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:** [Daraz.com ]

**Brief Company Description:** [Daraz is an online shopping platform that operates in various countries in South Asia and Southeast Asia. It offers a wide range of products including electronics, fashion, beauty, home appliances, and more. The platform provides a convenient and easy-to-use interface for customers to browse and purchase products, and also offers various payment options such as cash on delivery and online payment. Daraz also offers services such as same-day delivery, and often runs sales and promotions to attract customers

]

# 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:** [Recommendation System]

**Project 2:** [Customer segmentation using Gaussian Mixture Model]

**Project 3:** [Sales Prediction System]

**Project 4:** [Measuring Performances]

**Project 5:** [Predicting customer lifetime value for an online shopping business]

**Project 6:** [Fraud Deduction]

**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:** [Recommendation system]

**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: [In this digital age there are a lot of choices for someone if he want to shop online. So, there is a necessity to find out the most relevant product out of thousands of product available online which is the challenge for all online selling product business. There is a necessity of time to predict the behavior of customer and suggest him the product according to his taste and choice. So by making recommendation system for users it will help online selling business to forecast the behavior and make a special catalog for individual user which will enhance the growth of our business]

- Role of data science in addressing the business problem: [data science plays an important role to build the recommendation system for customer by preprocessing the customer purchase history, behavior and demographic information, we can make collaborative model. In Collaborative filtering the similar consumers are found on the internet and made a system to recommend the product to a specific customer, which is liked and bought by most of his similar customer. Other technique which is use to make a recommendation is content based filtering. By evaluating the attributes of every product similar products are found if someone is using one product he or she might also be interested to purchase its similar product. Data of the customer and product is most important entity which can be used to increase your sales by just suggesting relevant products to the customer. Recommendation models are trained on available data and predict the customer behavior. By visualizing sales data we can also evaluate the performance of our recommendation systems.]

- Targeted business objective(s): [By implementing recommendation system we can achieve following business objectives.

1. Increase in sales: recommendation system makes suggestion catalog for individuals, so the chances of sales increases. A company can generate higher revenue.

2. Improve customer engagement: Personalized catalog are built for individual customer so it can increase the interest of customer and make him a loyal to the company.

3. Improve efficiency: As by analyzing customer and products data an automated system is generated to recommend the product to best of their customer we can save a lot of sources used in making manual catalog and save a lot of money that would have spent on marketing.

4. Better inventory management: By determining the behavior of customer and popularity of product we can make more informed decisions about inventory of product and its distribution, which can assist to improve inventory levels and reduce waste.

]

**2. Data Science Classification**

- Approach: [In recommendation system data classification is organizing the data into different categories based on their attributes or characteristic. There are two main approaches to classify data which are following

1. Collaborative filtering: In collaborative filtering customer past purchase data is analyzed and similar customers are found with the help of this analysis and recommend the products to the similar users

2. Content based filtering: In this approach by analyzing the attributes and features of products, similar products are found and be recommended to users. The system generates recommendations by identifying the features of an item that customer has purchased in the past and recommend the similar products have more common features. ]

- Type of Model: [A recommendation system is characteristically a type of machine learning model. There are two common types of recommendation system

1. Collaborative filtering

2. Content based filtering

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

[The data needed for collaborative filtering comprises of the following:

1. User-Item Interactions: This contains information on how users have interacted with items i.e. ratings, views, clicks purchases, etc. This data will be gathered from the our site where our customers are placing order
2. User Characteristics: Information about the users, such as demographic information, interests, etc. as every customer will make his/her account on our site and we will ask some specific information like age, demographics, income so we can analyze that data
3. Rating or feedback data: data about the rating and feedback on the items they have purchased
4. Timestamps: Information about when the interactions happened can be used to weight latest interactions more greatly than older ones.

Data needed for content based filtering is Item Features. It comprises information about the items, i.e. product details, category, genre, keywords, description, etc.

]

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

[We can increase our sales by implementing the recommendation system in our business. A research told that recommendation engines market size will grow to USD 54 Billion in 2030 which is just USD 3 Billion dollar.

Generally, the deployment of a recommendation system can help a business to save on marketing costs by pursuing its marketing efforts more effectively, minimizing the need for customer acquirement, minimizing inventory costs, and minimizing costs of customer support.]

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

[The cost and complexity greatly depends upon some factors like available resources, complexity of system amount of data to be processed number of user for which we have to compute the recommendation system. For our company we are going to build in house recommendation system which will be costly as we need data scientists, processing devices and storage devices for data. We will need implementation of complex algorithm, maintenance and improvement of system continuously.

We are going to build medium level of recommendation system in start and will increase its complexity gradually.

So the cost of the development of recommendation system will be as follow.

Raw data of customer purchases and products attributes = free of cost

Development of minimum viable product = $ 110000

Deployment and release = $ 6000

Total cost = $ 17000]

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

[In start we will need a large amount of data of customer past purchases and we need to perform surveys to analyze the behavior of customers. So our recommendation system would be able to recommend the right product to certain degree and there will be a room for improvement which will be done with time and exact data and improvement after evaluation. With the passage of time we will be able to process data and build recommendation system which will recommend the right product to the customer with higher degree.

So the likelihood of value capture will gradually change from medium to high. ]

**7. Key Business Stakeholders**

[Key business stakeholders for recommendation systems are following.

1. Product owners are the stakeholder those make sure that recommendations which are made to the customers are relevant
2. Marketing teams which engage with customer trough recommendation and keep his/her interest in the company.
3. Data scientists are the stakeholders who build the recommendation system by processing purchase history data and keep improving system.
4. IT teams are also stakeholder for recommendation systems who make sure the integrity and deployment of recommendation system
5. Customer is the end user of recommendation system and most valuable stakeholder whose feedback is major source to improve the system
6. Executive are those who make decision for the company and provide required funds for the system

]

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

**Project 2 Name:** [Customer segmentation using Gaussian Mixture Model]

**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: [As population is growing rapidly and companies need to market their product to its customer. They make strategies to market their product but there is a problem there is problem they are marketing their product to everyone independent of the facts he/she is going to use or not that specific product. So companies are spending lot of resources to market their product now it is a time to know the exact customer domain or target audience for the product and there is need to optimize the marketing so that we can benefits ourselves a lot with minimum spending

So, in this project we are going to address the problem of finding right audience for the product and make decision by analyzing the behavior of the target audience. ]

- Role of data science in addressing the business problem: [Data science plays an important role to solve the segmentation problem by analyzing and evaluating large data of customers and find out specific features of customer which can help to classify customer into distinct segments. With customer segmentation we can optimize our marketing and make decisions according to the desires of each group.

Using data science visualization tools we can get the better and visual understanding of our customers segments. ]

- Targeted business objective(s): [There are following specific objectives of building a customer segmentation model.

1. Product improvement: You can improve your product quality in various ways but if you know the right segment of customer for your product you can understands the needs of your target customer and will improve the quality of your product according to the desire of your customer.

2. Price: If you have made the clusters of your customers according to their features you can easily determine that how much your target audience can pay for your product and if your customer can afford to pay more you can offer best quality and more features.

3. By computing the segments of customer you can optimize your marketing. You can be able to target the right audience and make promotion strategies according to the need of your right customer segment.

]

**2. Data Science Classification**

- Approach: [There are few classification approaches to compute the customer segmentation like clustering where we make clusters of customer according to the similarities of the customer and use K mean clustering. Others are using decision trees and neural networks. But we are going to use Gaussian Mixture Model (GMM) which uses probability techniques and utilize mean, and variance of data for clustering. GMM model use Expectation Maximization (EM) algorithm to determine these distribution parameters]

- Type of Model: [Gaussian Mixture Model is a probabilistic model which assumes that customer data is distributed according to Gaussian distribution and GMM model find the certain probability of belonging a customer to a certain cluster.

It is a soft clustering technique. ]

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

[We will need following types of data for making customer segmentation model.

1. Demographic data which includes customer age, gender, education level and income
2. Behavioral Data which include purchase history , reviews and feedback
3. And other data like lifestyles and personality characters.

The sources from where we can get that required data are following.

1. Surveys conducted to assess the characters and behaviors of customers.
2. Customer database systems provide customer data like purchase history.
3. Social media sites provide the customer interests and liking
4. Public available data like population and economics of your target customers.

]

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

[By implementing customer segmentation project you can optimize your marketing. You can run marketing campaigns for target and specific segment of customers. You can make better decision making to improve the quality of your product according to the needs of each group.

By computing segment of customer you can differentiate yourself from your competitors by choosing specific groups for selling your products. ]

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

[We need following to develop and implement the project

1. Customer surveys to collect data
2. Data scientists to develop the system
3. Software and tools to analyze and process the data
4. Once customer segmentation is done and need to launch a campaign for the target segment.

The cost of all to be done depends on the complexity of our project so we will start from spending few thousand dollars to develop this project and increase our budget for this project with passage of time ]

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

[We will get following benefits by implementing customer segmentation project.

1. Marketing efficiency.
2. Product quality improvement for targeted segment of customer.
3. Launch campaign and spend money in an optimized way.
4. Differentiate ourselves in the market by selecting the specific segment of customer.

Overall customer segmentation will provide efficiency in marketing, product quality determining customer behavior. The likelihood of value capture will be high for customer segmentation project.

]

**7. Key Business Stakeholders**

[Key Business stakeholders for customer segmentation projects are following

1. Marketing department who will launch campaign for right segment of customer.
2. Sales Department who will target that specific group of customer for selling your project
3. Product Development teams who will produce product according to the needs of specific group of customer
4. Data Scientist who will develop a system to compute customer segmentation
5. Surveys Teams who will conduct surveys and social media sites which will provide required data for customer segmentation
6. IT professional who will work for to handle software and tools for this project
7. Customer service department who will provide service for specific segment of customer
8. Executives Management who will use this customer segmentation information to make decision for company and allocate resources for this project.
9. Legal Teams who will be responsible for customer data privacy and security. ]

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

**Project 3 Name:** [Sales Prediction system using ARIMA]

**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: [In this era companies need to make decisions for future one cannot make decision without the knowledge how his company’s sales going to behave. You cannot set the optimized price for your product. There are may be two case first you manufactures more than the demand in future second is the inverse of first that you manufactures less than demand. In first situation you if you have manufactured more than demand you have made your product less worthy and your product will not be able to set a suitable price in the market. Second situation creates the problem of stock out and you will lose your customer loyalty towards your company both situation are critical and our company cannot afford any one of them.

To address above problem there is a need to build a model to predict the sales of your product in future. We are going to build a project to predict sales using Auto Regressive Integrated Moving Average (ARIMA) model]

- Role of data science in addressing the business problem: [Data science is used to address the problem of sales prediction problem in the following ways.

1- Data Preprocessing: we need to preprocess data like removing outlier, filling missing values of data and convert the data into a format on which we can implement ARIMA.

2- Data Science is used to visualize the graphs of sales and management can make better decisions by visualizing graphs of sales.

3. Fitting ARIMA model on the data

4. Evaluating model by comparing predictions with actual and finding accuracy using different measuring matrices like Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

]

- Targeted business objective(s): [Using Sales Prediction System we can achieve following objectives

1- Improves sales forecast accuracy: By using ARIMA model you can predict your future sales with more accuracy

2. Reduce cost: By predicting sales accurately you can reduce the cost of excessive inventory and can avoid stock out problem

3. Increase Revenue: A company can optimize the price of the product if we can predict sales and can launch different marketing strategies to improve the sales.

4. Inventory Management: With the help of sales predictions our company can manage the inventory level and reduce the holding costs of products.

5. Better Decision Making: Our management can make better decision with the knowledge of future predictions of sales]

**2. Data Science Classification**

- Approach: [In sales prediction model there are few different approaches one can adopt but we are going to use ARIMA model which will use Regression and Moving Average to predict the sales of our company]

- Type of Model: [Sales Prediction using ARIMA is a time series type of model which will use Auto Regression (AR) Integration (I) and Moving Average (MA) to find out the future trend and seasonality in sales of product]

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

[To build Sales Prediction Using ARIMA we will need customer sales data after every regular interval of time. To build ARIMA model we need time series data and it will be provided by the e commerce site which is collecting and maintaining the data of sales with time attribute. ]

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

[By predicting Sales we can avail the following opportunities

1. Inventory Management: We can manage our inventory levels by predicting the sales of our products.
2. Price Optimization: We can set the right price if we know the trends of sales in future and maximize our profit
3. Marketing: We can launch promotions if we know the sales are going down.

]

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

[To build sales prediction project we need the team of data science experts and tools to preprocess data and handle missing values. After that we need to implement ARIMA model

In start we will make a medium level of system and it will costs us few thousands of dollars]

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

[Likelihood of value capture of sales prediction system is very high the justifications are following

1. Improve forecasting accuracy with the help of ARIMA
2. We will be able to set right price for our product
3. We will be able to reduce cost by avoiding stock out problem and avoid excessive inventory

]

**7. Key Business Stakeholders**

[There are following stakeholders for our sale prediction system

1. Sales teams which will make strategies using the sales forecast
2. Marketing teams who will make marketing strategies using the forecast of sales
3. Operation Teams Who will use this system to manage inventory
4. Senior Management who will use these forecast for better decision making
5. Investors who will evaluate the performance of company with the help of these forecast and make investment decisions

]

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

**Project 4 Name:** [Measuring Performances]

**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: [there would be a great impact if we know our weak and strong performing departments, this would help us to not only improve the quality of our business but also reduce the dead weight throughout the different departments of the business and also awarding the more working employees which would lead to the boost of morale, in result increasing the efficiency and more opportunities for people in the business.

So by using decision tree we will build a model to predict the performances of employees and make decision to improve the quality]

- Role of data science in addressing the business problem: [Data of all employees containing the different features their performance and the like number of sales, work hours, team work behavior etc. Preprocess the data cleaning outlier and handling missing values.

Data science techniques are used to build a model to predict the performance of employees using decision tree.

Data science is used to evaluate the model and improve its performance.

]

- Targeted business objective(s): [By implementation of project we would be able to generate better revenue with effective cost reduction.

We will improve human resources by improving their performance

We will improve productivity.]

**2. Data Science Classification**

- Approach: [We will use decision tree approach to classify employees performance as “High performance” and “Low performance” by using the data of employee education, job tenure and previous performances. ]

- Type of Model: [We will use Decision tree model to build performance measure model]

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

[We will need the following data

1. Job tenure
2. Education level of employee
3. Age
4. Gender
5. Attendance (Work hours)
6. Productivity like number of sales if he/ she is in sales department
7. Employee feedback
8. Job role

And all above data be collected from HR department who is handling the employees]

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

[We can improve our decision making to do promotions in our company using this model.

We can improve efficiency of our employees by appreciating best performing employees

We can do great accountability of employees by using this model

]

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

[We can implement this project by using data science human resource which will collect data of employees we will need to build system in a way to store the performances of employees. So we can build this model using very minimal cost like very few thousand dollars]

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

[The likelihood of value capture is although low but it has a very effective use throughout the business and maybe used to maintain a better budget in the business and work effectively on making better business decisions]

**7. Key Business Stakeholders**

[There are following stakeholders for this project

1. Employees whose performances are going to be measured
2. HR department who will use this model to manage human resources in optimized way
3. Executives Management who will make decision by evaluating performances of employees]

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

**Project 5 Name:** [Predicting customer lifetime value for an online shopping business]

**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: [

Losing a high-value customer can have a significant impact on a business's bottom line. The loss of a high-value customer can result in a number of financial losses, including:

1-Direct revenue loss: The most obvious loss is the revenue generated by the customer's past purchases. If a high-value customer stops buying from a business, the revenue generated by that customer will be lost

2-Decreased future revenue: High-value customers tend to purchase more frequently and spend more money than low-value customers. Losing a high-value customer can result in a decrease in future revenue, as the business will miss out on the customer's future purchases

3-Increased marketing costs: A business may need to spend more on marketing to attract new customers to make up for the loss of a high-value customer. We are going to address above problem and make mechanism

]

- Role of data science in addressing the business problem: [Data science plays a crucial role in predicting the lifetime value (LTV) of customers for an online shopping business. By analyzing customer data, data scientists can identify patterns and trends that can be used to predict LTV. The following are some specific ways in which data science can be used to predict LTV:

1- Data Collection and Cleaning: Data scientists need to gather and clean the relevant data on customer demographics, purchase history, and other relevant information. This includes identifying and removing missing or duplicate data, and ensuring the data is in a format that can be easily analyzed

2- Predictive Modeling: Data scientists can use a variety of machine learning algorithms, such as linear regression, random forest, gradient boosting or neural networks to build models that predict LTV based on the collected data. These models can take into account factors such as customer demographics, purchase history, and web browsing behavior to make predictions.

3. Model Evaluation: Data scientists need to evaluate the performance of the model using metrics such as accuracy, precision, recall and F1-score to ensure it is accurate.

- Targeted business objective(s): [The targeted business objective of predicting the lifetime value (LTV) of a customer is to estimate the total revenue that a customer will generate over their lifetime with a company. This information can be used to make strategic business decisions such as allocating marketing resources, identifying high-value customers, and developing personalized retention strategies. To predict LTV, companies typically use historical data on customer behavior, such as purchase history, demographics, and engagement with marketing campaigns, to develop a predictive model. Machine learning techniques are often used to analyze this data and make LTV prediction]

**2. Data Science Classification**

- Approach: [We will be using Random forest model approach to predict LTV of customer]

- Type of Model: [Our project is supervised classification model which will predict the life time value of customer]

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

[To predict the lifetime value (LTV) of a customer, some of the data that might be needed for the project could include:

1. Demographic information
2. Purchase history
3. Engagement metrics
4. Customer behavior data

Sources for Data:

1. Customer surveys and questioners
2. Retail transactional data
3. Website analytic data social data
4. CRM Data

]

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

[The ability to accurately predict LTV can provide significant value for businesses by helping them to:

1. Identify high-value customers: By identifying which customers are likely to have the highest LTV, businesses can allocate resources to target and retain these customers, rather than wasting resources on lower-value customers
2. Optimize marketing and sales efforts: By understanding the drivers of LTV, businesses can develop more effective marketing and sales strategies to target and retain high-value customers
3. Improve customer retention: By identifying the factors that lead to customer churn, businesses can develop strategies to retain high-value customers and improve customer lifetime value.

]

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

[Overall, while the cost and complexity of developing and implementing a model to predict LTV can be significant, the potential benefits of such a model can justify the investment. It is important to weigh the potential benefits against the costs and complexity of the project before proceeding]

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

[The likelihood of value capture when using a model to predict the lifetime value (LTV) of a customer can vary depending on the specific context and industry, but in general, it can be classified as high because he likelihood of value capture when using a model to predict the lifetime value (LTV) of a customer can vary depending on the specific context and industry, but in general, it can be classified as]

**7. Key Business Stakeholders**

[Stakeholders are following

1. Business Owners or executive
2. Data Scientists
3. IT teams
4. Customers

]

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

**Project 6 Name:** [Fraud Deduction System]

**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: [There is a huge problem for online shopping company is that some people try to make some unauthorized transactions for example person A hack the credit card information of person B and make purchases fraudulently on his behalf due to these kind of transaction company reputation is damaged.

For the customer, unauthorized purchases can lead to unauthorized charges on their credit or debit card, which can be difficult and time-consuming to resolve. It can also put their personal information at risk if the thief uses it for other fraudulent activities.

So we are making system to detect the fraudulent transaction so that the above problem is addressed ]

- Role of data science in addressing the business problem: [Data science plays a crucial role in the development and implementation of a fraud detection system. The process of building a fraud detection system involves collecting and analyzing large amounts of data, identifying patterns and anomalies, and using this information to develop models that can detect fraudulent behavior.

1- Data collection and preprocessing: Data scientists collect and preprocess data from various sources such as transaction logs, customer information, and external data sources like social media and IP addresses. This data is then used to train and test the fraud detection models.

2- Feature engineering: Data scientists use feature engineering techniques to extract relevant information from the data and create new features that can be used to train the models. These features can include things like transaction amounts, customer demographics, and past transaction history.

3-Model development: Data scientists use a variety of machine learning algorithms and statistical techniques to develop models that can detect fraudulent behavior. These models are trained using the preprocessed data and features, and are then tested using a separate dataset to evaluate their performance.]

- Targeted business objective(s): [Targeted business objective for a fraud detection system in an online shopping company is to minimize financial losses due to fraudulent transactions. This includes reducing the number of chargebacks, which occur when a customer disputes a charge and requests a refund. Additionally, this objective could also include minimizing financial losses due to the costs associated with investigating and resolving fraudulent transactions.]

**2. Data Science Classification**

- Approach: [We are going to use Support Vector Machine Classification approach to build this fraud deduction system]

- Type of Model: [It is a supervised learning binary classification model]

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

[The data needed for a fraud detection project will depend on the specific problem the company is trying to solve and the features they want to use to train the model. Generally, the data needed for a fraud detection system includes:

1. Transaction data: This includes information about the transactions themselves, such as the transaction amount, date and time, merchant, and payment method. This data can be obtained from the company's own transaction logs.
2. Customer data: This includes information about the customer, such as their name, address, phone number, and email address. This data can be obtained from the company's customer database.
3. External data: This includes information from external sources such as IP addresses, social media profiles, and other publicly available data. This data can be obtained from various sources such as data brokers, social media platforms, and web scraping.
4. Labeled data: In order to train the model, it is important to have labeled data that indicate whether a transaction is fraudulent or not. This data can be obtained by manually reviewing a sample of transactions or by using a combination of manual review and machine learning algorithms to automatically label the data]

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

[The magnitude of opportunities for a fraud detection system in an online shopping company can be significant; as it can help the company minimizes financial losses, improve customer experience, and comply with relevant regulations.]

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

[The cost and complexity of developing and implementing a fraud detection system can vary depending on several factors, such as the size and complexity of the company's existing infrastructure, the amount of data that needs to be collected and processed, and the specific requirements of the system.

Complexity: The complexity of a fraud detection system can vary depending on the number of features and models needed, as well as the amount of data that needs to be collected and processed. For example, a system that uses a large number of features and models, or that needs to process a large amount of data, will be more complex and costly to develop and implement than a simpler system]

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

[The likelihood of value capture with a fraud detection system in an online shopping company is generally considered to be high. A well-implemented fraud detection system can help the company minimize financial losses due to fraudulent transactions, improve customer experience, and comply with relevant regulations

1. Minimizing financial losses: By detecting and preventing fraudulent transactions, a company can minimize financial losses due to chargebacks, penalties, and other costs associated with fraud. This can have a significant impact on the company's bottom line and can lead to significant cost savings.
2. Enhancing customer experience: A fraud detection system can help reduce the number of false declines, which occur when legitimate transactions are mistakenly flagged as fraudulent. This can improve the customer experience by reducing friction and frustration at checkout, leading to more loyal customers and repeat business.

]

**7. Key Business Stakeholders**

[1- Executive

2-IT teams

3. Data scientists

4. Customer Services

5. Risk Compliance

]

# 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:**  **[Recommendation System]** | [5] | [1] | [1] | [5] | [5] |
| **Project 2:**  **[Customer segmentation using Gaussian Mixture Model]** | [5] | [2] | [3] | [5] | [5] |
| **Project 3:**  **[Sales Prediction System]** | [5] | [3] | [2] | [5] | [5] |
| **Project 4:**  **[Measuring Performances]** | [4] | [3] | [3] | [4] | [4] |
| **Project 5:**  **[Predicting customer lifetime value]** | [4] | [3] | [3] | [4] | [4] |
| **Project 6:**  **[Fraud Deduction system]** | [4] | [4] | [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 | [Recommendation system] | [It is more important for online shopping company to implement its recommendation system] |
| 2 | [Customer Segmentation using GMM ] | [Optimized marketing after finding out target audience is very important for business today] |
| 3 | [Sales Prediction System using ARIMA] | [To meet demand and supply it important for a company to predict its sales] |
| 4 | [Fraud Deduction system using SVM] | [In today’s world its more important for online company to detect fraudulent transactions] |

**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 Model]

**Justification:**

[As we are working on 6 different project of data science so there will be 6 different teams following the embedded system reporting to their heads but heads of each team sharing the information with the chief data scientist. So there will be a central data team of heads and every head is leading his functional team. I this environment there will be cross functional collaboration. ]

**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 ] | [heads of teams ] |
| 2 | [Marketing Analyst ] | [1] |
| 3 | [HR Data Analyst ] | [2] |
| 4 | [Sales Analyst ] | [3] |
| 5 | [Sales Analyst] | [4] |
| 6 | [ Data Analyst ] | [5] |
| 7 | [Senior Data Scientist ] | [6] |
| 8 | [Machine Learning Engineer] | [4] |
| 9 | [Machine Learning Engineer] | [5] |
| 10 | [Machine Learning Engineer] | [6] |

**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 | [Business Intelligence Expert/ Business Analyst ] | [In Modern days we need a BI expert who can represent the data in graphs by analyzing the graphs one can understand quickly and make decision ] |
| 2 | [Data Engineer ] | [Data engineer will analyze which attributes of data should be pipelined to carry data together from different sources ] |
| 3 | [Machine Learning Engineer] | [In this digital age every consumer analyzes the reviews of other buyer of the product. So our ML expert will design a recommendation system by analyzing the reviews of consumer and will recommend and encourage other consumer ] |
| 4 | [Data Scientist] | [in online shopping there is a danger of fraud so we will hire a data scientist who will analyze the data of each buyer and will build a model to predict that the buyer is a fraud or genuine ] |

**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 | [Make a system to analyze the reviews and comments of consumer to make good decisions for the company ] |
| 2 | [Saving and optimizing marketing budget by just market the right customer by doing customer segmentation it could be only done by visualizing your sales data. ] |
| 3 | [Use modern Warehousing technique because in today world data is new electricity so we need to save and use the data in best way ] |
| 4 | [We have to implement the system to forecast our sales so we could make good decision in advance for that we need to analyze the marketing trends and our sales to forecast the sales ] |
| 5 | [We need to use data driven techniques to evaluate our human recourse we have to build a system where we can measure the performance of each employee on the bases of data ] |
| 6 | [It is most important to make your business fraud free for that we need to record and save every transactions and its customer data so our data scientists could be able to detect fraud in upcoming transaction ] |

**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? | [Your data strategy should include be a review process in your data gathering. So by reviewing the data you can improve your data managements. Asses your data after a regular interval of time and compare its results and with your competitors ] |
| Data Governance | How will we promote data availability? (provide at least two ideas) | [1. There should be multiple sources of data means you should always have a backup of your data if one source fails but your data availability must not stop  2. There should be an automate failover system if your one component fails to provide data other should automatically replace it ] |
| How will we promote usability? (provide at least two ideas) | [1. By making attractive and useful dashboard using BI tools  2. You should use the review and comment data to predict the behavior of your consumer] |
| How will we guarantee integrity? (provide at least two ideas) | [1. You should periodically authenticate the sources of data.  2. You should uniform the data format and applicability which is being reported to the organization ] |
| How will we guarantee security? (provide at least two ideas) | [1. You need to use encryption strategy to secure your data  2. You need to implement data loss prevention technique to stop the loss of your important data ] |
| Technology | Identify the components of your Data Architecture | [There should be proper pipelines of data to collect and process data in a beneficial way.  You can also use cloud storage to store and process your data] |
| Skills and Capacity | How will we promote development of data literacy skills and capacity throughout the organization (provide at least three ideas) | [1. You are supposed to motivate your teams to play with data and get insights and hidden information from the data.  2. you should promote the questions culture about data you need to encourage your team to asks question about data.  3. Provide environment and tools to do experiment with data and should arrange workshops and trainings to enhance your team skill ] |
| Support for Machine Learning | Give a brief description of the machine learning architecture and how it will interface with the data architecture | [Using machine learning on your data you can forecast the sales and behavior of your customer. You can generate missing data using machine learning technique you can classify fraud using machine learning techniques ] |

**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.