Transactions on Engineering and Computing Sciences - Vol. 13, No. 02
Publication Date: April 25, 2025
DOI:10.14738/tmlai.1302.18536.

Habtemariam, S., Yihdego, D., & Lambert, E. (2025). Managing AI and Data Science Projects: Unveiling Secrets from Industry Experts for Real-world Success. Transactions on Engineering and Computing Sciences, 13(02). 116-133.

SERVICES FOR SCIENCE AND EDUCATION

Managing AI and Data Science Projects: Unveiling Secrets from Industry Experts for real-world Success

Samuel Habtemariam
School of Science and Engineering, Atlantic International University,
Pioneer Plaza, 900 Fort Street Mall 905, Honolulu, Hawaii 96813, USA

Daniel Yihdego
Microsoft, Azure Data Science & ML, Bellevue, Washington, USA

Edward Lambert
School of Science and Engineering, Atlantic International University,
Pioneer Plaza, 900 Fort Street Mall 905, Honolulu, Hawaii 96813, USA

ABSTRACT

Data science projects are becoming increasingly critical in today's data-driven landscape, yet effectively managing their complexities remains challenging, and this often results in the project's underperformance and or failure. This study provides comprehensive guidelines for critical aspects of data science project management - from initiation and implementation to scaling. Both primary and secondary data were gathered using mixed research methodology. Twenty-six (26) experts from international companies were engaged with a well-structured questionnaire, and selective companies successful in previous data science projects were approached to gather primary data. For secondary data, papers published in reputable journals were researched systematically to support claims from the primary data. However, based on an extensive literature review and critical sampling of opinions, significant challenges for data science projects identified include a lack of feedback-constrained timeframes, project complexity, the inability to meet high stakeholder expectations, the inability to enable informed decisions, and the iterative nature of data science is underscored, along with nuanced expositions on navigating the potential challenges of science data projects. Finally, other best practices identified through literature searches and case studies include regular audits, effective metadata management, data privacy and security, optimal cloud storage for cost-effectiveness, and computation acceleration. Further, recommendations were provided, which contributed to positioning this paper as a valuable guide for practitioners, researchers, organizations, and individuals in the field.

Keywords: Data science, artificial intelligence (AI), project management

INTRODUCTION

According to Foote (2021), the term "data science" was introduced in the early 1960s. It was specifically formulated to refer to a new profession meant to aid the interpretation and comprehension of the large amounts of data that were being compiled at the time. Thereafter, Chinthamu and Karukuri (2023) claim, data science has continued to thrive as a discipline using computer science and statistical methodology to gain essential insights into different other fields. The indelible impact of data science on astronomy and medicine is prominent as it has significantly helped to deduce meaningful amounts of data from raw and unstructured information through modern technological software like machine learning algorithms and other statistical methods (Hey et al., 2009; Obermeyer & Emanuel, 2016).As further explained by Medeiros et al. (2020), the employment of data sciences by industries and businesses is essential to Data science; the end goal—such as the accuracy of a model—might be a moving target. Iteration is more frequent in data science, with projects often looping back to earlier stages based on newfound insights. Moreover, the success metrics in data science projects can be subjective, based on stakeholder interpretation rather than clear-cut benchmarks.

Managing data science projects presents a plethora of challenges, many of which stem from the very nature of the data and tasks involved. Raw data, often termed as 'messy,' requires substantial preprocessing and cleaning, tasks that can be time-consuming and uncertain in terms of outcomes. Sculley et al. (2015) further elaborate on the hidden technical debts in machine learning systems, discussing the challenges that are not immediately apparent but can derail projects. Moreover, the stakeholder expectations in data science projects often revolve around tangible business outcomes, creating pressure to deliver actionable insights irrespective of the uncertainties involved.

The lifecycle of data science takes a business through every stage of a data science project, from the initial problem identification to the point where the solution provided yields consistent profit to the business (Joel, 2022). Furthermore, Panda (2023) claims that most data science projects fail due to their inability to establish a solid foundation that should guide and hold the project to be successful. Before a data science project begins, the problem needs to be identified. This can be achieved in several ways today, such as online forums, surveys, social media, company websites, and so on. The next step includes data investigation, which is a continual business process. The company must dive into the enterprise's data collection methods and repositories to complete this step. Next is the pre-processing of data, which is where all the data processing takes place. Data science operations are carried out following data collection and ETL process completion. Understanding the role of the ETL in the data process is very important. Following this is the exploration of data. At this stage, different plots are to be deployed to visualize the data for a comprehensive study that improves the overall knowledge about the data. Following this is data modelling, which is simply the act of converting raw data into a form that can be transverse into other applications. Lastly, model evaluation and monitoring should be jointly done with the other stages of the data science life cycle because it helps in the early detection of problems regarding data analysis (Mukamwiza & Hakizimana, 2021). For instance, this step helps to know if the model the analyst is working on is accurate or in need of amendments.

However, this paper aimed to explore and give apt attention to providing comprehensive guidelines covering critical aspects of data science project initiation, implementation, management, and scaling, relying on an extensive search of credible resources and sampled opinions of experts in the field of data science through questionnaire literature review, survey, and case studies. Purpose is to manage AI and Data Science projects in order to unveil secrets from industry experts for real-world success.

The Need for Project Management in Data Science

In recent years, data science has cemented its position as a cornerstone of technological innovation. Due to its expansive range from analytics to machine learning, the potential of data science in technological innovation is undeniable. However, the pathway to unlocking this potential is filled with challenges, the source of which lies in the unique nature of data science projects. Unlike traditional projects, these projects often start with a vague problem statement, where the solution is not predefined but discovered in the process. According to Saltz and Shamshurin (2016), data science projects frequently involve uncertainties in data, algorithms, and expected outcomes, making them more complex than conventional software development. Existing frameworks like Agile, while effective for software development, might not be directly applicable to data science (Gupta et al., 2018). This creates a void in terms of best practices. Thus, structured project management becomes not just beneficial but essential, providing the framework to navigate these uncertainties and ensuring that the project remains on course, meets its objectives, and delivers value.

Existing frameworks like Agile, effectively functional for software development, might not be directly applicable to data science. This creates a void in terms of best practices and structured approaches tailored for data science projects.

Transitioning Data Science Projects to Production

Machine Learning Operations (MLOps) is a discipline that fuses the world of machine learning and operations (Luz, 2023). The growth of the data science field has helped to expand its focus beyond just creating models to making them function seamlessly. MLOps is the reason for this transition, as it facilitates the end-to-end lifecycle of machine learning projects from development to deployment and monitoring.

MLOps closes the gap between the theory and the practice by ensuring the accuracy of machine learning models. In the context of this discussion, the concept of MLOps is explained in-depth to provide better insight into why MLOps is a significant part of the data science toolkit. According to Makinen et al. (2021), the knowledge of the foundational principle of MLOps is crucial to effectively navigating its landscape. The usefulness of MLOPs is in streamlining the lifecycle of machine learning models, which includes development, deployment, monitoring, and governance. MLOps aims to automate as many processes as possible, reducing human intervention and potential errors (Makinen et al., 2021). Another feature of MLOps is its seamless data handling ability. Unlike traditional software, ML depends heavily on data quality, volume, and consistency, which is why MLOps places significant emphasis on data management (Ruf et al., 2021).

Harnessing Innovative Strategies for Advanced Projects

In the data science field, effective adoption of innovative strategies is paramount for tackling advanced and complex projects. According to Resnikov & Svitiana (2024), for organizations to unlock new frontiers of insight, drive transformative solutions, and gain competitive advantage, it is essential to adopt cutting-edge approaches that will help in achieving seamless fusion of technical prowess. Below are relevant approaches that can be harnessed for advanced projects.

1. **Cloud**: In data science, the cloud offers unparalleled advantages. Traditional on-premise solutions are more likely to fail in terms of scalability, flexibility, and cost-effectiveness, especially when handling massive datasets (Sether, 2016). However, cloud storage solutions, like Amazon S3 or Google Cloud Storage, offer scalable, secure, and cost-effective data storage options, ensuring that data scientists always have the necessary data at their fingertips.
2. **Transparency:** Tools like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive Explanations) have been developed to shed light on model decisions. These tools, while model-agnostic, provide insights into how different features influence model predictions (Wang et al., 2024). Transparency in models also aids in debugging. If a model delivers skewed results, understanding the reason allows data scientists to trace back to potential data biases or model training issues, ensuring the model can be refined and improved (Morandini et al., 2023).
3. **Data version control and collaboration tool**: This integration ensures that experiments are reproducible and results are traceable to specific data and code versions. Collaboration tools, such as GitHub and GitLab, further enable team members to review each other's work, discuss changes, and ensure that best practices are followed consistently throughout the project's lifecycle (Beckman et al., 2020).
4. **Environment development and deployment tools:** Tools like Docker and Conda facilitate this consistency, allowing teams to create, share, and replicate environments with specific library versions and dependencies (Martin-Santana et al., 2018). Regarding deployment, platforms like Kubernetes and tools like MLflow aid in serving models, scaling based on demand, and monitoring performance. These tools not only streamline the deployment process but also facilitate the transition of models from development to production, bridging the often-challenging gap between these phases (Moiner & Gullen, 2020).

Overall, Grabis et al. (2019) posit that advanced techniques are not only about mastering complex algorithms; they also involve when and how to apply them effectively. The field of data science is dynamic, with new techniques and innovations emerging regularly. Embracing a culture of curiosity and adaptability ensures that data professionals remain at the pinnacle of their craft, ready to leverage the latest advancements for transformative results.

THEORETICAL FRAMEWORK IN DATA SCIENCE

Traditional project management, embedded in well-established practices and methodologies, has long been the backbone of industries ranging from construction and finance IT, and other industries (Ozsoy et al., 2024). Its structured frameworks, from Waterfall to Agile, have been regarded for their ability to deliver projects timely and affordably. However, how do these time-tested methods perform when introducing them to the dynamic and exploratory world of data science?

The Agile methodology, which focuses on collaboration and responsiveness, has its roots in software development. It emerged due to the need to adapt quickly to changes and deliver incremental value (Al-Saqqa et al., 2020). Additionally, CRISP-DM (Cross-Industry Standard Process for Data Mining) was developed specifically for data mining and later became a standard framework for broader data science projects (Martinez-Plumed et al., 2019). KDD (Knowledge Discovery in Databases) was one of the earliest systematic processes to outline the steps for extracting knowledge from large volumes of data (Maimon & Rokach, 2005).

Agile thrives on adaptability. It accepts the inevitability of change and encourages rapid response to such changes. *Agile* is broad and used in software development, product management, and even beyond tech industries (Hotz, 2024). *CRISP-DM* is more structured, laying a clear path from understanding business objectives to deploying models. *CRISP-DM* is more niche-focused and explicitly tailored for data-driven tasks. *CRISP-DM*'s general structure can be adapted for various data projects (Cazacu & Titan, 2020). However, *KDD* focuses on the tangible outcome of extracting knowledge from data, emphasizing the actual discovery process. KDD is particularly suited for projects that extract actionable insights from large datasets (Rahman et al., 2014). However, the hybrid framework has been adopted as the practical data science framework. This hybrid framework involves an integration of the best practices from multiple established frameworks that are customized for modern data science endeavors. While established frameworks like *CRISP-DM* and *KDD* are effective in specific scenarios, the world of data science is vast, and there is no one-size-fits-all. Drawing from traditional methodologies like Agile, the hybrid approach emphasizes iterative development, frequent communication, and adaptability.

The hybrid framework hinges on its emphasis on understanding the business problem, data preparation, and model evaluation from Data Science-specific models such as CRISP-DM. These stages are crucial as they form the basis of the project in its real-world context and ensure that the developed models are accurate and relevant. However, the real strength of the hybrid model lies in its flexibility (Hotz, 2024). In other words, the flexibility of the hybrid model allows project managers to adjust the phases based on the project's size, complexity, and objectives. For instance, for a smaller project with a tight deadline, managers might prioritize rapid prototyping and iterative feedback. On the other hand, for larger projects, project managers can put more emphasis on data exploration and model fine-tuning.

Research Aim & Questions

In order to explore the key aspects of data science projects and to identify successful practices in the field, critical attention is placed on seeking genuine answers to real-world problems that form the basis for this study. This paper explores the intricacies of handling data science projects and is able to provide solutions that are recent and accurate. Thus, the following research questions were formulated:

1. What are the best approaches and tools to effectively manage data science projects?
2. What are the major challenges faced in managing data science projects, and how can they be effectively mitigated?
3. How can the present and future data science projects be more effectively managed using recommendations from experts and case studies?

METHODOLOGY

Research methodology is a crucial aspect of any study as it provides a systematic approach to answering the research questions and the collection and analysis of the data (Igwenagu, 2016). Accordingly, a mixed research approach was adopted for this study to provide a comprehensive, valid, and reliable understanding of the research problem (Wasti et al., 2022). Thus, this research employed primary and secondary data gathered from industry experts. Also, an extensive literature review was carried out to investigate the effective measures to handle AI and data science projects, and the information gathered through primary data industry case studies also provided practical insight into the recommended practices.

Overview of Participants

Participants in the research survey include industry experts who have gathered experience working on data science projects in renowned companies across the world. In total, twenty-six (26) participants were approached by companies such as Apple, Walmart, LinkedIn, Amazon, Google, Microsoft, Dallas Rail Road, Wellsfargo Bank, Chase Bank, University of Toronto hospital Research Center, Toronto based Fintech, Blue Cross Blue Shield, Navis, Sales Force and Square. However, 6 of the participants who hold significant positions at Apple, such as the Tech Lead Software Engineer at Apple, Data Science Manager at Walmart, Senior Data Scientist at LinkedIn, Senior Data Scientist at Amazon, Engineering Manager at Amazon, and Project Manager at Google were administered questionnaires as provided in this section.

The statistics of the participants in Figure 1 below show that they have a wealth of experience in data science management. Data gathered shows that most of the participants in the research survey have years of experience, having worked on different successful projects in the field of data science.

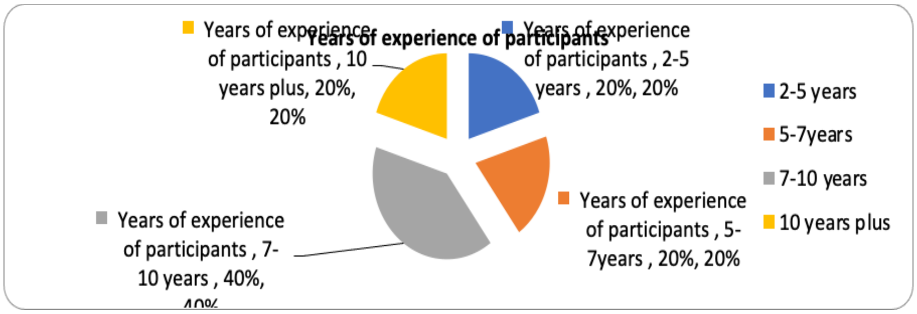

Figure 1: Showing the years of experience of the participants

Instrument

For this study, a questionnaire survey was designed such that the research participants were able to share their thoughts on specific questions on the effectiveness of managing AI and data science projects (Check Appendix, A). According to Kuphanga (2024), the adoption of a research questionnaire helps researchers to gather valuable insights, perspectives, and experiences from a larger sample of data science professionals and stakeholders. This approach helped to gather both qualitative and quantitative data from experts using both open-ended and closed-ended questions. Using a questionnaire provides anonymity and confidentiality to research participants, thus boosting their courage to provide candid responses (Kang & Hwang, 2023).

More so, the inclusion of the questionnaire enhances the representativeness and the quantitative rigor of the study.

Data Analysis Approach

Data analysis methods employed are inferential analysis for the quantitative data and thematic analysis for the qualitative data gathered through the research questionnaire survey. According to Singh and Jassi (2023), adopting inferential analysis helps the researcher draw conclusions and make inferences about the broader population based on sample data. More so, it helps to easily identify the relationships between variables and their impacts on the project outcome. For this project, the inferential statistics enhance the numerical representation of the research data and give a picture of the significant information at a glance. In addition, the use of thematic analysis helps deal with the open-ended questions of the survey. This approach helps researchers to identify, analyze, and report patterns or themes within qualitative data (Rasairo, 2023). Specifically, this approach helped to provide insight into the challenges of data science project management. More so, sample case studies are used to analyze the challenges further and recommend strategies for overcoming challenges within data science project management.

FINDINGS & DISCUSSION

The findings of this study are to provide a comprehensive overview of the intricate landscape of data science project management by unveiling the practical strategies, challenges and recommended best practices from stakeholders in the field. Using a mixed research approach, key insights emerge, offering a roadmap for practitioners to navigate their data science projects effectively.

Effective Method for Data Science Projects

The research findings provided in Figure 2 below indicate that experts who participated in the survey employ the Waterfall program method more than others like *Kumban, Agile,* and Scrum. The popularity of this approach is linked to the perception of better control and predictability following a structured and well-defined process (Senarath, 2021). However, Hotz (2024) claims that the rigidity and lack of flexibility of the method may not align with the iterative and dynamic nature of data science projects, which often require frequent adjustments and adaptation based on emerging insights and changing requirements. Following the Waterfall method, the Agile method is the next most adopted by the participants. This result aligns significantly with the growing trend in the industry toward adopting more flexible and adaptive approaches. According to Daraojimba et al (2024), the Agile methodology is more suited for projects with high levels of uncertainty, rapid change, and the need for frequent course corrections. Also, the research findings show that Kamban is averagely popular and adopted by industry experts.

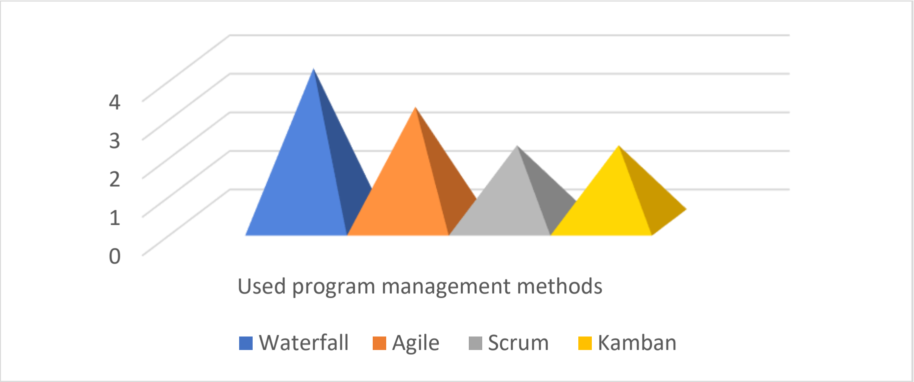

Figure 2: Showing the selection and preferences of industry experts on the most effective method

More so, the opinions of experts in the industry indicated that the Agile method is not only popular among experts in data science projects but is also affirmed to be the most reliable method. As indicated in Figure 3 below, there is incremental delivery, continuous collaboration, and adaptability to changing requirements making the Agile methodology more suitable for exploratory data analysis, model development, and iterative refinement. According to Omonije (2024), the effectiveness of the Agile method is attributed to how the approach facilitates rapid feedback loops, increases effective communication, and enhances the sharing of knowledge. Also, it is identified that the approach is flexible, making it possible for data science teams to pivot and adjust their approach for more effective results.

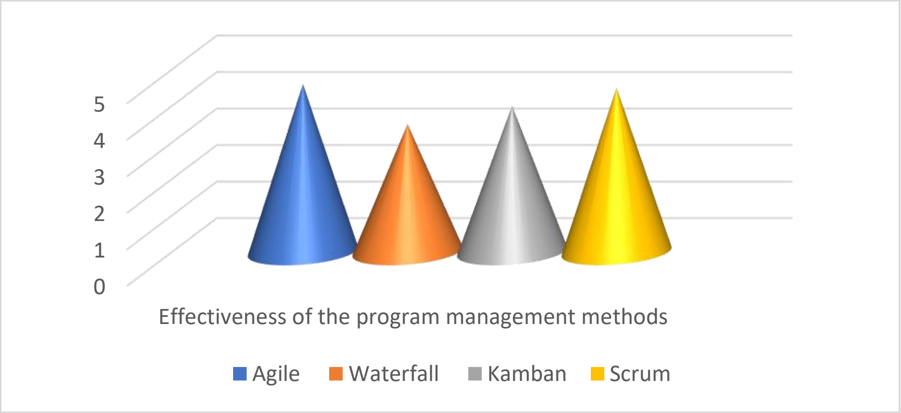

Figure 3: Effectiveness of the program management methods rated by experts.

Approaches to Data Science Project Management

The research findings in this aspect highlighted several best approaches to managing data science projects, as presented in Figure 4 below. Moreover, the survey's results are aligned with the existing literature on data science project management and other industry best practices. Significant emphasis is placed on establishing a well-defined data science workflow or methodology to ensure a consistent and reputable process throughout the project's lifetime. According to Martinez et al. (2021), adopting a clear and well-defined project methodology is vital for promoting collaboration and project tracking for effective results. More so, equally high emphasis is placed on the importance of establishing a well-defined, documented, and clear objective as it helps to entrench shared understanding and alignment among shareholders. Furthermore, experts support the idea that project resources, timelines, and budgets should be planned for, and that there must be appropriate monitoring to ensure that things go as planned. In addition, they also agree that the project team has to be first identified and engaged while testing the data science solution to be applied before project execution.

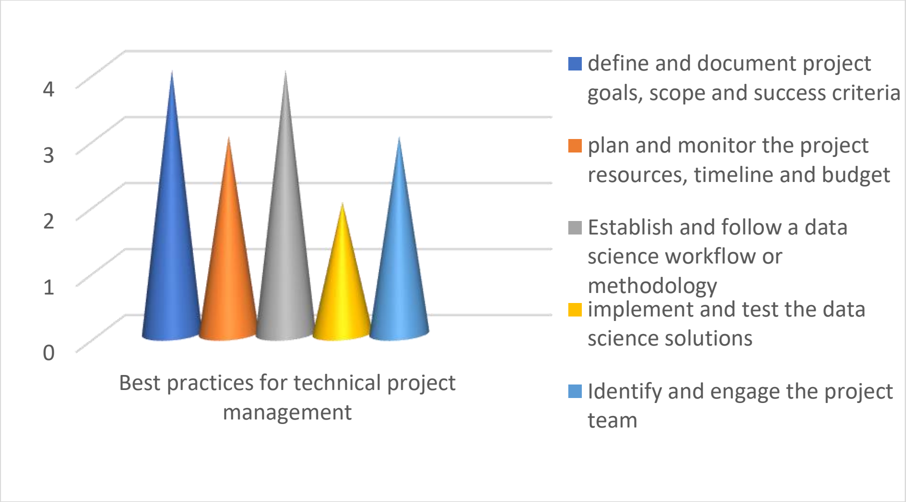

Figure 4: Showing experts' support on the best approaches for technical project management

Step-by-step Approach Using a Capstone Case study of Fictitious Clothing:

Experts' opinions and observations of the activities of the company are used to provide a nuanced understanding of the approach(es) used to manage the intricacies of data science projects. These step-by-step approaches, backed up by facts from the literature, are the strategies employed by Fictitious Clothing with proven effectiveness over the years.

Step 1: Ideation and Objectives:

Clear ideas and defined objectives are primal to the success of a data science project. Tracing it down to the history of the Fictitious e-commerce clothing platform, the problem of the company is first identified, and then a clear goal is set, showing that the company understands its goals. Ideation is not only about identifying the problem and solution, but it extends into knowing the possibility of such innovation in the business now and in the future. The ideation phase also demands in-depth market research by analyzing competitor platforms, user reviews, and the latest trends in e-commerce personalization (Runco & Jaeger, 2013). This stage has a significant impact on the chance of success of the project (Heising (2012).

Step 2: Data Collection, Cleaning, and Preparation:

Steps taken by Fictitious Clothing in this aspect include tapping into the existing user database, cleaning the data gathered to make them useable, integrating of external data for overcoming the cold start problem, where recommendations are needed for new users with no prior behavior on the platform (Bozic, 2024), ensuring data privacy, effective storage and management protocol, and integration of continuous data updating mechanism.

Step 3: Model Development, Validation, and Deployment:

This process starts with data splitting, which includes setting aside a portion of the data set for validation and testing to ensure that the model's real world, once trained, could be evaluated on unseen data. At this stage, data is validated through cross-validation, the model's hyperparameters are tuned to optimize its performance, and various metrics are used to evaluate the model's performance in terms of its accuracy, precision, and F1 score (Schulz et al., 2022). After successfully developing the model, the company moves to the next stage: deployment. This stage accounts for integrating the model into existing infrastructure, monitoring the deployed model, maintaining the system through periodic updates and retraining to maintain the effectiveness of the model (Hawkins, 2022).

Step 4: Monitoring and Continuous Improvement:

The Company was able to monitor the project through several key metrics that allow for immediate feedback on the model's performance in the real world. Also, actions such as the exploration of advancements in AI and machine learning help the company to continually improve its project activities.

Step 5: Feedback Loop and Iterative Enhancement:

The Fictitious e-commerce platform placed great emphasis on creating an efficient feedback mechanism. The primary source of feedback was the users themselves. The feedback loop and iterative enhancement process are the backbone of a successful AI-driven e-commerce platform because it constantly listens to users, acts on feedback, and makes regular improvements. By analyzing metrics and feedback, businesses can optimize their recommendation systems to enhance user experience and drive sales (Zhu & Lü, 2012)

Challenges and Considerations in Data Science Projects

Based on the results of the research survey, the challenges identified by industry experts when dealing with data science projects are highlighted in this section. As identified in Figure 5 below, the most challenging parts of data science projects pinpointed include the ability of data scientists to effectively manage complex data due to the high demand for expertise and experience. According to Somasundaram (2023), considering the diversity of data sources, varying data formats, and data quality issues that threaten data science project management, effective data management and integration are essential for ensuring the integrity and reliability of insights derived from data science projects. Thus, the expertise and experience required to navigate the data-related challenges of handling data science projects can be a significant hurdle.

Also, concerns were raised about the ability of the scientist to make appropriate decisions that will contribute positively to their projects. For instance, one of the industry experts states that "it may be challenging choosing the right technologies and tools for data science projects". As acclaimed by Mori & Uchihira (2019), data science projects require balancing trade-offs between model accuracy, interpretability, and operational constraints. Effective decision-making in this process requires a combination of technical expertise, business acumen, and an understanding of the project's broader objectives and stakeholder expectations.

Other challenges identified by experts include the burden of meeting customers' high expectations. Data science initiatives often hold the promise of transformative insights and competitive advantages, which can lead to unrealistic expectations or misaligned goals (Medeiros et al., 2020). Aside from this, data integration is another serious challenge for data scientists because this is an important process that directly influences the outcome of the project. Scientists need to be careful not to process the wrong data. As indicated in Nguyen et al. (2019), the choice of tools and technologies can significantly impact the project's efficiency, scalability, and overall success. The iterative nature of the data science project makes it challenging to estimate timelines and adhere to strict deadlines.

Finally, buy-in from stakeholders poses serious challenges for the successful management of data science projects. According to Ibraheem (2018), successful data science projects often require collaboration and buy-in from diverse stakeholders, including business leaders, subject matter experts, and end-users, so as to align objectives and easy embrace of change.

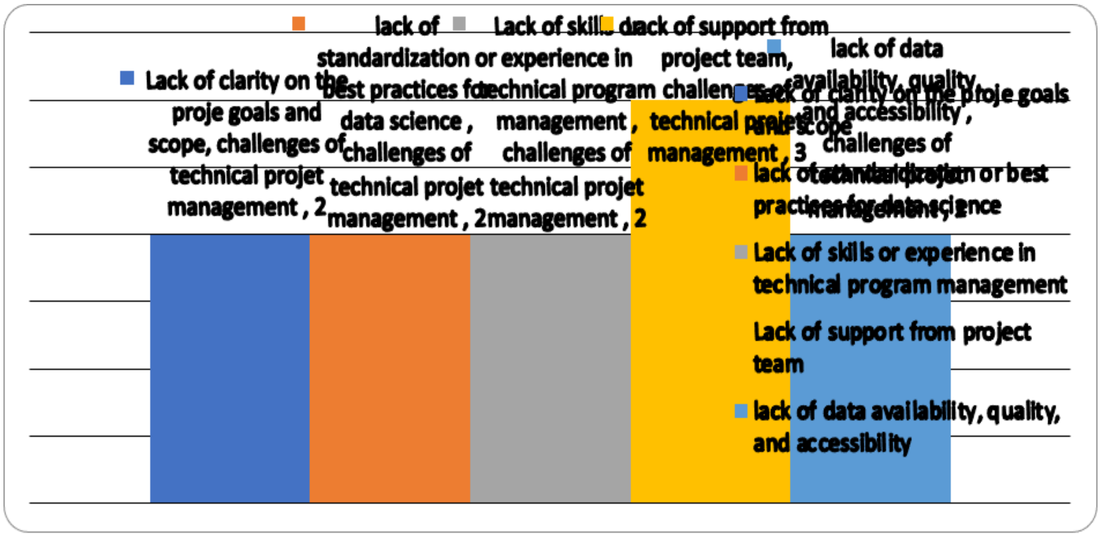

Figure 5: Showing project management challenges as identified by industry experts

Addressing Ethical Issues in Data Science Projects:

Advanced data science techniques bring forth ethical considerations regarding privacy, transparency, and bias mitigation. Implementing ethical frameworks and conducting regular reviews of models for bias are critical in ensuring ethical integrity in data science projects. Data scientists should strongly consider ethical practices so they do not fail in their contribution to society and, at the same time, not lose the trust of the people (Garaczarek & Steuer, 2019). Ethical AI is not just about adhering to rules but about ensuring that technology uplifts and benefits humanity without causing unintentional or intentional harm.

Bias:

Bias in AI has led to significant concerns. According to Zafar (2021), ethical AI should not be taken as optional; it should be strictly entrenched to ensure that algorithms and models are designed and implemented in a way that avoids bias and discrimination against individuals or groups. Models trained on biased data can perpetuate and amplify societal biases (Mensah, 2023). A practical instance of such may include a facial recognition system that has not been trained on diverse datasets, which might misidentify individuals from underrepresented groups, leading to potential misjudgments. As contributed by Weinberg (2022), prioritization of ethical consideration in data science projects is crucial to ensure that there is fairness in decision-making processes that rely on AI, such as lending, hiring, and criminal justice.

Privacy:

In the age of data, where AI-driven solutions permeate every aspect of human life , privacy concerns cannot be an afterthought. Regulations such as GDPR emphasize the importance of informed consent, where users are aware of how their data will be used and have the right to retract this consent at any point. Beyond consent, there's the aspect of data minimization. Just because a company can collect vast amounts of data does not mean they should. Ethical data practices involve collecting only data that are necessary for the relevant task, ensuring that individual privacy is not unnecessarily infringed upon. Mainly, data anonymization and pseudonymization techniques play a pivotal role in protecting individual privacy (Majeed & Lee, 2020). By ensuring that data, once collected, is stripped of personally identifiable information, companies can use the data for insights without directly tying it back to individuals.

Measures for building privacy

- Anonymization and pseudonymization
- Right over data delete
- Continuous monitoring and audits

Transparency:

The transparency of AI models further ties into ethics. Black-box models, while potentially more accurate, raise ethical concerns (Beltramin et al., 2022). If an AI model impacts an individual's life – be it a loan application, medical diagnosis, or job application – there is a moral imperative to explain that decision. But ethical AI is not just about the models; it extends to the entire lifecycle of the data science project.

Recommendations for Data Science Projects

The recommendations provided in this section are based on industry experts' opinions and the relevant case studies considered in the study.

Resolutions on Ethical Issues:

Following the case studies considered in this study, there is a profound insight into the complexities of ethical decision-making in data science projects. Technological tools, such as data anonymization softwares and AI fairness assessment platforms, can aid in maintaining ethical compliance. These tools are crucial in addressing privacy concerns and mitigating biases in AI models. These tools help protect privacy and mitigate biases such as GDPR and HIPAA. Also, privacy-preserving analytics tools and ethical AI frameworks provide directions for designing and deploying ethical AI systems. Compliance management platforms offer comprehensive solutions for managing regulatory requirements and promoting trust and accountability in data-driven decision-making (Lurie & Mark, 2015).

Building Ethical Team and Culture:

Building ethical teams and cultures is about ensuring that every team member, from data scientists to project managers, operates with an ethical mindset. An ethical culture starts at the top. Leadership teams need to prioritize ethics, ensuring that it is not just a bullet point in company values but a lived experience. This involves setting clear ethical guidelines, providing necessary training, and ensuring that there are consequences for breaches.

Measures for Building Ethical Team and Culture:

1. Collaboration is key for building ethical teams and cultures. Ethical considerations often are not black and white. Mao et al. (2019) emphasize that data scientists and engineers collaborate on different levels, such as model deployment and data preparation, and monitor and maintain any issues that arise in the project. The creation of a culture of collaboration is vital in ensuring that there is cross-functional learning and development. Through collaboration, team members share expertise and gain a better understanding of one another's roles and responsibilities (Cross et al., 2021).
2. Ethical considerations should be embedded in every stage of the project's lifecycle. From data collection to model deployment, teams should be trained to continuously ask about the ethical implications of their actions and decisions.
3. Diverse teams, with members from various backgrounds, cultures, and experiences, bring in a plethora of perspectives, ensuring that ethical considerations are not myopic.

Recommendations from Case Studies:

Case Study 1: A Toronto-Based Manager's Insight into Successful Remote Data Science Management:

The company of the manager, specializing in healthcare analytics, faced the abrupt transition to remote work amidst global disruptions. With a dedicated team of data scientists, engineers, and analysts spread across various locations, the following strategies were adopted to create a remote work environment:

- Adoption of Robust Communication Infrastructure such as Zoom for daily stand-ups and Slack for continuous communication.
- Leveraging Cloud-Based Collaboration using platforms such as Microsoft Azure
- Adopting an Agile methodology tailored to remote work, the team implemented structured sprints, clear deliverables, and regular retrospectives.
- The manager detailed the implementation of enhanced security protocols, including VPNs and multi-factor authentication, to protect data integrity and privacy.
- Navigating Time Zone Differences by establishing "overlap hours" where the entire team was available, coupled with flexible scheduling to accommodate personal work preferences.
- Maintaining Team Morale through flexible work policies and an open-door policy for mental health discussions.

Case Study 2: The Tech Lead Software Engineer at Apple: Building AI for User Experience:

Apple's ethos revolves around delivering seamless user experiences, and AI is no exception. Their Tech Lead, whom we interviewed, shed light on how they interweave AI into the user journey. Starting with identifying potential touchpoints, they gather user behavior data to train models that predict user preferences. Ensuring user privacy is paramount, so the data processing happens locally on devices. Continuous learning is crucial. As users interact more with their Apple devices, the models refine predictions, further enhancing user experience. However, managing such AI projects requires collaboration, clear goals, and iterations to ensure that the user always remains at the core of every decision.

Case Study 3: Data Science Manager at Walmart: AI for E-Commerce Recommendations:
E-commerce thrives on personalization, and Walmart's AI recommendation system is a testament to that. Our discussion with their Data Science Manager revealed the intricacies of creating a system that curates a personalized shopping experience. Walmart's vast data reservoir feeds its models, which then analyze purchase histories, browsing habits, and more to generate product suggestions. A key challenge is ensuring real-time responsiveness. Thus, models undergo rigorous stress tests. Moreover, the Agile approach facilitates regular feedback, allowing for continuous model enhancements. Managing such projects demands a balance between technical proficiency and understanding the nuances of online shopping behaviors.

Case Study 4: Senior Data Scientist at LinkedIn: AI Notifications Tailored to User:
With millions of users, LinkedIn's challenge lies in presenting relevant notifications. During our conversation with their Senior Data Scientist, it became evident that crafting such a personalized system requires a multi-faceted approach. By analyzing user activity, connections, interests, and job changes, LinkedIn's AI crafts notifications that resonate. However, avoiding notification fatigue is crucial. Therefore, models are trained to prioritize and space out alerts. Project management in this context involves rigorous model evaluations, A/B testing, and ensuring that users find value in every notification they receive.

Case Study 5: Senior Data Scientist at Amazon: AI's Role in Enhancing User Journey:
Amazon's e-commerce dominance is supported by its robust AI systems. Their Senior Data Scientist shared insights into the behind-the-scenes workings. From product recommendations to Alexa's responses, AI powers myriad touchpoints. The challenge is to make these interactions feel intuitive. User feedback loops are integral. They inform model tweaks, ensuring that the AI's predictions align with user expectations. Additionally, managing such expansive projects requires a holistic view, understanding where AI can truly augment the user experience, and ensuring that it remains a helpful assistant rather than a hindrance.

Case Study 6: Engineering Manager at Amazon: Orchestrating AI's Backend Operations:
The magic of AI often lies in its invisible backend operations. Amazon's Engineering Manager provided a peek into this realm. Every AI-powered feature on Amazon is backed by a robust infrastructure, ensuring its smooth operation. From managing server loads to ensuring minimal latency, the challenges are manifold. Projects often involve interdisciplinary teams, from data scientists crafting models to engineers ensuring their seamless deployment. Effective project management here requires an understanding of both AI's potential and the technical intricacies of making it work flawlessly in real time.

Case Study 7: Project Manager at Google: From Ideation to Deployment of AI Feature:
Google, a tech behemoth, integrates AI across its services. Their Project Manager shared the journey from ideating an AI feature to its deployment. It starts with identifying a user need. Once established, interdisciplinary teams collaborate, with data scientists designing models and engineers integrating them. Iterative testing is key, ensuring that the feature aligns with Google's standards. Moreover, post-deployment, feedback is continuously gathered and analyzed. Managing such projects requires a keen eye for detail, ensuring timelines are met and the final product resonates with users.

Case Study 8: Agile Project Management with Jira at Data Analytics:
Data Analytics, a mid-sized predictive analytics firm, faced challenges in managing multiple data science projects simultaneously. The implementation of Jira transformed their project management approach, enabling customizable workflows and real-time communication among team members. The result was a 40% improvement in project delivery times and a significant reduction in budget overruns.

Case Study 9: Environment Management with Docker and Kubernetes at Fintech:
A fintech startup leveraged Docker and Kubernetes to overcome their model deployment challenges. This strategic move enabled consistent environment management across development and production, reduced deployment times from weeks to hours, and facilitated automated scaling of services. The startup experienced enhanced reliability and efficiency in deploying machine-learning models.

Future Trends

1. AI experts have predicted the future invaluable features of AI or businesses. Although the influence of AI software in project management dates back to 1987, it is not until recently that it is notable. Experts are convinced that AI will be a distinctive accelerator and game changer for project managers and thereby help increase project success rates (Lahmann, 2018).
2. 3.5 Quintillion bytes of data are created every day, which is almost impossible to fathom. As data develops, so do the challenges of managing it and harnessing it to carry out business tasks. These tools enable data science teams to automate various aspects of their workflow, allow data science teams to deliver updates and improvements to their models more rapidly, and allow for monitoring and quality assurance (Green, 2023).
3. Augmented analytics combines AI and machine learning with user independence, allowing users to drive analytics without expert assistance. These platforms offer fast, deep insights, transforming business decisions (Jagreet, 2023).
4. Cloud-based solutions like Delta Lake technology make data more accessible and affordable, catering to businesses of all sizes. Businesses that invest in it now are more likely to be better equipped for the future, which will help them with a significant competitive advantage.
5. Maturation of Microsoft Fabric and Databricks Integration: Microsoft Fabric and Databricks’ profound framework integration will reshape the DaaS market by offering unified data platforms with enhanced security and governance.

6. Businesses' investments in AI and ML will help unlock better insights and drive data-driven decision-making, which will positively transform various industry sectors. Increased Investment in AI and Machine Learning Models.

CONCLUSION

The global nature of data science requires managers to consider diverse perspectives and challenges. Understanding cultural differences and regulatory environments and fostering international collaboration can enhance project outcomes and drive innovation. The case studies in this context illustrate how organizations navigate diverse global perspectives by implementing region-specific privacy measures, addressing cultural sensitivities, overcoming infrastructure challenges, and complying with industry regulations. The questionnaire's primary data also provided insight into practical issues that surround data science projects. However, throughout this paper, the importance of collaboration is emphasized, giving more prominence to the important potential contribution of human factors to the perfect and successful execution of data science projects. Collaboration should be across teams, and it should reflect the perfect incorporation of external feedback from users. In like manner, ethical consideration also appeared at the top of the requirement for fair integration of AI models. Emphasis is placed on the responsible deployment of AI models and genuine concern for users' data and privacy. Also, the findings cover the unending aspect of data science projects, emphasizing the need to monitor and continually improve the ever-changing landscape of the field of data science. However, the data science field is rapidly evolving, with new skills and roles emerging as technology advances. Preparing for the future involves investing in education, encouraging interdisciplinary collaboration, and promoting a culture of flexibility and innovation. By staying adaptable and forward-thinking, data science professionals can navigate the changes and continue to make significant contributions to the field.

References

Al-Saqqa, S., Sawalha, S., & Abdel-Nabi, H. (2020). Agile software development methodologies and trends. *Computer Science and IT Research Journal, 14*(11), 246-270.

Beckman, M., Cetinkaya, M., Horton, N. J., Rundel, C. W., Sullivan, A. J., & Tackett, M. (2020). Implementing version control with Git and GitHub as a learning objective in statistics and data science courses. *Journal of Statistics and Data Science Education, 29*(suppl. 1), 1–35. https://doi.org/10.1080/10691898.2020.1790417

Beltramin, D., Lamas, E. V., & Bousquet, C. (2022). Ethical issues in the utilization of black boxes for artificial intelligence in medicine. *Advances in Informatics, Management, and Technologies in Healthcare*, 249-252.

Bozic, V. (2024). The impact of artificial intelligence on business intelligence. *International Journal of Business Intelligence, 18*(2), 55-68.

Cazacu, M., & Titan, E. (2020). Adapting CRISP-DM for social sciences. *Broad Research in Artificial Intelligence and Neuroscience, 11*(2), 99-106.

Chinthamu, N., & Karukuri, M. (2023). Data science and applications. *Journal of Data Science and Intelligence Systems, 1*(2), 81–91.

Daraojimba, C., Nwasike, C. N., Adegbite, A. O., Ezeigweneme, C. A., & Gidiagba, J. O. (2024). A comprehensive review of Agile methodologies in project management. *Computer Science and IT Research Journal, 5*(1), 190-218.

Foote, K. D. (2021). A brief history of data science. *Data Governance*. Retrieved from https://www.dataversity.net

Garczarek, U., & Steuer, D. (2019). Approaching ethical guidelines for data scientists. *Journal of Data Ethics, 4*(1), 55-68.

Hawkins, J. (2022). *Getting data science done*. Business Expert Press.

Heising, W. (2012). The integration of ideation and project portfolio management – A key factor of sustainable success. *International Journal of Project Management, 30*(5), 582–595. https://doi.org/10.1016/j.ijproman.2011.11.004

Hotz, N. (2024). Agile data science. *Data Science Process Alliance*. Retrieved from https://www.datascience-pm.com/agile-data-science/

Ibraheem, I. F. (2018). The effects of stakeholder engagement and communication management on project success. *MATEC Web of Conferences, 25*(3), 1-12.

Igwenagu, C. (2016). *Fundamentals of research methodology and data collection*. Research Gate, 1–47.

Joel, A. (2022). Data science life cycle: Detailed explanation. *Data Science*. Addis Ababa University.

Kang, E., & Hwang, H. (2023). The importance of anonymity and confidentiality for conducting survey research. *Journal of Research and Publications Ethics, 4*(1), 33-45.

Kuphanga, D. (2024). Questionnaires in research: Their role, advantages, and main aspects. *Journal of Research Methodology, 12*(4), 1–8.

Lurie, Y., & Mark, S. (2015). Professional ethics of software engineers: An ethical framework. *Science and Engineering Ethics, 22*, 1245-1263. https://doi.org/10.1007/s11948-015-9665-x

Luz, A. (2023). *Machine learning operations*. University of Melbourne Press.

Maimon, O., & Rokach, L. (2005). Chapter 1: Introduction to knowledge discovery in databases. In *Data Mining and Knowledge Discovery Handbook* (pp. 1-17). Springer.

Majeed, A., & Lee, S. (2020). Anonymization techniques for privacy-preserving data publishing: A comprehensive survey. *IEEE Access, 4*, 1-25.

Makinen, S., Skogstrom, H., Laaksonen, E., & Mikkonen, T. (2021). Who needs MLOps: What data scientists seek to accomplish, and how can MLOps help? *1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, 55-62.

Mao, Y., Wang, D., Muller, M., & Varshney, K. R. (2019). How do data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right questions? *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1-25.

Martinez, I., Viles, E., & Olaizola, I. G. (2021). Data science methodologies: Current challenges and future approaches. *Big Data Research, 24*(3), 102-115.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N. J. A. H., Ramírez-Quintana, M. J., & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2019.2962680

Martin-Santana, S., Perez-Gonzalez, C. J., Colebrook, M., Roda-Gracia, J. L., & Gonzalez-yanes, P. (2018). Deploying a scalable data science environment using Docker. *Journal of Cloud Computing, 15*(4), 45-57.

Medeiros, M., Hoppen, N., & Macada, C. G. (2024). Data science or business: Benefits, challenges, and opportunities. *Data Science for Business*. Emerald Publishing Limited.

Mensah, C. B. (2023). Artificial intelligence and ethics: A comprehensive review of bias mitigation, transparency, and accountability in AI systems. *AI Ethics, 5*(1), 33-49.

Nguyen, G., Dlugolinsky, S., Bobak, M., Tran, V., Garcia, A. L., Heredia, I., Malik, P., & Hluchy, L. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey. *Artificial Intelligence Review, 52*(1), 77–124.

Omonije, A. (2024). Agile methodology: A comprehensive impact on modern business operations. *International Journal of Science and Research, 13*(2), 55-70.

Ruf, P., Madan, M., Reich, C., & Abdeslam, D. O. (2021). Demystifying MLOps and presenting a recipe for the selection of open-source tools. *Applied Science, 11*(19), 39.

Schulz, M., Neuhaus, U., Kaufmann, J., & Kuhnel, S. (2022). DASC-PM v1.1: A process model for data science projects. *NORDKADEMIE gAG Hochschule der Wirtschaft*.