A Review of Predictive Analytics: Techniques, Advantages, Related Works and Applications

Madhuri S. Desai
Department of Computer Science & Engineering,
Shreeyash College of Engineering & Technology,
Chatrapati Sambhajinagar, India

P.S. Deshpande, PhD
Department of Computer Science & Engineering,
Shreeyash College of Engineering & Technology,
Chatrapati Sambhajinagar, India

ABSTRACT
Predictive analysis is the type of data analysis that help organisation predict future event by identifying patterns and its historical data. Predictive analytics uses data-mining, predictive modelling and machine learning techniques that an analysis of historic data and based on that analysis to predict the future. A model can be created to predict using Predictive Analytics modeling techniques. The form of these predictive models varies depending on the data they are using. Classification & Regression are the two main objectives of predictive analytics. Predictive Analytics is composed of various statistical & analytical techniques used to develop models that will predict future occurrence, events or probabilities. Predictive analytics is able to not only deal with continuous changes, but discontinuous changes as well. Classification, prediction, and to some extent, affinity analysis constitutes the analytical methods employed in predictive analytics.

Keyword
Predictive analytics Techniques, Predictive analytics process, Opportunities, Advantages, Disadvantages, Application, Benefits.

1. INTRODUCTION
The term predictive analytics refers to the use of statistics and modeling techniques to make predictions about future outcomes and performance. It analyses the current and historical data in order to make predictions about the future by employing the techniques from statistics, data mining, machine learning, and artificial intelligence. It brings together the information technology, business modeling process, and management to make a prediction about the future. Industries and disciplines, such as insurance and marketing, use predictive techniques to make important decisions.

Predictive Analytics Process
Predictive analytics involves several steps through which a data analyst can predict the future based on the current and historical data. This process of predictive analytics is represented in figure 1 given below

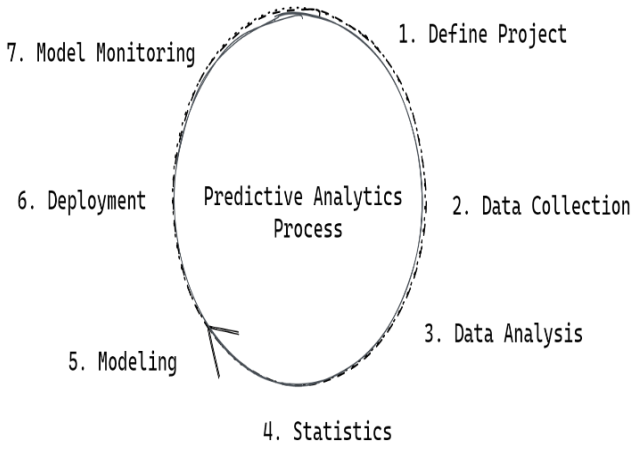

Figure 1: Predictive Analytics Process

1. Define Project
Define the project's outcomes, deliverables, scope of the effort, and business objectives, as well as the datasets that will be employed.

2. Data Collection
Predictive analytics data mining prepares data from multiple sources for analysis, providing a detailed picture of customer interactions.

3. Data Analysis
It is the process of examining, cleaning, converting, and modeling data to extract useful information and draw conclusions.

4. Statistics
Statistical analysis validates assumptions and hypotheses by utilizing standard statistical models to test them.

5. Modeling
Predictive modeling develops accurate predictive models about the future automatically. There are additional alternatives for evaluating multiple models to find the optimal answer.

6. Deployment
Predictive model deployment integrates analytical results into day-to-day decision-making processes, providing results, reports, and outputs by automating decisions based on the modeling.

7. Model Monitoring:
Model performance is reviewed and monitored to ensure that the model is delivering the expected results.

2. PREDICTIVE ANALYTICS OPPORTUNITIES

Though there is a long history of working with predictive analytics and it has been applied widely in many domains for decades, today is the era of predictive analytics due to the advancement of technologies and dependency on data. The most common opportunities in the field of predictive can be listed as:

- **Detecting Fraud**

Detection and prevention of criminal behavior patterns can be improved by combining the multiple analysis methods. The growth in cybersecurity is becoming a concern. The behavioral analytics may be applied to monitor the actions on the network in real time. It may identify the abnormal activities that may lead to a fraud. Threats may also be detected by applying this concept.

- **Reduction of Risk**

Likelihood of default by a buyer or a consumer of a service may be assessed in advance by the credit score

applying the predictive analytics. The credit score is generated by the predictive model using all the data

related to the person's creditworthiness. This is applied by credit card issuers and insurance companies to identify the fraudulent customers.

- **Marketing Campaign Optimization**

The response of customers on purchase of a product may be determined by applying predictive analytics. It may also be used to promote the cross-sale opportunities. It helps the businesses to attract and retain the most profitable customers.

- **Operation Improvement**

Forecasting on inventory and managing the resources can be achieved by applying the predictive models. To set the prices of tickets, airlines may use predictive analytics. To maximize its occupancy and increasing the revenue, hotels may use predictive models to predict the number of guests on a given night. An organization may be enabled to function more efficiently by applying the predictive analytics

- **Clinical Decision Support System**

Expert systems based on predictive models may be used for diagnosis of a patient. It may also be used in the development of medicines for a disease.

3. PREDICTIVE ANALYTICS TECHNIQUES

List of the important techniques below which are used popularly in developing the predictive models.

3.1 Decision Tree

A decision tree is a classification model but it can be used in regression as well. It is a tree-like model which relates the decisions and their possible consequences. The consequences may be the outcome of events, cost of resources or utility. In its tree-like structure, each branch represents a choice between a number of alternatives and its every leaf represents a decision. Based on the categories of input variables, it partitions data into subsets. It helps the individuals in decision analysis. A typical model of the decision tree is represented in figure 2.

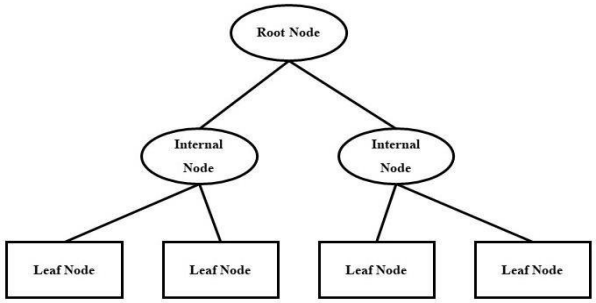

Figure 2: Decision Tree

A decision tree is represented in figure as a tree-like structure. It has the internal nodes labeled with the questions related to the decision. All the branches coming out from a mode are labeled with the possible answers to that question. The external nodes of the tree called the leaves, are labeled with the decision of the problem. This model has the property to handle the missing data and it is also useful in selecting the preliminary variables.

3.2 Regression Model

Regression is one of the most popular statistical techniques which estimates the relationship between variables. It models the relationship between a dependent variable and one or more independent variables. Regression analysis is used for one of two purposes: predicting the value of the dependent variable when information about the independent variables is known or predicting the effect of an independent variable on the dependent variable. This modeled relation between dependent and independent relation is represented in figure 3.

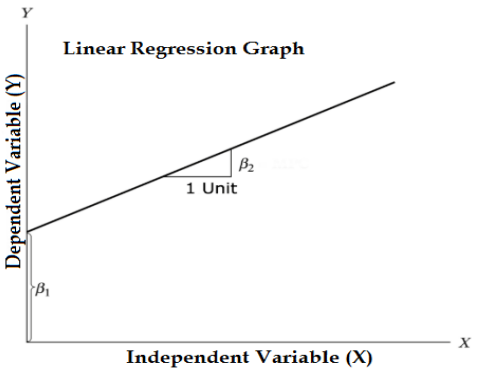

Figure 3: Regression Model

In the context of the continuous data, which is assumed to have a normal distribution, the regression model finds the key pattern in large datasets. It is used to find out the effect of specific factors influence the movement of a variable. In regression, the value of a response variable is predicted on the basis of a predictor variable. In this case, a function known as regression function is used with all the independent variables to map them with the dependent variables. In this technique, the variation of the dependent variable is characterized by the prediction of the regression function using a probability distribution. There are two types of regression models are used in predictive analytics, the linear regression model, and the logistic regression model.

3.3 Bayesian Statistics

This technique belongs to the statistics which takes parameters as random variables and use the term "degree of belief" to define the probability of occurrence of an event. The Bayesian statistics is based on Bayes' theorem which terms the events

priori and posteriori. In conditional probability, the approach is to find out the probability of a posteriori event given that priori has occurred. On the other hand, the Bayes' theorem finds the probability of priori event given that posteriori has already occurred. It is represented in figure 4.

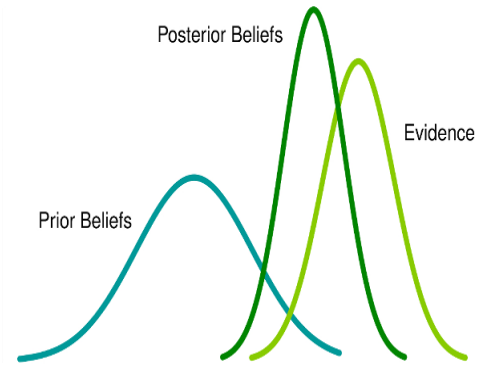

Figure 4: Bayesian Statistics

It uses a probabilistic graphical model which is called the Bayesian network which represents the conditional dependencies among the random variables. This concept may be applied to find out the causes with the result of those causes in hand. For example, it can be applied in finding the disease based on the symptoms.

3.4 Artificial Neural Network

Artificial neural network, a network of artificial neurons based on biological neurons, simulates the human nervous system capabilities of processing the input signals and producing the outputs. This is a sophisticated model that is capable of modeling the extremely complex relations. The architecture of a general purpose artificial neural network is represented in figure 5.

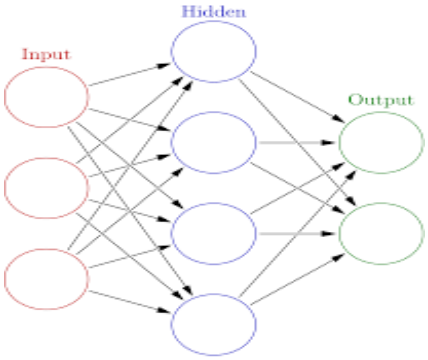

Figure 5: Artificial Neural Network

Artificial neural networks are used in predictive analytics application as a powerful tool for learning from the example datasets and make a prediction on the new data. Through the input layer of the network, an input pattern of the training data is applied for the processing and it is passed to the hidden layer which a vector of neurons. Various types of activation functions are used at neurons depending upon the requirement of output. The output of one neuron is transferred to the neurons of next layer. At the output layer, out is collected that may be the prediction on new data.

3.5 Support Vector Machine

It is supervised kind of machine learning technique popularly used in predictive analytics. With associative learning algorithms, it analysis the data for classification and regression. However, it is mostly used in classification applications. It is a discriminative classifier which is defined by a hyperplane to classify examples into categories. It is the representation of examples in a plane such that the examples are separated into categories with a clear gap. The new examples are then predicted to belong to a class as which side of the gap they fall. The example of separation by a support vector machine is represented in figure 6.

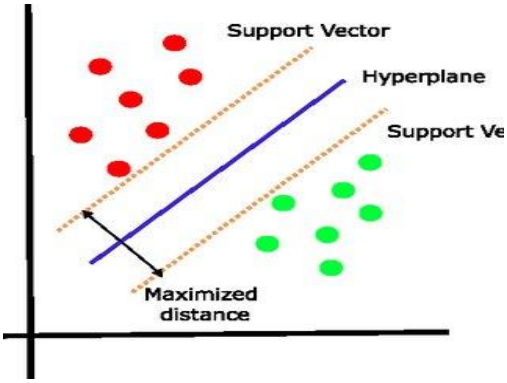

Figure 6: Support Vector Machine

Table 1: Advantages and disadvantages of different predictive techniques

Sr. NO.	Technique	Advantages	Disadvantages
1.	**Decision tree**	1) Compared to other algorithms decision trees requires less effort for data preparation during pre-processing. 2) A decision tree does not require normalization of data.	1) Decision tree often involves higher time to train the model. 2) Decision tree training is relatively expensive as the complexity and time has taken are more.
2.	**Regression Model**	1) Regression models are easy to understand as they are built upon basic statistical principles, such as correlation and least-square error. 2) The output of regression models is an algebraic equation that is easy to understand and use to predict.	1) The predictive power of regression models matches with other predictive models and sometimes performs better than the competitive models. 2) Regression models can include all the variables that one wants to include in the model.
3.	**Bayesian Statistics**	1) Extremely efficient when the dataset is tiny. 2) Particularly well-suited for online learning as opposed to batch learning, when we know the complete dataset before we begin training the model. This is so that Bayesian Regression can be used without having to save data.	1) The model's inference process can take some time. 2) The Bayesian strategy is not worthwhile if there is a lot of data accessible for our dataset, and the regular probability approach does the task more effectively.
4.	**Artificial Neural Network**	1) Robustness to incomplete or noisy data 2) Feature extraction capabilities 3) Require less formal statistical training 4) Ability to detect complex nonlinear relationships between dependent and independent variables.	1) Lack of transparency in decision-making 2) Potential overfitting without proper regularization 3) Complexity and difficulty in model tuning.
5.	**Ensemble Learning**	1) Improved accuracy: Ensemble voting can improve the accuracy of predictions by combining the strengths of multiple models and reducing the impact of individual model weaknesses. 2) Versatility: Ensemble voting can be used with a wide variety of machine learning models, making it a versatile approach to prediction.	1) Increased complexity: Ensemble voting can be more complex to set up and manage than individual models, as it requires coordination and integration of multiple models. 2) Computationally expensive: Ensemble voting can be computationally expensive, as it requires training and running multiple models and combining their predictions.
6.	**Gradient boosting**	1) Interpretable models: Gradient Boosting algorithms such as XGBoost and LightGBM produce models that are easy to interpret, as they can provide feature importance scores and decision trees. 2) Faster training: Gradient Boosting algorithms generally require less training time and computational resources than deep learning algorithms, making them a good choice for smaller datasets or when rapid model development is needed.	1) Limited scalability: Gradient Boosting algorithms can struggle with large datasets and may require more memory and computational resources than deep learning algorithms. 2) Less flexibility: Gradient Boosting algorithms can be less flexible than deep learning algorithms when it comes to handling complex tasks like natural language processing and computer vision.
7.	**Support Vector Machine**	1) SVM works relatively well when there is a clear margin of separation between classes. 2) SVM is more effective in high dimensional spaces. 3) SVM is effective in cases where the number of dimensions is greater than the number of samples. 4) SVM is relatively memory efficient.	1) SVM algorithm is not suitable for large data sets. 2) SVM does not perform very well when the data set has more noise i.e. target classes are overlapping. 3) In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
8.	**k-nearest neighbors (k-NN)**	1) Simple and Easy to Understand 2) No Training Required 3) Can Handle Large Datasets 4) Non-parametric 5) Accurate and Effective	1) Sensitive to Outliers 2) Computationally Expensive 3) Requires Good Choice of K 4) Limited to Euclidean Distance 5) Imbalanced Data

4. BENEFITS OF PREDICTIVE ANALYTICS

- **Scalability:** Data science and data engineering tasks can be automated. Models can be trained, tested, and deployed across multiple enterprise applications in real-time. Extend data science skills in hybrid and multi-cloud systems.
- **Simplicity:** To handle the complete data science lifecycle, use a central platform. Develop and deploy consistent processes. Create unified data governance and security framework for the entire enterprise.
- **Speed:** Utilize pre-built applications and models that have already been trained. With state-of-the-art and open-source technologies, data scientists and business teams may collaborate and streamline model construction.
- **Fraud Detection:** Predictive analytics analyses all network actions in real-time to spot anomalies that could signal fraud or other vulnerabilities.
- **Operations Improvement:** Businesses employ predictive analytics models to estimate inventory, manage resources, and run more efficiently.
- **Risk Reduction:** Predictive analytics is used to examine and determine the risk of future defaults in credit scores, insurance claims, and debt collections.
- **Predictive Maintenance:** Data is used by organizations to predict when routine equipment maintenance is needed and to plan it before a problem or malfunction occurs.

5. APPLICATION OF PREDICTIVE ANALYTICS

There are many applications of predictive analytics in a variety of domains. From clinical decision analysis to stock market prediction where a disease can be predicted based on symptoms and return on a stock, investment can be estimated respectively. We will list out here below some of the popular applications.

- **Banking and Financial Services:** In the banking and financial services industry, predictive analytics and machine learning are used in conjunction to detect and reduce fraud, measure market risk, identify opportunities and much, much more.
- **Retail:** Retailers are using predictive analytics and machine learning to better understand consumer behaviour; who buys what and where? These questions can be readily answered with the right predictive models and data sets, helping retailers to plan ahead and stock items based on seasonality and consumer trends – improving ROI significantly.
- **Health and Insurance:** The pharmaceutical sector uses predictive analytics in drug designing and improving their supply chain of drugs. By using this technique, these companies may predict the expiry of drugs in a specific area due to lack of sale. The insurance sector uses predictive analytics models in identifying and predicting the fraud claims filed by the customers. The health insurance sector using this technique to find out the customers who are most at risk of a serious disease and approach them in selling their insurance plans which be best for their investment.
- **Government and Public Sector:** The government agencies are using big data-based predictive analytics techniques to identify the possible criminal activities in a particular area. They analyse the social media data to identify the background of suspicious persons and forecast their future behavior. The governments are using the predictive analytics to forecast the future trend of the population at country level and state level. In enhancing the cybersecurity, the predictive analytics techniques are being used in full swing.
- **Transportation and Logistics:** Predictive analytics techniques help logistics companies and ride-sharing platforms optimize routes based on real-time traffic data, weather conditions, and historical trends, reducing fuel costs and delivery times. In transportation networks, predictive analytics techniques can provide dynamic routing solutions based on live data, such as traffic flow, congestion, and incidents. This helps drivers avoid delays, ensuring on-time deliveries, especially for last-mile delivery in urban areas.
- **Education:** The education sector is used predictive analytics techniques to identify students at risk of falling behind academically or dropping out based on early indicators like grades, attendance, and engagement. Schools and universities use predictive analytics techniques to forecast enrollment trends based on factors like historical enrollment patterns, local demographics, and economic conditions. This helps institutions plan resources, adjust teaching loads, and predict future infrastructure.
- **Agriculture:** The agriculture sector is undergoing a transformation with the help of advanced technologies like predictive analytics techniques**.** Predictive analytics techniques help forecast crop yields by analysing weather patterns, soil moisture, temperature, and other environmental factors. Predictive analytics is used to monitor soil health through sensors that track soil pH, nutrient levels, and moisture. This data helps farmers determine the right time for irrigation, fertilization, and planting, leading to more efficient use of resources and better crop growth.

6. RELATED WORKS

The table in the below section states the work done using various algorithms and modeling techniques in different areas such as education, finance-driven models, weather forecasting, medical research, neuroscience, agriculture, text-mining etc. The applied techniques enhance the simplicity, effectiveness and makes the system more robust.

Table 2: Different sectors, applications, and benefits of each prediction techniques

Sector	Techniques used	Applications	Benefits
Education	Decision Tree	Governance and management of educational institution	Simplicity
Banking, financial services and insurance	Regression Techniques, KNN	Stock prices, banking	Risk Reduction, Fraud Detection,
Weather forecasting	Decision Tree, Naïve Bayes, KNN	Estimate the weather	Accuracy

Agriculture	KNN	Climate forecasting And estimating soil and water parameters	Simplicity, effectiveness, robust
Marketing, Business Model	Support Vector Machine, Decision trees, Neural networks	Bankruptcy prediction model	Accuracy, Fraud Detection
Production and Manufacturing	Support Vector Machine, Neural networks	Predictive maintenance schedules and reduce equipment downtime	Improve production efficiency

7. CONCLUSION AND FUTURE SCOPE

There has been a long history of using predictive models in the tasks of predictions. Earlier, the statistical models were used as the predictive models which were based on the sample data of a large-sized data set. Predictive analytics techniques are revolutionizing industries by providing organizations with the tools to make data-driven decisions, optimize processes, and foresee future trends. By leveraging a variety of methods such as regression analysis, time series forecasting, machine learning neural networks, predictive analytics helps businesses and industries predict outcomes, improve operational efficiency, mitigate risks, and gain deeper insights into customer behavior, market conditions, and potential disruptions. These techniques are being applied in diverse areas like finance, healthcare, marketing, manufacturing, and agriculture, With the improvements in the field of computer science and the advancement of computer techniques, newer techniques have been developed and better and better algorithms been introduced over the period of time. The developments in the field of artificial intelligence and machine learning have changed the world of computation where intelligent computation techniques and algorithms are introduced. This paper opens a scope of development of new models for the task of predictive analytics. There is also an opportunity to add additional features to the existing models to improve their performance in the task.

8. REFERENCES

[1] Anitha K, Bhoomika C, J Andrea Kagoo, Kruthika K, Aruna Mg, "Student Grade Prediction Using Multiclass Model", August 2022, Ijirt, Volume 9 Issue 3, Issn: 2349-6002.

[2] Eric Siegel, 2016, “Predictive Analytics”, John Willey and Sons Ltd.

[3] Rustagi, Mitanshi; Goel, Neha Predictive Analytics: A study of its Advantages and Applications IARS' International Research Journal, vol. 12, núm. 1, 2022.

[4] Raheja, S. and Jain, D. (2021) “AN EMPIRICAL STUDY OF SELECTED INDIAN COMPANIES’ TALENT MANAGEMENT”, Amity Journal of Professional Practices. Florida, USA, 1(1). doi: 10.55054/ ajpp. v1i1.461.

[5] John Edwards,” Predictive analytics: Transforming data into future insights | CIO”, article - 16 AUGUST 2019 15:30 IST.

[6] Peter M Lee, 2012, “Bayesian Statistics: An Introduction, 4th Edition”, John Willey and Sons Ltd.

[7] Charles Elkan, 2013, “Predictive analytics and data mining”, University of California, San Diego.

[8] B Kaminski, M Jakubczyk, P Szufel, 2018, “A framework for sensitivity analysis of decision trees”, Central European Journal of Operations Research, Vol26, Issue-1, Pages-135-159.

[9] Ben Hur et al, 2001, “Support Vector Clustering”, Journal of Machine Learning Research, Vol-2, Pages125-137.

[10] Kam Cheong Li, Hoi Kuan Lam, Sylvia Siu-Yin LamA Review of Learning Analytics in Educational Research. Technology in Education. Technology-Mediated Proactive Learning, 559 (2015), pp. 173-184.

[11] Vaibhav Kumar Predictive Analytics: A Review of Trends and Techniques July 2018 International Journal of Computer Applications 182(1):31-37.

[12] Charles Nyce, 2007, “Predictive Analytics White Paper”, American Institute of CPCU/IIA.