WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Survey on Statistics and Ml in Data Science and Effect in Businesses

G Swajan Reddy [1], Revanth Yanduri [2], Ch Arun Reddy [3]

[1,3] IV BTech Students, Department of Computer Science and Engineering, ACE Engineering College, Hyderabad, Telangana

[2] IV BTech Student, Department of Computer Science and Engineering, Neil Gogte Institute of Technology, Hyderabad, Telangana

DOI: https://doi.org/10.55248/gengpi.2022.3.11.38

ABSTRACT

Data science is the most important area of any companies to improve their business by many analytical techniques and drawing conclusions that can be used to increase their company's revenues. This paper contains a review of statistics and ML algorithms involved in data science and some of their business use cases like stock price analytics, product analytics and real estate valuations.

Keywords: Data Science, Machine Learning, Statistics, Data Analytics, Business Analytics.

1.Introduction

Data science can be mostly used as a business intelligence module to make decisions that can generate profits or revenues for the company by analytics of their process, products, performance or even customer behaviour such as an online streaming application can analyze the pattern of their customer watching movies and try to provide recommendations accordingly. Another example is that to analyze customerpattern of buying things together in a store and making some deals or give them as a combo can actually make their sales to increase in number. Data science can be used to forecast the data. It can be used to predict things like weather conditions, stock prices in share market, etc. Analysing time series data and discovering patterns can be useful for various data predictions.

While most business decisions are taken to make profits, with the help of Data Science, these decisions are perfectly planed and tailored for business profits and needs. Stakeholders can understood the underlying problems by the explanations given by data analysts or data scientists, and can enforce new changes in their businesses to generate revenue.

2.Data Science Evolution

In recent years data science evolved rapidly and gained popularity and has become most popular field. This tremendous growth happened because of powerful languages like python for data collection, data interpretation and data analysis.

In 1963, an American mathematician, John W.Tukey stated the idea of data science, he was too far ahead of his time. A Danish computer engineer, Peter Naur was another person with accomplishments in data science. One of the first definations in data science is "The establishment of the data science field is further followed by the data relationship and its representation assigned to other sciences and fields."

In 1977, In order for linking traditional statistical methodology, the knowledge of domain experts and modern computer technology, to transform data into knowledge and information, The International Association for Statistical Computing (IASC) was founded and theories and predictions of scientists are put into practice.

In 1980's & 1990's, improvements are being made in data science by "Knowledge Discovery in Databases (KDD)" workshop and the foundation of the "International Federation of Classification Societies" (IFCS). These organizations trained professionals in data science.

For the first time in 1994, Database Marketing appeared. Business companies started gathering data to understand patterns in consumers, competitors and wound up with more data than they could handle. They found new ways and hired more employees.

The data science has evolved more in late 1900's and in early 2000's and William S Cleveland contributed to modern data science conception.

In 2000's, technology had hit huge peaks by making internet connectivity, communication and widespread data collection. Work of action plan laid out by William S Cleveland improved technical skills and knowledge of data analysts.

Big Data entered the market in 2005 with data processing methods like Hadoop, Cassandra and spark to handle massive data generated by top tech companies.
Until 2014, As the data increased, demand for data scientists also increased and need to analyse patterns and business decisions.
As in 2015, Artificial Intelligence (AI), Machine Learning, and Deep learning entered the market. These technologies have created various advancements in past years to implements these technologies into our real life.
In 2018, new regulations have been introduced and until 2020 AI and ML and Big Data seen to be ever increasing.

3.Domains Involved

Fraud and Risk Detection

Data analysis hold a different department in financial industry. The data science is used to reduce and protect from loses. It works by collecting previous data from customers and find probability of risk.

Healthcare

Disease identification can be difficult even for a trained professional, data science is used to identify and predict diseases from symptoms of patients previous medical records and various other patients records for high probable result.

Targeted Advertising

Data generated by tech industries is huge, it can be used to predict user pattern and can be used by advertising agencies for specified users for high chances of success.

Internet Search

In our current world there are many search engines which make use of data science algorithms to provide fast search results and accurate results.

Advanced Image Recognition

When photo is uploaded to platforms with face recognition algorithms it makes automatic suggestions with people of similar photos and also used in reverse image search.

4.Data Science Methods

Data Science involves many methodologies like data collection, extraction, data pre-processing, data analysis, data modelling, etc., but main core things in Data Science are statistics and machine learning models. Statistics is used to clean, analyze, interpret and data visualization while machine learning models can take the data science into new level by building models that are appropriate for the problems to predict or make decisions.

Statistics in Data Science

Importance of Statistics for Data Science:

1. Identify the importance of features by using various statistical tests.
2. Finding the relationship between features to eliminate the possibility of duplicate features.
3. Converting the features into the required format.
4. Normalizing and scaling the data. This step also involves the identification of the distribution of data and the nature of data.
5. Taking the data for further processing by using required adjustments in the data.
6. After processing the data identify the right mathematical approach/model.
7. Once the results are obtained the results are verified on the different accuracy measurement scales.

Statistical Techniques for Data Science

1.Random Variable:A random variable can be any variable where you are storing a value. Every feature that is present in a dataset is basically a random Variable.

Types of Random variable:

1. Numerical Random Variable
2. Categorical Random Variable

2.Measures of Central Tendancy

--MEAN: Arithmetic Mean is the average of all data values which you work with. Mean is used to find the average value around which your data values range. Generally, when working with data, you may want to know the average data value. This will give you a term that incorporates every data value from the dataset.

--MEDIAN: The median refers to the middle value of your data. You can use the median to figure out the point around which your data is centered. It divides the data into two halves and has the same number of data points above and below.The median is especially useful when

you have skewed data. That is, it has high data distribution towards one side.

--MODE: Using the mode, you can find the most commonly occurring point in your data. This is helpful when you have to find the central tendency of categorical values, like the flavour of the most popular chip sold by a brand. You cannot find the average based on the orders; instead, you choose the chip flavour with the highest orders.

3.Standard Deviation: Standard deviation is a number that describes how spread out the observations are. A mathematical function will have difficulties in predicting precise values, if the observations are "spread". Standard deviation is a measure of uncertainty. A low standard deviation means that most of the numbers are close to the mean (average) value. A high standard deviation means that the values are spread out over a wider range.

4.Covariance: The covariance is a measure of the joint variability of two random variables and describes the relationship between these two variables. It is defined as the expected value of the product of the two random variables' deviations from their means.

5.Corrleation: The correlation is also a measure for relationship and it measures both the strength and the direction of the linear relationship between two variables. If a correlation is detected then it means that there is a relationship or a pattern between the values of two target variables.

6. Probability Distribution Functions:A function that describes all the possible values, the sample space, and the corresponding probabilities that a random variable can take within a given range, bounded between the minimum and maximum possible values, is called *a probability distribution function (pdf)* or probability density.
Types:

- Bernoulli
- Possion
- Normal

Machine Learning for Data Science

Machine learning normally experiences information pre-processing, learning, and assessment stages. Data pre-processing gets ready crude information into the "right form" for resulting learning steps. The raw data is probably going to be unstructured, loud, fragmented, and conflicting. The pre-processing step changes such information into a frame that can be utilized as contributions to learning through information cleaning, extraction, change, and combination. The learning stage picks learning calculations and tunes display parameters to produce wanted yields utilizing the pre-processed input
information. Some learning techniques, especially authentic learning, can likewise be utilized for information pre-processing. The assessment takes after to decide the execution of the educated models. For example, execution assessment of a classifier includes dataset determination, execution measuring, mistake estimation, and factual tests .

The assessment results may prompt changing the parameters of picked learning calculations and additionally choosing diverse calculations. Designing a learning
system, i.e. an application of machine learning, involves four design choices.

1. Choosing the training data.
2. Choosing the target function.
3. Choosing the representation.
4. Choosing the learning algorithm.

5.Some Ml Algos in Data Analytics:

Linear regression: Linear regressionis an equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (B).

Decision Trees: Decision tree model is a binary tree representation model. Each node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. Predictions are made by walking the splits of the tree until arriving at a leaf node and output the class value at that leaf node. Trees are fast to learn and very fast for making predictions .

Naive Bayes : Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling. The model is comprised of two types of probabilities that can be calculated directly from training data: The probability of each class. The conditional probability for each class is given for each x value. Once calculated, the probability model can be used to make predictions for new data using Bayes

Theorem.

K-Nearest Neighbour: K-Nearest Neighbour predictions are made for a new data point by searching through the entire training set for the K most similar instances (the neighbours) and summarizing the output variable for those K instances. For regression problems, this might be the mean output variable, for classification problems this might be the mode (or most common) class value.

Linear Vector Quantization: Linear Vector Quantization is an artificial neural network algorithm that allows choosing how many training instances to hang onto and learns exactly the final instances.

Support Vector Machines:Support Vector Machines learning algorithm finds the coefficients those results in the best separation of the classes by the hyper plane. The distance between the hyper plane and the closest data points is referred to as the margin.

Table 1: ML Algorithm Comparison

	INPUT IMPORTANCE	OVERFITTING PREVENTION	EASY TO IMPLEMENT	COMPUTATIONAL COST	APPLICATIONS
MULTIPLE LINEAR REGRESSION	YES	NOT PRONE	SIMPLE	LOW	MODEL RELATIONS AMONG INDEPENDENT VARIABLES
ARTIFICIAL NEURAL NETWORK	NO (sensitivity Analysis Possible)	DEPENDS	MANUAL IMPLEMENTATION REQUIRED	HIGH	RESOURCE EFFICIENCY, RISK MANAGEMENT
REGRESSION TREE	YES	PRUNING REQUIRED	SIMPLE	DEPENDS ON TRAINING FUNCTION	FINDING OPTIMAL MODEL FOR PREDICTIONS
KERNAL REGRESSION ANALYSIS	YES	NOT PRONE	MODERATE	LOW	ESTIMATE CONDITIONAL EXPECTATION OF RANDOM VARIABLE
SUPPORT VECTOR MACHINES	YES	NOT PRONE	MODERATE	HIGH FOR BIG DATA	CLASSIFICATION, DETECTION, ANALYSIS, REGRESSION
K-NEAREST NEIGHBORS	NO	REQUIRES SMOOTHING	SIMPLE	EXPENSIVE	RECOMMENDATION, PREDICTION

6. Data Visualization

Understanding the data which is in a table is quiet difficult. Data visualization refers to making the data into graphs, pie-charts, etc., so that it is easy to read and understand the data. It is a process of translating data into a meaning full or visual context so that we can draw up insights and conclusions from them. Many tools are available like tabluea, power-bi, seaborn python libraries, etc., to make this step happen.

Some data visualization forms are :

1. Graphs: Bar graphs, Line graph, Scatter plot, Bin Graphs, etc.,
2. Pie-charts
3. Maps

With the help of these, we can make dashboards so that insights from the data are clearly understood by stakeholders in businesses.

7.Data Science for Businesses

Stock Price and time series Forecasting

Stock prices in share market are subjected to so much volatility. Prediction by human means becomes difficult and time consuming. They can be forecasted to predict important values of stock prices like Open, High, Low, Close (OHLC) based on various parameters. Some methods include Long short-term memory (LSTM) showed an efficiency of 94% accuracy. This model is built using integrating LSTM with neural networks.

Time series analysis in various sectors is another important thing that data science is involved. It requires recurrent neural networks to analyze and interpret missing patters in time series data. There are three important classes that are mostly used in time series models, stochastic, neural networks and support vectors.

Product Matching

Matching products to increase the sales is an another important data science application. We use different statistics to know about customer behaviour or what types of products are being sold in a single purchase. By big data analytics we can correlate the products which are being sold in a single order. For example, most of the customers bought lighting cable type - c along with their new mobile phone. Another example is that most customers want to buy bread with milk. By observing these patterns by data visualization techniques from big data analytics, company can identify what their customers want to buy more and they can take decisions accordingly for instance they might give a combo offer for a certain period on certain products to increase the number of sales. Thus, data science can help in increasing revenues for online stores by product recommendations and offline stores by providing a plan to organize the products in a store so that customers have a high chance of buying them.

Real Estate Values

Real estate is a form of property rather than a personal possession, the value of real estate depends on different factors such as if the property is near to city or country side, or if basic facilities like school, hospital, groceries store are near, these are some of factors involved in price prediction of real estate. This is a very huge task for humans and technology can help solve this problem by use of data science. It can help mitigate risks, forecast customer behaviour, increases employee performance and offers more solutions to clients in very short period. With this process clients feel happier and safe and projects themselves to have a suitable property.

Vehicle Prices

Purchasing a brand new vehicle right off the showroom is no hassle because vehicles in showroom have fixed prices set by the company but at the time of selling a used vehicle or purchasing a used vehicle is a tricky choice because many people sell used cars at different prices at different locations, data science is used to predict prices of different vehicles based on cost price, when it was purchased, vehicle condition, and many factors can used and data can be collected from respective areas for better predictions and used to help people to buy or sell vehicles for a fair price. Predicting model s for car prices can be build by random forest, regression techniques.

8.Conclusion

In conclusion data science has various fields that focus on different attributes based on the data and generate patterns by understanding the data, statistics plays a major part in data science, as it helps solve problems and generate valuable data which can be further used to train machine learning models and are also capable of predicting accurate data. Patterns and data generated by using data science and predicting using machine learning models helps understand customer purchasing pattern in respective areas and can be used by hospitals to predict diseases based on symptoms and their family tree or other patients data, which can help save their life ahead of time also save money, these predictions and implementation of these collected information to our daily life helps us to overcome many hurdles and makes our life easier. In future by increase in technology, data cleaning can be automated, but for now it is a manual process.

REFERENCE

1. Data Science: the impact of statistics: Weihs, C., Ickstadt, K. Data Science: the impact of statistics. *Int J Data Sci Anal* 6, 189–194 (2018). https://doi.org/10.1007/s41060-018-0102-5
2. Machine Learning and Data Science: An Introduction to Statistical Learning methods with R: By Daniel D. Gutierrez, Technics Publications (2015)
3. Supervised and Unsupervised Learning for Data Science pp 3–21DOI: 10.1007/978-3-030-22475-2_1
4. Fama EF (1965) The behavior of stock-market prices. J Bus 38(1):34–105
5. G. Bathla, "Stock Price prediction using LSTM and SVR," *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 2020, pp. 211-214, doi: 10.1109/PDGC50313.2020.9315800.
6. Ristoski, Petar et al. 'A Machine Learning Approach for Product Matching and Categorization'. IOS press content library, 1 Jan. 2018 : 707 – 728.
7. S. Samadani and C. J. Costa, "Forecasting real estate prices in Portugal : A data science approach," *2021 16th Iberian*

Conference on Information Systems and Technologies (CISTI), 2021, pp. 1-6, doi: 10.23919/CISTI52073.2021.9476447.

8. Budiharto, W. Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM). *J Big Data* **8**, 47 (2021). https://doi.org/10.1186/s40537-021-00430-0