Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 04, April-2018

Python Libraries, Development Frameworks and Algorithms for Machine Learning Applications

Dr. V. Hanuman Kumar,
Sr. Assoc. Prof.,
NHCE, Bangalore-560 103

Abstract- Nowadays Machine Learning (ML) used in all sorts of fields like Healthcare, Retail, Travel, Finance, Social Media, etc., ML system is used to learn from input data to construct a suitable model by continuously estimating, optimizing and tuning parameters of the model. To attain the stated, python programming language is one of the most flexible languages and it does contain special libraries for ML applications namely SciPy, NumPy, etc., which great for linear algebra and getting to know kernel methods of machine learning. The python programming language is great to use when working with ML algorithms and has easy syntax relatively. When taking the deep-dive into ML, choosing a framework can be daunting. The most common concern is to understand which of these frameworks has the most momentum in ML system modelling and development. The major objective of this paper is to provide extensive knowledge on various python libraries and different ML frames works to meet multiple application requirements. This paper also reviewed various ML algorithms and application domains.

Keywords: Machine Learning, Python, Framework, Algorithm, Parameters, Model, Libraries

I. INTRODUCTION:

What exactly is *'Machine Learning (ML)'*? ML is in fact a lot of things. The ML-field is quite vast and is growing rapidly, being continually partitioned and sub-partitioned ad nauseam into different sub-specialties and types [1]. Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed and solves problems that cannot be solved by numerical means alone. Among the different types of ML tasks, a crucial distinction is drawn between supervised and unsupervised learning: *Supervised ML*, the program is "trained" on a pre-defined set of "training examples", which then facilitate its ability to reach an accurate conclusion when given new data. *Unsupervised ML*- The program is given a bunch of data and must find patterns and relationships therein.

Nowadays, the strength of a company is measured by the amount of data it has. These companies analyse data and extract useful information. For instance E-Marts keeps on suggesting products based on your purchasing trends and Facebook & Twitter always suggest friends and posts in which you might be interested. Data in raw form is like crude oil, you need to refine crude oil to make petrol and diesel. Similarly, ML comes handy to process data to get useful insights in which.

ML has different models such as regression, classification and clustering. *The linear regression* is studied as a model to understand the relationship between input and output numerical values and is of the form $y=A0+A1*x$, for input value x and A0, A1 are the coefficients used in the representation with the data that we have. *The classification model* helps to identify the sentiment of a particular post. For instance a user review can be classified as positive or negative based on the words used in the comments. Emails can be classified as spam or not, based on these models. *The clustering model* helps when we are trying to find similar objects. For instance If you are interested in read articles about chess or basketball, etc.., this model will search for the document with certain high priority words and suggest article about chess or basketball respectively.

ML algorithms are being applied in lots of places in interesting ways. It's becoming increasingly ubiquitous with more and more applications in places where we cannot even think of, out of some application areas *Healthcare*, *Education* and *Data Centre Optimization.*

In *Healthcare* ML plays an essential role and is being increasingly applied to medical image segmentation, image registration, multimodal image fusion, computer-aided diagnosis, image-guided therapy, image annotation, and image database retrieval, where failure could be fatal [2].

In Education instructors need to prepare teaching materials, manually grade students' homework, and provide feedback to the students on their learning progress. Students, on the other hand, typically go through an extremely burdensome "one-size-fits-all" learning process that is not personalized to their abilities, needs, and learning context [3]. In recent advances in ML provide new opportunities to tackle with challenges in education system by collecting and analyzing the student's data and generate when they interact with a learning system.

The *Data Centres (DC)* can be larger than a football field because the virtual world is built on physical infrastructure and every search that gets submitted, email sent, page served, comment posted, and video loaded passes through DC. Those thousands of racks of humming servers use vast amounts of energy; together, all existing data centres use roughly 2% of the world's electricity, and if left unchecked, this energy demand could grow as rapidly as Internet use. So making data centres to run as efficiently as possible is a very big deal. ML is well suited for the DC environment given the complexity of plant operations and the abundance of existing monitoring data and some of the tasks handled by ML are Simulating Process Water Supply Temperature

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 04, April-2018

Increases, Catching Erroneous Meter Readings and DC Plant Configuration Optimization [4].

II. MACHINE LEARNING ARCHITECTURE:

Machine Learning, simply put is the process of making a machine, automatically learn and improve with prior experience. Recently, Machine Learning has gained a lot of popularity and is finding its way through wide areas such as medicine, finance, entertainment also.

This section discusses the architectural components shown in figure-1 involved in solving a problem using machine learning.

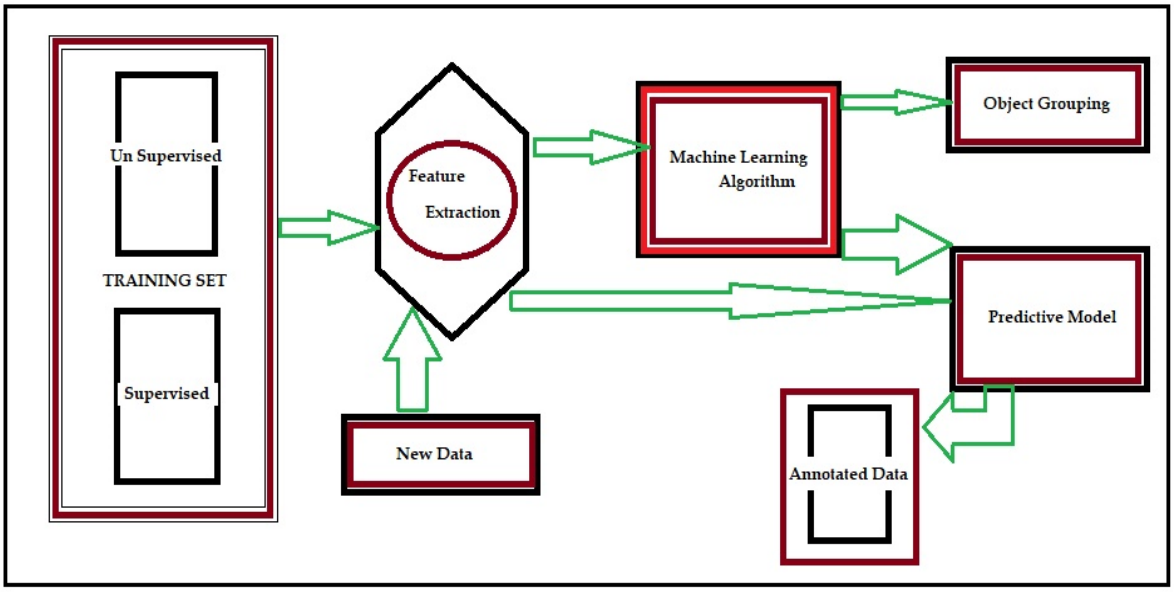

Figure-1: Architecture of Machine Learning

Training-In this step we tune our algorithm based on the data we already have. This data is called training set as it is used to train our algorithm. This is the part where our machine or software learn and improve with experience.

Domain knowledge, Feature Extraction - we really need to understand what type of data we are dealing with and what eventually we want to get out of it. Essentially we need to understand how and what *features* need to be *extracted* from the data. For instance assume we want to build software that distinguishes between male and female names. All the names in text can be thought of as our raw data while our features could be number of vowels in the name, length, first & last character, etc of the name.

Feature Selection-In many scenarios we end up with a lot of features at our disposal. We might want to *select* a *subset* of those based on the resources and computation power we have. In this level we select a few of those influential features and separate them from the not-so-influential features. There are many ways to do this, information gain, gain ratio, correlation etc.

Choice of Algorithm-There is wide range of algorithms from which we can choose based on whether we are trying to do prediction, classification or clustering. We can also choose between linear and non-linear algorithms. Naive Bayes, Support Vector Machines, Decision Trees, k-Means clustering are some common algorithms used.

Predictive modelling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation or it may be a complex neural network, mapped out by sophisticated software. As additional data becomes available, the statistical analysis model is validated or revised.

Testing-Lastly, we test how our machine learning algorithm performs on an unseen set of test cases. One way to do this is to partition the data into training and testing set. Understand the architecture of the network as a part of the testing process. Few if any will be able to actually follow a set of inputs through the network of algorithms, but understanding how the network is constructed will help testers determine if any other architecture might produce better results.

III. MACHINE LEARNING FRAMEWORK:

ML engineers are part of the engineering team who build the product and the algorithms, making sure that it works reliably, quickly, and at-scale. They work closely with data scientist to understand the theoretical and business aspect of it. ML engineers build, implement, and maintain production machine learning systems and Data scientists conduct research to generate ideas about machine learning projects, and perform analysis to understand the metrics impact of machine learning systems. In this section some of the ML frameworks are described and general ML framework as shown in figure-2.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 04, April-2018

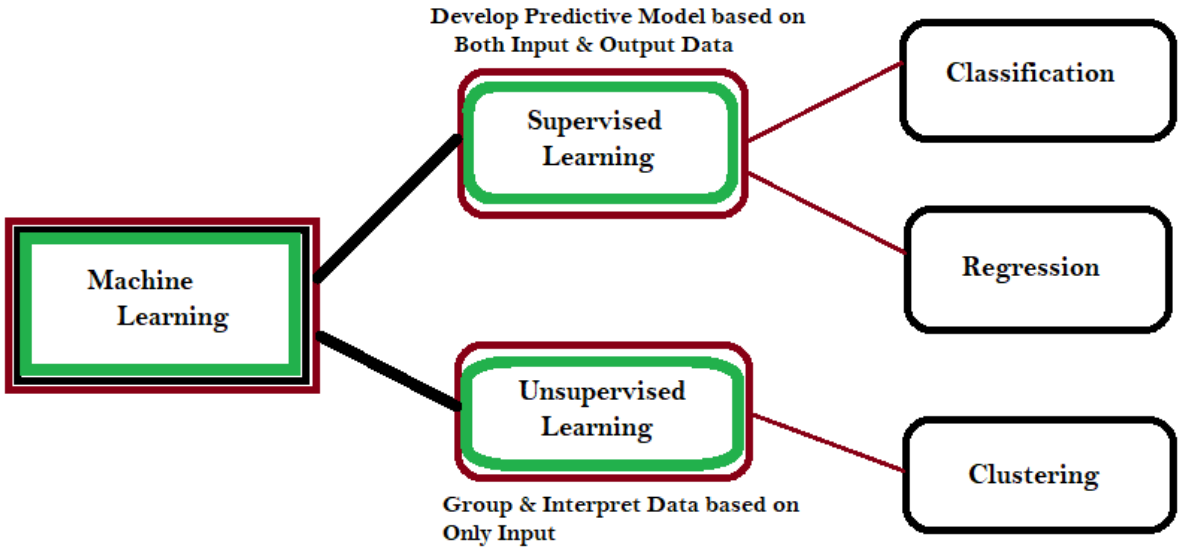

Figure-2: Machine Learning Framework

Amazon ML is a service that makes it easy for developers of all skill levels to use machine learning technology. Amazon Machine Learning provides visualization tools and wizards that guide you through the process of creating ML models without having to learn complex ML algorithms and technology. It connects to data stored in Amazon S3, Redshift, or RDS, and can run binary classification, multiclass categorization, or regression on said data to create a model.

Azure ML Studio allows Microsoft Azure users to create and train models, then turn them into APIs that can be consumed by other services. Users get up to 10GB of storage per account for model data, although you can also connect your own Azure storage to the service for larger models. A wide range of algorithms are available, courtesy of both Microsoft and third parties. You don't even need an account to try out the service; you can log in anonymously and use Azure ML Studio for up to eight hours.

Shogun is among the oldest, most venerable of machine learning libraries, Shogun was created in 1999 and written in C++, but isn't limited to working in C++. Shogun can be used transparently in such languages and environments: as Java, Python, C#, Ruby, R and Matlab. Shogun is designed for unified large-scale learning for a broad range of feature types and learning settings, like classification, regression, or explorative data analysis.

Spark /ML Lib-is Apache Spark's machine learning library. Its goal is to make practical machine learning scalable and easy. It consists of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as lower-level optimization primitives and higher-level pipeline APIs.

Tensor Flow -is an open source software library for numerical computation using data flow graphs. Tensor Flow implements what are called data flow graphs, where batches of data ("tensors") can be processed by a series of algorithms described by a graph. The movements of the data through the system are called "flows" — hence, the name. Graphs can be assembled with C++ or Python and can be processed on CPUs or GPUs.

H2O -makes it possible for anyone to easily apply math and predictive analytics to solve today's most challenging business problems. It intelligently combines unique features not currently found in other machine learning platforms including: Best of Breed Open Source Technology, Easy-to-use Web-UI and Familiar Interfaces, Data Agnostic Support for all Common Database and File Types. With H2O, you can work with your existing languages and tools. Further, you can extend the platform seamlessly into your Hadoop environments.

Massive Online Analysis (MOA) is the most popular open source framework for data stream mining, with a very active growing community. It includes a collection of machine learning algorithms (classification, regression, clustering, outlier detection, concept-drift detection and recommender systems) and tools for evaluation, relate to the WEKA project, MOA is also written in Java, while scaling to more demanding problems.

ML Pack- is a C++ based ML library originally rolled out in 2011 and designed for "scalability, speed, and ease-of-use," according to the library's creators. Implementing ML Pack can be done through a cache of command-line executables for quick-and-dirty, "black box" operations, or with a C++ API for more sophisticated work. ML pack provides these algorithms as simple command-line programs and C++ classes which can then be integrated into larger-scale machine learning solutions.

Pattern is a web mining module for the Python programming language. It has tools for data mining (Google, Twitter and Wikipedia API, a web crawler, a HTML DOM

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 04, April-2018

parser), natural language processing (part-of-speech taggers, n-gram search, sentiment analysis, Word Net), machine learning (vector space model, clustering, SVM), network analysis and <canvas> visualization.

Theano is a Python library that lets you to define, optimize, and evaluate mathematical expressions, especially ones with multi-dimensional arrays. Using Theano it is possible to attain speeds rivalling hand-crafted C implementations for problems involving large amounts of data. It was written at the LISA lab to support rapid development of efficient machine learning algorithms. Theano is named after the Greek mathematician, who may have been Pythagoras' wife. Theano is released under a BSD license.

Torch- is a scientific computing framework with wide support for machine learning algorithms that puts GPUs first. It is easy to use and efficient, thanks to an easy and fast scripting language, LuaJIT, and an underlying C/CUDA implementation. The goal of Torch is to have maximum flexibility and speed in building your scientific algorithms while making the process extremely simple. Torch comes with a large ecosystem of community-driven packages in machine learning, computer vision, signal processing, parallel processing, image, video, audio and networking among others, and builds on top of the Lua community.

Veles -is a distributed platform for deep-learning applications, and it's written in C++, although it uses Python to perform automation and coordination between nodes. Datasets can be analyzed and automatically normalized before being fed to the cluster, and a REST API allows the trained model to be used in production immediately. It focuses on performance and flexibility. It has little hard-coded entities and enables training of all the widely recognized topologies, such as fully connected nets, convolution nets, recurrent nets etc.

IV. Machine Learning Algorithms:

Machine learning poses with a wide collection of algorithms. Each algorithm may be more proficient than others with respect to few parameters. Only with advance knowledge of the dataset, an efficient algorithm can be selected. This section deals with few algorithms in classification, clustering and association.

A. Classification

1) Naive Bayesian classifier: It can be used for sentiment classification, text classification, medical data mining, and app categorization and so on. For an instance it can evaluate sentiments of movie reviews to rank the people opinion and support for those movies. When people purchase any product, they tweet about their satisfaction level. Viewer's opinions may differ according to their locations. Naive bayes is considered as a standard model in this task [9]. Facebook status updates are analyzed to reveal both positive and negative emotions of people. Naive bayes executes well when it is used for text classification. It offers higher classification accuracy when compared with the other classification techniques. Multinomial naive bayes event model is more suitable for large datasets when compared to the multi-variate bernoulli naive bayes model. Google uses document classification to index documents and find relevancy scores i.e. the Page Rank. Accuracy and computational efficiency are the best outputs of this algorithm when compared with neural networks, decision tree, and support vector machines. Google mail uses naive bayes algorithm to classify emails as spam or not.

2) Support vector machine: It can be used for asset identification and stock market forecasting. A hybrid method, combing support vector machine and decision tree was applied for stock market forecasting. This approach achieves better average precision rate when compared with bootstrap support vector machine, bootstrap decision tree and back propagation neural network. It was used to identify assets based on their connectivity. Naive bayes, k-nearest neighbor, random forest and support vector machines were applied for a give dataset and the performance of these algorithms were assessed in terms of F-score. Support vector machine based classifier was found to be good in terms of accuracy and computational expensiveness.

3) *Artificial neural networks:* It can be used for prophesying protein localization sites, cytotoxic effect in breast cancer, and forecasting electricity generation. Neural network with five hundred hidden neurons and scaled conjugate gradient algorithm was used to predict protein localization sites [10]. It performs better than probabilistic classification, decision tree and naïve bayesian. It provides an appropriate approach for drug cytotoxicity modelling and anticipation. Power generation forecasting system predicts the amount of power required at a rate closer to the power consumption [11].

4) Decision tree: It was used to identify at risk patients and disease trends, speech recognition and estimating users pose in video conference. "Rush University Medical Centre", has developed a software tool named "Guardian". It uses a decision tree machine learning algorithm to identify at-risk patients and disease trends [12]. It is used with large vocabulary speech recognition. In estimating users pose in video conference, only the user's pose data are extracted from photographed images using a binary decision tree. This is done to reduce data traffic.

B. Clustering

1) *Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH):* It can be used for filtering real images, generating the initial codebook for image compression and outlier detection. It performs very well on several huge datasets and is suggestively superior to Clustering Large Applications based on RANdomized Search (CLARANS) in terms of quality, speed and order-sensitivity for filtering real images [13].

2) *K means:* It can be used to derive meaningful relationships between students' eating habits and tendency of getting cold, customer segmentation and image segmentation [9]. K means was applied to extract meaningful rules from big data. In customer segmentation, this algorithm has achieved purity measure of 0.95 indicating 95% accurate segmentation of the customers. Image segmentation was performed using adaptive k-means clustering algorithm for 3- dimensional and multi-valued images.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 04, April-2018

3) *Clustering LARge Applications* (*CLARA):* Compared to Partitioning Around Mediods (PAM), CLARA can deal with much larger data sets in extracting interesting spatial patterns and features, capturing intrinsic relationships between spatial and non-spatial data, presenting data regularity concisely and at higher conceptual levels and helping to reorganize spatial databases to accommodate data semantics [14].

4) Clustering Using Representatives (CURE): It can be used to filter outliers. It employs a combination of random sampling and partitioning that allows it to handle large database efficiently [15].

C. Association

1) *Apriori algorithm:* It can be used for personalized marketing promotions and customer management, product placement strategies in store and smarter inventory management, medical data mining, insurance client analysis, packet signature mining, social network analysis and game analysis [9]. Apriori discovers hidden patterns in the data. Those interesting can be used to recognize variables in the data and the co-occurrences of different variables that appear with the greatest frequencies.

2) *Frequent Pattern (FP) growth:* This algorithm can be used for an instance lottery analysis and prediction, web usage mining, gene ontology and query recommendations. It was used to analyze the past winning numbers and predict the next phase of the lottery number in order to improve the probability of winning [16]. It can be utilised to find the most frequent access pattern generated from the web log data. It is scalable for mining both long and short frequent patterns. FP growth algorithm was used to constructs the description of gene groups. Query recommendation is a method to improve search results in web. It can be applied on a large dataset for mining search engine query logs to achieve fast query recommendation [9].

V. MACHINE LEARNING APPLICATIONS:

Machine learning is the latest buzzword sweeping across the global business landscape. It's captured the popular imagination, conjuring up visions of futuristic self-learning Artificial Intelligence and robots. As we move forward into the digital age, one of the modern innovations we've seen is the creation of Machine Learning. This incredible form of artificial intelligence is already being used in various industries and professions some of them are Healthcare, Finance, Retail, Travel, Social media, etc.., shown in figure-3. In 2014, it has been reported that a machine learning algorithm has been applied in Art History to study fine art paintings, and that it may have revealed previously unrecognized influences between artists [19].

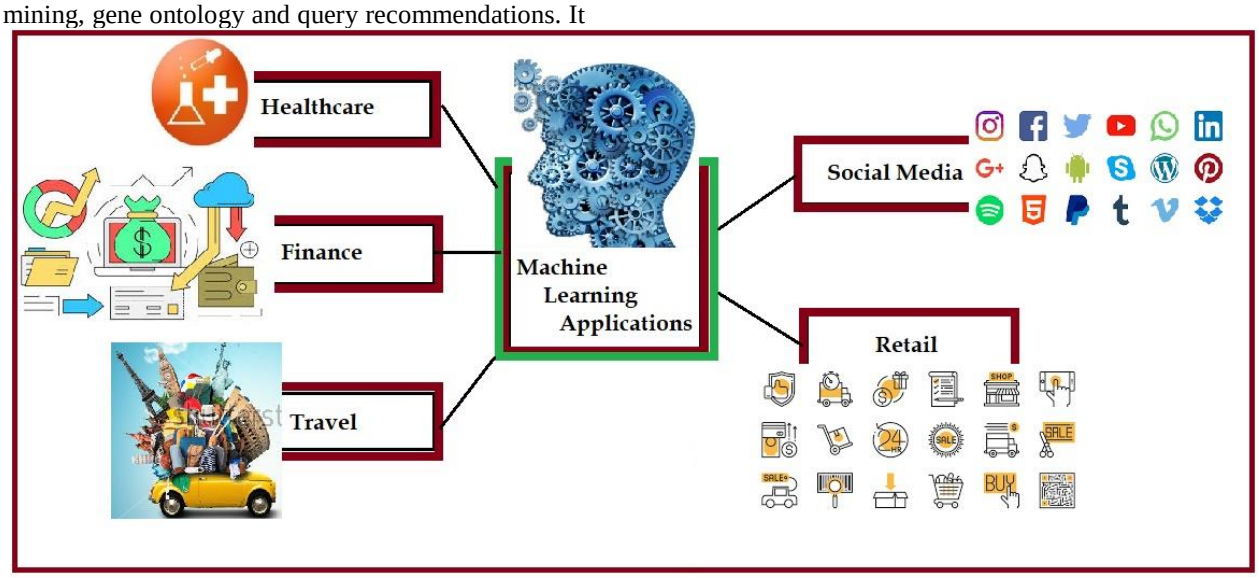

Figure-3: Machine Learning Applications

A. Healthcare- In 2012, Vinod Khosla, co-founder of Sun Microsystems predicted that 80% of medical doctor's jobs would be lost in the next two decades to automated machine learning medical diagnostic software [18]. Medical systems will learn from data and help patients save money by skipping unnecessary tests. Manufacturing or discovering a new medicine is costly and extensive process as thousands of compounds need to be subjected to a series of tests, and only a single one might result in a usable drug. Machine learning can speed up one or more of these steps in this lengthy multi-step drug manufacturing or discovering process.

Pfizer is using IBM Watson on its immuno-oncology (a technique that uses body's immune system to help fight cancer) research. This is one of the most significant uses of IBM Watson for *drug discovery* [20]. *Personalized treatment* has great potential for growth in future, and machine learning could play a vital role in finding what kind of genetic makers and genes respond to a particular treatment or medication.

B. Finance- Nowadays more than 90% of the top 50 financial institutions around the world are using machine learning and advanced analytics for their business. The application of machine learning in Finance domain helps banks offer personalized services to customers at lower cost, better compliance and generate greater revenue. One of the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 04, April-2018

core machine learning use cases in banking/finance domain is to combat fraud. Machine learning is best suited for this use case as it can scan through huge amounts of transactional data and identify if there is any unusual behaviour. In 2010 The Wall Street Journal wrote about the firm Rebellion Research and their use of Machine Learning to predict the financial crisis [17].

Citibank has collaborated with Portugal based fraud detection company *Feedzai* that works in real-time to identify and eliminate fraud in online and in-person banking by alerting the customer and *PayPal* is using machine learning to fight money laundering.

C. Retail- According to The Realities of Online Personalisation Report, 42% of retailers are using personalized product recommendations using machine learning technology. Machine learning in retail is more than just a latest trend; retailers are implementing big data technologies like Hadoop and Spark to build big data solutions and quickly realizing the fact that it's only the start. They need a solution which can analyse the data in real-time and provide valuable insights that can translate into tangible outcomes like repeat purchasing. Machine learning algorithms process this data intelligently and automate the analysis to make this supercilious goal possible for retail giants like Amazon, Target, Alibaba and Walmart.

D. Travel- According to Alvin Chin, BMW Technology Corporation, by 2030, there will be a solution for each unique travel purpose. Instead of commuting to work and stressing about finding parking, you can take a ride sharing service. Dynamic Pricing is an example of machine learning in travel.

According to Amadeus IT group, 90% of American travellers with a Smartphone share their photos and travel experience on social media and review services. Trip Advisor analyses this information to enhance its service.

E. Social Media- Machine learning offers the most efficient means of engaging billions of social media users. From personalizing news feed to rendering targeted ads, machine learning is the heart of all social media platforms for their own and user benefits.

Nowadays the social networks artificial neural networks machine learning algorithm identifies familiar faces from contact list. The ANN algorithm mimics the structure of human brain to power facial recognition and the professional network LinkedIn knows where you should apply for your next job, whom you should connect with and how your skills stack up against your peers as you search for new job by machine learning techniques.

Machine Learning does have some frightening implications when you think about it, these applications are just several of the numerous ways this technology can improve our lives.

VI. MACHINE LEARNING PYTHON LIBRARIES:

Python is becoming popular day by day and has started to replace many popular languages in the industry. The simplicity of python has attracted many developers to build libraries for Machine learning and Data Science, because of all these libraries, Python is almost popular as R for Data Science. Some of the best Machine Learning libraries for Python shown in Figure-4 and are explored in this section.

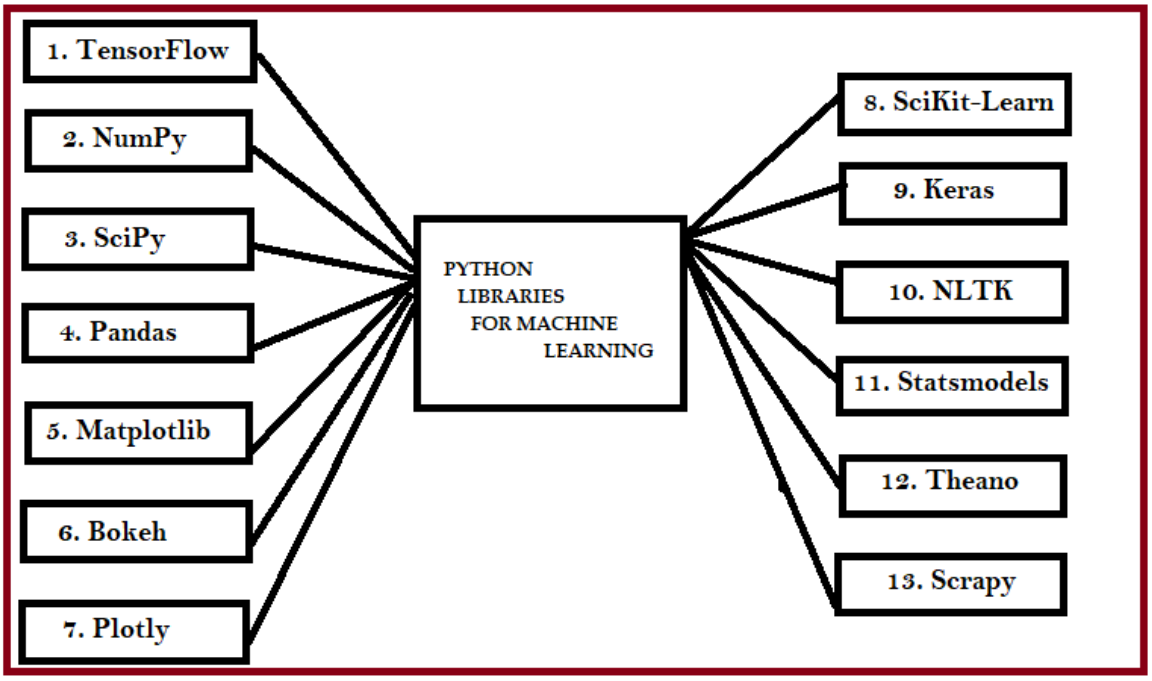

Figure-4: Python Libraries for Machine Learning

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 04, April-2018

1. TensorFlow-Coming from developers at Google, it is an open-source library of data flow graphs computations, which are sharpened for Machine Learning. It was designed to meet the high-demand requirements of Google environment for training Neural Networks and is a successor of DistBelief, a Machine Learning system, based on Neural Networks. However, TensorFlow isn't strictly for scientific use in border's of Google—it is general enough to use it in a variety of real-world application.
The key feature of TensorFlow is their multi-layered nodes system that enables quick training of artificial neural networks on large datasets. This powers Google's voice recognition and object identification from pictures.

2. NumPy- The most fundamental package, around which the scientific computation stack is built, is NumPy (stands for Numerical Python). It provides an abundance of useful features for operations on n-arrays and matrices in Python. The library provides vectorization of mathematical operations on the NumPy array type, which ameliorates performance and accordingly speeds up the execution.

3. SciPy- SciPy is a library of software for engineering and science; here you need to understand the difference between SciPy Stack and SciPy Library. SciPy contains modules for linear algebra, optimization, integration, and statistics. The main functionality of SciPy library is built upon NumPy, and its arrays thus make substantial use of NumPy. It provides efficient numerical routines as numerical integration, optimization, and many others via its specific sub modules. The functions in all sub modules of SciPy are well documented.

4. Pandas - Pandas is a python package designed to do work with "labelled" and "relational" data simple and intuitive. Pandas is a perfect tool for data wrangling. It designed for quick and easy data manipulation, aggregation, and visualization.

5. Matplotlib- Another SciPy Stack core package and another Python Library that is tailored for the generation of simple and powerful visualizations with ease is Matplotlib. It is a top-notch piece of software which is making Python (with some help of NumPy, SciPy, and Pandas) a cognizant competitor to such scientific tools as MatLab or Mathematica.
However, the library is pretty low-level, meaning that you will need to write more code to reach the advanced levels of visualizations and you will generally put more effort, than if using more high-level tools, but the overall effort is worth a shot.
With a bit of effort you can make just about any visualization:

- Line plots;
- Scatter plots;
- Bar charts and Histograms;
- Pie charts;
- Stem plots;
- Contour plots;
- Quiver plots;
- Spectrograms.
-

There are also facilities for creating labels, grids, legends, and many other formatting entities with Matplotlib. Basically, everything is customizable. The library is supported by different platforms and makes use of different GUI kits for the depiction of resulting visualizations. Varying IDEs (like IPython) support functionality of Matplotlib.

6. Bokeh-Another great visualization library is Bokeh, which is aimed at interactive visualizations. The main focus of Bokeh is interactivity and it makes its presentation via modern browsers in the style of Data-Driven Documents.

7. Plotly- It is a web-based toolbox for building visualizations, exposing APIs to some programming languages (Python among them). There is a number of robust, out-of-box graphics on the plot.ly website. In order to use Plotly, you will need to set up your API key. The graphics will be processed server side and will be posted on the internet, but there is a way to avoid it.

8. SciKit-Learn- Scikits are additional packages of SciPy Stack designed for specific functionalities like image processing and machine learning facilitation. In the regard of the latter, one of the most prominent of these packages is scikit-learn. The package is built on the top of SciPy and makes heavy use of its math operations. The scikit-learn exposes a concise and consistent interface to the common machine learning algorithms, making it simple to bring ML into production systems. The library combines quality code and good documentation, ease of use and high performance and is de-facto industry standard for machine learning with Python.

9. Keras- It is an open-source library for building Neural Networks at a high-level of the interface, and it is written in Python. It is minimalistic and straightforward with high-level of extensibility. It uses Theano or TensorFlow as its backends, but Microsoft makes its efforts now to integrate CNTK (Microsoft's Cognitive Toolkit) as a new back-end. The minimalistic approach in design aimed at fast and easy experimentation through the building of compact systems. Keras is really eased to get started with and keep going with quick prototyping. It is written in pure Python and high-level in its nature. It is highly modular and extendable. The general idea of Keras is based on layers, and everything else is built around them. Data is prepared in tensors, the first layer is responsible for input of tensors, the last layer is responsible for output, and the model is built in between.

10. NLTK -The name of this suite of libraries stands for Natural Language Toolkit and, as the name implies, it used for common tasks of symbolic and statistical Natural Language Processing. NLTK was intended to facilitate teaching and research of NLP and the related fields (Linguistics, Cognitive Science Artificial Intelligence, etc.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 04, April-2018

and it is being used with a focus on this today. The functionality of NLTK allows a lot of operations such as text tagging, classification, and tokenizing, name entities identification, building corpus tree that reveals inter and intra-sentence dependencies, stemming, semantic reasoning. All of the building blocks allow for building complex research systems for different tasks, for example, sentiment analytics, automatic summarization.

11. Statsmodels- Statsmodels is a library for Python that enables its users to conduct data exploration via the use of various methods of estimation of statistical models and performing statistical assertions and analysis. Among many useful features are descriptive and result statistics via the use of linear regression models, generalized linear models, discrete choice models, robust linear models, time series analysis models, various estimators. The library also provides extensive plotting functions that are designed specifically for the use in statistical analysis and tweaked for good performance with big data sets of statistical data.

12. Theano- Theano is a Python package that defines multi-dimensional arrays similar to NumPy, along with math operations and expressions. The library is compiled, making it run efficiently on all architectures. Originally developed by the Machine Learning group of Université de Montréal, it is primarily used for the needs of Machine Learning. The important thing to note is that Theano tightly integrates with NumPy on low-level of its operations. The library also optimizes the use of GPU and CPU, making the performance of data-intensive computation even faster.

13. Scrapy - Scrapy is a library for making crawling programs, also known as spider bots, for retrieval of the structured data, such as contact info or URLs, from the web. It is open-source and written in Python. It was originally designed strictly for scraping, as its name indicate, but it has evolved in the full-fledged framework with the ability to gather data from APIs and act as general-purpose crawlers. The architecture of Scrapy is built around Spider class, which encapsulates the set of instruction that is followed by the crawler.

VII. CONCLUSIONS:

In this paper, we have discussed various python libraries and development frameworks used to build up a Machine Learning system. We have reviewed the machine learning algorithms on classification, clustering and association. Each algorithm is better than the other for different applications. We also depicted extensive knowledge associated to different application domains of machine learning with a systematic architectural view. This paper definitely reduces the survey time of a machine learning researcher to develop a model for a selected application domain.

VIII. REFERENCES:

[1] https://www.toptal.com/machine-learning/machine-learning-theory-an-introductory-primer
[2] Machine Learning Theory and Applications for Healthcare, by Ashish Khere, Moongu Jeon, Ishwar K Sethi, Benlian Xu, Journal of Healthcare Engineering, Volume 2017 (2017), Article ID 5263570, September 2017
[3] https://www.quora.com/What-are-some-applications-of-machine-learning-in-education, post by Lawrence Wright
[4] Machine Learning Applications for Data Center Optimization Jim Gao, Google
[5] https://techbeacon.com/moving-targets-testing-software-age-machine-learning
[6] Open Source for U, Magazine, January-2018
[7] https://www.kdnuggets.com/2016/04/top-15-frameworks-machine-learning-experts.html
[8] https://www.sciencedirect.com/science/article/pii/S2210832717301485
[9] Survey of Machine Learning Methods for Big Data Applications by A. Vinathini, Dr. S. Bhagavathi Priya, 2017 International Conference on Computational Intelligence in Data Science(ICCIDS), 2-3 June 2017
[10] ”Predicting the protein localization sites using artificial neural networks”,by V Arulmozhi , Rajesh Reghunadhan, Springer. m 8th German Conference on Chemoinformatics: 26 CIC-Workshop Goslar, Germany ,vol.5,issue 1,pp.11-13, 2013.
[11] Hamidreza Mansouri”Modeling and prediction of cytotoxicity of artemisinin for treatment of the breast cancer by using artificial neural networks”, by Abdolhossein Qaderi , Neda Dadgar , SpringerPlus ,vol. 2 ,pp.340, 2013.
[12] “Decision tree-based acoustic models for speech recognition “,by Masami Akamine and Jitendra Ajmera, Springer Journal on Audio, Speech, and Music Processing, vol. 1,issue 10,2012.
[13] M“BIRCH: an efficient data clustering method for very large databases”, by Zhang.T,Ramakrishnan R and Livny, In Proceedings of the ACM SIGMOD Conference, vol.25,issue 2,pp.103-114,June 1996.
[14] “A Survey on Data Mining using Clustering Techniques” by T.Revathi, Dr.P.Sumathi, International Journal of Scientific & Engineering Research, vol. 4, Issue 1, January-2013.
[15] ”CURE: An Efficient Clustering Algorithm for Large Data sets” , by Sudipto Guha , Rajeev Rastogi , Kyuseok Shim , Proceedings of the ACM SIGMOD Conference,vol. 27,issue 2,pp.73-84,June 1998.
[16] ”Research on Application of FP-growth Algorithm for Lottery Analysis” , by Jianlin Zhang, Suozhu Wang, Huiying Lv, Chaoliang Zhou , Springer, Proceedings of 3rd International Conference on Logistics, Informatics and Service Science,pp.1227-1231,2015.
[17] https://www.wsj.com/articles/SB10001424052748703834604575365310813948080
[18] Vonod Khosla (January 10, 2012). "Do We Need Doctors or Algorithms?". Tech Crunch.
[19] When A Machine Learning Algorithm Studied Fine Art Paintings, It Saw Things Art Historians Had Never Noticed, *The Physics at ArXiv blog*
[20] https://www.dezyre.com/article/top-10-industrial-applications-of-machine-learning/364
[21] https://towardsdatascience.com/best-python-libraries-for-machine-learning-and-data-science-part-1-f18242424c38
[22] http://stackabuse.com/the-best-machine-learning-libraries-in-python/
[23] https://medium.com/activewizards-machine-learning-company/top-15-python-libraries-for-data-science-in-in-2017-ab61b4f9b4a7