JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DATA SCIENCE IN BUSINESS FOR CUSTOMER ACQUISITION, BETTER MARKETING, INNOVATION AND ENRICHING LIVES

[1.] **Dr. R. Palaniappan, Associate Professor, Department of Computer Science, V.H.N. Senthikumara Nadar College (Autonomous), Virudhunagar**

[2.] **Dr.P. Sundara Pandian, Principal, V.H.N.Senthikumara Nadar College (Autonomous), Virudhunagar**

1.1 INTRODUCTION

The 21st century will be ruled by data. Data Science has become an indispensable part of many businesses and industries. It provides valuable insights into customer behavior that can lead to increased conversions, more detailed market analysis for competitive advantage in pricing strategies or product development, improved operational efficiency, and minimized risk exposure through accurate forecasting models. The emergence of disruptive technologies like IoT, digital media platforms, smartphones, artificial intelligence, big data analytics, blockchain, and quantum computing has ushered in an era where Data Science will be central to organizational success. Data Science has critical applications across most industries. Organizations, big and small, need Data Science to make decisions, analyze market trends, minimize losses and maximize profits. Data-driven insights cannot only radically transform businesses but also help target new markets, address customer pain points, boost revenue and much more. As such, an ever-increasing number of businesses are focusing on capturing, interpreting, and being informed by data.

Research shows that companies and organizations are heavily investing in data-driven businesses. A part of their investment is directed towards technology. The next generation of employees are based on data-driven culture. For example, Lockheed Martin has launched data literacy workshops to teach employees across their U.S. operations. The aerospace and global security company plans to roll out their classes to people working in other non-traditional analyst roles, including manufacturing. As a result, the analytics team has seen shifts in how employees treat data concerning their roles and, equally important, the added value they create as data literate professionals.

1.2 Developments in Data Science

1.2.1 Healthcare sector- The biggest application of Data Science is in healthcare. The accessibility of large datasets of patients can be used to build a Data Science approach to identify the diseases at a very early stage. Healthcare is one of the biggest sectors for providing opportunities for the professional who can use their medical expertise with Data Science and provide immediate help to the suffering patients.

1.2.2 Arms and Weapons- Data Science can help in building various automated solutions to identify any attack at a very early phase. Other than that data science can help in constructing automated weapons that will be smart enough to identify when to fire and when not to.

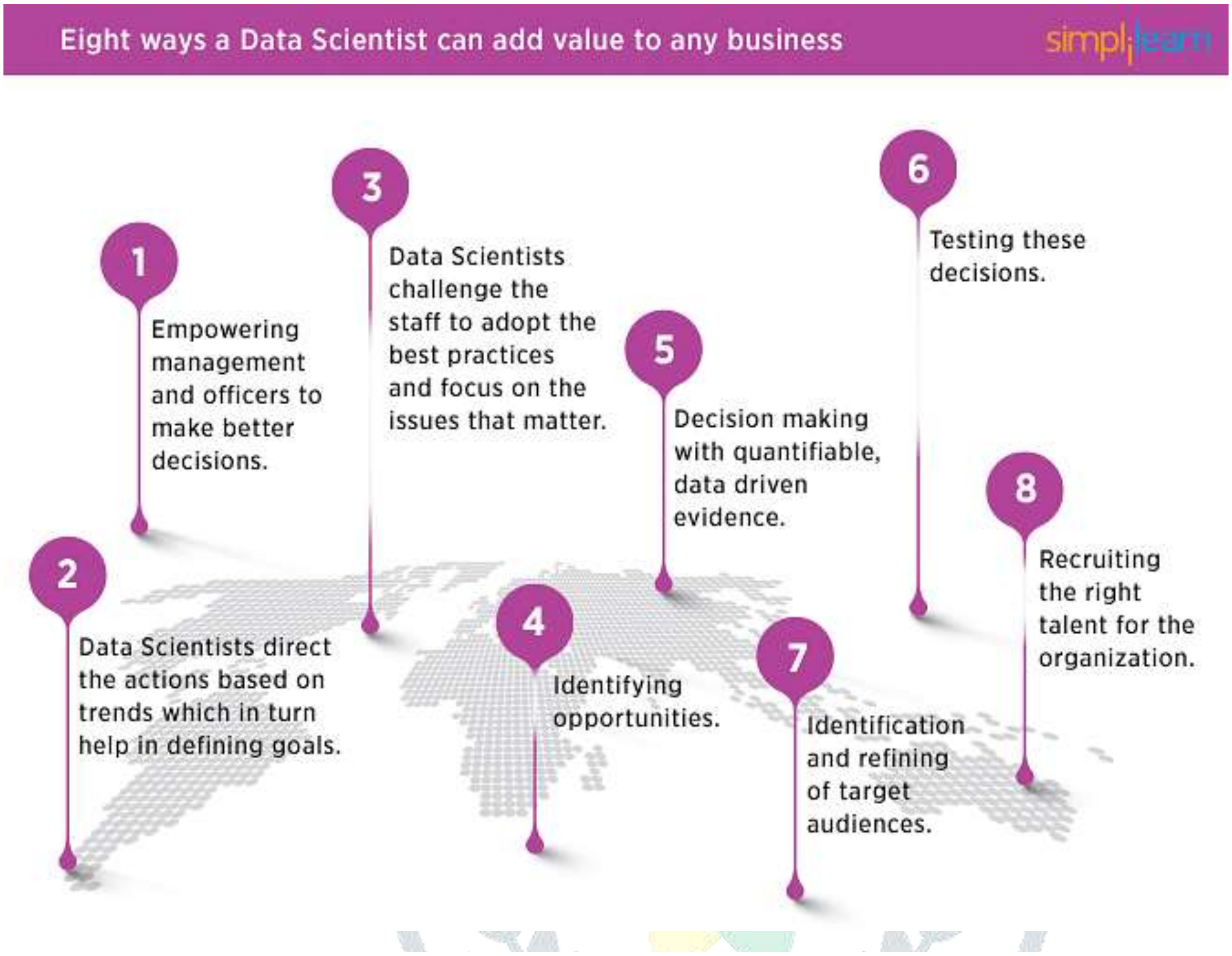
1.2.3 Banking and Finance- Data Science in the Banking and Finance sector can be used in managing the money effectively to invest in the right places based on data science predictions for best results.

1.2.4 Automobile Industry:Other than the above sectors data science is also applied in Automobile Industry like self-driving cars. Fixed destination cabs as well as in power and energy data science can predict the maximum safest potential and can help in building AI bots that can easily handle enormous power sources.

The implementation of Data Science cannot be ignored as it is already in action in the present stage. When one looks for something in Myntra or Flipkart and then onegets similar recommendations or similar advertisements for whatever onehas searched on the internet is all about Data Science.

1.3 Eight ways a data scientist can add value to any business.

The whole world is operated by Data Science. For every single search in Google, the process of data science is activated. The future of data science is growing. Data will stipulate modern health care, finance, business management, marketing, government, energy, and manufacturing. The scale of big data is truly staggering as it has already entwined itself in the fundamental aspect of business as well as personal life. The following image shows eight ways a data scientist can add value to any business.


Eight ways a Data Scientist can add value to any business
simplilearn
1
Empowering management and officers to make better decisions.
2
Data Scientists direct the actions based on trends which in turn help in defining goals.
3
Data Scientists challenge the staff to adopt the best practices and focus on the issues that matter.
4
Identifying opportunities.
5
Decision making with quantifiable, data driven evidence.
6
Testing these decisions.
7
Identification and refining of target audiences.
8
Recruiting the right talent for the organization.

1.3.1 Empowering Management and Officers to Make Better Decisions

An experienced data scientist is likely to be a trusted advisor and strategic partner to the organization's upper management by ensuring that the staff maximizes their analytics capabilities. A data scientist communicates and demonstrates the value of the institution's data to facilitate improved decision-making processes across the entire organization, through measuring, tracking, and recording performance metrics and other information.

1.3.2. Directing Actions Based on Trends—which in Turn Help to Define Goals

A data scientist examines and explores the organization's data, after which they recommend and prescribe certain actions that will help improve the institution's performance, better engage customers, and ultimately increase profitability.

1.3.3. Challenging the Staff to Adopt Best Practices and Focus on Issues

One of the responsibilities of a data scientist is to ensure that the staff is familiar and well-versed with the organization’s analytics product. They prepare the staff for success with the demonstration of the effective use of the system to extract insights and drive action. Once the staff understands the product capabilities, their focus can shift to addressing key business challenges.

1.3.4Identifying Opportunities

With the arrival of data scientists, data gathering and analyzing from various channels has ruled out the need to take high stake risks. Data scientists create models using existing data that simulate a variety of potential actions—in this way, an organization can learn which path will bring the best business outcomes.

1.3.5 Testing These Decisions

Half of the battle involves making certain decisions and implementing those changes. It is crucial to know how those decisions have affected the organization. This is where a data scientist comes in. It pays to have someone who can measure the key metrics that are related to important changes and quantify their success.

1.3.6 Identification and Refining of Target Audiences

From Google Analytics to customer surveys, most companies will have at least one source of customer data that is being collected. The importance of data science is based on the ability to take existing data that is not necessarily useful on its own and combine it with other data points to generate insights an organization can use to learn more about its customers and audience.A data scientist can help with the identification of the key groups with precision, via a thorough analysis of disparate sources of data. With this in-depth knowledge, organizations can tailor services and products to customer groups, and help profit margins flourish.

1.3.7 Recruiting the Right Talent for the Organization

Data scientists can work their way through all these data points to find the candidates who best fit the organization’s needs. By mining the vast amount of data that is already available, in-house processing for resumes and applications—and even sophisticated data-driven aptitude tests and games—data science can help the recruitment team make speedier and more accurate selections.

1.4 Problems faced by Data Scientist

There are many problems faced by data scientist. Data scientist usually works with a team which includes stakeholders like a product Manager. Often the business problem statement can be framed poorly by either party. Other main problems are multiple data sources, imbalanced data, overfitting and version control.

1.4.1 Multiple Data Sources

Big companies use various software and mobile applications like ERPs and CRM to collect and manage information related to their customers, sales details and employees' details. Data consolidation is a complex process which leads to non- uniformed formats. Moreover, there are a variety of sources to handle and extract data from.Heterogeneous sources often make it difficult for data scientists to understand and gather meaningful insights. Hence, they end up spending more time on filtering it, which leads to errors and unreliable decision-making. In such cases, it is crucial to standardize data for accurate analysis.

1.4.2 Imbalanced Data

Most Data Scientists encounter issues of imbalanced data. An example is when a problem is regarding classification problem, and the data scientist is using logistic regression to either assign a 0 or 1 to new data. The target variable would hopefully be 50% of 0, and 50% of 1. However, this nearly does not happen as the data scientist would expect it to be. If the data scientist is trying to classify a new animal as either a dog or a cat. What would you want, for example, 1,000 cat and 1,000 dog rows of training data. This way, the model would be confident in identifying the difference between the two. If you had 1,900 cats and 100 dogs in your training data, then you would misleadingly have high confidence that most new animals should be a cat. This is a common problem.

1.4.3 Overfitting

The overfitting problem occurs when the data scientist builds a Data Science model that learns the training data too well. The model becomes too specific, absorbing detailed information about the training data, as well as noise in the data — which is not useful in predicting new, real data. However, the model will, as a result, not infer or generalize properly. The goal of a model is to perform well on unseen data, so one will want to find and use solutions that perform well on new data.

1.4.4 Version Control

Data Scientists can get so used to working alone, they get comfortable having 20 Jupyter Notebooks that are just different versions of the same, main project. Data scientist end up using some naming convention that they forget the next day. Then, they try to share and reenact what they created, and it does not work and everything becomes a mess.

1.4.5 Business Problems

Before performing data analysis and building solutions, data scientists must first thoroughly understand the business problem. Most data scientists follow a mechanical approach to do this and get started with analyzing data sets without clearly defining the business problem and objective.

1.4.6Optimization Problems

These are problems that can be characterized as maximising or minimising factors such as costs, revenues, risks, time or pollution, within a well-defined quantitative framework and with a given set of constraints.

1.4.7 Fraud Analytics

As organizations transition into cloud data management, cyberattacks have become increasingly common. This has caused two major problems –

1. Confidential data becoming vulnerable
2. As a response to repeated cyberattacks, regulatory standards have evolved which have extended the data consent and utilization processes adding to the frustration of the data scientists.

Organizations should utilize advanced machine learning enabled security platforms and instill additional security checks to safeguard their data. At the same time, they must ensure strict adherence to the data protection norms to avoid time-consuming audits and expensive fines.

1.5 PROMINENT FACTORS/ PROBLEMS FACED BY DATA SCIENTIST

The problems faced by data scientist are analyzed using Principal Component Factor Analysis, and are depicted in Table 1. To ensure the suitability of the instrument and to increase its validity and reliability, the 14 statements were subjected to pretest. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity were also implemented to test the fitness of the data.

Table: 1

Data Science Problems

ROTATED COMPONENT MATRIX

Variables	**Factor 1**	**Factor 2**	**Factor 3**
Heterogenous sources	**.854**		
Time filtering the data	**.754**		
Unreliable decision making	**.697**		
Insufficient data	**.875**		
Overfitting	**.742**		
Version Control	**.682**		
Supply chain risk optimisation		**.789**	
Specialist Algorithms		**.546**	
Path Dependent		**.659**	
Chain Optimization		**.568**	
Counter-fraud analytics			**.894**
Automated product matching			**.456**
Data Requirements			**.654**
Risk Taking			**.547**

Source: Computed Data

The above table exhibits the rotated factor loadings for the 14 statements (factors) which influence the problems faced by the data scientist. It is clear from the table that all the 14 statements had been extracted into three factors, namely F1, F2, F3. These factors were identified with new names such as Business-Related Problems, Optimization Problems and Fraud Analytics and had been presented in the following tables.

1.5.1 Business Problems

Retail Businesses and marketing or advertising firms are facing problems based on many factors including increasing revenue by improving product recommendations, upselling, cross selling, reducing churn and improving retention rates, personalizing the user experience, improving target marketing, sentiment analysis, product or service personalization and pricing optimization. The data scientist should understand the needs, motivation, likes and dislikes of all the customers with the available data. In creating data bases the data scientist face problems like heterogenous sources, time filtering data, unreliable decision making, imbalanced data, overfitting and version control.

FACTOR - 1

Business Related Problems

Variables	Factor Loadings	Eigen Value	% of Variance
Heterogenous sources	**.872**	9.355	24.620
Time filtering the data	**.854**		
Unreliable decision making	**.754**		
Insufficient data	**.697**		
Overfitting	**.875**		
Version Control	**.742**		

Source: Computed data

The variables such as heterogenous sources, time filtering the data, unreliable decision making, insufficient data, overfitting and version control. loaded in Factor 1. Hence F1 is termed as Business Related Problems. The eigen value for the above Factor 1 was 9.355 and the percentage of variance was 24.620. It would be concluded that business related problems rank as the first important factor/ problem facedby the data scientist.

1.5.2 Optimization Problems

These are problems that can be characterized as maximising or minimising factors such as costs, revenues, risks, supply chain risk optimisationwithin a well-defined quantitative framework and with a given set of constraints. The problems involves modelling graphs or networks and solving them heuristically using specialised algorithms. Typically, this is complex because solutions are ‘path dependent’. Some examples of optimization problems are supply chain optimisation, logistics and transportation. It also includes a retailer optimizing staffing levels per store within a shift pattern of work or an airline wishing to optimize its route network.

FACTOR - 2

Optimization Problems

Variables	**Factor Loadings**	**Eigen Value**	**% of Variance**
Supply chain risk optimization	**.789**	6.0488	15.916
Specialist Algorithms	**.546**		
Path Dependent	**.659**		
Chain Optimization	**.568**		

Source: Computed data

The variables such as supply chain risk optimization, specialist algorithms, path dependent and chain optimization are loaded in Factor 2. Hence F2 is termed as optimization Problems. The eigen value for the above Factor 2 was 6.0488 and the percentage of variance was 15.916. It would be concluded that Optimization problems ranks as the second important factor/ problem facedby the data scientist.

1.5.3 Fraud Analytics

Counter fraud can be among the most challenging data science problems for several reasons. Counter-fraud analytics is chasing an 5ever-moving target. Secondly, the counter-fraud analyst does not know the true extent of fraud. This make statistical generalizations difficult, which in turn must be incorporated into the model building process.Thirdly, and perhaps most importantly, fraud is the quintessential needle in a haystack problem. 99.9% of banking transactions are not fraudulent. Therefore, the number of fraud data points from which to generalize is very small. This is, again, critical when considering statistical models.

FACTOR - 3

Fraud Analytics

Variables	Factor Loadings	Eigen Value	% of Variance
Counter-fraud analytics	**.894**	4.798	12.625
Automated product matching	**.456**		
Data Requirements	**.654**		
Risk Taking	**.547**		

Source: Computed data

The variables such as counter fraud analytics, automated product matching, data requirements and risk taking are loaded in Factor 3. Hence F3 is termed as fraud analytics. The eigen value for the above Factor 3 was 4.798 and the percentage of variance was 12.625. It would be concluded that fraud analytics ranks as the third important factor/ problem faced by the data scientist

The result of KMO measure and Barlett's test of sphericity clearly indicate the appropriateness of the use of factor analysis. The factor loadings of all accepted statements are greater than 0.5, and eigen values of all dimensions/ factors are higher than 1.0. This confirms the report of Hair, Black, Babin, Anderson and Tatham (2005) regarding the appropriateness of factor analysis. As predicted by the factor analysis,the major problems faced by data analyst are Business Related Problems, optimization problems and fraud Analytics.

1.6 Suggestions

- ✓ The data scientist should use a centralized platform that allows integrating data from those sources. Next step is to create a data strategy and quality management plan as the data gathered from these sources will be dynamic. Prioritizing and integrating datasets in a centralized system saves time and effort as well as it helps in aggregating data at a single location in real-time.

- ✓ Data scientist should use Machine learning algorithms to create new synthetic data. Specific methods such as ADASYN (*Adaptive Synthetic*)BorderlineSMOTE, KMeansSMOTE, RandomOverSampler, SMOTNC andSVMSMOTE (*Nominal and Continuous*) can be used.
- ✓ For overfitting problems, the data scientist can use k-fold cross-validation, removing duplicate or similar features, early stopping, regularization, ensemble approach, non-parametric Machine Learning algorithms and training on more data.
- ✓ The data scientist can use Git, GitHubalso, other more homemade ways, like versioning with special numeric methods (*e.g.*, *Notebook.1*, *Notebook.2*) for version control problems
- ✓ Identifying key objectives, deliverables, and data requirements, and realistically estimating resources will go a long way to ensuring that data scientist can substantially mitigate the risk.

1.7 Conclusion

Data is a new electricity in the age of fourth industrial revolution. There is a massive data explosion which have resulted in the culmination of new technologies and smarter products.Its importance reflected in the many products designed to boost customer experiences. Despite all the challenges, data scientists are the most in-demand professionals in the market. With the data world changing at a rapid pace, being successful data scientists is not just about having the right technical skills but also about having a clear understanding of the business requirements, collaborating with different stakeholders, and convincing business executives to act upon the analysis provided. Data scientist will help the business to take right business decisions and maximize their profits.

References

1. Tukey JW (1962) The future of data analysis. Ann Math Stat 33:1–67.
2. Gelman A, et al. (2014) Bayesian Data Analysis (CRC, Boca Raton, FL), 2nd Ed.
3. Murphy K (2013) Machine Learning: A Probabilistic Approach (MIT Press, Cambridge, MA).

4. Barber D (2012) Bayesian Reasoning and Machine Learning (Cambridge Univ Press, Cambridge, UK). 11 Hastie T, Tibshirani R, Wainwright M (2015) Statistical Learning with Sparsity: The Lasso and Generalizations (CRC, Boca Raton, FL).
5. Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. Science 349:255–260.
6. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444.
7. Pearl J (2009) Causality (Cambridge Univ Press, Cambridge, UK), 2nd Ed.
8. Imbens G, Rubin D (2015) Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction (Cambridge Univ Press, Cambridge, UK).
9. Morgan S, Winship C (2015) Counterfactuals and Causal Inference (Cambridge Univ Press, Cambridge, UK), 2nd Ed.
10. Sra S, Nowozin S, Wright S (2012) Optimization for Machine Learning (MIT Press, Cambridge, MA).
11. Efron B, Tibshirani R (1993) An Introduction to the Bootstrap (Chapman & Hall/CRC, Boca Raton, FL).
12. Robert C, Casella G (2004) Monte Carlo Statistical Methods, Springer Texts in Statistics (Springer, New York), 2nd Ed.
13. Green PJ, Łatuszy ´nski K, Pereyra M, Robert CP (2015) Bayesian computation: A summary of the current state, and samples backwards and forwards. Stat Comput 25:835–862.
14. Dean J, Ghemawat S (2008) MapReduce: Simplified data processing on large clusters. Commun ACM 51:107–113. 22 Bekkerman R, Bilenko M, Langford J, eds (2011) Scaling Up Machine Learning: Parallel and Distributed Approaches (Cambridge Univ Press, Cambridge, UK).
15. Jordan MI (2013) On statistics, computation and scalability. Bern 19:1378–1390.
16. Cleveland WS (2001) Data science: An action plan for expanding the technical areas of the field of statistics. Int Stat Rev 69:21–26.
17. Hardin J, et al. (2015) Data science in statistics curricula: Preparing students to “think with data.” Am Stat 69:343–353. 26 Goodman A, et al. (2014) Ten simple rules for the care and feeding of scientific data. PLOS Comput Biol 10:e1003542.
18. Borgman CL, et al. (2015) Knowledge infrastructures in science: Data, diversity, and digital libraries. Int J Digit Libr 16:207–227