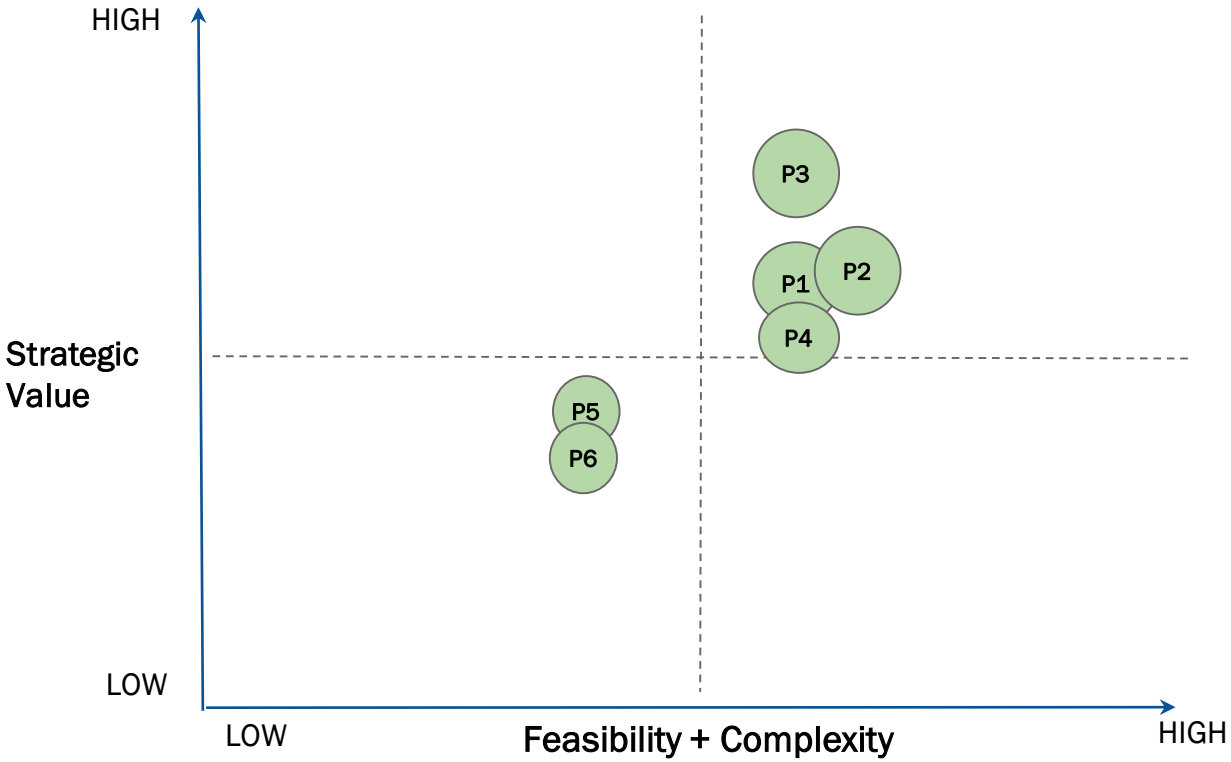
OLIST

Data Science Adoption Strategy

Candidate Data Science Projects

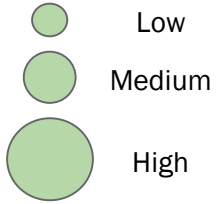
	Functional Area	Project Description
Project 1:Delivery Date Prediction	Delivery	The logistics team at Olist uses heuristics to provide an estimated delivery date for the orders placed. It is very conservative about the delivery dates. As a result, it is able to deliver the products much in advance. Although this is beneficial for the 'on time delivery' KPI of the logistics team, it is not favorable for the CMO. He found that on average, the estimated time to deliver products that is 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.
Project 2: Sentiment Analysis	Customer service	The Chief Marketing Officer at Olist wanted to understand the experience of the customers based on the reviews received after the delivery of the orders. He also wanted to identify the areas of improvement based on these reviews. He had heard that NLP can be used for sentiment analysis and topic modeling, which will be useful in finding topics in customer reviews. However, he was also cognizant of the fact the customer reviews are in Portuguese, whereas the NLP algorithms are not so sophisticated in Portuguese.
Project 3: Customer Churn	Marketing	Customer churn is a critical metric for a CMO at an e-commerce company. Olist wants to develop customer churn models to identify 'at-risk' customers so that appropriate retention strategies can be built. This will provide insights into the factors driving customer churn, thus reinforcing its retention efforts.
Project 4: Customer Acquisition Cost Optimisation	Cost Optimisation/Sales/Marketing	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.
Project 5: Fraud Detection	Fraud prevention	Fraud is one the most challenging areas to deal with in an e-commerce industry, as it can result in huge financial losses. There can be fraud in the areas of merchant identity, advanced fee, and wire transfer scams, chargeback fraud, etc. The CFO wants to use the power of analytics to identify fraudulent transactions so as to help guard the organization against such actions.
Project 6: Price Optimisation	Sales/marketing	Pricing is one of the most important piece of business for an e-commerce organization. It has a direct and profound impact on revenue, sales, profit and demand. Price optimization 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 make a response to the change of price. OLISTs sales team wants to build a price optimization algorithm so as to maximize the sales and revenue.

Complete the Data Science Opportunity Matrix below by modeling each of the six projects in terms of feasibility, complexity, strategic and business value impact.



Project 1: Delivery Date Prediction
Project 2: Sentiment Analysis
Project 3: Customer Churn
Project 4: Customer Acquisition Cost Optimisation
Project 5: Fraud Detection
Project 6: Price Optimisation

Business Value



Highest-Priority Data Science Projects

Order	Project	Data feasibility	Infrastructure feasibility	Complexity	Strategic Value	Business Value
		1=Low; 5=High	1=High; 5=Low	1=High; 5=Low	1=Low; 5=High	1=Small; 5=Large
First	Project 3: Customer Churn	3.6	4	3	4	3
Second	Project 1: Delivery Date Prediction	4.2	4	4	3	4

Complete the “Data Science Road Map” below with the first four data science projects chosen for implementation.

Order	Project	Order Justification
1	**Project 3:** Customer Churn	The proposed solutions to develop customer churn models include both ML and non-ML solutions. ML-based solutions can provide accurate and automated predictions, while non-ML solutions can provide cost-effective and interpretable solutions. Both approaches can help companies develop targeted retention strategies for at-risk customers and gain insights into the factors driving customer churn.
2	**Project 1:** Delivery Date Prediction	The estimation of accurate delivery dates is a regression problem to be solved. You use various data to estimate the time needed for delivery, then add the time to order date to get the right delivery date. Delivery date prediction is also a kind of balancing act between competitiveness and accuracy. You can always have long delivery dates and always be accurate, but might lose on sale to some competitor who can deliver quickly
3	**Project 2:** Sentiment Analysis	To perform sentiment analysis on Portuguese customer reviews, we can use either ML-based or non-ML-based approaches. The ML-based approach involves training and fine-tuning a machine learning model on a labelled dataset, while the non-ML approach is based on rule-based techniques.
4	**Project 6:** Price Optimisation	solutions can help optimize pricing and maximize sales and revenue. The ML solution can provide real-time optimization and adapt to changing market conditions, while the non-ML solution can be effective in consistent pricing strategies.

Technical Infrastructure Needed to Support the Data Science Organization

Data Requirements	What data should be included in the Data Strategy?	Business data: This includes data on customers, products, sales, revenue, expenses, and other business-related metrics. Operational data: This includes data on internal processes and systems, such as inventory management, supply chain, and logistics. Financial data: This includes data on budgets, financial statements, and other financial metrics. Marketing data: This includes data on advertising campaigns, website traffic, social media engagement, and other marketing-related metrics. External data: This includes data from external sources, such as government data, industry benchmarks, and market research.
Data Governance	Data Availability	• Data Replication: Ensure that data is replicated across multiple locations to ensure availability in case of system failures or disasters. • Data Recovery: Establish procedures for data backup and recovery in case of system failures, disasters, or other disruptions • Data Resiliency: Ensure that data systems and infrastructure are designed to withstand outages and maintain data availability.
	Usability	• Metadata Management: Ensure that data is properly labeled and described with metadata to make it easier to understand and use. • Data Dictionary: Develop a data dictionary that defines data elements and their attributes, providing a common understanding of the data across the organization. • Data Catalog: Develop a data catalog that provides a searchable inventory of available data assets, including information on their purpose, format, and usage.
	Integrity	• Data Access: Define who has access to what data and under what conditions. This includes establishing appropriate access controls and permissions for each user or role. • Data Security: Ensure that the data is secure and protected against unauthorized access or theft. This includes implementing appropriate security measures such as encryption, firewalls, and intrusion detection systems.
Skills and Capacity	Data literacy skills and organizational capacity	• Training and education: Provide training and education programs to help employees develop data literacy skills, including data analysis, data visualization, and data communication. • Role-specific training: Provide role-specific training for employees who use data in their work, such as data analysts or business intelligence professionals. • Data documentation: Ensure that data is properly documented with metadata, data dictionaries, and data catalogs to make it easier for users to understand and interpret the data. • Data communication: Develop effective data communication techniques to make it easier to share and communicate data insights with others.
Support for Machine Learning	Machine learning	• Required proper skill sets of DS professionals.. Licences of tool is required. Make sure we have right tool to execute the ML solution, regression models are ok but NN, RNN, NLP etc.. Might require proper tool and environment execute the model algorithm. • Ensure that the data used to train machine learning algorithms is accurate, complete, and reliable. This includes establishing data quality standards and monitoring data quality over time. • Develop processes to label data that is used to train machine learning algorithms. This includes creating a set of ground-truth labels that can be used to evaluate the accuracy of the machine learning models.