formula for linear regression equation is given by:

$$y_l = a + bx_l$$

Where, y_l is y-variable of linear regression equation

and x_l is x-variable of linear regression equation.

a and *b* are calculated by following formulas …..

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{(\sum x^2) - (\sum x)^2}$$

In above two formulas ------

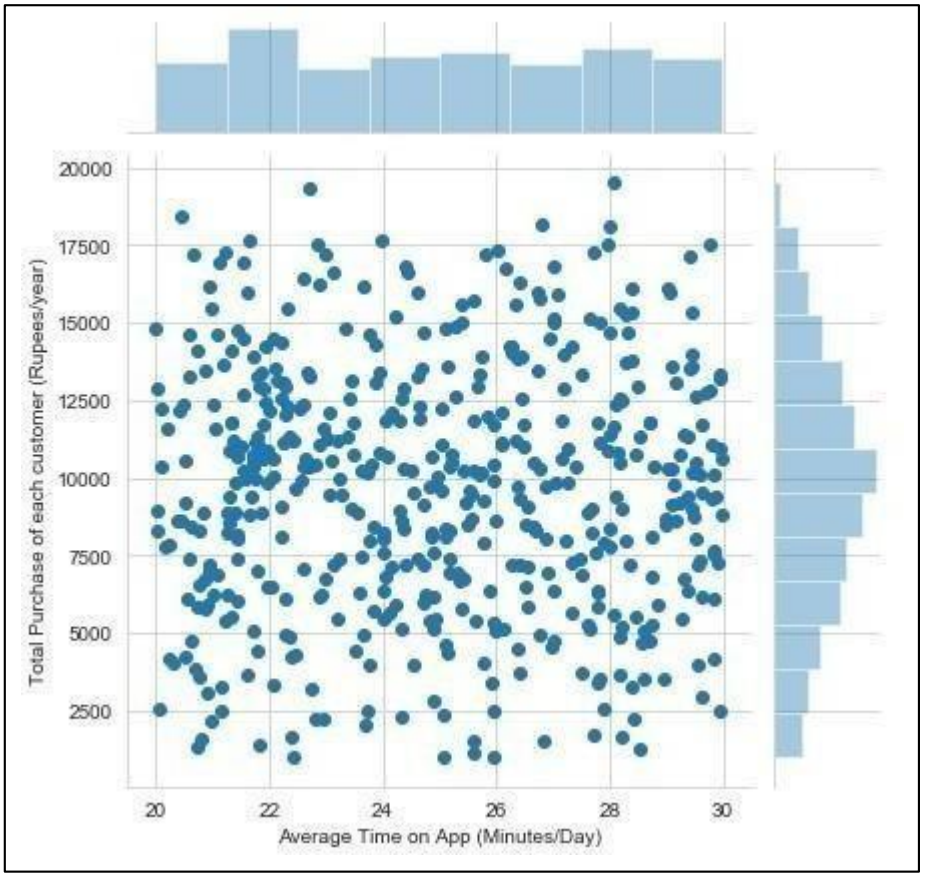
x = Values of one column of given dataset
y = Values of the another column of given dataset
n = Number of total input

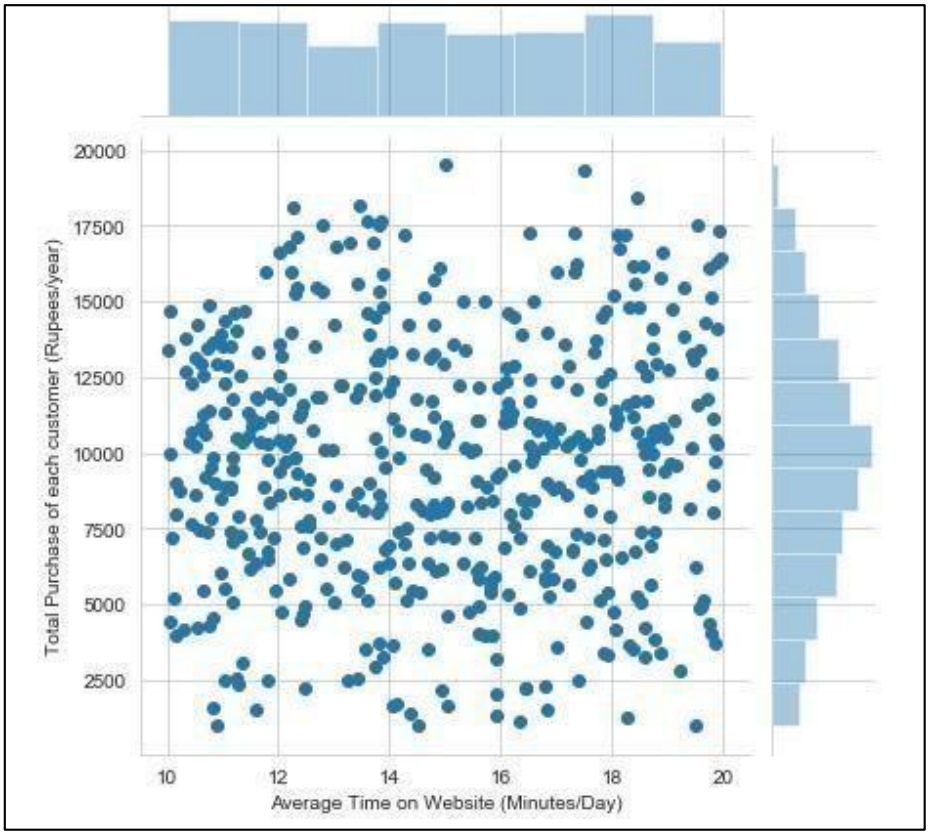
RESULTS & DISCUSSION

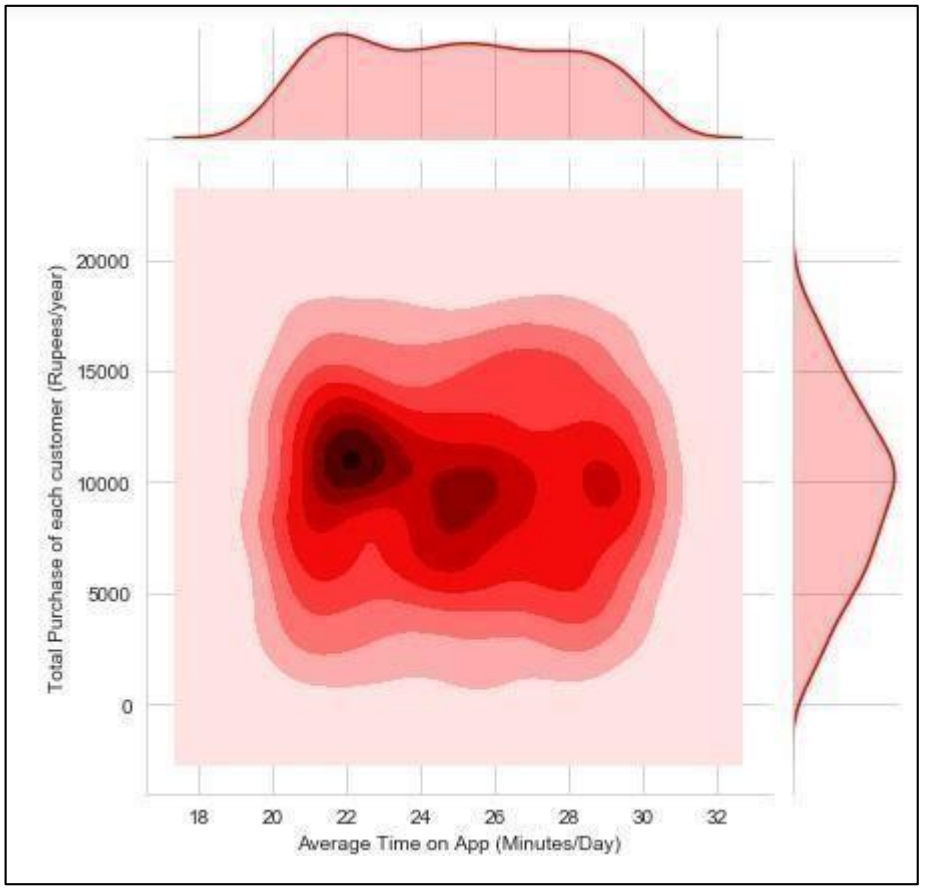
We have used one dataset. Dataset A contains information about customers. It includes ……

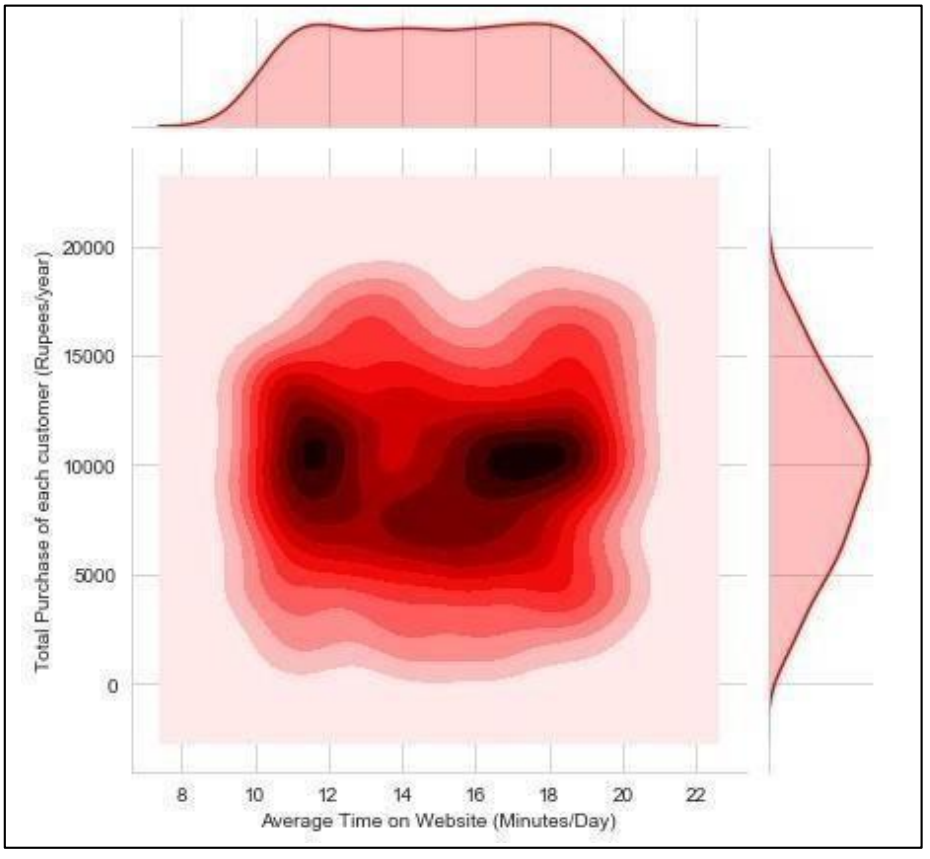
a) Name of Customers
b) Email ID of Customers
c) Phone number of Customers
d) Average time spent on App by each Customer (Minutes / Day)
e) Total Purchase from App by each Customer (Rupees / Year)
f) Average time spent on Website by each Customer (Minutes / Day)
g) Total Purchase from Website by each Customer (Rupees / Year)
h) Length of Membership of each Customer (Year)
i) Yearly Amount Spent by each Customer (Minutes / Year)
j) Total Purchase of each Customer (Rupees / Year)
k) Ratings given by Customer (Out of 5)

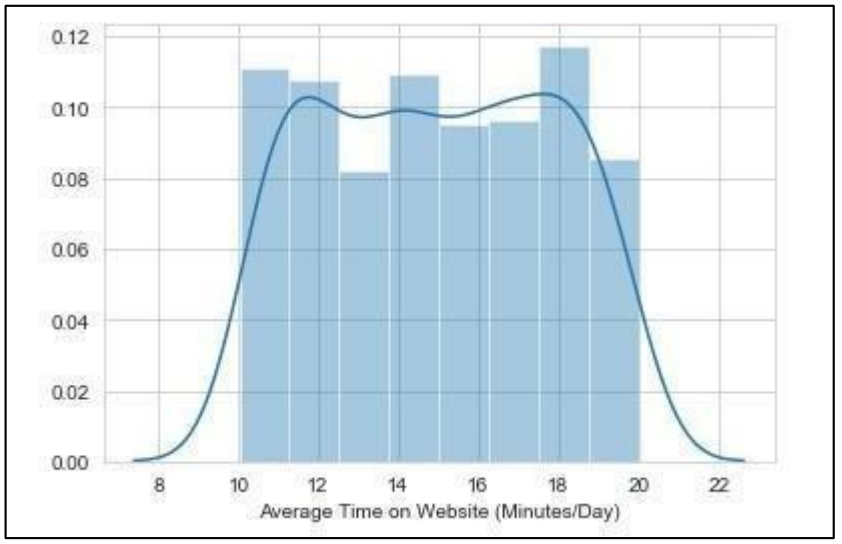
We have drawn some graphs on this dataset …… but it is not possible to draw conclusion from them.

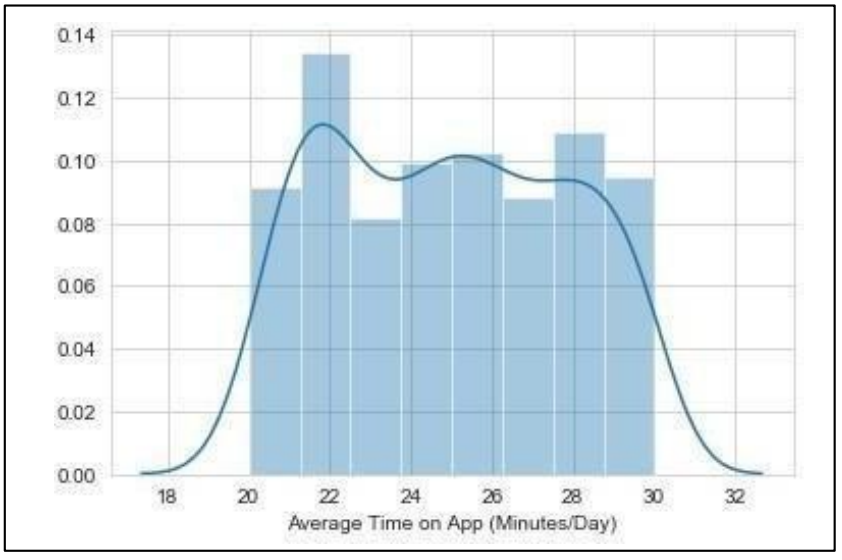


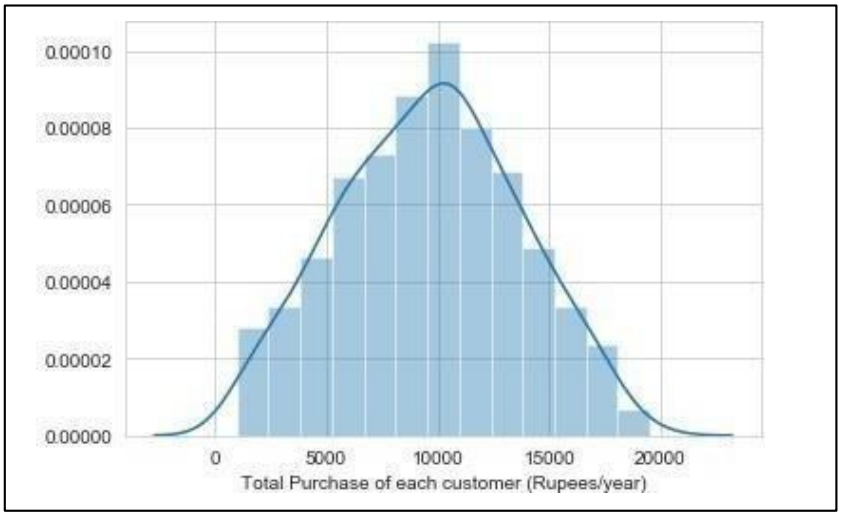



Total Purchase of each customer (Rupees/year)
20000
15000
10000
5000
0
18
20
22
24
26
28
30
32
Average Time on App (Minutes/Day)


Total Purchase of each customer (Rupees/year)
20000
15000
10000
5000
0
8
10
12
14
16
18
20
22
Average Time on Website (Minutes/Day)


0.12
0.10
0.08
0.06
0.04
0.02
0.00
8
10
12
14
16
18
20
22
Average Time on Website (Minutes/Day)


0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
18
20
22
24
26
28
30
32
Average Time on App (Minutes/Day)


0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
5000
10000
15000
20000
Total Purchase of each customer (Rupees/year)

So we have applied logistic regression and linear regression to our dataset. Then we built predictive model, confusion matrix and classification report. From those predictive models, we have calculated the coefficients of important attributes. The attribute with highest coefficient has highest importance.

	Coefficient
Average Time on App (Minutes/Day)	20.213146
Average Time on Website (Minutes/Day)	59.398727

Predictive Model of Given Dataset

```
[[30  5 12]
 [22  7  9]
 [28  6  6]]
```

Confusion Matrix of Given Dataset

	precision	recall	f1-score	support
3	0.38	0.64	0.47	47
4	0.39	0.18	0.25	38
5	0.22	0.15	0.18	40
accuracy			0.34	125
macro avg	0.33	0.32	0.30	125
weighted avg	0.33	0.34	0.31	125

Classification Report of Given Dataset

Our Data Science based analysis shows that Average Time on Website (Minutes/Day) has highest coefficient (59.398727).

So it is convenient that Average Time on Website is most important.

Therefore to maximize profit, business strategy of next year should be based on "Average Time on Website".

FUTURE PLAN

We can apply complex regression formulas to analyze this type of problems more efficiently.

We can apply other important concepts of mathematics, statistics or data science to analyze this type of problems more efficiently.

We can collect more information about customers and offers to enlarge our dataset.

We can create new offers for customers to maximize profit.

We should also concentrate in shopping mall based sales and calculate profits according to them.

Customers from different countries may have different preferences in shopping. We may construct new business strategy based on this information.

Customers from different religions may have different preferences in shopping. We may construct new business strategy based on this information.

REFERENCES

Internet Links →

a) https://en.wikipedia.org/wiki/E-commerce

b) https://en.wikipedia.org/wiki/Data_science

c) https://en.wikipedia.org/wiki/Predictive_modelling

d) https://www.investopedia.com/terms/d/data-science.asp

e) https://en.wikipedia.org/wiki/Regression_analysis

i) https://en.wikipedia.org/wiki/Logistic_regression

j) https://en.wikipedia.org/wiki/Linear_regression

k) https://en.wikipedia.org/wiki/Simple_linear_regression

Books→

a) Business Analytics: The Science of Data - Driven Decision Making, Wiley Publications

b) Predictive Analytics for Dummies, Wiley Publications

c) Applied Machine Learning, Tata McGraw Hills Publications