Role of machine learning in Data Science: A detailed study

Revathi S

Researcher/Data Scientist

***Abstract*- The machine learning empowers data science to reduce human efforts and become a most valuable asset for business needs through pattern recognition, prediction, analysis and efforts. Now-a-days, organizations really emphasize using data to improve their product needs, where machine learning makes the day of Data Scientist easier by automating the task, and by analyzing enormous amount of data which proves that Data scientist should have in-depth knowledge of Machine learning to improve their prediction process. Machine learning is a subset of Artificial Intelligence, a set of algorithms which trains machine or computers the ability to predict the data on their own. In this paper, a detailed overview of different structures of Data Science and address the impact of machine learning on steps such as Data Collection, Data Preparation, Training the model, Model Evaluation and Prediction. Also, a study on detailed 3 keys on machine learning algorithms such as Classification, regression and clustering is been discussed in this paper.**

***Keywords*- Structure of Data Science, Machine learning, Classification, Regression and Clustering**

I. INTRODUCTION

In recent years the phrase Data Science has become buzz word in all industry. Data science is field of study that involves processing large data to obtain insights and valuable information from data. It comprises different fields of expertise and skills to solve and optimize the process. Data Science is a multi-faceted interdisciplinary field of study with Computer/IT, Mathematics/Statistics and Business need/Domain Knowledge [1]. Further, these three domains separately result in a variety of careers as Software (combining computer Science and Business need), Research (Combining Business need and Mathematics) and Machine learning (combining Computer Science and Mathematics). With these areas Data Scientist can maximize their performance by interpreting data and providing innovative solution and achieves improvements in prediction [2].

Machine learning is the field of intersecting computer Science, mathematics and statistics. It is used to identify patterns, recognize behaviors, and make decisions from data with minimal human intervention. It is a method of data analysis that automates data collection, data preparation, feature engineering, training the model, and eventually model evaluation and prediction [3]. Machine learning allows data scientists to implement very complex models, such as neural networks or support vector machines, and an ensemble of simple models like gradient boosting, random forests and decision trees. These complex models can be captured between the independent variables (input) and the dependent variable (target/output). The intersection of domain knowledge and mathematics/statistics is the research field. Research skills enable data scientists to apply and develop a new technique with very complex model that are accurate and build model with less functional form [4]. It is used to speed up the development process with fewer assumption between the relationship on dependent and independent variable. Software skills is the field refer to the intersection of computer science and domain knowledge. It has familiarity in open-source languages and other world-class software languages that help data scientists to create new models. The combination of these three skills helps data scientists to solve the business problem and improve a specific business process.

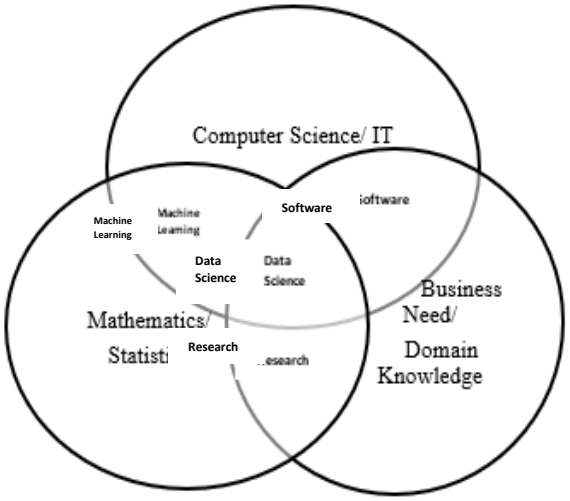

Fig.1 Fields of Data Science

Our future technology like machine learning and artificial intelligence algorithms are depends on data sources for training and building models by using designed algorithms likes naive Bayes, Supervised and unsupervised, clusters and regression models etc. The data scientist has to fetch potential information from the massive unstructured dataset [5]. They can modify, visualize through statistics, interrelate, predict the model using machine learning methods. By fetching small information from the hidden data, Machine learning can be used by the organization in terms of improving process throughput, optimization, maintenance and many more related tasks.

II. ROAD MAP FOR MACHINE LEARNING IN DATA SCIENCE

In today's era, the word Machine learning, Deep learning and Artificial Intelligence have dominated the industry by overshadowing the Data Science aspects like Data Analytics, Extract-Transform and Load (ETL) and Business Intelligence. Machine learning is a subset of Artificial Intelligence [6]. It consists of set of Algorithms used to analyzes large chunks of Data, that automatically process the Data Analysis and makes data prediction in real time without human aid. A Data model is built automatically using Machine learning procedure and the system is been trained for real time prediction. The below figure shows the roadmap of where the machine learning algorithms are used in the Data Science lifecycle.

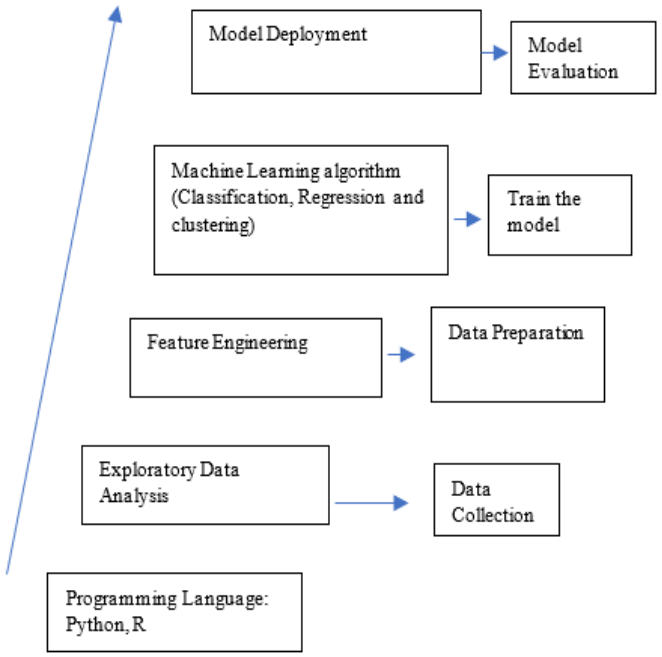

Fig. 2 Roadmap for Machine learning in Data Science

The idea behind the machine learning is to teach and train the machine by feeding the data and defining features. Our system is been able to learn, grow, train and develop by themselves based on the dataset we used. It also observes and identifies pattern from the data and automatically makes predictions.

A. Role of Python Programming Language

Python is a renowned programming language which plays important role in machine learning. It is an extraordinary interpreted and interactive programming language that results in easy composition and identification of data [7]. Python is open-source programming language, that can easily incorporate with other necessary tools. The key feature of python programming language in machine learning are as follows:

1) It has various libraries and frameworks: Python is widely known for its large number of libraries and frameworks that make coding easy. Its most popular libraries, such as NumPy, SciPy, and Scikit, are used for scientific calculations, advanced computations, and data analysis respectively. Even its most popular frameworks, such as TensorFlow and Apache Spark, are quite likeable among the engineers who are working on machine learning and deep learning projects [8]. Its libraries and frameworks are highly beneficial for the new developers.

2) Its coding is quite simple to understand and apply: For new developers, Python offers an easy coding feature that is concise and easily readable. Python comes with a simple syntax that offers quick development of applications compared to other types of programming languages available.

3) It receives great online support: Python is the programming language that is accepted worldwide by developers. Because of this, this language receives great support from the large community available online.

4) It offers flexible integrations: Python is a language that comes with a syntax that is quite easy to understand. Developers find this language best for developing prototypes. It also helps in boosting productivity. Even the Python projects can easily be integrated with other programming languages.

5) It offers visualization benefits: Since machine learning and deep learning projects require presentation of data in a human-readable format, Python comes with such frameworks that offer great visualization support.

B. Exploratory Data Analysis

Collecting data is considered as initial step of machine learning. Collecting relevant and reliable data becomes very important as the quality and extend of data directly impact the outcome of our machine learning model. As first step, we need to segregate data based on its datatypes as

1) Numerical: This represents some sort of quantitative measurement. Example: height of people, stock price etc. It can be further broken up into 2 parts: *Discrete data:* This is integer based, often counts of some event. Example: "How many songs do a user like? *Continuous data:* It has an infinite number of possible values. Example: "House price"

2) Categorical: This represents qualitative data with no apparent inherent mathematical meaning. Example: Yes or No, Gender, Marital Status etc.

3) Ordinal: It has features of both numerical and categorical data. But the numbers placed in categories have mathematical meaning. Example: movies rating on 1–5.

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics by plotting them visually. This step is very important especially when we arrive at modeling the data in order to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter plot and many more. It often takes much time to explore the data.

Through the process of EDA, we can ask to define the problem statement or definition on our data set which is very important. The steps in EDA are [9]:

C. Dropping the duplicate rows

This is often accessible to do in huge dataset, that may contain some duplicate rows and columns which might be disturbing, so before going with machine learning modeling we need to eliminate duplicate values from the dataset. For example, suppose we have 10000 rows of data but after removing the duplicates 950 data meaning that we have 50 of duplicate data.

D. Dropping the missing or null values

This is mostly similar to the previous step but in here all the missing values are detected and are replaced with statistical approaches like filing the missing values with mean, median or mode for numerical data, fillna() method or chi square test for categorical data or we can use simple imputation methods for complex data.

E. Detecting Outliers

Outliers in the dataset are one of the primary reasons for resulting less accurate model, so it's often good practice to remove outliers before training the dataset. To identify outliers, some visualization techniques like Box plots are used to find the Interquartile score (IQR) for the data. The data points outside the IQR are identified as outliers and are removed

F. Plotting techniques

1) Histogram: Histograms is one of the best form of visualizations when working with single continuous variable. It plots the relative frequencies of the variables as like the bar chart.

2) Heat Maps: Heat Maps is a type of plot which is necessary when we need to find the dependent variables. One of the best ways to find the correlation between the features.

3) Scatterplot: Scatterplot is used to study the relationship between the two variables which vary one along the x-axis and the other along the y-axis. It is most widely used form of plot and is used to study the regression models.

4) Boxplot: Boxplot is a plot which is used to get a sense of data spread of one variable. The top line of box represents third quartile, bottom line represents first quartile and middle line represents median. The top line above the box represents 1.5 times the inter-quartile range (difference between third and first quartile). The dots above and below the lines are outliers.

III. FEATURE ENGINEERING

Machine learning fits mathematical notations to the data in order to derive some insights. The models take features as input. A feature is generally a numeric representation of an aspect of real-world phenomena or data. Just the way there are dead ends in a maze, the path of data is filled with noise and missing pieces.

Mathematical formulas work on numerical quantities, and raw data isn't exactly numerical. Feature Engineering is the way of extracting features from data and transforming them into formats that are suitable for Machine Learning algorithms [10].

It is divided into 3 broad categories: -

A. Feature Selection

All features aren't equal. It is all about selecting a small subset of features from a large pool of features. We select those attributes which best explain the relationship of an independent variable with the target variable. There are certain features which are more important than other features to the accuracy of the model. It is different from dimensionality reduction, because the dimensionality reduction method does so by combining existing attributes, whereas the feature selection method includes or excludes those features. The methods of Feature Selection are Chi-squared test, correlation coefficient scores, LASSO, Ridge regression etc.

B. *Feature Transformation*

It means transforming our original feature to the functions of original features. Scaling, discretization, binning and filling missing data values are the most common forms of data transformation. To reduce right skewness of the data, we use log.

C. Feature Extraction

When the data to be processed through an algorithm is too large, it's generally considered redundant. Analysis with a large number of variables uses a lot of computation power and memory, therefore we should reduce the dimensionality of these types of variables. It is a term for constructing combinations of the variables. For tabular data, we use PCA to reduce features. For image, we can use line or edge detection.

IV. MACHINE LEARNING ALGORITHM

The dataset can be classified under 4 major categories:

- Classification
- Regression
- Clustering
- Time Series Analysis

A. Classification

When the output variables are **discrete** values, Classification is used. If we want to find which category the data belongs to, then it is a Classification problem. Classification Algorithms look at existing data to predict the Class or Category of the new data. Classification is more like finding curves that separate the data points into different Classes/Categories [11]. Labelling an Email as Spam is a Classification problem. Some famous Classification Algorithms are Support Vector Machines, Neural Networks, Naive Bayes, Logistic Regression, and the K Nearest Neighbor.

B. *Regression*

When the output variable is in continuous values, Regression is used. Regression works on the Curve-Fitting Techniques or slope intercept formula as "y=mx+c" where the slope value of y when x=0. The data points fall in the curve are used to predict the output values. Regression is useful for Financial Predictions like Stock Market Prediction and Housing Price Prediction. Some famous Regression Algorithms are Linear Regression, Perceptron, and Neural Networks.

C. *Clustering*

Clustering is to group the data based on the similar characteristic, without labels. Ideally, the similar data points are grouped together in the same Cluster based on different definitions of similarity. The points in different Clusters should be as dissimilar as possible. The Clustering Algorithms try to find a pattern in a dataset without associating labels with it. Buying behaviors of customers is Clustered using the clustering Algorithm. Some famous Clustering Algorithms are K-Means Clustering **and** Agglomerative Clustering**.** Regression and Classification come under the Supervised Learning Model of Machine Learning while Clustering comes under the Unsupervised Learning Model.

D. *Time Series Analysis:*

A time-series contains sequential data points mapped at a certain successive time duration. Predicting future variables in the datasets, known as time series forecasting [12]. Several real-world applications, including weather forecasting, earthquake prediction, statistics, and most business — market forecasting use time series analysis. Broadly specified time-series models are Autoregressive (AR)**,** Integrated (I)**,** Moving Average (MA)**,** and some other models are the combination of these models such as Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA) models

V. MODEL DEPLOYMENT

Machine learning deployment is the process of deploying a machine learning model in a live environment. The model can be deployed across a rang of different environments and will often be integrated with apps through API. Deployment is a key step in an organization gaining operational value from machine learning. Machine learning models will usually be developed in an offline or local environment, so will need to be deployed to be used with real time data. Deployment is the final step for an organization to start generating a return om investment for the organization. However, deployment is the final step for an organization to start generating a return on investment for the organization. However, deployment from a local environment to a real- world application can be complex. Models may need specific infrastructure and will need to be closely monitored to ensure ongoing effectiveness. For this reason, machine learning deployment must be properly managed so its efficient and streamlined.

VI. MACHINE LEARNING USE CASE IN DATA SCIENCE

Machine Learning finds its application in almost every sector, from Finance Institutions to Entertainment Industry. It is Machine Learning that goes behind the Apps you use on a regular basis to make your life easier such as Google Maps, Microsoft Cortana, and Alexa. Given below are some of the most popular real-life applications of Machine Learning in Data Science:

A. *Fraud Detection*

Banks use Machine Learning for fraud detection to keep their customers safe. Machine Learning Models are Trained to flag transactions that appear suspicious based on the defined features and transaction patterns. Machine Learning can ensure the safety of consumers not just to Banks but to Private Enterprises as well.

B. *Speech Recognition*

Ever wondered what goes behind **Siri**? The Voice Assistants on Smartphones also leverage Machine Learning to recognize what you say and craft a response accordingly. Machine Learning Models are Trained on human languages and various accents to convert the speech into words, and then make a response a smart response.

C. *Online Recommendation Engines*

As already discussed in the previous sections, Online Recommendation Engines make use of Machine Learning to suggest relevant recommendations to their users. Amazon often lists Recommended Products for its customers, YouTube provides personalized Video Recommendations to its users, and similarly, Facebook suggests Friends Recommendations. Machine Learning Models are Trained on **Customer Behaviors**, **Past Purchases**, **Browsing History**, and any other behavioral information about consumers.

The framework of machine learning cycle is been depicted in below figure

Fig. 3 Framework of machine learning cycle

VII. DATASCIENCE USECASES

One of the most common use cases for Data Science perspective is customer segmentation. Customer segmentation is a process of grouping customers into segments according to the coincidences of particular criteria in their characteristics.

There are three significant segmentation types that are the most often used. These are:

- segmentation based on touchpoint engagement
- segmentation based on purchase patterns.

Application of micro-segmentation appears to be a rising trend in marketing. Micro-segmentation is far more advanced. It helps to segment people into more precise categories especially concerning behavioral intentions. Thus, marketing actions may be tailored to the preferences even of the least numerous customer groups.

A. Predictive analytics for customers' behavior

Cluster models, predictions, collaborative filtering, regression analysis are all applied to spot the correlation patterns in the customers' behavior to predict future tendencies in purchasing.

VIII. CONCLUSION

Nowadays, organizations really emphasize using data to improve their products. Data Science is just Data Analysis without Machine Learning. Data Science and Machine Learning go hand in hand. Machine Learning makes the life of a Data Scientist easier by automating the tasks. In the near future, Machine Learning is going to be used prominently to analyze a humongous amount of data. Therefore, Data Scientists must be equipped with in-depth knowledge of Machine Learning to boost their productivity.

This paper gave a study on Data Science and Machine Learning with the help of real-life examples. Also, to understand how Machine Learning is being used in Data Science for Data Analysis and the extraction of valuable insights from data. The paper also briefed the workflow of Machine Learning in Data Science with the most popular Machine Learning Algorithms that are used in Data Science. In addition, concluded with a peek into the real-life applications of Machine Learning in Data Science.

REFERENCES

[1] Cao, L.: Data science: a comprehensive overview. ACM Computing Survey (2017). https://doi.org/10.1145/3076253

[2] Matthew J. Graham. 2012. The art of data science. In Astro statistics and Data Mining, Springer Series in Astro statistics, Vol. 2. 47–59.

[3] Donoho, D.: 50Years of Data Science. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience .pdf (2015)

[4] Dyk, D.V., Fuentes, M., Jordan, M.I., Newton, M., Ray, B.K., Lang, D.T., Wickham, H.: ASA Statement on the Role of Statistics in Data Science. http://magazine.amstat.org/blog/2015/10/01/asastatement-on-the-role-of-statistics-in-data-science/ (2015)

[5] Nathan Yau. 2009. Rise of the Data Scientist. Retrieved from http://flowingdata.com/2009/06/04/rise-of-thedata-scientist/

[6] Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science 349(6245), 255–260 (2015).

[7] Guo, P.: Python is Now the Most Popular Introductory Teaching Language at Top U.S. Universities, July 2014. http://cacm.acm.org/blogs/blog-cacm/176450- python-is-now-the-most-popular-introductory-teaching-language-at-top-u-s - universities

[8] McKinney, W.: Python for data analysis. O'Reilly (2012).

[9] Komorowski, Matthieu & Marshall, Dominic & Salciccioli, Justin & Crutain, Yves. (2016). Exploratory Data Analysis. 10.1007/978-3-319-43742-2_15.

[10] Galli, Soledad. (2021). Feature-engine: A Python package for feature engineering for machine learning. Journal of Open-Source Software. 6. 3642. 10.21105/joss.03642.

[11] Behera, Rabi & Das, Kajaree. (2017). A Survey on Machine Learning: Concept, Algorithms and Applications. International Journal of Innovative Research in Computer and Communication Engineering. 2.

[12] Nokeri, Tshepo. (2021). Time-Series Analysis. 10.1007/978-1-4842-6870-4_3.