International Journal of Computer Sciences and Engineering
Vol.**12**, Issue.**4**, pp.**55-60**, April **2024**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Impact of Near Real Time Data on Data Science Model Predictions

Ankush Ramprakash Gautam[1*], Ritu Sharma[2]

[1]Senior Manager Engineering, Datastax, Frisco, Texas, USA
[2]Lead Data Scientist, JPMorgan Chase & Co, Plano, Texas, USA

**Corresponding Author: ankush.gautam@yahoo.com*

Received: 04/Mar/2024; **Accepted**: 05/Apr/2024; **Published**: 30/Apr/2024. **DOI**: https://doi.org/10.26438/ijcse/v12i4.5560

Abstract: The article delves into an exploration of how the integration of almost real time data streams impacts the accuracy, strength and effectiveness of models, in the ever changing field of data science. groups go beyond boundaries to examine sectors carefully analyzing the effects of data velocity on model performance in industries like finance, healthcare and transportation. Through an investigation the article reveals a story that highlights not the many benefits but also examines the complex challenges involved in utilizing almost real time data for modeling purposes. Additionally the article takes a look at the details discussing the necessary setup requirements and explaining the various methodological approaches needed to seamlessly integrate rapidly updating data streams into existing modeling frameworks. The paper also covers considerations and privacy requirements, for handling data responsibly emphasizing the importance of preserving individual privacy and data integrity. In the end this research acts as a signal emphasizing the importance of utilizing nearly real time data to enhance predictive abilities and drive a significant change in how decisions are made in various fields. This pushes us towards a future of opportunities and transformative possibilities.

Keywords: Data Science, Data Quality, Real Time

1. Introduction

In the changing world of data science, the pursuit of accuracy and relevance in modeling is more important than ever. With the abundance of data streams producing amounts of information rapidly incorporating near real time data has become a game changer that is reshaping how predictive analytics work. [1] Businesses are striving to leverage data driven insights for making informed decisions highlighting the role that near real time data plays in enhancing the precision, dependability and flexibility of data science models. While traditional models have traditionally relied on data to predict outcomes and provide insights into past trends they are limited in capturing the dynamic nature of real world situations. [2] This limitation has led to the integration of real time data sources into analytics frameworks to overcome this challenge. Whether its media updates, sensor readings from devices, market fluctuations or weather reports incorporating up to the minute information allows data scientists to grasp the subtle details of evolving scenarios with exceptional detail. This article explores how near real time data influences predictions made by data science models, from angles. The article delves into how the quickness and freshness of real time information can boost the abilities of machine learning algorithms allowing organizations to predict and react to emerging trends, irregularities and occurrences instantly.

Additionally the article discusses the difficulties posed by the speed, quantity and diversity of real time data streams shedding light on methods to reduce interference, ensure data accuracy and promote clarity in models. Furthermore the article analyzes the impacts of integrating near real time data in sectors such, as finance, healthcare, transportation and cybersecurity. By showcasing case studies and recommended approaches the article illustrates how organizations use real time data to optimize resource distribution, improve risk management tactics, customize user interactions and strengthen resilience in a changing and interconnected environment. While exploring the power of real time data in predicting data science models outcomes the article addresses ethical concerns related to data privacy protectionism, security issues and bias. The surge in real time data raises issues regarding consent,transparency and responsibility when making decisions based on gathered information; therefore it calls for an approach to ethical leadership governance practices and regulatory compliance. Essentially incorporating data, in time brings about a fresh era of flexibility and insight in the field of data science enabling businesses to extract practical knowledge from the constant flow of information circulating in the digital world. This article seeks to shed light on the benefits, obstacles and ethical aspects of using real time data for analysis with the goal of showing ways to tap into the impact of data driven

insights, in a world that is becoming more intricate and ever changing.

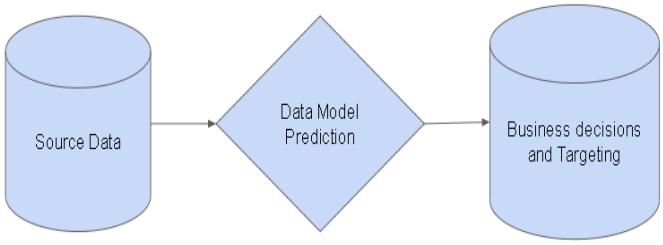

Figure 1. Data Science prediction batch flow

2. Advantages of near real time data

In today's dynamic business landscape companies that incorporate data into their data analysis frameworks gain a notable edge over their competitors. Utilizing near real time data allows organizations to make informed decisions based on the recent insights, react promptly to evolving situations, predict future trends, offer tailored experiences to clients and streamline resource management effectively. Furthermore real time data facilitates identification and mitigation of irregularities fostering learning and enhancement of data analysis models. By harnessing the power of information companies can enhance prediction accuracy, enhance responsiveness, gain foresight and preparedness personalize experiences, for individuals optimize resource distribution, identify and prevent irregularities proactively while fostering continuous learning and improvement. These efforts ultimately lead to success and competitiveness in a data focused world.

2.1 Accuracy Improvement

Incorporating near real time data into models represents a progress that boosts the precision of forecasts and decision making processes. By offering models, with the information they can accurately mirror developments and tendencies. The utilization of real time data guarantees that predictions and decisions are grounded on the knowledge leading to an elevated level of accuracy. Integrating near real time data into models stands as an enhancement that allows organizations to make well informed decisions with increased assurance.

2.2 Model Responsiveness

Utilizing models powered by, up to the minute data boosts how quickly organizations can respond, allowing them to adjust swiftly to changing conditions. This flexibility empowers companies to react promptly to market changes, evolving customer needs and operational challenges. This ability provides an edge by supporting decision making, proactive actions and improved operational effectiveness, in a fast evolving landscape.

2.3 Risk Management

In the field of risk management the use of, up to the minute data has transformed how sectors, like finance and healthcare function. Being able to track data flows empowers companies to spot risks promptly. This proactive strategy lets organizations act quickly to lessen risks reducing losses and bolstering their resilience overall. The immediacy of the data enables organizations to anticipate dangers and react promptly safeguarding assets, reputation and operational consistency.

2.4 Personalization

[3] Businesses can use real time data to tailor their products, services and experiences to match each customer's preferences and behaviors. By analyzing data companies can offer suggestions, deals and assistance enhancing customer satisfaction and building loyalty.

2.5 Financial impact of near real time data to America

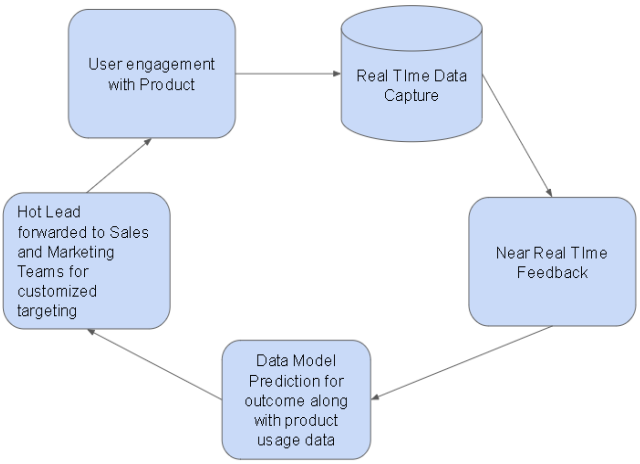

Figure 2. Data Science prediction batch flow

The increasing use of, up to the minute data in the United States is making an impact on data science models in ways. To start with it boosts accuracy and efficiency by allowing models to learn from updated information resulting in dependable forecasts and improved decision making. This can lead to advantages in fields like fraud detection, targeted marketing and risk management. Secondly real time data enables responses empowering businesses to take advantage of opportunities and address risks. For instance a retail store that leverages real time customer behavior data can adjust prices or promotions on the go to maximize sales and profits. Moreover real time data helps enhance efficiency by streamlining processes and allocating resources effectively leading to cost savings. As an illustration factories can monitor equipment performance. Anticipate maintenance requirements using real time sensor data to prevent downtimes. Additionally real time data creates revenue opportunities by supporting the creation of data driven services and products that generate income streams. While the financial rewards are significant there are challenges to navigate such as the costs associated with establishing and maintaining real time data infrastructure as the necessity for robust systems, for processing and managing data accurately to maintain quality standards and avoid mistakes. In general the influence of data on data science models in the United States is beneficial giving companies an edge in competition and opening up fresh financial possibilities.

2.6 Impact to finance data science models

In the realm of finance the incorporation of, up to the minute data into data science models has sparked a transformation. [4] This combination has revolutionized how financial institutions function, analyze markets and handle risks. It's akin to having a connection to the heartbeat of the market providing an accurate grasp of market trends, investor attitudes and economic indicators. Armed with this newfound insight, data science models can make informed decisions. Craft efficient risk management plans. Real time data also facilitates the fine tuning of trading strategies offering an advantage in the moving realm of finance. Automated trading platforms can promptly respond to shifting market conditions by executing trades with heightened accuracy and speed. Algorithms for frequency trading leverage real time market information to identify opportunities and capitalize on inefficiencies enhancing efficiency and liquidity in markets. The integration of up to the minute data extends beyond trading to encompass risk management and fraud detection well. By monitoring data and customer behaviors data science models can identify irregularities and fraudulent activities in near real time. This proactive approach helps protect assets and strengthen risk management frameworks. Furthermore real time data enables dynamic pricing adjustments, portfolio management enhancements, as personalized customer experiences. Financial organizations have the ability to modify pricing, premiums and investment distributions in response to market conditions. They can also gain an understanding of customer preferences customizing products and services to suit needs. To sum up, the advent of real time data has completely transformed the sector. It has enabled institutions to analyze markets, conduct trades with greater efficiency, manage risks more adeptly and cater to customers with enhanced intelligence. By leveraging real time analytics financial institutions are able to thrive in today's intricate environment.

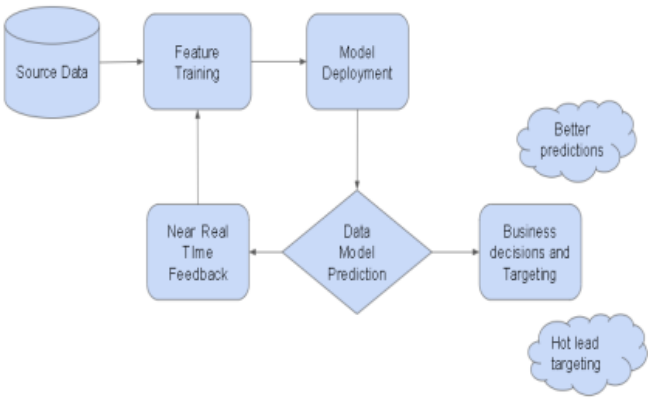

Figure 3. Personalized customer targeting using real time data in data science models

2.7 Impact to healthcare data science models

In the field of healthcare the incorporation of, up to the minute data into healthcare data science models is causing a stir. This innovation is reshaping patient care, medical research and overall healthcare services. Just picture having access to data from wearable gadgets, digital health records and medical scans. [5] It's akin to possessing an insight tool that empowers healthcare professionals to spot diseases, forecast patient outcomes and tailor treatments with precision. It's like having a hero with the ability to save lives and enhance results! From disease identification to customized treatment strategies, community health monitoring to decision aids, up to the minute data is an influential factor. It's revolutionizing how healthcare is provided by making it more personalized, preventive and effective.

2.9 Impact to Transportation Data Science Models

[6] In the field of transportation the introduction of data has had a significant impact on data science models changing how transportation systems are managed, optimized and experienced. This data enables transportation agencies to monitor traffic conditions in time allowing them to adjust traffic signals, modify toll prices and implement routing systems to reduce congestion and improve traffic flow. Using near real time data for maintenance helps save costs and enhance reliability by identifying infrastructure issues early on. Leveraging near real time vehicle data for public transit optimization improves route efficiency, schedules and capacity management for services. Managing pricing and demand dynamically in transportation services helps balance supply and demand to boost profitability for providers. Additionally real time monitoring of traffic conditions and incidents enhances safety by enabling responses to emergencies and improved incident management. Integrating near real time data into transportation models enhances efficiency, reliability and sustainability of systems leading to congestion, increased safety standards and improved mobility, for people as well as goods transport

3. Challenges of near real time data

Real time data can offer insights, for predicting outcomes in data science. It's not without its challenges. It's important to acknowledge the issues that come with it. Firstly the quality of the data may be inconsistent containing gaps, inaccuracies or discrepancies. Secondly, working with real time data requires tools and expertise compared to batch processing methods. This adds complexity to managing and analyzing the data streams effectively. Thirdly there is a cost associated with handling real time data due to the need for infrastructure and resources. Moreover there may be delays in accessing or processing the data impacting the accuracy of predictions. Privacy and security concerns are also considerations to prevent access or tampering with sensitive information. Lastly there is a risk of models being overly tuned to patterns in data or failing to adjust to evolving trends, in data distribution over time.

3.1 Data modeling challenges

[7] Dealing with data modeling for almost real time data brings about a range of obstacles because of the need for processing and analysis of changing information. These hurdles include the speed at which data flows, meaning models must be able to handle input, processing and analysis. The volume of data is also a concern in near real time systems requiring modeling methods. Different. Structures of real

time data call for adaptable modeling techniques. Managing complexity involves using algorithms to keep up with changing data patterns. Minimizing processing delays is crucial due to latency issues. Ensuring data quality is tricky because of the flow of information, necessitating validation, cleaning and error correction processes. Scalability involves expanding capabilities to manage growing amounts of data and user activity. Processing streaming data requires utilizing frameworks designed for stream processing. Taking into account time related factors involves considering event timestamps and time sensitive information in datasets. Tackling these challenges calls for an effort involving experts, in data modeling, engineering, science and IT specialists.

3.2 Data privacy challenges

In the realm of processing data, instantly prioritizing data privacy stands as a concern, due to the swift pace at which data is received, processed and analyzed. [8] Various significant challenges have surfaced concerning data privacy in real time systems posing barriers to effectively safeguarding information. These challenges encompass a variety of areas such as securing data through encryption, controlling access concealing and anonymizing data, managing consent minimizing data collection, complying with regulations on where data can be stored and who has jurisdiction over it, preventing data leaks and ensuring auditing and compliance reporting. Tackling these obstacles calls for an approach that includes technical measures, clearly defined policies and standardized procedures. It demands collaboration among stakeholders like experts in data privacy, cybersecurity professionals, data specialists and business leaders to comply with laws on data privacy regulations effectively safeguard private information from unauthorized use or access while maintaining trust, among individuals and organizations.

3.3 Infrastructure costs

When dealing with almost real time data the expenses for the infrastructure can change depending on factors such as the amount of data, its speed, diversity and complexity. Let's break it down. [9] Initially data ingestion involves bringing in data from sources to the processing environment, which may entail setting up connectors or APIs. Real time data processing is facilitated by streaming platforms that could be expensive based on software licenses and cloud charges. Analyzing and transforming the data requires compute resources like machines or containers. Data storage solutions are used to retain the data with costs varying according to data volume and storage type. Networking costs arise from moving data between parts of the pipeline. Monitoring tools and management aids ensure uptime and responsiveness with associated expenses. Costs also stem from scalability and redundancy measures, like instances or disaster recovery setups. To effectively handle these expenses organizations must carefully evaluate their needs. Select infrastructure choices.

4. Considerations

Ensuring the efficacy of models critically depends on the quality of the data utilized. Real-time data streams frequently present challenges such as incomplete or noisy data, posing risks of biases and inaccuracies. Therefore, employing data cleansing techniques, imputation methods, and outlier detection algorithms becomes imperative to address issues such as missing values, anomaly identification, and preserving data integrity. Latency, denoting the temporal delay between data generation and its preparedness for processing, poses a significant obstacle. To tackle this issue, streaming architectures like Apache Kafka and Apache Flink are employed. These frameworks facilitate the ingestion and processing of data in real-time. Methods such as micro-batching and incremental model updates play a pivotal role in mitigating latency without compromising model performance.

Within the domain of real-time data analysis, feature engineering assumes a crucial role in uncovering substantive insights. Unlike static datasets, real-time data manifests temporal patterns and dynamics that require capture. In response, feature engineering methodologies such as time series analysis, sliding windows, and change point detection become instrumental. These methodologies facilitate the identification of pertinent features and extraction of temporal patterns, thereby fostering a more profound comprehension of the data. Evaluating model performance and safeguarding against data leakage stand as pivotal facets within the realm of machine learning. Employing cross-validation techniques, such as holdout sets and time series cross-validation, aids in assessing a model's ability to generalize while detecting potential instances of overfitting or underfitting. Data leakage, a prevalent concern where forthcoming information is inadvertently utilized in model training, demands meticulous consideration. Rigorous data partitioning and temporal validation are imperative measures to counteract data leakage effectively and uphold the integrity of model evaluation.

In machine learning and data science workloads, evaluating model performance and preventing data leakage are essential aspects. Cross-validation techniques, such as holdout sets and time series cross-validation, allow for the assessment of a model's generalization capabilities and help detect overfitting or underfitting. Data leakage, a common issue where future information is unintentionally included in model training, requires close attention. Careful data partitioning and temporal validation are crucial measures to mitigate data leakage and ensure the integrity of model evaluation. As the volume of real-time data streams expands, ensuring scalability emerges as an imperative for proficient processing and inference. Distributed computing frameworks such as Apache Spark and Apache Hadoop have garnered acclaim for managing extensive data processing tasks in real-time environments. Moreover, model architectures equipped with incremental data processing capabilities, such as online learning algorithms and deep neural networks, are utilized to attain scalability.

Sustaining model performance over time necessitates ongoing monitoring and consistent maintenance efforts. Essential monitoring metrics such as accuracy, precision, and recall serve to identify any deviations in model behavior. In instances where performance deteriorates, alerts are activated to prompt corrective measures. To accommodate evolving data patterns and uphold optimal performance levels, regular retraining sessions and model updates are deemed indispensable. In the realm of real-time data management, privacy, security, and interpretability pose significant challenges. Preserving data privacy requires employing data encryption, access control, and anonymization techniques. Interpretable models like decision trees and linear regression are essential for understanding how a model makes decisions. This promotes comprehension of model behavior and adherence to fairness and transparency principles.

The integration of real-time data with domain expertise is paramount for the formulation of substantive and applicable models. Specialists in the subject matter domain offer profound insights into its intricacies, facilitating the identification of pertinent features, optimization of model parameters, and alignment of models with overarching business objectives. By capitalizing on these insights, data scientists can effectively utilize real-time data to develop dependable and efficient models. These models empower businesses to make informed decisions, streamline operations, and stimulate innovation.

5. Conclusion

In the changing field of data science the use of real time data has become a game changer transforming how predictive models perform in different areas. We've looked into how incorporating near real time data significantly influences the accuracy and effectiveness of predictions made by data science models. One key advantage of using near real time data is that it adds flexibility to models helping them quickly adapt to changing trends and patterns. Unlike datasets that can quickly become outdated near real time data ensures models are continuously updated with the recent information. This continuous updating feature allows models to capture changes in behavior, market dynamics or other relevant factors ultimately improving their accuracy and applicability.

[10] Additionally integrating near real time data promotes decision making processes. Organizations can use real time insights to make well informed decisions giving them an advantage in dynamic settings. Whether it involves adjusting marketing strategies based on consumer feelings or optimizing supply chain operations in response to shifting demand patterns, having access to real time data enables stakeholders to act decisively and proactively. Moreover incorporating, up to the minute data enhances the development of more adaptable models. By blending data with real time information streams data experts can craft models that not only provide accurate predictions but also adjust to unexpected situations. This adaptability is especially vital in situations where sudden disruptions or irregularities can have an impact on results, such as in financial forecasting or risk management.

Furthermore the use of real time data supports the implementation of strategies for mitigating risks and detecting anomalies. Through monitoring of incoming data streams organizations can promptly spot deviations from patterns and take proactive measures to reduce potential risks. Whether it involves identifying activities in systems or pinpointing irregularities in network traffic for cybersecurity purposes the capability to analyze real time data empowers organizations to anticipate and address emerging threats effectively.

Nevertheless it is important to recognize that integrating near real time data also presents challenges and considerations. Factors, like data accuracy, time delays and scalability must be handled thoughtfully to ensure the dependability and effectiveness of models. Additionally ethical concerns related to collecting, storing and utilizing real time data demand attention to protect privacy and uphold data security.

In summary the influence of data, on the forecasts of data science models is unquestionable. By leveraging information companies can discover avenues for creativity, flexibility and staying ahead in the competition. Yet achieving the benefits of integrating real time data demands an approach to tackling technical, ethical and organizational obstacles. As time moves through this era governed by data, making efficient use of almost real time data will remain an element in molding the future of data science and predictive analysis.

The next steps for the article will be to conduct experiments by collaborating with industry professionals and measure true cost impact to the business when near real time data is leveraged.

Conflict of Interest
No known competing or financial interests are reported in this paper

Funding Sources
No funding was received by the authors

Authors' Contributions
Ritu Sharma has contributed to the data science aspects of the paper and Ankush Ramprakash Gautam has added insights from near real time data point of view and stitched the article together. Ritu and Ankush's collaboration was not limited to their individual contributions; it extended to a synergistic integration of diverse perspectives and methodologies. They recognized the complementary nature of their expertise and achieved a seamless fusion of technical rigor with real-time relevance. Through iterative discussions and collaborative brainstorming sessions, they navigated complex issues and refined their research framework with coherence and robustness. The result of their collaboration was a scholarly paper that not only elucidated intricate data science concepts but also resonated with practical insights and applicability.

References

[1] The Art of Data Science & Big Data Analytics Inspecting & transforming data , Akella Subhadra , Survey Paper | Journal Paper, Jun, Vol.**8**, Issue.**6**, pp.**91-100**, **2020**, CrossRef-DOI: , https://doi.org/10.26438/ijcse/v8i6.91100

[2] LONGBING, University of Technology Sydney, Australia, Data Science: A Comprehensive Overview, ACM Computing. Vol.**50**, Issue.**3**, pp.**1–42**, **2017.**

[3] Dmitry Ivanov, Omer Ben-Porat, Real-Time Customer Journey Personalization using Online Machine Learning. Published: August **2021**.

[4] Pankaj Gupta, "Leveraging Machine Learning and Artificial Intelligence for Fraud Prevention," SSRG International Journal of Computer Science and Engineering, Vol.**10**, No.**5**, pp.**47-52**, **2023.** Crossref, https://doi.org/10.14445/23488387/IJCSE V10I5P107

[5] Rahnuma Mahzabin, Fahim Hossain Sifat, Sadia Anjum, Al-Akhir Nayan, Muhammad Golam Kibria, Blockchain associated machine learning and IoT based hypoglycemia detection system with auto-injection feature, Vol.**27**, No.**1**, pp.**447-455, 2022**. Cross Ref-DOI https://doi.org/10.48550/arXiv.2208.02222

[6] Niklas Christoffer Petersen a,b,*, Filipe Rodrigues b, Francisco Camara Pereira b, Multi-output Deep Learning for Bus Arrival Time Predictions, Cross Ref-DOI, Vol.**41**, pp.**138-145, 2019**. https://doi.org/10.1016/j.trpro.2019.09.025

[7] Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy, Mining Data Streams: A Review , ACM SIGMOD Record, Vol.**34,** Issue.**2**, pp.**18–26**, **2005.** https://doi.org/10.1145/1083784.1083789

[8] Xinjun Qi, Mingkui Zong, An Overview of Privacy Preserving Data Mining, Data Privacy challenges, Vol.**12**, pp.**1341-1347**, **2012**. https://doi.org/10.1016/j.proenv.2012.01.432

[9] J.D. Kelleher, B.M. Namee, A. D'Arcy, "Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies" The MIT Press, **2015.**

[10] Damiano Perri,Marco Simonetti,Osvaldo Gervasi , Deploying Efficiently Modern Applications on Cloud, Electronics, Vol.**11**, Issue.**3**, pp.**450**, **2022.** https://doi.org/10.3390/electronics11030450

AUTHORS PROFILE

Mr. Ankush Ramprakash Gautam earned his Bachelor's of Engineering from University of Mumbai in 2007 and Masters of Science in Information Technology and Management from The University of Texas at Dallas in 2011.. He is currently working as a Senior Manager Engineering at Datastax from 2023. He has experience in delivering enterprise wide Data projects with direct customer impacts. He also volunteers time for Diversity and Inclusion Initiatives, acts as a volunteer judge for STEM competitions and reviews Data related books for authors. He has published journals in International Journal of Computer Sciences and Engineering and International Journal of Computer Applications. His main research work focuses on Data Engineering, Data Science and Analytics. He has 15 years of work experience.

Ms.Ritu Sharma earned her Bachelors of Technology from Dr. B. R. Ambedkar National Institute of Technology Jalandhar in 2011 and Masters of Science in Business Analytics from The University of Texas at Dallas in 2020. She is currently working as a Lead Data Scientist for JPMorgan Chase and Co since 2019. Her main research work focuses on Data Science and Data Analytics. She has 10 years of work experience spanned across large corporations such as Larsen and Toubro Infotech Ltd, Infosys and HCL Global Systems Inc.

