An Overview of Machine Learning Techniques and Tools for Predictive Analytics

Darshan Labhade[1], Nikhil Lakare[2], Aniket Mohite[3], Siddhesh Bhavsar[4], Sushma Vispute[5], Govind Mahajan[6]
[1,2,3,4,5]Department of Computer, Engineering, Pimpri, Engineering, Chinchwad College Engineering, Pune, India
[6]Director Redivivus Technology Pvt. Ltd, Pune, India
[1]labhadedarshan@gmail.com, [2] nikhil.lakare4@gmail.com, [3]aniket91mohite@gmail.com, [4]siddhesh099@gmail.com,
[6]govind.mahajan@redivs.com

***Abstract* — Predictive analytics is the use of raw facts or data, algorithms of statistics and techniques of machine learning to identify what is the possibility of future outcomes based on historical data. Our main goal is to get the knowledge of what has happened in the past and predict future scenarios. This paper gives a brief introduction of various machine learning techniques and tools which use these machine learning techniques to accurately predict the outcomes based on the given data and business requirement. Furthermore, this paper is aimed help beginners in the field of predictive analytics to choose between various tools and techniques available in the market which can maximize the accuracy and outcomes.**

Keywords* — *Prediction; Analytics; Machine Learning; Techniques; Tools; Data Mining;

I. INTRODUCTION

Predictive analytics is composed of two words predict & analysis, but it works in reverse viz. first analyse then predict. It is human nature to want to know and predict what the future holds. Predictive analytics deals with the prediction of future events based on previously observed historical data by applying sophisticated methods like machine learning. The historical data is collected and transformed by using various techniques like filtering, correlating the data, and so on. Prediction process can be divided into four steps: collect and pre-process raw data; transform pre-processed data into a form that can be easily handled by the (selected) machine learning method; create the learning model (training) using the transformed data; report predictions to the user using the previously created learning model [1].

II. PREDICTIVE ANALYTICS AND DATA MINING

The advance data mining leads to predictive data analytics. Data mining and data extraction terms are almost seeming very similar; but there is a significant difference. Obtaining data from one data source and loading it integrated database is main task in data extraction.

Thus, one may 'extract' data from a source or legacy system to put it into a standard database or data warehouse. On the other hand, extraction of obscure or hidden predictive information from large databases or data warehouses is data mining. Data mining is searching for patterns in stores of data from database. Data mining uses computational techniques from statistics and pattern recognition from historical data. According to patterns in data thus defines the nature of data mining. [1]

III. PREDICTIVE ANALYTICS TECHNIQUES

Various risks and opportunities are determined by predictive models which analyze identify patterns in historical and transactional data. Relationships between many factors captured by forecasting models. These captured relationships give valuation of the risks or potential associated with a particular set of conditions, guiding decision making for candidate transactions. Predictive analytics is carried out by three basic techniques which are Data profiling and Transformations, Sequential Pattern Analysis and Time Series Tracking. In Data profiling and transformation technique, there are functions that data formats, change the row and column attributes and analyses dependencies, aggregate records, and make rows and columns, merge fields. Identifying relationships between the rows of data is done in sequential pattern analysis technique and it involves identifying frequently observed occurrence of items across ordered transactions over time which are sequential. Time series data is usually created by tracking corporate business metrics or monitoring industrial processes. This analysis gives the fact that the data points taken over time or timestamp.

Classification-Regression, Association analysis, Time series forecasting are some advanced Predictive analytic techniques. Classification uses attributes in data. That attributes used to assign an object to a predefined class or predict the value of a numeric variable of interest. Regression analysis is predictive modelling technique which scrutinizes the relationship between a dependent and independent variable. Association analysis describes significant associations between data elements. Time series analysis is employed for forecasting the future value of a measure based on past values. [1]

IV. PREDICTIVE MODELS

Making highly creative and artistic model requires great skill in predictive analytics so it is very skilled task and also accepted by experts. To develop predictive model, we need some basic steps. Steps are as follows:

1. Project Definition: Define the desired outcomes and business objectives for the project and convert outcomes into predictive analytic objectives and tasks;
2. Exploration: Analyze source data to determine the most efficient data and model building approach;
3. Data Preparation: Select, extract, and transform data upon which to create models;
4. Model Building: Create, test, and validate models, and evaluate whether they will meet project

metrics, objectives and goals;
5. Deployment: Apply model results to business decisions or processes;
6. Model Management: Manage models to improve performance (i.e., accuracy), control access, promote reuse, standardize toolsets, and minimize redundant activities.

According to most experts, the data preparation phase of creating predictive models is the most time-consuming part of the process. [1]

V. MODELING PROCESS

Modeling Process contain various stages; some of the stages discussed over here as follows:

Purpose: To describe the objective of the project

Obtain the data: Gathering datasets from various sources regarding the project.

1. Explore through all data, clean and pre-process the data: Exploration can be performed by describing the variables, attributes, tokens and other terms which is used in project quite general. Sometimes these terms are in cryptic form or may be in short form, for which we have to tell the full explanation and the places where it can be used.
2. Partitioning the data for training after reducing the data and partition them into training, validation and test partitions: In this stage we try to reduce the variables or terms for the sake of simplicity. We can reduce number of variables by making the small group of similar purpose variable. i.e. redundant variables.
3. To see how well the model does, we will partition the data into a training set to build the model and a validation set. This technique is in classification and prediction problem which is a part of supervised learning process. To develop other models these problems can be used for developing other model and the value of outcome variables can be used in unknown places. At this stage we can partition the data into training and validation or testing. Training will build the model and to see how well the model does partition will apply model on data. A Data mining endeavor involves testing multiple models, perhaps with multiple settings on each model. Starting from one model and testing one validation data might give us an idea about the performance of that model on such data. However, the validation data no longer provide an unbiased estimate of how the models might do with more data, until and unless we choose the best performing model. When we choose the best model the validation data becomes the part of the model itself.
4. Determining the data mining task: To find the objective data mining task is in building the model.
5. Choosing the technique: The data which is divided into training and validation partitions can be used for creating the model by various techniques.
6. Use the algorithm to perform the task: In this stage to find fitted value (by applying algorithm on training data) and predicted value (by applying algorithm on validation data) we will apply some of the algorithm. Since predicted values are for the records to which the model was fit would often be called the fitted values.
7. Interpret the results: In this stage we try other prediction algorithms and see how they perform error-wise. We might also try several settings on the various models. After choosing the best model (typically, the model with the lowest error on the validation data while also recognizing that ¨simpler is better¨), we use that model to predict the output variable in fresh data.
8. Deploy the model: After the best model is chosen, it must apply to new data.

Prediction error can be measured in various ways.

1. Average error
2. Total sum of squared errors
3. RMS error (Root mean squared error)

VI. MACHINE LEARNING ALGORITHMS FOR PREDICTIVE ANALYTICS

A. Naïve Bayes:

In the domain of machine learning, Naïve bayes is an algorithm which is mainly used for separating or classifying variety of objects depending upon the specific parameters. Naïve bayes algorithm works upon the bayes theorem. It considers strong independence between different parameters. It basically thinks that each individual feature has its own part in classifying the data, even if there might be some kind of relation between different parameters. Naïve bayes classifier works on bayes theorem which is given by

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

Where, P(A|B) mean given that B has occurred what is the probability of happening of A. The Naïve Bayes classifier is not a single machine learning algorithm but the group of many algorithms who believe that parameter or feature is (naïve) independent of other. Types of Naïve Bayes Classifier algorithm are Multinomial Naïve Bayes Bernoulli Naïve Bayes and Gaussian Naïve Bayes. Multinomial Naïve bayes is mostly used in the areas where classification of documents is required.

Bernoulli Naïve Bayes is same to Multinomial Naïve bayes but parameters are only Boolean values i.e. yes or no. Gaussian Naïve Bayes is used where parameter have continuous value. [2]

B. Support Vector Machines:

Support Vector Machine also known as SVM is supervised leaning algorithm which can be used for both classification analysis and regression analysis. In SVM we plot data point in m dimensional space where m stands for the number of parameters in the dataset. Each feature or parameter have value of a specific co-ordinate. This algorithm basically separates data points by finding out the

hyperplanes which separate different classes neatly. Support vectors means co-ordinates of individual point and support vector machine are at front who separates different classes best.

There can be many hyperplanes separating two classes but only hyperplane who has biggest margin or the distance between two support vector gets selected. While separating two classes we can make use of linear hyperplane but problem arises when we have to separate classes in multi dimension. Here, SVM makes use of the Kernel trick. Kernel is basically a function which takes lower level dimension and transforms them into higher dimension. After that SVM algorithm finds suitable hyperplanes. [3]

C. Regression Analysis:

Regression analysis consists of group of machine learning algorithms which enables us to predict value of a variable (y) depending upon the value of multiple predictor values.

It basically builds a mathematical equation which defines y as of function of predictor variables. There are many different types of regression algorithms but mostly used are linear regression and logistic regression. Linear regression algorithm tries to fit a line through the given dataset plotted on the graph such that for any data value x it can predict the value of y. Logistic regression is used when we have to find probability of any event which is binary i.e. success or failure. It is widely used in classification problems. Linear regression is used in prediction problems or forecasting problems.

D. K-Means Clustering:

K-means clustering is an unsupervised machine learning, which is mostly used to group(cluster) unlabelled data. Unlabelled data means data which do not have defined categories. This algorithm form group of data and the number of group form is given by variable k. This algorithm works iteratively to attach different data points to one of the groups. This assignment is based on the features given.

This algorithm gives us the result as centroids of the k clusters formed. These centroids can be used as label for new data. This algorithm initialises k means with some random values. It then iterates over different data points and assigns them with nearest mean and then update value of mean. To find closeness of mean to any point algorithm uses techniques such as Cosine distance, Manhattan distance or Minkowski distance [5]

Accuracy and efficiency of some of the machine learning techniques on a student dataset is given below [6]

TABLE I.

	PNN	Random Forest	Decision Tree	Naïve Bayes	Logistic Regression
Correct Classified	475	485	477	478	493
Accuracy	85.89%	87.70%	87.85%	86.43%	89.15%
Cohen's Kappa (k)	0.767	0.799	0.803	0.782	0.823
Wrong Classified	78	68	66	75	60
Error	14.105 %	12.297 %	12.155%	13.562 %	10.85%

VII. TOOLS FOR PREDICTIVE ANALYTICS

For implementing machine learning algorithms or making application using machine learning algorithm one does not have to write code from scratch. There are various tools available which enables us to implement machine learning algorithms without writing much code or no code at all. One can build complete machine learning application using UI/UX facilities provided by some machine learning tools.

Some of the most popular tools:

TABLE II.

Tool Name	Available on Platforms
Sci-Kit Learn	Windows, Linux, Mac OS
PyTorch	Windows, Linux, Mac OS
TensorFlow	Windows, Linux, Mac OS
WEKA	Windows, Linux, Mac OS
KNIME	Windows, Linux, Mac OS
Collab	Cloud-Service
Apache Mahout	Cross-Platform
Keras.io	Cross-Platform
Rapid Miner	Cross-Platform

KNIME Analytics:

KNIME stands for Konstanz Information Miner. Knime is an open source and free tool which is used for data analytics. Knime works using modular data pipelining. It provides a graphical users interface through which user can create a workflow by just drag and dropping the nodes. These nodes can be configured for different functionalities such as data reading, data pre-processing, processing, modelling, visualization. KNIME is written in java and makes use of JDBC.

Workflow:

Fig. 1.

Nodes in figure:

1. File Reader: This node is useful for reading dataset which is stored in file.
2. Partitioning: While Implementing machine learning algorithm we never use whole data for training we split data as training and testing data. This node lets us do that.
3. Simple Regression Tree learner: This node lets us train simple regression tree using training data
4. Simple Regression Tree Predictor: It is used for predicting the testing data.
5. Column Filter: It is used to remove some unwanted columns from result.
6. Line Plot: It is used for plotting a line graph for visualization purpose.
7. Numeric scorer: It is used for finding the accuracy of model.

VIII. CONCLUSION

Thus, we took an overview of different machine learning techniques and tools which uses these machine learning techniques for accurate prediction of outcomes based on the given data which will lead us to choose the right machine learning technique and tool for one's business requirement.

REFERENCES

[1] Nishchol Mishra and Dr.Sanjay Silakari, "Predictive Analytics: A Survey, Trends, Application, Opportunities and Challenges," International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4434- 4438

[2] I.Rish, "An Empirical Study of Naïve Bayes Classifier", IJCAI 2001 Empir Methods Artif Intell. 3.

[3] Sasan Karamizadeh, Shahidan M. Abdullah et al, "Advantages and Drawbacks of Support Vector Functionality," 2014 IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014), September 2-4, Langkawi, Kedah, Malaysia.

[4] Jin Huang, Jingjing Lu and Charles X. Ling, "Comparing Naïve Bayes, Decision Trees, and SVM with AUC and Accuracy," Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)0-7695-1978-4/03

[5] Arpit Bansal, Mayur Sharma, and Shalini Goel, "An Improved K-Means Clustering for Prediction Analysis using Classification Technique in Data Mining," International Journal of Computer Applications (0975 – 8887) Volume 157 – No 6, January 2017

[6] Aderibigbe Israel Adekitan and Odunayo Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," Heliyon 5 (2019) e01250

[7] S. R. Vispute, S. Kanthekar, A. Kadam, C. Kunte and P. Kadam, "Automatic Personalized Marathi Content Generation," 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), Mumbai, 2014, pp. 294-299.

[8] S. R. Vispute and M. A. Potey, "Automatic text categorization of marathi documents using clustering technique," 2013 15th International Conference on Advanced Computing Technologies (ICACT), Rajampet, 2013, pp. 1-5.

[9] S. R. Vispute, S. Patil, S. Sangale, A. Padwal and A. Ukarde, "Parallel Processing System for Marathi Content Generation," 2015 International Conference on Computing Communication Control and Automation, Pune, 2015, pp. 575-579.

[10] Sandeep Kumar, Deepak Kumar, and Rashid Ali, "Factor Analysis Using Two Stages Neural Network Architecture", International Journal of Machine Learning and Computing, Vol. 2, No. 6, December 2012

[11] Abhay Kumar, Ramnish Sinha, Daya Shankar Verma, "Modeling using K-Means Clustering Algorithm", 1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012 |

[12] J. Han and M. Kamber, "Data mining Concepts and techniques", 2nd edition, Morgan Kaufmann Publishers, pp. 401-404, 2007.

[13] Stephen J. Redmond, Conor Heneghan, "A method for initialising the K-means clustering algorithm using kd-trees", Pattern Recognition Letters 28 (2007) 965-973.

[14] O. Dekel, O. Shamir, and L. Xiao. Learning to classify with missing and corrupted features. Machine Learning, 81(2):149–178, 2010.

[15] M.Reza, F.Derakhshi, M.Ghaemi, "Classifying Different Feature Selection Algorithms Based on the Search Strategies", International Conference on Machine Learning, Electrical and Mechanical Engineering (ICMLEME'2014), Dubai (UAE)