DSBL Capstone

# Step 0 - Introduction. 100-day Data Science Plan: Build a Data Science Strategy

Upon assuming a new leadership role within a company (whether from an internal move or joining the company anew), it is common for an executive to be asked to prepare a plan for their first 100 days in the job.

As part of this project, you will build/create the following:

1. Identification of six data science opportunities for the organization
   1. Opportunities must be spread across three different functional areas
   2. Detail the risks, challenges, and key factors for success for each of these opportunities
2. Prepare a roadmap for executing these six data science opportunities.
   1. Rack and stack evaluation of these opportunities
3. Prepare a Human Capital plan for your data science organization
4. Prepare a Technical plan for your data science organization
   1. Data and Data Architecture Strategy
   2. Machine Learning Architecture

The work product for this Capstone project will be a detailed presentation to the CEO, detailing your plan and the rationale behind your decisions.

This project asks you to prepare that 100-day data science plan for a company of your choosing; this could be your current company or some other existing company.

**Name of Company Chosen:** TechSolutions Inc.

**Brief Company Description:** TechSolutions Inc. is an advanced technology company specializing in providing software solutions and IT services for small and medium-sized businesses. We focus on delivering innovative products that help customers improve business performance and enhance information security. With our team of experienced professionals, TechSolutions Inc. is committed to delivering high value and meeting all the technological needs of our clients.

# Step 1 - Identify Data Science Opportunities in the Business

Throughout the course, you have been exposed to multiple examples of data science projects implemented in a business setting. Now, based on your knowledge of your specific business context, you will generate six potential projects to be considered by the executive leadership team. These projects must span three unique functional areas of the business, with any one functional area representing no more than 3 projects:

Acceptable Project Mixes

\* 2 marketing + 2 supply chain + 2 finance

\* 2 marketing + 1 human resources + 1 procurement + 1 product + 1 manufacturing

\* 3 finance + 1 legal + 2 marketing

Unacceptable Project Mixes:

\* 3 marketing + 3 finance

\* 4 marketing + 1 product + 1 manufacturing

**Please identify your six projects here:**

**Project 1:** Customer Churn Prediction [marketing]

**Project 2:** Sales Forecasting [marketing]

**Project 3:** Inventory Optimization [Supply Chain]

**Project 4:** Supplier Performance Analysis [Supply Chain]

**Project 5:** Financial Fraud Detection [Finance]

**Project 6:** Credit Risk Assessment [Finance]

**Note: You may choose to represent this information on slide 5 of the CEO Presentation Template**

**For each candidate project, please provide the following detail:**

**Project 1 Name:** Customer Churn Prediction

**Business Functional Area:** Marketing

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

- Business Problem Addressed: Many customers are leaving the service without a clear reason, which negatively impacts revenue and increases the cost of acquiring new customers.

- Role of data science in addressing the business problem: Use machine learning models to predict which customers are likely to churn.

- Targeted business objective(s): Enhance customer retention strategies.

**2. Data Science Classification**

- Approach: Predictive analysis

- Type of Model: Logistic regression or decision tree model

**3. Data needed for project and sources for that data**

Customer behavior data, transaction data, customer feedback data from the CRM system.

**4. Magnitude of opportunity (with justification)**

Reducing the customer churn rate by 10% can increase revenue by 20%.

**5. Cost and complexity of development and implementation**

Average cost, medium complexity due to the need for integrating multiple data sources.

**6. Likelihood of value capture (Low/Medium/High) with justification**

High, because retaining existing customers is more cost-effective than acquiring new ones.

**7. Key Business Stakeholders**

Marketing department, customer service department, IT department.

**For each candidate project, please provide the following detail:**

**Project 2 Name:** Sales Forecasting

**Business Functional Area:** Marketing

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

- Business Problem Addressed: Difficulties in predicting sales revenue, affecting production planning and inventory management.

- Role of data science in addressing the business problem: Forecast future sales revenue based on historical data, thereby helping to plan sales strategies and manage inventory.

- Targeted business objective(s): Optimize sales planning, reduce inventory costs, increase revenue.

**2. Data Science Classification**

- Approach: Time forecasting

- Type of Model: ARIMA model or LSTM neural network

**3. Data needed for project and sources for that data**

Historical revenue data, marketing campaign data, seasonal data.

**4. Magnitude of opportunity (with justification)**

Increasing forecast accuracy by 15-20% can reduce inventory costs and enhance the effectiveness of marketing campaigns.

**5. Cost and complexity of development and implementation**

Average cost, high complexity due to the need for advanced machine learning models and multiple influencing factors.

**6. Likelihood of value capture (Low/Medium/High) with justification**

Medium to high, as accurate forecasting helps minimize unnecessary inventory and boosts revenue through effective marketing campaigns.

**7. Key Business Stakeholders**

Marketing department, sales department, inventory management department.

**For each candidate project, please provide the following detail:**

**Project 3 Name:**  Inventory Optimization

**Business Functional Area:** Supply Chain

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

- Business Problem Addressed: High storage costs due to inefficient inventory management.

- Role of data science in addressing the business problem: Use optimization models to manage inventory, reduce storage costs, and ensure the availability of necessary products.

- Targeted business objective(s): Reduce storage costs, increase the ability to meet customer demand, and decrease stockouts.

**2. Data Science Classification**

- Approach: Optimization

- Type of Model: Mixed-Integer Linear Programming (MILP) model

**3. Data needed for project and sources for that data**

Inventory data, order data, sales data, delivery time data.

**4. Magnitude of opportunity (with justification)**

Reduce inventory costs by up to 15-25% by optimizing warehouse management processes.

**5. Cost and complexity of development and implementation**

Average cost, high complexity due to the need to build optimization models and integrate multiple data sources.

**6. Likelihood of value capture (Low/Medium/High) with justification**

Medium, as inventory optimization not only helps reduce costs but also improves the ability to meet customer demand.

**7. Key Business Stakeholders**

Inventory management department, purchasing department, finance department.

**For each candidate project, please provide the following detail:**

**Project 4 Name:** Supplier Performance Analysis

**Business Functional Area:** Supply Chain

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

- Business Problem Addressed: Difficulties in evaluating supplier performance, affecting product quality and delivery time.

- Role of data science in addressing the business problem: Analyze supplier performance based on key performance indicators, aiding strategic partnership decisions and improving supply quality.

- Targeted business objective(s): Improve product quality, enhance supply chain efficiency, and reduce incidental costs.

**2. Data Science Classification**

- Approach: Descriptive analysis and predictive analysis

- Type of Model: Regression analysis, graph analysis

**3. Data needed for project and sources for that data**

Data from the ERP system, delivery data, supplier evaluation data.

**4. Magnitude of opportunity (with justification)**

Increase supply chain efficiency and productivity by up to 20% by utilizing more effective suppliers.

**5. Cost and complexity of development and implementation**

Low cost, medium complexity due to the need to focus on analyzing available data.

**6. Likelihood of value capture (Low/Medium/High) with justification**

High, as selecting better suppliers directly improves product quality and delivery time.

**7. Key Business Stakeholders**

Supply chain management department, purchasing department, finance department.

**For each candidate project, please provide the following detail:**

**Project 5 Name:** Financial Fraud Detection

**Business Functional Area:** Finance

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

- Business Problem Addressed: Financial fraud causes severe losses.

- Role of data science in addressing the business problem: Use machine learning to detect fraudulent transactions in the company's financial system, minimizing financial risk

- Targeted business objective(s): Minimize losses from fraud, protect company assets.

**2. Data Science Classification**

- Approach: Predictive analysis and anomaly detection

- Type of Model: Machine learning models such as decision trees, neural networks, and k-nearest neighbors (K-NN) algorithm.

**3. Data needed for project and sources for that data**

Financial data, transaction data, user behavior data.

**4. Magnitude of opportunity (with justification)**

Reduce fraud-related losses by up to 30-40%, protecting the company's reputation and assets.

**5. Cost and complexity of development and implementation**

High in cost and complexity due to stringent data security requirements and complex machine learning models.

**6. Likelihood of value capture (Low/Medium/High) with justification**

High, as early detection and prevention of fraud directly improve the company's profitability.

**7. Key Business Stakeholders**

Finance department, internal control department, IT department.

**For each candidate project, please provide the following detail:**

**Project 6 Name:** Credit Risk Assessment

**Business Functional Area:** Finance

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

- Business Problem Addressed: Difficulties in assessing customers' credit risk, leading to losses when customers are unable to repay.

- Role of data science in addressing the business problem: Assess customers' credit risk based on financial and consumer behavior data, enabling more effective credit risk management.- Targeted business objective(s): Minimize lending risk, enhance credit risk management.

**2. Data Science Classification**

- Approach: Predictive analysis

- Type of Model: Logistic regression, neural networks

**3. Data needed for project and sources for that data**

Credit data, customer behavior data, historical financial data.

**4. Magnitude of opportunity (with justification)**

Reduce the bad debt ratio by up to 20-30% through more accurate risk assessment.

**5. Cost and complexity of development and implementation**

Average cost, high complexity due to data requirements and complex machine learning models.

**6. Likelihood of value capture (Low/Medium/High) with justification**

High, as accurate credit risk assessment directly reduces financial risks and losses.

**7. Key Business Stakeholders**

Finance department, risk management department, credit department.

# Step 2 - Developing a Roadmap: Prioritizing Data Science Opportunities in the Business

A strategic approach to data science requires the business to consider the relative opportunities, costs, and risks of potential projects to identify the best order to carry out the projects. What should be tackled first? What is best pushed off until later? Completing the Data Science Roadmap requires stepping through key considerations to determine which project(s) should be considered ‘top priority’ and at what pace these and subsequent projects should be initiated.

**1. Complete this “Rack and Stack Exercise” worksheet to determine the relative strategic alignment, cost, complexity of implementation, certainty of value capture, and magnitude of benefit for each of the six projects**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Direct Alignment with Strategic Goals?** | **Cost** | **Complexity of Implementation** | **Certainty of Value Capture** | **Magnitude of Benefit** |
|  | 1=Low; 5=High | 1=High; 5=Low | 1=High; 5=Low | 1=Low; 5=High | 1=Small; 5=Large |
| **Project 1:**  **Customer Churn Prediction** | 5 | 3 | 3 | 5 | 5 |
| **Project 2:**  **Sales Forecasting** | 4 | 3 | 2 | 4 | 4 |
| **Project 3:**  **Inventory Optimization** | 4 | 3 | 4 | 4 | 4 |
| **Project 4:**  **Supplier Performance Analysis** | 3 | 4 | 3 | 4 | 3 |
| **Project 5:**  **Financial Fraud Detection** | 5 | 2 | 2 | 5 | 5 |
| **Project 6:**  **Credit Risk Assessment** | 5 | 3 | 2 | 5 | 5 |

**Note: You may choose to represent this information on slide 8 of the CEO Presentation Template**

**Please complete Step 2, Part 2, the Data Science Opportunity Matrix, using slide 1 of the CEO Presentation Template (You may or may not decide to include this slide as part of your CEO presentation)**

**Step 2, Part 3: Complete the table below by referencing the first four data science projects chosen for implementation. Include your justification for each project's order of implementation (e.g., how will the third project benefit from being implemented after the completion of the first two projects?)**

|  |  |  |
| --- | --- | --- |
| **Project Order** | **Project Title** | **Order Justification** |
| 1 | Financial Fraud Detection | This project has the highest strategic alignment, low cost, low implementation complexity, high certainty of value capture, and large magnitude of benefit. It will immediately reduce financial losses due to fraud. |
| 2 | Credit Risk Assessment | This project also has high strategic alignment, relatively low cost and complexity, and high certainty of value capture. It complements the Financial Fraud Detection project by further reducing financial risk. |
| 3 | Customer Churn Prediction | This project has high strategic alignment and a large magnitude of benefit, but has medium complexity and cost. It should follow once the more straightforward financial projects are underway. |
| 4 | Sales Forecasting | This project has high strategic alignment and a large magnitude of benefit. It is less complex and costly compared to Inventory Optimization, making it optimal to implement at this stage. |

**Note: You may choose to represent this information on slides 6 and 7 of the CEO Presentation Template**

# Step 3 - Establishing a Data Science Human Capital Strategy for your Data-driven Business

Now that we have established a roadmap for carrying out data science projects, our attention must turn to building and configuring the organization we will leverage to carry out this roadmap. The Data Science Human Capital Plan completed in this step will cover the organizational structure and talent configuration best suited to carry out the business’s roadmap, as well as the activities that the organization in particular -- and business more broadly -- must complete in order to promote a data-driven culture throughout the business.

**1. Identify the organizational model best suited for the data science organization that your business will need to deliver on the roadmap completed in Step 2. Provide justification for your selection based on the needs, scope, and timing of projects to be implemented in the Data Science Roadmap. If your organization should start with one model and evolve toward a different model, you may provide that detail and justification in your response.**

**Organizational Model:** Hybrid Organizational Model

**Justification:**

The hybrid organizational model combines the strengths of both the centralized and decentralized models, offering flexibility and alignment with organizational goals. This model is particularly suited to deliver on the roadmap outlined in Step 2 due to the following reasons:

**Scalability and Flexibility:**

* Scalability: As the organization scales, the hybrid model allows for easy adaptation and incorporation of new data science projects without disrupting the overall structure.
* Flexibility: This model provides flexibility to allocate resources dynamically based on project requirements, timelines, and strategic importance.

**Balancing Centralized Control and Decentralized Expertise:**

* Centralized Control: The core data science team can maintain standardized processes, tools, and methodologies, ensuring consistency in data analytics practices across projects.
* Decentralized Expertise: Embedding data scientists within business units allows for domain-specific expertise and closer collaboration with business stakeholders, ensuring that solutions are tailored to specific needs.

**Resource Optimization:**

* Core Team Efficiency: A central data science team focuses on foundational tasks like data governance, model development, and maintaining best practices, optimally utilizing expert resources for these crucial tasks.
* Distributed Efficiency: Business unit-specific data scientists can leverage their unique insights and domain knowledge, ensuring efficient problem-solving and faster implementation of solutions.

**Promoting a Data-Driven Culture:**

* Central Hub for Data Initiatives: A central data science team can serve as a hub for data initiatives, promoting a data-driven culture across the organization through standardized training and development programs.
* Localized Impact: Data scientists embedded within business units can champion data-driven decision-making, encouraging their colleagues to rely on data insights for strategic and operational decisions.

**Support for Roadmap Projects:**

* Financial Fraud Detection and Credit Risk Assessment: The central data science team can develop robust models for financial projects, ensuring consistency in methodologies and tools. Domain-aligned data scientists within finance and risk departments can provide valuable insights to refine these models and ensure their practical applicability.
* Customer Churn Prediction and Sales Forecasting: Centralized resources can provide the foundational support for data processing and algorithm development. Decentralized experts within marketing and sales teams can tailor these predictive models to align closely with customer behavior and market dynamics.

**Evolution Toward a More Mature Model**

Initially, the organization may start with a more centralized model to establish a strong foundation in data science practices, resource management, and governance policies. As the organization matures and employees gain confidence and experience with data-driven approaches, transitioning to a hybrid model will enable optimizing both centralized efficiencies and decentralized domain-specific insights.

**2. Complete the “Human Capital Plan” Worksheet for your data science organization.**

**- Identify the first ten professional roles for which you would recruit. How would you organize these roles into teams within the organization?**

For example, if you had 4 data scientists split evenly into two teams, your response would look like this:

|  |  |  |
| --- | --- | --- |
|  | **Position** | **Team** |
| 1 | Data Scientist | 1 |
| 2 | Data Scientist | 1 |
| 3 | Data Scientist | 2 |
| 4 | Data Scientist | 2 |

Identify your roles and teams below:

|  |  |  |
| --- | --- | --- |
|  | **Position** | **Team** |
| 1 | Chief Data Scientist | Leadership |
| 2 | Data Engineer | Technology |
| 3 | Data Engineer | Technology |
| 4 | Data Engineer | Project Team 1 |
| 5 | Data Engineer | Project Team 1 |
| 6 | Data Analyst | Project Team 1 |
| 7 | Data Scientist | Project Team 2 |
| 8 | Data Scientist | Project Team 2 |
| 9 | Data Analyst | Project Team 2 |
| 10 | Machine Learning Engineer | Innovation |

**Note: You may choose to represent this information on slide 9 of the CEO Presentation Template**

**Assume that leadership will allocate four new FTE’s for your data science organization during the current fiscal year. How would you prioritize your organizational buildout?**

|  |  |  |
| --- | --- | --- |
| **Order of Hire** | **Position** | **Justification** |
| 1 | Chief Data Scientist | A strategic leader is essential to provide vision, direction, and oversight for data science initiatives, ensuring alignment with business goals and driving project execution. |
| 2 | Data Engineer | Critical for establishing and maintaining the data infrastructure, ensuring data availability, quality, and integrity which are foundational for any data science work. |
| 3 | Data Scientist | Necessary for developing predictive models and analytics solutions, focusing first on high-impact projects like Financial Fraud Detection and Credit Risk Assessment. |
| 4 | Data Analyst | Supports the data scientist by handling data preparation, exploratory analysis, and visualization, enabling quicker and more efficient model development and insights generation. |

**Craft a “Data-Driven Transformation Strategy” by identifying six specific initiatives that you would recommend the data science organization and/or the business undertake in order to promote a data-driven culture across the business.**

|  |  |
| --- | --- |
|  | **Strategy** |
| 1 | Establish a Data Governance Framework |
| 2 | Implement Data Literacy Programs |
| 3 | Create Data and Analytics Centers of Excellence (CoE) |
| 4 | Develop a Self-Service Analytics Platform |
| 5 | Align Data Initiatives with Business Objectives |
| 6 | Promote Data-Driven Success Stories |

**Note: You may choose to represent this information on slide 10 of the CEO Presentation Template**

# Step 4 - Establishing the Technical Infrastructure to Support the Data Science Organization

With a completed Data Science Roadmap and a Human Capital Plan for executing the data science strategy, we turn our attention to the technological capabilities that must be built to support the new Data Science organization.

Complete the table on the next page by entering strategic aspects your business might consider to meet its Data and Data Architecture needs.

**Data and Data Architecture Strategy for the business**

|  |  |  |
| --- | --- | --- |
| **Component** | | **Strategy** |
| Data Requirements | What data should be included in the Data Strategy? | **Customer Data**: Demographics, behavioral data, transactional history.  **Financial Data:** Revenue and expenses, transactions, investment data. **Operational Data:** Supply chain, manufacturing, human resources data. **Market and Competitive Data:** Market trends, sales volumes, pricing models. **Product and Service Data:** Product specifications, service performance, R&D information. **External Data Sources:** Social media metrics, third-party data, regulatory information. **IoT and Sensor Data:** Machine performance, environmental data. |
| Data Governance | How will we promote data availability? (provide at least two ideas) | **Centralized Data Repository**: Implement a centralized data lake or data warehouse where data from various sources is aggregated and stored. **Data Catalog and Documentation:** Develop and maintain a comprehensive data catalog that includes metadata, data lineage, and documentation for each dataset. |
| How will we promote usability? (provide at least two ideas) | **Self-Service Analytics Tools:** Deploy user-friendly self-service analytics tools that enable employees to explore, visualize, and analyze data without needing advanced technical skills. **Standardized Data Formats and Clear Documentation:** Ensure that all data is stored in standardized formats and provide clear, comprehensive documentation on how to access and use the data, including guidelines and best practices. |
| How will we guarantee integrity? (provide at least two ideas) | **Data Validation Rules:** Implement stringent data validation and cleansing rules to ensure data quality before it is ingested into the centralized repository. **Audit Trails:** Maintain detailed audit trails that track all data transformations, movements, and usage. This includes logging who accessed the data, what changes were made, and when. |
| How will we guarantee security? (provide at least two ideas) | **Role-Based Access Control (RBAC):** Implement role-based access control mechanisms to ensure that only authorized users have access to sensitive data. Define roles based on job functions and grant permissions accordingly. **Data Encryption:** Use strong data encryption methods for both data at rest and data in transit. Ensure that encryption keys are managed securely and access to them is restricted. |
| Technology | Identify the components of your Data Architecture | **Data Sources:** The origin points of data, which can include databases, APIs, third-party data providers, IoT devices, web applications, and social media. **Data Ingestion:** The processes and tools used to collect, import, and integrate data from various sources into the data architecture. **Data Storage:** Repositories where data is stored for processing and analysis. This includes both raw and processed data. **Data Processing:** The transformation and preparation of data for analysis and reporting. This can include data cleaning, enrichment, merging, and aggregation. **Data Analytics**: Tools and platforms used to analyze data, generate insights, and support decision-making. This can involve data mining, machine learning, and statistical analysis. **Data Access and Consumption**: Interfaces through which users can access and interact with the data. **Data Governance:** Strategies and policies to manage data quality, security, privacy, and compliance. **Metadata Management:** Systems that manage metadata to ensure data is easily discoverable and usable. **Data Security**: Measures to protect data integrity, confidentiality, and availability. |
| Skills and Capacity | How will we promote development of data literacy skills and capacity throughout the organization (provide at least three ideas) | **Training and Workshops:** Organize regular training sessions and workshops that focus on data literacy, covering topics such as data analytics, data visualization, interpreting data, and using data tools. **Mentorship and Support Programs:** Establish mentorship programs where experienced data professionals mentor less-experienced staff. Provide continuous support through dedicated help desks or internal forums for data-related queries. **Access to Online Learning Platforms:** Provide access to online learning platforms like Coursera, edX, Udacity, and LinkedIn Learning, offering courses on data science, data analytics, and other related fields. |
| Support for Machine Learning | Give a brief description of the machine learning architecture and how it will interface with the data architecture | **Data Collection and Ingestion**: This component involves gathering raw data from various sources and ingesting it into a central repository for processing. **Data Processing and Preparation:** Involves cleaning, transforming, and normalizing data to make it suitable for training ML models. **Model Training:** The process of training machine learning models using historical data to predict future outcomes or classify data. **Model Validation and Testing:** Evaluates the performance of ML models using validation datasets and various performance metrics.  **Model Deployment**: Deploys trained models to production environments where they can be used for real-time predictions. **Model Monitoring and Maintenance:** Continuously monitors the performance of deployed models and retrains them as necessary to maintain accuracy over time. |

**Note: You may choose to represent this information on slide 11 of the CEO Presentation Template**

# Step 5 (OPTIONAL) - Record a short video of you presenting your final slide deck to your CEO or Executive Committee (5 minutes)

You may wish to submit a short video of you presenting your final presentation to your CEO; while this is not a formal requirement for the Capstone project, it does provide an outstanding way to gain practice with communicating about data science in business contexts.