# **Capstone: Industry Level DS (Data Science) Adoption Strategy Development**

## **Problem Statement Background**

Today, almost all companies are trying to adopt DS solutions and techniques into their business programs. The adoption of DS offers certain unique capabilities and opportunities that impact both sides of the balance sheet, increase revenue and reduce costs. For some businesses like fleet management and OTT platforms, DS adoption is easy. However, it is difficult for some other businesses like manufacturing and healthcare.

Regardless of how difficult it is to implement DS projects, the benefits they offer outweigh the efforts one needs to put in. For instance, in manufacturing companies, workers are required to wear safety equipment. Now, it can be extremely labour intensive for security personnel to look at all the CCTV footage at once to keep a watch on the workers. However, a DS solution can do this quite easily. It can use a computer vision model to consume all the CCTV footage, identify employees who are not wearing safety equipment and highlight these employees to security personnel. The security personnel can then decide the action that needs to be taken.

Usually, the companies that are at an advanced stage in DS adoption have in-house data science teams, while other companies start by outsourcing projects to consultants.

## Problem Statement

OLIST is a Brazilian e-commerce giant similar to Amazon. The company is looking to leverage the power of analytics to improve its processes and improve its product offering. You need to step into the shoes of the DS transformation in charge. The leadership team at OLIST has identified some projects that they think will create value in the business.

Now, there are certain constraints in the company; the company has set up the data science team recently. The team is small, and resources are limited. Given such a situation, you have to choose the right roadmap for DS adoption based on the size, feasibility, complexity and value of the use case.

You are required to create a presentation that will demonstrate the road to DS adoption and the monetary benefits that will accompany the adoption. You may use the following points to create your presentation:

* Company objective
* Impact of the project on the company objective
* Feasibility check of the project
* Data and skill requirements for the project
* DS solution approach (Refer to the module on Data Strategy)
* Changes in the current process on account of adopting the solution
* Developing a proof of concept
* Suitable success metric
* Estimated monetary benefits of the project

**Some pointers before you begin the solution**

1. Apart from the prioritisation framework covered in the earlier session, you can refer to the Data Architecture and Data Strategy modules to come up with a road map.
2. Calculating the monetary benefits:   
     
   Quantifying the benefits of a DS solution will be a difficult task because the current process and metrics of the company are not known. Hence, it is not expected that you will be able to precisely calculate the monetary benefits of a solution. Instead, you are expected to come up with an approach to quantify the solution.   
     
   The monetary benefits of a DS solution are a critical deciding factor for the prioritisation of use cases in real life as well. When you perform this calculation, you will understand the current metrics and process of the company. This is when you can come up with an approach to estimate the benefits of a solution. And finally, you will be required to make judgments and certain assumptions (based on facts) to estimate the monetary benefits of the DS solution. After you execute your project, you can compare the monetary benefits you gained (actual case) from the project to the estimates you had made earlier.   
     
   To calculate the monetary benefits, you are free to assume any metric you like. However, there are two conditions you must always keep in mind. You must be able to calculate the number if you work for any real e-commerce company. For instance, if you assume that the customer is looking at four other e-commerce websites for the same product, this will be an invalid assumption. This is because it will be extremely difficult for your team to capture this data. So, you will need to change your approach to rely on data that can be easily available.  
   Hence, in the capstone project, the focus will be on the approach and the assumptions, and not on the actual number you calculate.
3. You are required to invest a significant amount of time in researching this topic, which every consultant normally does before creating a strategy presentation. This strategy presentation is the first step in the journey of working as a DS consultant. You will be required to present your strategic plan to the senior management to propose a DS adoption roadmap.

## Brazilian e-commerce company: OLIST

OLIST is an e-commerce company similar to Amazon; it is a marketplace where sellers and buyers come together. For the immediate short term, the company leadership has come up with the following four business goals:

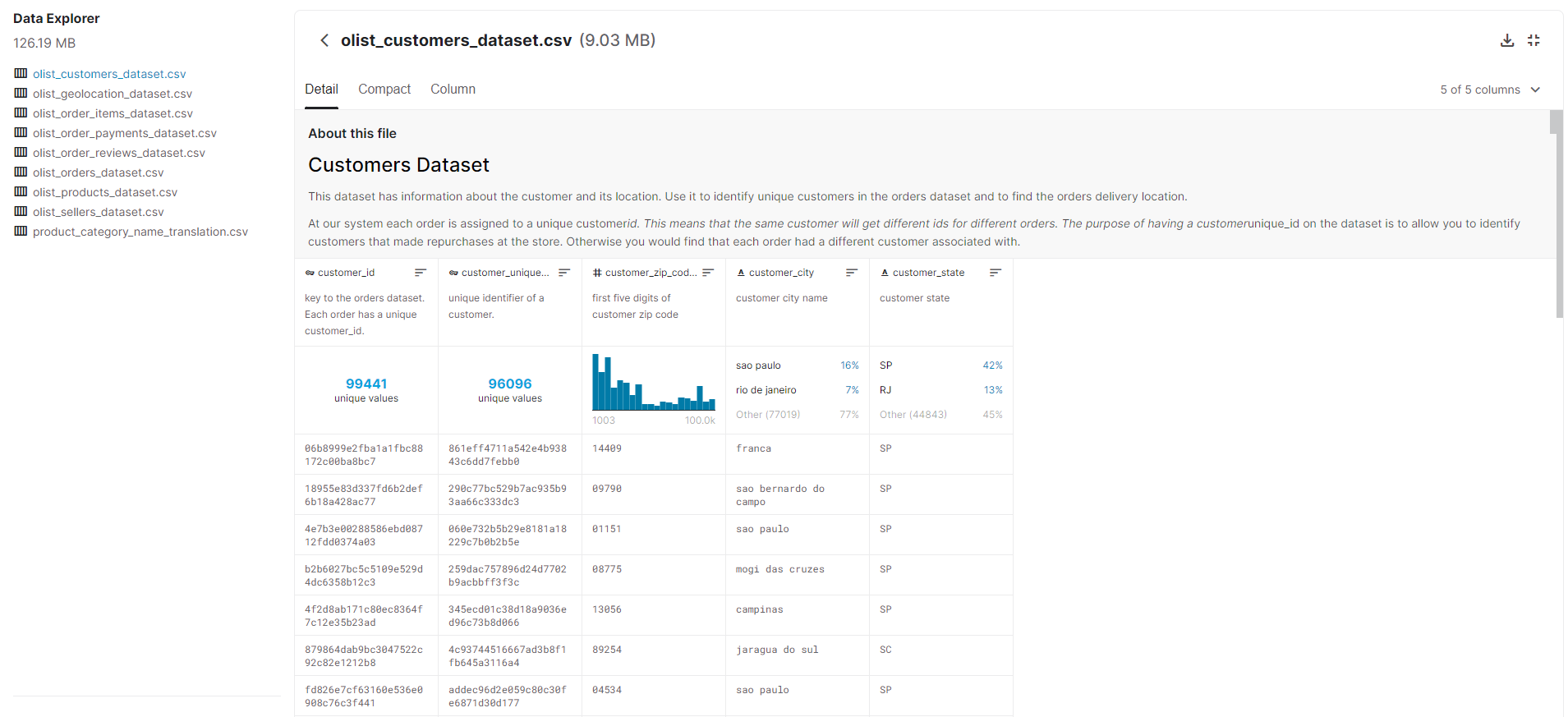
1. Increasing the number of active customers
2. Increasing revenue
3. Increasing the efficiency of services
4. Improving customer experience

You are expected to create proposals of DS/Analytics projects to help achieve company strategic goals. The submission will be in the form of a presentation explaining the strategy to execute the projects you have chosen.

**OLIST Data Set**

You can find the OLIST data set attached below. It contains the company data for approximately 100k orders from 2016 to 2017. The diagram given below shows the different tables included in the data set and how they are connected to each other.

You can visit the Kaggle page of the data set [here](https://www.kaggle.com/olistbr/brazilian-ecommerce). The Kaggle page will allow you to look at the overall summary of each table. As an example, a screenshot of the ‘OLIST\_customers\_dataset.csv’ table is given below.



You will have statistics like the total number of unique values in a given feature and the distribution of unique values. You can also view a few data points to get a sense of the data. This information will help you make assessments of the data.

You can also find the actual data set in the CSV format linked below. You can use this for further analysis if you are interested. However, it is not necessary to visit the data set. Most of the information needed to make your decisions is given in the data dictionary.

[Dataset link](https://drive.google.com/file/d/1n7i7ARFWodBHL25JP4TfiGE_XIGHCeNy/view?usp=sharing)

On the Kaggle page, you must have noticed that there are multiple files of data. To make sense of all the files in the data set, you can refer to the diagram given below.

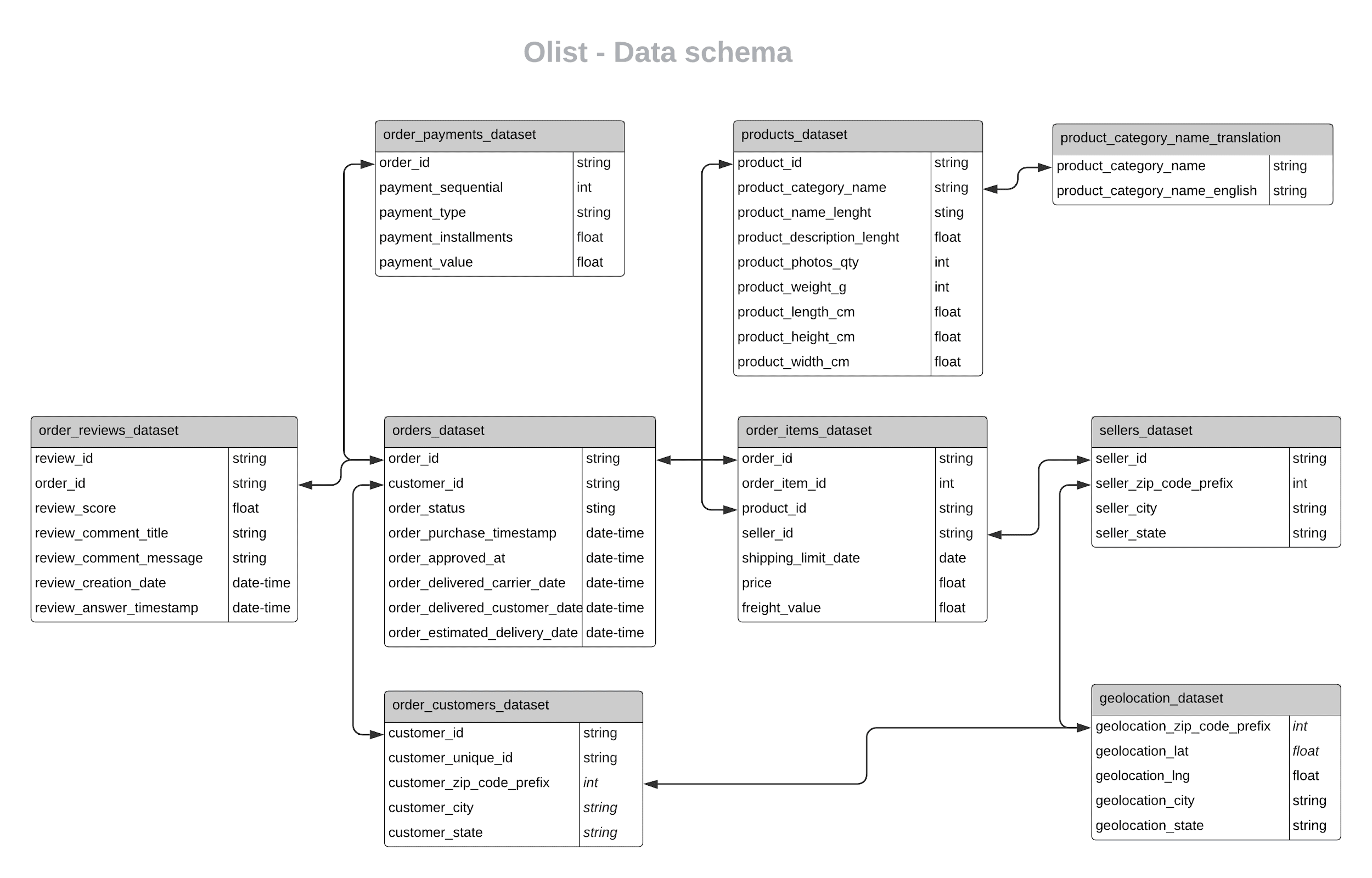
Each table contains certain attributes that are being recorded. For instance, the ‘sellers\_dataset’ table contains the following four attributes: seller\_id, seller\_zip\_code\_prefix, seller\_city and seller\_state.

Additionally, the seller\_data table is attached to the order\_items table set via seller\_id, and the geolocation\_data set via seller\_zip\_code\_prefix. You can use the seller\_id attribute to join order\_items and seller\_data and so on.

If you combine the information in the diagram given above with the information on the Kaggle page and the given data dictionary, you can answer a lot of questions about the data.

For instance, can you find the city where the highest proportion of customers of OLIST live?

To answer this question, you first need to find the table that contains the names of cities where customers live. The order\_customers\_dataset has the unique customer\_id and the cities where the customers live. You can look at the Kaggle table to get the distribution. From the Kaggle table, you can find that the highest number of OLIST customers live in Sao Paulo.



**Data Dictionary**

You can find the details of all the attributes in all the tables in the sheet linked [here](https://docs.google.com/spreadsheets/d/1s9buJ2ssz5HHU9uuXTIhebCQMltMQJgPn6PVNFJbc0Q/edit?usp=sharing).

**Identified Use Cases**

The leadership team at OLIST has come up with the following six use cases:

1. Delivery Date Prediction  
      
   The logistics team at OLIST uses heuristics to provide an estimated delivery date for the orders placed by the customers. It is very conservative about the delivery dates. As a result, the team is able to deliver the products much in advance. Although this is beneficial for the logistics team’s 'on time delivery' KPI, it is not favourable for the Chief Marketing Officer (CMO). The CMO has found that on average, the estimated time to deliver products that are given to customers is twice that of the actual delivery time. Such a high expected delivery time is driving away OLIST's customers. So, the CMO is looking to use ML to get a far more accurate expected delivery date.  
     
   Accurately predicting delivery dates has many benefits; the first and the most obvious one is customer delight. Receiving a delivery on the expected data not only gives customers happiness but also keeps the business competitive. Making accurate date predictions will also set the right expectation with the delivery team, to have tighter control over inventory management and last mile delivery.
2. Sentiment Analysis  
     
   The Chief Marketing Officer at OLIST wants to understand the experience of the customers based on the reviews received after the delivery of the orders. He also wants to identify the areas of improvement based on these reviews. He has heard that NLP can be used for sentiment analysis. However, he is also cognizant of the fact that the customer reviews are written in Portuguese, whereas NLP algorithms are not so sophisticated in Portuguese.  
     
   Customer sentiment is an intangible but important resource any business has. Like any other resource sentiment, it needs to be managed. There are multiple ways of performing sentiment analysis, but this use case relates to using customer reviews to identify the customers that are happy with OLIST and the ones that are not happy with it. Once the customers are identified, you can think of different ways to manage their sentiment or exploit it.
3. Customer Churn  
     
   Customer churn is a critical metric for the CMO of any e-commerce company. OLIST wants to develop customer churn models to identify 'at-risk’ customers so that an appropriate retention strategy can be built. This will provide insights into the factors driving customer churn, thus reinforcing its retention efforts.  
     
   Maintaining a large customer base is an important way of increasing revenue. However, as it happens in many businesses, customers tend to move between e-commerce companies. To prevent customers from constantly migrating, the company has built a churn model. The model is used to identify the customers who are likely to migrate. Now, the company wants to come up with a strategy to prevent churn.
4. Customer Acquisition Cost Optimisation  
     
   The Marketing team at OLIST runs multiple promotional campaigns to acquire new customers. However, the CFO believes that the marketing team is burning significant cash by offering deep discounts on products and other benefits, which is inflating the customer acquisition cost. The CFO wants to initiate a new process to measure the effectiveness of the acquisition campaigns by comparing them against the lifetime value of customers.  
     
   Another way of increasing revenue is to gain more customers. The money that a company spends on getting one customer is called the acquisition cost. For instance, suppose OLIST has to spend 30 BR to acquire one customer. In this case, 30 Brazilian Real (BR) is the acquisition cost of the customer. Obviously, it would be worth spending the 30 BR only if the customer generates more than 30 BR of lifetime revenue. So, the company wants to solve this optimisation problem.
5. Fraud Detection  
     
   Fraud is one of the most challenging areas to deal with in the e-commerce industry, as it can result in huge financial losses. There can be fraud in the areas of merchant identity, advanced fee, wire transfer scams, chargeback transactions, etc. The CFO wants to use the power of analytics to identify fraudulent transactions so as to help guard the organisation against such actions.  
     
   E-commerce marketplaces are a platform that brings together sellers and buyers. Any fraud that happens between independent sellers and buyers will harm the company’s image. The harm caused will have a direct impact on the revenue of the company.
6. Price Optimisation   
     
   Pricing is one of the most important aspects of business for an e-commerce organisation. It has a direct and profound impact on revenue, sales, profit and demand. Price optimisation is performed using a number of factors such as the location, the attitude of the customer, competitor’s pricing, etc. And, the data science algorithm predicts the customer’s segmentation to develop a response to the change in price. The OLIST sales team wants to build a price optimisation algorithm so as to maximise sales and revenue.  
     
   Similar to acquisition cost optimisation, price optimisation is also a balancing act. There are multiple factors that go into deciding the price of a product such that a customer is most likely to buy it. If the product is priced high, then the probability of selling the product is low but the profit generated is high. On the other hand, if the price is low, then the probability of selling the product is high but the profit generated is low. Moreover, the probability of selling a product is dependent on multiple factors such as customer segments and special occasions.

Evaluation Rubrics

| **Partition** | **Topic** | **Meets Expectations** | **Does Not Meet Expectations** |
| --- | --- | --- | --- |
| Part 1  (35%) | 1. Propose high-level solutions for the use cases (ML and non-ML) 2. Estimate the benefits of the proposed solutions over the current workflow (Process and economics) 3. Summarise the DS technique used 4. Define appropriate success metrics and KPIs | Multiple solutions are proposed, including non-ML solutions as well (wherever applicable).  The benefits of the proposed solutions are quantified appropriately. All the necessary assumptions are clearly stated.  The limitations and assumptions of the DS technique being used are explored.  Appropriate success metrics and KPIs are selected. | Obvious and straightforward solutions to problems are missing.  Various streams of process and economic impact are not thought about.  Any potential risk in the DS solution that can have an impact on adoption is not thought about.  Selected KPIs are not the right measure of success of the project. |
| Part 2  (40%) | 1. Prioritise use cases strategically. (Factors to be considered:  Feasibility check, Complexity rating, Strategic alignment, and  Business value) 2. Define a project road map | Prioritised use cases are selected based on the framework.  The road map entails the sequence in which the use cases need to be picked up. | Factors that might have an impact on prioritisation are not considered.  The proposed road map is not in the best interest of the company. |
| Part 3  (5%) | 1. Cite adoption examples from other markets 2. Conclusion and references | There are sufficient examples from other domains similar to the project at hand.  The document contains a well-written conclusion, and all references are cited appropriately. | Parallel projects exist in the industry but are not considered.  References are not mentioned, and plagiarism is found. |
| Part 4  (20%) | 1. Create a presentation | All the factors affecting the prioritisation of use cases are communicated clearly through the presentation. | The final presentation is lengthy and contains irrelevant insights. |

Submission Guidelines

**List of documents to be submitted**

You need to submit two files as a part of the solution:

1. Create strategy presentations for the DS adoption roadmap in a PDF format.   
   Please remember that the presentation is supposed to be made to a leadership team, so make it crisp and to the point. You can use [this](https://docs.google.com/presentation/d/1LCc6E_PfwO3Q8ucKj9kekxnw_URvezyoQR63uK-jayo/edit?usp=sharing) template for reference.
2. Create word documents for solutions of each use case. You can use [this](https://docs.google.com/document/d/1YIyuM7fx6EatQs_w-dXKLH_jz3g6w4kk31NSkdAHSP4/edit) template for reference.
3. Create an Excel file for prioritising use cases. You can use [this](https://docs.google.com/spreadsheets/d/1EGoVubzdetsv8YhrKeenSq_uaBSubtsgZm2WydL1U1I/edit#gid=1473855288) template for reference.   
     
   Please ensure the following:
4. Your document is well structured.
5. All references/citations used in the document are mentioned in the Appendix section.
6. There is no plagiarism in your document.

Best of luck!