ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(38), DOI: 10.17485/ijst/2019/v12i38/145564, October 2019*

Predictive Analysis: Role in Big Data

Palak Mittal[1], Mansi Sharma[1], Nidhi Garg[1] and Dr. Prateek Jain[2,*]

[1]Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad 121003, India; palakmittal2109@gmail.com, mansi261198@gmail.com, nidhi.fet@mriu.edu.in
[2]Accendere CL Educate Ltd, New Delhi 110044, India; prateek.jain@accendere.co.in

Abstract

Background/objectives: Predictive analysis not only gives us a better insight into customer and user's likes and dislikes but can also help in improving and increasing revenues, profits and reach. Predictive analysis is something that everyone should be aware of as it can bring many advantages to any field be it social media, e-commerce, IT sector, business, startups etc. In this paper, some techniques are discussed that are used in predictive analytics. **Methods/statistical analysis:** It is important to explore the new possibilities and opportunities that a huge collection of data can bring us and one of the best ways of doing is to analyze and improve decision-making processes. We have discussed various Regression techniques. **Findings:** With the increase in technologies and everything shifting towards the online market and the need for analytics increased too. **Improvements/applications:** In this study importance and need for predictive analysis in various field is discussed.

Keywords: Predictive Analysis, Predictive Models, Machine Learning, Analytics.

1. Introduction

Big data is a huge collection of data gathered and collected; the source of this data can be social media, online surveys, e-commerce sites, websites and many more. These days with the growth of the online market and social media sites the amount of data produced has also grown hence resulting in the huge availability of data.

There are four kinds of Big Data:

1. Prescriptive – This investigation covers the details regarding the moves that ought to be made. This is called as the most important investigation aspect and for many of the parts bring about guidelines and suggestions for subsequent stages.
   The prescriptive investigation is extremely important, yet generally not utilized. Where enormous information investigation, by and large, reveals insight into an area, this sort of investigation provides a laser-like focus to answer particular inquiries. For instance, in the medical line, one can deal with the patient populace by means of using this technique of investigation in order to identify the number of patients who are fat. At the same time, various factors for the same are also considered including diabetes, LDL cholesterol levels for figuring out about the treatment of the same.
2. Predictive – An investigation of likely situations of what may happen. The expectations are generally a predictive figure. Predictive analytics utilizes enormous information to recognize past examples to foresee what's to come. For instance, a few organizations are utilizing predictive analytics for deals lead scoring. A few organizations have gone above and beyond utilize predictive analytics for the whole deals process, investigating the lead source, number of correspondences, sorts of interchanges, web-based social networking, records, CRM information, and so forth. Legitimately tuned predictive analytics can be utilized to help deals, showcasing, or for different sorts of complex figures.

*Author for correspondence

3. Diagnostic – A glance at past execution to figure out what happened and why. The consequence of the examination is regularly an analytic dashboard.
   Diagnostic analytics are utilized for disclosure or to decide why something happened. For instance, for an online networking advertising effort, one can utilize graphic examination to survey the number of posts, notices, adherents, fans, site visits, audits, pins, and so forth. There can be a great many online notices that can be refined into a solitary view to perceive what worked in the past and what didn't.
4. Descriptive – This gives an idea of what is occurring now in light of approaching information. To mine the examination, a continuous dashboard or potentially email reports is regularly utilized.

Descriptive analytics or information mining area at the base of the huge information esteem chain. However, they can be important for revealing examples that offer knowledge. A straightforward case of descriptive analytics would evaluate credit hazard; utilizing past monetary execution to foresee a client's imaginable money related execution. Expressive investigation can be valuable in the business cycle, for instance, to order clients by their imaginable item inclinations and deals cycle.

The web today is to a great extent a content driven system. Beginning from basic information exchange between two PCs directly associated by a wire, the multifaceted nature of content delivery over the Internet has made some amazing progress to incorporate a few complex applications, for example, adaptive video streaming, peer-to-peer file-sharing, massively multiplayer online gaming, cloud storage, and cloud-based computation.

Predictive analytics is made of predict & analysis words, but on the contrary, it operates in the reverse fashion i.e. it first analyzes the given data and then the prediction is performed. Since the human tendency of having the future verdict, this analysis technique can help in predicting future events by comparing with the earlier observed historical data. This is done by using the approach of machine learning. The historical data is collected and is being transformed by means of the utilization of other methods called as correlating and filtering.

Predictive modeling depends on at least one data instance for which we need to anticipate the value of an objective variable. Information driven predictive modeling, for the most part, initiates a model from training information, for which the value of the objective (the name) is known. These cases ordinarily are depicted by a vector of features from which the predictions will be made.

Traditional Predictive Analysis will describe the instances with a small to medium numbers – for example, dozens up to hundreds of features compressing attributes of the examples. An important aspect in the traditional setting is that each instance in the dataset has a non – trivial value for every or for most of the features that it contains.[1–3]

Big Data thinking opens our view to nontraditional information for predictive investigation—datasets in which every data point may contain less data, yet when taken in totality may give substantially more information. Organizations progressively are taking advantage of such information. For instance, information on people's visits to massive quantities of specific website pages is utilized as a part of predictive analytics for focusing on an online show of advertisements. Data on individual geographic areas are utilized for focusing on versatile advertisements. Data on the individual vendors with which one executes are utilized to target keeping banking advertisements.

As indicated by a few assessments, Walmart gathers around 2.5 petabytes of data consistently about transactions, client behavior, area and devices. An IT investigator firm Gartner suggests that there will be 20 Billion devices associated with the "Internet of Things". The amount of data generated by these devices is very large. The data collected can give a full view of customer purchasing behavior. These information sources will include "columns" to our databases that give an expanded capacity to anticipate client behavior and the ramifications of marketing on it.

The predictive process can be separated into four stages:

Gathering& pre-processing of raw data

- Transforming the data already being processed into a form that can deal in an efficient manner using the machine learning technique.
- Utilizing the learning model for information transformation.
- Predictions reporting to the client on the utilization of the earlier learning model (Figure 1).

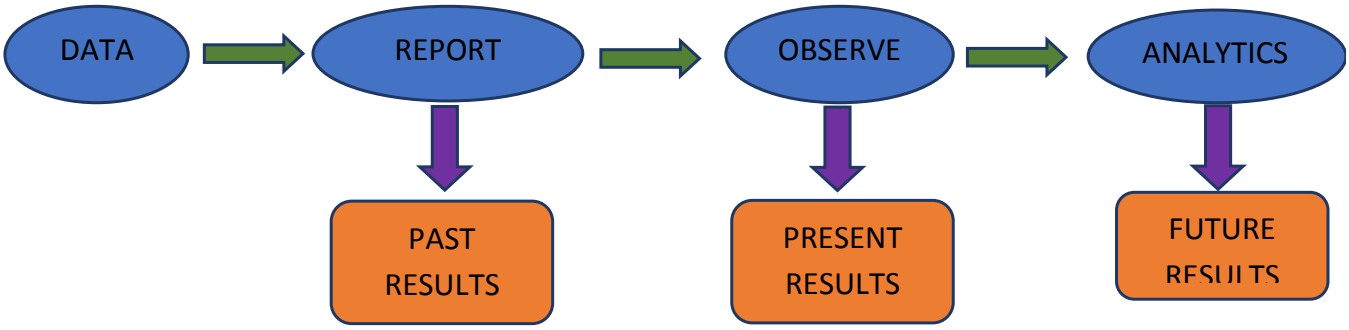

Figure 1. Steps of analytics.6

2. Need for Predictive Analysis

After introducing ourselves to predictive analysis, it's time to know what makes it interesting, important and so trending. Let us just start with a few cases where analysis can save time and improve the quality and strength of work.

- Sam and Josh own business; being spent only a few years in the business field, knowing their customer needs will give their business uplift. But it would be quite a difficult task to reach each customer and know their needs. Well, in this case, tracking the pattern of the customer's search says searching a product on their website can help to figure out the customers.
- John owns a stationary; he tries hard to keep track of his customer's choices so that they always find the right stationery. So, john decides to pen down the most liked and chosen products, but soon he realizes that continuing this process is time and energy-consuming as well as not efficient. If only, John had a software tool so that he could keep his details not only organized but also are able to provide him the desirable result.
- A group of friends launched their own social networking sites. In order to make their site more popular and trending among different age group people, they need to be more interactive with the users. By only wishing users on their birthdays, letting them know the latest information about his region, recommending them movies, music, meets based on their interest would bring them closer to them. But how?
- For any online-based company like e-commerce fraud detection becomes a really important thing to consider as one fraud attack can weaken the structure of that company. But how is it possible to know about an attack that may happen or may not.

In all the above cases, knowing the customers, users and visitors in some way boosts the purpose of our work. Predictive analytics is a pack of information examination technologies and strategies moved up under one flag. It lets you know the result of correlated data, variables and objects. It lets you know the probabilities and possibilities of any work. Collecting huge chunks of data is useless unless we find some creative and efficient way to utilize it. So, someone with huge data need not worry about doing all the manual work as there exist some software tools to lighten the workload. These tools can be open source/freeware or proprietary. Some of them are listed following:

Open-source/Freeware: R, RapidMiner, PredictionIO, SciPy, Orange, NumPy, GraphLabCreate, Octave, KNIME, HP Distributed, Weka and many more.

Proprietary Tools: TIBCO Analytics, SAP Infinitelnisight, IBM Analytics, SAS Predictive Analytics, Advanced Miner, Alteryx Analytics, STATISTICA, Alpine chorus, TIMi Suite, H2O, Analytics Solver and many more.

2.1. Real-World Example

2.1.1. Amazon Online Shopping

These days online shopping has become so convenient, and when we see a list of products recommended by Amazon, it definitely makes us more comfortable shopping. These recommendations are made using machine learning which learns through your preferences, purchase pattern and behavior (Figure 2).

Figure 2 is an example of Amazon's recommendations. Here, when searched for a particular book, amazon

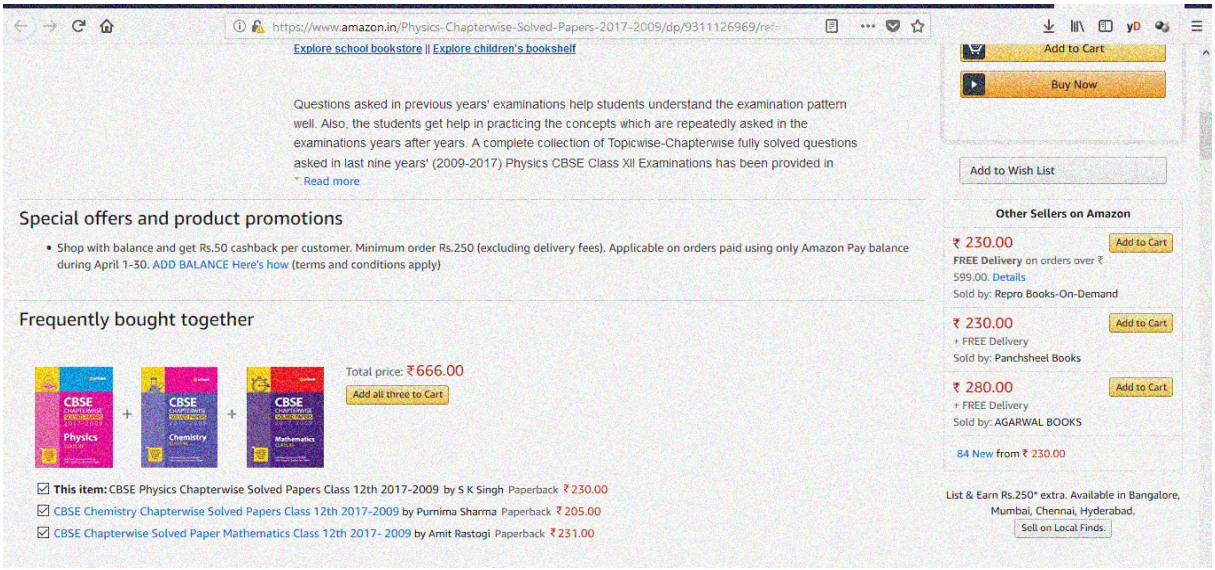

Figure 2. Amazon's recommendation system.

suggests a buyer to purchase a set of books. This idea of getting all books together can attract a customer into buying three books instead of one. This recommendation is the result of analyzing the frequent past purchases of customers. This can a) make the person buy it or b) ensure that the buyer may revisit the site to get such offers. Imagine if this offer had a discount offer too, this would have benefited too. Predictive analytics + marketing techniques can prove to be a good combination.

2.1.2. Google Search

With the introduction of Google suggest, Google has made searching easy. Google suggest provides frequently asked questions and displays them matching the user's requirement. Kevin Gibbs, who started out trying to build a URL predictor that auto-completes the URL as a user types it in. This idea was then used by Google search which soon proved to be saving user's time[4] (Figure 3).

3. Predictive Models

After learning about what predictive analysis and its need in today's world let us discuss certain models and techniques that make predicting results possible. In predictive modelling we try to create, test and validate a model that can predict an outcome with high probability.[5] In many cases, any model is selected based on the detection theory that tries to guess the probability of an outcome with a setoff input data, for example, the probability of an email being spam.[6] Predictive models can be broadly classified as parametric, non-parametric and semi-parametric (Figures 4–6).

3.1. Stages of Predictive Modeling

Albeit most experts concur that predictive analysis requires extraordinary ability and some go so far as to propose that there is a masterful and profoundly imaginative side to making models, for the most part, predictive models require some fundamental steps of creating them. These steps are shown in Figure 7.

3.2. Techniques of Predictive Analysis

Predictive analytics techniques can be broadly classified into Regression techniques and Machine Learning techniques. Further Regression technique includes a Linear Regression model, discrete choice model, Logistic regression, Multinomial logistic regression and many more. Similarly, the Machine Learning technique includes neural networks, Radial basis functions, Support vector machines, Naïve Bayes among many. Let's discuss some of them in brief but first, let's understand the term "Predictive Model" as it will be used many times in this paper.

The predictive model is like a procedure or system that with the help of individual input (variables) predicts the behavior and as an output, it may provide a score which on being higher or less tells about the predicted behavior.

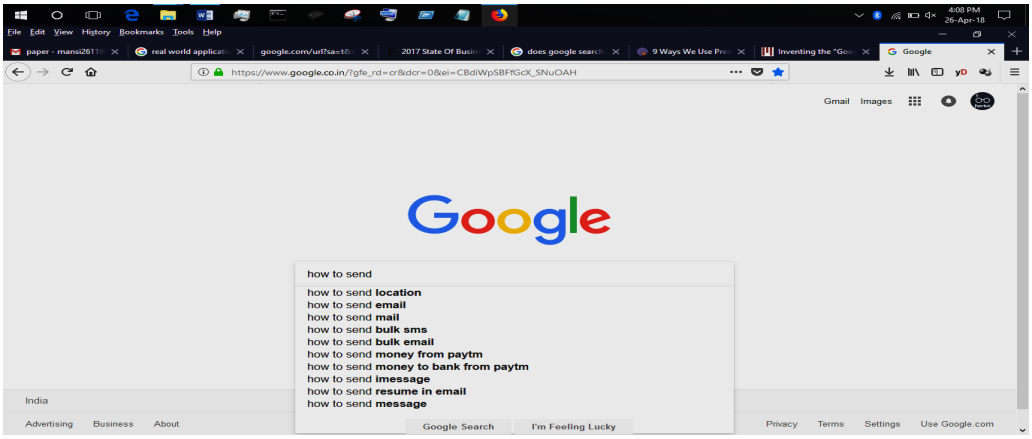

Figure 3. Google search.

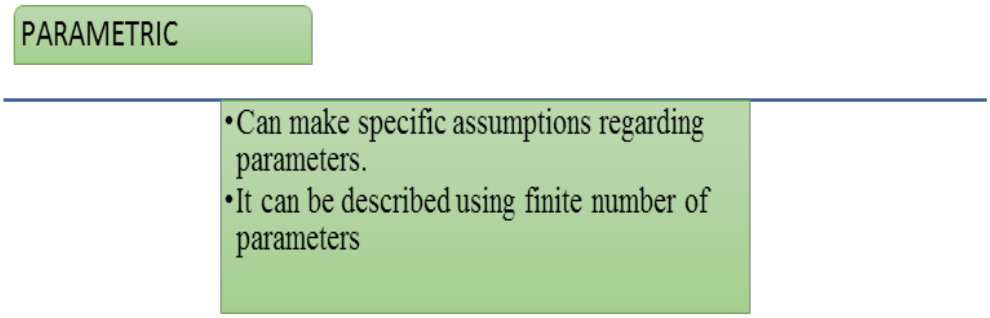

Figure 4. Parametric model.

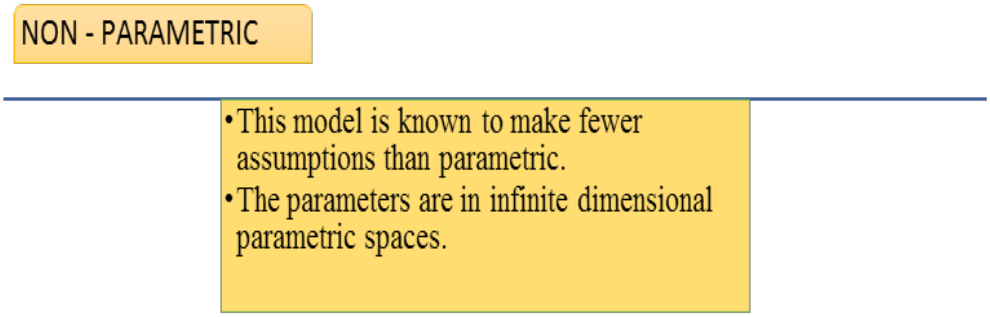

Figure 5. Non-parametric model.

3.2.1. Regression Techniques

Regression models are the pillars of predictive analysis. The attention is on building up a mathematical equation as a model to represent the interactions between the different variables in consideration. For example, the relation between drink driving and a number of car accidents. Regression analysis is one of the most popular techniques.

1. **Linear Regression model**: This model tends to analyze the relation among the dependent variable as well as the set of dependent variables. Thereafter this relationship is expressed in terms of equations that are responsible for predicting the responsive variable as a parameter with a linear function. The said parameters are thereby adjusted so as to optimize his measure of fit accurately.[7] Dependent variable is called a continuous variable while the independent variable seems to be discrete. The relationship between these

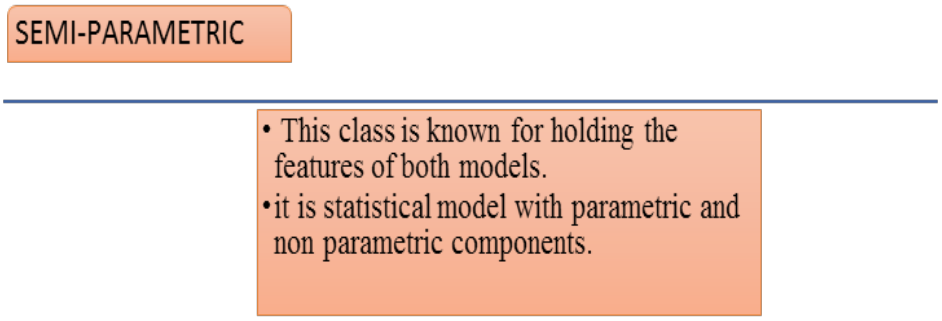

Figure 6. Semi-parametric model.

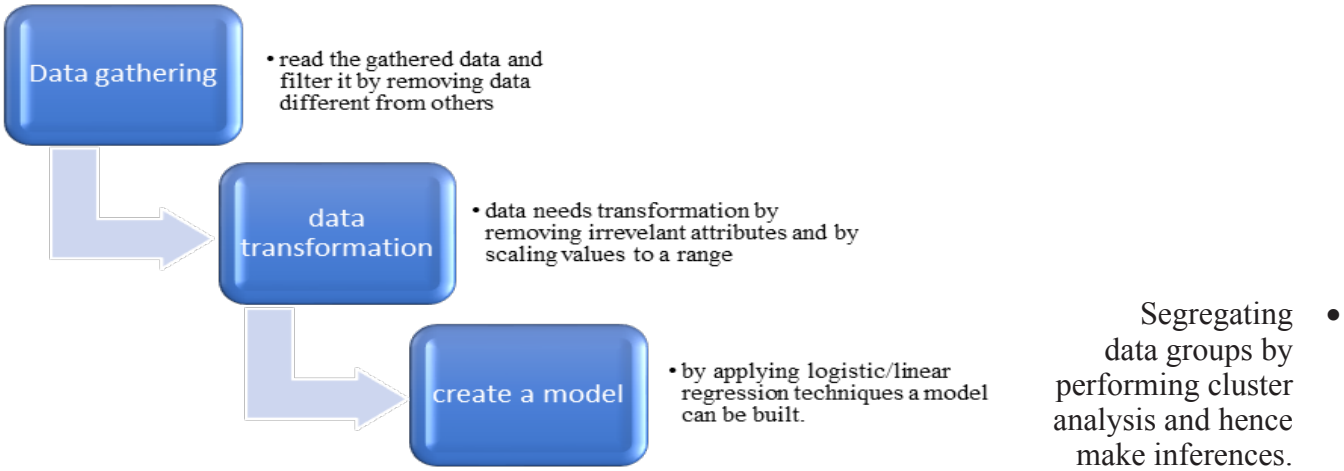

Figure 7. Stages of predictive modeling.

variables is established by the regression lines. In this model of regression, the nature of the line is called linear.

2. **Discrete choice model**: Generally, the multiple regressions are used in case of response variable being continuous. On the other hand, the response variable is not always continuous and can be discrete or discontinuous too. Here multiple regressions mean that the result or value a variable is based on more than one variable.
3. **Logistic Regression:** Logistic Regression is a characterization calculation. It is utilized to anticipate a paired result (1/0, Yes/No, True/False) given an arrangement of autonomous factors. To represent the binary/categorical result, we utilize dummy variables. You can likewise consider logistic regression as an uncommon instance of linear regression when the result variable is clear cut, where we are utilizing log of chances as the dependent variable. In straightforward words, it predicts the likelihood of the event of an occasion by fitting information to a logit function.
4. **Stepwise Regression**: This type of regression is utilized when we manage numerous free factors. In this strategy, the determination of autonomous factors is finished with the assistance of a programmed procedure, which includes no human intervention. This accomplishment is accomplished by watching factual esteems like R-square, t-details and AIC metrics to perceive huge factors. Stepwise regression essentially fits the regression model by including/dropping co-variates each one in turn in light of a predefined measure.
5. **Ridge Regression**: Ridge Regression is a procedure utilized when the information experiences multi-collinearity (autonomous factors are much corresponded). In multi-collinearity, despite the fact that the minimum squares estimates (OLS) are unbiased, their changes are extensive which goes amiss the

observed value is a long way from the true value. By adding a level of inclination to the regression estimates, edge relapse decreases the standard blunders. In a linear eqn, the prediction blunders are decayed into multiple subcategories.

3.2.2. Machine Learning Techniques

Machine learning is the utilization of Artificial Intelligence (AI) that gives frameworks the capacity to consequently take in and enhance for a fact without being unequivocally modified. Machine learning centers around the improvement of PC programs that can get to information and utilize it learn for themselves.

1. Neural Networks

Neural systems are appropriate for recognizing non-linear patterns, as in designs where there isn't a direct, balanced connection between the input and the yield. Rather, the systems recognize designs between mixes of input sources and a given yield.

What separates neural systems from other machine learning algorithms is that they make utilization of an architecture inspired by the neurons in the human brain. These systems end up being appropriate to displaying high-level deliberations over a wide cluster of disciplines and businesses.

2. Radial Basis Functions

A radial basis function (RBF) is a term that portrays any genuine esteemed capacity whose yield depends solely on the separation of its contribution from some starting point.

In machine learning, radial basis functions are most regularly utilized as a portion for characterization with the help vector machine (SVM). For this use of RBFs, the Gaussian radial basis function is practically constantly chosen.

RBFs can likewise once in a while be utilized as initiation works in neural systems to frame what is known as a Radial Basis Function Network.

3. Support Vector Machines

This algorithm is utilized for the classification and regression challenges. This is used for most of the parts which are utilized as a part of problems based on classification. Every data item is plotted as a point in n-dimensional space having the estimation of every feature being the estimation of a specific co-ordinate. At this movement of time, arrangements are performed by the help of searching the hyperplane which is useful in separating the two sorts of classes extremely well.

4. Naïve Bayes'

This gives us a chance to look at the likelihood of an occasion in light of the earlier information of any occasion that identified with the previous occasion. So, for instance, the likelihood that the cost of a house is high can be better surveyed if we know the facilities around it, contrasted with the evaluation made without the learning of the area of the house. Bayes' theorem does precisely that.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Naïve Bayes' equation[8]

Above equation gives the basic representation of Bayes' theorem. Here A and B are two events and,

P(A/B): the conditional probability that event A occurs, given that B has occurred. This is also known as the posterior probability.

P(A) and P(B): the probability of A and B without regard to each other.

P(B/A): the conditional probability that event B occurs, given that A has occurred.

4. Applications of Predictive Analysis

1. **Medical Decision Support System**: Doctors can learn a lot from previous case results. Knowing and having some analyzed data of previous related works can provide a helping hand in creating a solution for new problems.[9] Experts utilize predictive analysis in health care principally to figure out which patients are in danger of building up specific conditions like diabetes, asthma, coronary illness and other lifetime diseases.
2. **Fraud Detection**: Fraud is generally spread crosswise over enterprises. Instances of fraud show up in different fields, for example, Visa initiations, solicitations, government forms, online exercises, protection cases and telecom call exercises. Predictive modeling may be used to detect financial statement fraud in companies. Fraud detection can help companies to expand their revenue potential. Preventing any fraud not only saves the company from being under the financial crisis but also improves quality and vulnerability.

3. **Insurance**: Similar to fraud, surprisingly high and suspicious cases are the most despicable aspect of insurance agencies. They might want to avoid paying such claims. In spite of the fact that the goal is sufficiently basic, predictive modeling has had just partial accomplishment in wiping out this source of high loss to organizations. For example, life insurers are known to be one of the early users of statistics and data analysis. Various methods give information about relative risk selection and are also accepted broadly by this industry.[10]
4. **Health**: While the precise uses of predictive modeling in human services are generally new, the principal applications are like those in alternate areas. After all limiting customer risk is the target. In healthcare, this is the danger of readmission, which can be decreased by recognizing high-risk patients and checking them. Medical cases from different places are analyzed and the goal is to find something that can be of use for future or present cases.[11]
5. **Client Retention**: By a successive examination of a client's past service use, performance, spending and other conduct designs, predictive models can decide the probability of a client needing to end service at some point.

5. Conclusion

In this study, we have discussed the predictive analysis which means predicting a result and analyzing that result. As we discussed in some cases we came to realize the need for such analysis and the benefits it gives us and this is the reason behind why predictive analytics is such a trending and popular topic today. Today every startup, company is it small or big, social media site, e-commerce site wants to let their customers and user know that they know them. This may seem like magic that how an organization can know us so better and predictive models, algorithms and techniques make this magic happen hence it becomes important to understand these techniques. Learning about different techniques behind it is really important to understand how the results are predicted and used for the sfuture. Today predictive analytics finds its application in a number of fields like health, fraud detection, insurance, financial prediction and many more. The predictive analysis provides the true use of big data. It is a way to utilize all that data and convert it into useful and profitable results. The future scope of predictive analytics is vast. The use of self-driving cars (driverless cars, autonomous cars) is the future and its increased usage and availability can prove to be a blessing for disabled people as such cars make use of predictive analytics and machine learning process. Just like Google Flu trends many new ways can be discovered to predict the outbreak of any virus or disease, prior knowledge of any such situations can help and give time to prepare ourselves and find a cure. IT can be used to know the probability of success and failure for an approach or procedure to be performed on patients. In the near future, we can expect this approach to be so widely used that the predictive analytics products will be easily accessible to everyone at a cheaper price. Its integration with IoT can prove to be really successful and profitable for many organizations as interconnected devices can provide so much information and data.

References

1. Balachandran A. Large scale data analytics of user behavior for improving content delivery. No. CMU-CS-14-142. Carnegie-Mellon Univ Pittsburgh PA School of Computer Science; 2014. P. 1–121.
2. Mishra N, Silakari S. Predictive analytics: a survey, trends, applications, opportunities & challenges. Int J Comp Sci Inf Technol. 2012;3(3):4434–38.
3. Bradlow ET. The role of big data and predictive analytics in retailing. J Retailing. 2017;93(1):79–95.
4. 9 ways we use predictive analytics without even knowing it – predictive analytics part 6. [cited 2013 Oct 18]. https://www.fusioncharts.com/blog/9-ways-we-use-predictive-analytics-without-even-knowing-it-predictive-analytics-part-6/.
5. Hill T, Smith ND, Mann MF. Role of efficacy expectations in predicting the decision to use advanced technologies: the case of computers. J Appl Psychol. 1987;72(2):307–13.
6. Predictive modeling. [cited 2019 Apr 26]. https://en.wikipedia.org/wiki/Predictive_modelling.
7. Ostrom CW. Time series analysis: regression techniques. Vol. 9. Sage; 1990.
8. Eubank RL, Spiegelman CH. Testing the goodness of fit of a linear model via nonparametric regression techniques. J Am Stat Assoc. 1990;85(410):387–92.
9. Shmueli, Galit and Otto R. Koppius. "Predictive Analytics in Information Systems Research." *MIS Quarterly* 35 (2011): 553-572.
10. Predictive modeling for life insurance. [cited 2017 Dec 01]. https://uwaterloo.ca/advances-in-predictive-analytics/sites/ca.advances-in-predictive-analytics/files/uploads/files/kevin-pledge.pdf.
11. Banumathi S, Aloysius A. Predictive analytics concepts in big data – a survey. Int J Adv Res Comp Sci. 2017;8(8):1–4.