OLIST – Data Strategy

Data Strategy Candidates Evaluation

Revision 1.01

Date: 22nd November 2022

Authors

Jagdish Junur

Prayag Sanjay

© Copyright 2022 OLIST

**Proprietary Data**

This guide is the property of OLIST , Sao Palo, Brazil

No part of this document may be reproduced by any means whatsoever, or communicated to any third party without the prior permission of OLIST.

OLIST believes that the information contained in this document is accurate and reliable and much care has been taken in its preparation. No responsibility, however, financial or otherwise, can be accepted for any of the consequences arising out of the use of this material, including loss of profit, indirect special or consequential damages.

It is the Customer's responsibility to ensure that the use of the software described does not breach the laws, rules and regulations of the jurisdiction with respect to which it is used.

OLIST has a policy of continual review and improvement of their products and hence the information contained within this guide is subject to change. Reviews may be issued from time to time to advise of changes and/or additions.

**Contents**

[1. Introduction 5](#_Toc120126665)

[2. Delivery date prediction 6](#_Toc120126666)

[2.1. Problem statement: 6](#_Toc120126667)

[2.2. Proposed Solutions: 6](#_Toc120126668)

[2.3. Benefits of Proposed Solution 7](#_Toc120126669)

[2.3.1. Process improvements – 7](#_Toc120126670)

[2.3.2. Monetary benefits – 7](#_Toc120126671)

[2.4. Summary of the DS approach 7](#_Toc120126672)

[2.5. Conclusion 8](#_Toc120126673)

[2.6. Project Roadmap 8](#_Toc120126674)

[2.7. Success Metrics 8](#_Toc120126675)

[2.8. Prioritising Use Case 8](#_Toc120126676)

[2.9. Adoption Examples from Similar Domain 9](#_Toc120126677)

[3. Sentiment Analysis 10](#_Toc120126678)

[3.1. Problem statement: 10](#_Toc120126679)

[3.2. Proposed Solutions: 10](#_Toc120126680)

[3.3. Benefits of Proposed Solution 11](#_Toc120126681)

[3.3.1. Process Benefit 11](#_Toc120126682)

[3.3.2. Monetary benefits. 11](#_Toc120126683)

[3.4. Summarize the Solution 11](#_Toc120126684)

[3.5. Project Roadmap 12](#_Toc120126685)

[3.6. Limitations 12](#_Toc120126686)

[3.7. Success Metrics 12](#_Toc120126687)

[3.8. Prioritising Use Case 12](#_Toc120126688)

[3.9. Adoption Examples from Similar Domain 13](#_Toc120126689)

[4. Customer Churn 14](#_Toc120126690)

[4.1. Problem statement: 14](#_Toc120126691)

[4.2. Proposed Solutions: 14](#_Toc120126692)

[4.3. Benefits of Proposed Solution 15](#_Toc120126693)

[4.3.1. Process improvements 15](#_Toc120126694)

[4.3.2. Monetary benefits 15](#_Toc120126695)

[4.4. Summarise the DS approach 15](#_Toc120126696)

[4.5. Project Roadmap 15](#_Toc120126697)

[4.6. Limitation 16](#_Toc120126698)

[4.7. Prioritising Use Case 16](#_Toc120126699)

[4.8. Adoption Examples from Similar Domain 16](#_Toc120126700)

[5. Customer Acquisition Cost Optimisation 17](#_Toc120126701)

[5.1. Problem statement: 17](#_Toc120126702)

[5.2. Proposed Solutions: 17](#_Toc120126703)

[5.3. Benefits of Proposed Solution 18](#_Toc120126704)

[5.3.1. Process improvements – 18](#_Toc120126705)

[5.3.2. Monetary benefits – 18](#_Toc120126706)

[5.4. Summarise the Solution 19](#_Toc120126707)

[5.5. Project Roadmap 19](#_Toc120126708)

[5.6. Limitations 20](#_Toc120126709)

[5.7. Prioritising Use Case 20](#_Toc120126710)

[5.8. Adoption Examples from Similar Domain 20](#_Toc120126711)

[6. Fraud Detection 21](#_Toc120126712)

[6.1. Problem statement: 21](#_Toc120126713)

[6.2. Proposed Solutions: 21](#_Toc120126714)

[6.3. Benefits of Proposed Solution 22](#_Toc120126715)

[6.3.1. Process improvements – 22](#_Toc120126716)

[6.3.2. Monetary benefits - 22](#_Toc120126717)

[6.4. Summarize the DS approach 23](#_Toc120126718)

[6.5. Project Roadmap 23](#_Toc120126719)

[6.6. Limitation 23](#_Toc120126720)

[6.7. Success Metrics 23](#_Toc120126721)

[6.8. Prioritising Use Case 23](#_Toc120126722)

[6.9. Adoption Examples from Similar Domain 24](#_Toc120126723)

[7. Price Optimisation 25](#_Toc120126724)

[7.1. Problem statement: 25](#_Toc120126725)

[7.2. Proposed Solutions: 25](#_Toc120126726)

[7.3. Benefits of Proposed Solution 26](#_Toc120126727)

[7.3.1. Process improvements - 26](#_Toc120126728)

[7.3.2. Monetary benefits - 27](#_Toc120126729)

[7.4. Summarise the DS approach 28](#_Toc120126730)

[7.5. Project Roadmap 28](#_Toc120126731)

[7.6. Limitation 28](#_Toc120126732)

[7.7. Success Metrics 29](#_Toc120126733)

[7.8. Prioritising Use Case 29](#_Toc120126734)

[7.9. Adoption Examples from Similar Domain 29](#_Toc120126735)

[8. Appendix 30](#_Toc120126736)

# Introduction

As part of OLIST larger initiative to use data science, this document examines various candidate projects where data science could be leveraged to meet company’s strategic goals. The document describes the problem statement, give both machine and non-machine learning solutions. It goes into detail of machine learning solution, it benefits, limitation and roadmap.

The company think tank, management and other stake holders have come up with following six candidates

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

This document will delve into each of above use cases in detail.

# Delivery date prediction

## Problem statement:

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 logistics team’s 'on time delivery' KPI, it is not favourable for the CMO. He 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.

## Proposed Solutions:

1. ML solutions –
   1. Use one regression model using features like destination location and source location, and historic delivery date, item size, proximity to distribution hubs, seller, etc to predict the estimated time of delivery. Add the delivery time to the order date and calculate the estimated delivery.
   2. Divide the entire delivery process into individual stages of transport.
      1. Estimate time for getting the goods from the vendor.
      2. Estimate the time that the item will stay in the warehouse.
      3. Estimate the time required to send the item from the warehouse to the delivery location.  
           
         Build individual models for estimating these times, add the time to get the overall delivery time and add it to the order time to get the estimated delivery date.   
         The advantage of building separate models is more fine control over the processes and better prediction.
2. Non - ML solutions
   1. A rule-based approach to predicting delivery times. There are set times for each step of the delivery process such as set times for getting goods from the warehouse to the transport hub, from the transport hub to the next transport hub etc.
   2. Each of these times is mapped to the nearest shipping cutoffs wherever appropriate. Fixed additional processing times are added for weekends and holidays. Thus, the rule-based model doesn’t adapt based on recent performance changes and is designed based on heuristics.
3. Selecting First ML solution - because the data points for individual steps are not present.

## Benefits of Proposed Solution

### Process improvements –

1. Predicting delivery dates more accurately will not impact any process in the e-commerce setup.
2. Reduction in Inventory costs.

### Monetary benefits –

1. The monetary benefits of accurate delivery can be divided into two streams
   1. Reducing customer churn because of accurate delivery date prediction. Hence, increasing the revenue of the company.
      1. Reduction in the number of people who drop out - 10% (Assumed)
      2. Number of customers daily - 272.44

(total number of orders in 1 year = 99442 / number of days in a year 365)

* + 1. Average order value - 100 Brazilian real (Assumed)
    2. Revenue increased by = 0.1 \* 272.44 \* 100 = 2724.4 per day.

1. Inventory management
   1. Predicting the delivery date correctly will reduce the pressure on the supply chain. The through out rate of deliveries of the entire system increases but predicting the delivery date correctly. The warehouses will have to hold lesser inventory.
      1. Reduction of time items stay in warehouse = 40% (assumed)
      2. In the same time frame warehouse will be able to hold more items = 40%

(assumed)

* + 1. Warehouse cost per item delivered will reduce = 25% (assumed)
    2. Initial warehouse cost per item delivered = 5 BR (assumed)
    3. Total savings per day = 272.44 \* 0.75 \* 5 BR = 1021.65 BR

## Summary of the DS approach

1. 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.
2. Delivery date prediction is also a kind of balancing act between competitiveness and accuracy.
3. You can always have long delivery dates and always be accurate but might lose on sale to some competitor who can deliver quickly.
4. Or you can have extremely short delivery time promises and disturb the customer sentiment. Hence add a buffer

[There is a constant trade-off between being accurate and being competitive and, of course, we would aim to optimize both.](https://towardsdatascience.com/delivery-date-estimation-5aff1a0ff8dc)

## Conclusion

This project presents a unique problem of unavailable data for features needed in the solution. Generating a non-ML solution is counterproductive. The DS team will be approaching the project using a purely ML solution. The expected cost saving to the company’s exchequer is approx R$3740/-.

## Project Roadmap

1. Goals and initiatives – The aim of the project is to achieve efficiency in delivery date which in turn would increase revenue and customer experience.
2. Releases and Deliverables – this project will have the following releases and deliverables: -
   1. Proof of Concept (PoC).
      1. Timeline – one month
      2. Deliverables – process setup and preliminary design, review, mock-up and trials.
   2. Beta release – Beta trials in staging area
      1. Timeline – three months
      2. Deliverables – Testing of project in staging area, vulnerability assessment , performance testing, stress testing, bug removal
   3. Final release – In production area
      1. Timeline – one month
      2. Deliverables – deployment to production area, performance monitoring, drift monitoring
3. RMSE might be high
4. High complexity model - non linearity , thus non linear model might be needed - is the team capable or not.
5. Think of edge cases - data for special cases is not available

## Success Metrics

1. Early delivery rates
2. Late delivery rates
3. Churn rate decrease after optimized delivery date

## Prioritising Use Case

Refer to the use case prioritisation framework placed at the excel sheet.

## Adoption Examples from Similar Domain

1. Online Fashion Commerce: Modelling Customer Promise Date

<https://sdm-dsre.github.io/pdf/online_fashion_commerce.pdf>

1. Estimating the Time, Cost, and Deliverables of an ML App Project

<https://appinventiv.com/blog/machine-learning-app-project-estimate/>

# Sentiment Analysis

## Problem statement:

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.

## Proposed Solutions:

The company could hire a Portuguese language expert to categorise user reviews in order to understand client sentiments because this challenge needs sentiment analysis. For greater comprehension, the expert can offer a ranking of the various parameters based on the review comment message input feature. Both ML and non-ML methods can use this attribute as datapoints.

1. ML Solution
   1. Use logistic regression on customer reviews of products, coupled with price, destination, review time delay, language expert ranking of categories, and real delivery time. The time interval between ordering and actual delivery can be used to compute the delivery time. The discrepancy between the review date and the actual delivery date can be used to calculate the review time delay.
   2. Use logistics regression and Decision Tree classifier to build a sentiment analysis model with following features.
      1. Product price
      2. Destination – use geolocation of cities data
      3. Product origin – use geolocation data. This would provide insight on the acceptability of products from different locations.
      4. Review time delay – calculate the time difference between the product ordered date and time and actual delivery date and time.
      5. Frequency of products ordered – higher frequency means fast-moving and other frequencies can be accordingly binned for use in the model.
      6. Review rating
2. Non-ML solution
   1. Create groups of data points depending on the seller, price, seller location, and delivery time. Then, check the impact of each factor individually or in combination on the review score. Understanding the main factors that influence review score can also be aided by a correlation matrix.
   2. *Group customers on Geolocation* – The customer city should be used to widely categorise the customer evaluations for related products. This would give a sense of the merchandise at that location.
   3. *Delivery time* – Product review can be verified against the duration taken by the company to deliver the product for understanding if there is any link.
3. Selecting ML Solution – since data points available and could provide better insights than non-ML solution.

## Benefits of Proposed Solution

### Process Benefit

It will give insight into improved products and consumer acceptance. In this approach, product preferences can be recognised, and the supply chain can be improved to enable prioritising of faster-moving products as well as storage needs. The benefits listed below are anticipated: -

* 1. Information on the suitability of sellers and products for various areas would make it possible to pre-stock favourite goods.
  2. A seller's reputation based on the number of positive customer reviews
  3. Streamlining the warehousing process to facilitate quicker processing of popular or quickly moving products.
  4. Lowering storage and transportation expenses as well as preventing product waste and expiration.
  5. Different product kinds are impacted by delivery time.

### Monetary benefits.

1. The monetary benefits could be derived from the assumed savings in the following ways: -
   1. Increase in customer number
      1. Additional customers – 10% (assumed)
      2. Additional customer (in numbers) – 272.44 X 0.10 = 27.24
      3. *Profit increase =* 27.24*X 100 = BR 2724 per day*
   2. Cross selling / Upselling opportunities.
      1. *Number of customers daily - 272.44   
         (total number of orders in 1 year = 99442 / number of days in a year 365)*
      2. *Average order value - 100 Brazilian real (Assumed)*
      3. *Revenue increased by = 0.10 \* 272.44 \* 100 = 2724.4 BR per day.*

## Summarize the Solution

Without NLP models for the target language, sentiment analysis is a challenging task. We can only try to improve our understanding of customer sentiment based on location, cost, seller, and frequency of product movement as a potential workaround. With the agreed limits, the ML solution would produce the anticipated review score. However, due to the fundamental structure of the models, the impact of characteristics employed in ML models might go unexplored. A non-ML solution would explain customer behaviour, show acceptability, and so give a hint about potential sentiment.

## Project Roadmap

1. Goals and initiatives – The project will be able to achieve the company’s short-term goals of increasing the customer experience leading to increasing revenue.
2. Releases and Deliverables – this project will have the following releases and deliverables: -
   1. Proof of Concept (PoC).
      1. Timeline – one month
      2. Deliverables – process setup and preliminary design, review, mock-up and trials.
   2. Beta release – Beta trials in staging area
      1. Timeline – three months
      2. Deliverables – Testing of project in staging area, vulnerability assessment , performance testing, stress testing, bug removal
   3. Final release – In production area
      1. Timeline – one month
      2. Deliverables – deployment to production area, performance monitoring, drift monitoring

## Limitations

1. NLP model for Portuguese language is yet to mature to the required levels needed for machine learning. In absence of such a capability, the client needs to identify patterns based on available and derived metrics and guidelines set by the language expert.
2. Information available in the dataset and other metrics that aid in data interpretation can only supplement available information in the data. Interpretation based purely on such data might be incorrect.

## Success Metrics

1. Daily sales volume increase
2. Increase in customer numbers

## Prioritising Use Case

Refer to the use case prioritisation framework placed at the excel sheet.

## Adoption Examples from Similar Domain

1. Sentiment Analysis Based Product Selection for Enhancing E-Commerce

<https://www.ijitee.org/wp-content/uploads/papers/v9i3s/C10830193S20.pdf>

1. Sentiment Analysis in Product Reviews using Natural Language Processing and Machine Learning

<http://www.warse.org/IJISCS/static/pdf/file/ijiscs35822019.pdf>

1. Sentiment Analysis in Product Reviews using Natural Language Processing and Machine Learning

<http://www.warse.org/IJISCS/static/pdf/file/ijiscs35822019.pdf>

# Customer Churn

## Problem statement:

1. 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.
2. 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.

## Proposed Solutions:

1. ML solutions –
   1. Use one regression model using features like Customer Id, Customer zip code, Customer city, Customer state, Order Id, Rating, Review Comments and historic delivery date etc to predict the customer churn data. As per the customer churn data, strategic measures should be implemented for various clusters of customers based on the order history, geographic location, demographics etc.
   2. Divide the entire customer churn into individual stages of assessment & prediction.
      1. Estimate the number of customers belonging to low rating, delayed delivery date
      2. Cluster the customers as per their order history (spent behaviour), geographic locations
      3. Build individual models for estimating/segmentation the customers, add the results together and predict the customer churn and based on the outcome, decide the various measures to prevent the customers from churning
      4. The advantage of building separate models is more fine control over the processes and better prediction.
2. Non - ML solutions
   1. A rule-based approach to predict customer churn. Analyse the review rating and review comments manually.
   2. A rule-based approach to predict delivery times. There are set times for each step of the delivery process such as set time for getting goods from the warehouse to the transport hub, from the transport hub to the next transport hub etc.
   3. Every time, new review arrives, the analysis has to be done manually & which might not be efficient to predict the customer churn.
3. Selecting First ML solution – An automatic process will be in place for predicting the accurate customer churn in the company

## Benefits of Proposed Solution

### Process improvements

1. Predicting customer churn accurately may involve addition of new processes to act on the insights provided by the ML solution to improve customer experience & to minimize customer churn.
2. Improved profits for the company.

### Monetary benefits

1. The monetary benefits of predicting customer churn which will result in reduction of the customer churn can be divided into two streams
   1. Reducing customer churn will increase the revenue of the company.
      1. Reduction in the number of people who drop out - 12% (Assumed)
      2. Number of customers daily - 272.44   
         (total number of orders in 1 year = 99442 / number of days in a year 365)
      3. Average order value - 100 Brazilian real (Assumed)
      4. Revenue increased by = 0.12 \* 272.44 \* 100 = 3269.28 per day.

## Summarise the DS approach

1. The estimation of accurate customer churn is a regression problem to be solved. You use various data to estimate the factors for customer churn.
2. Customer churn prediction is also a kind of preventive approach to retain the customer.
3. The ML solution can predict more number of false positives which will result in company spending unnecessary time on the customers which are not at the risk of churn
4. Or the ML solution can predict more number of false negatives which will result in company not taking any action on the customers who are likely to churn

## Project Roadmap

1. Goals and initiatives – The project will be able to achieve the company’s short-term goals of increasing the number of active customers thereby increasing revenue.
2. Releases and Deliverables – this project will have the following releases and deliverables: -
   1. Proof of Concept (PoC).
      1. Timeline – one month
      2. Deliverables – process setup and preliminary design, review, mock-up and trials.
   2. Beta release – Beta trials in staging area
      1. Timeline – three months
      2. Deliverables – Testing of project in staging area, vulnerability assessment , performance testing, stress testing, bug removal
   3. Final release – In production area
      1. Timeline – one month
      2. Deliverables – deployment to production area, performance monitoring, drift monitoring

## Limitation

1. There are no actual labels for customer churn which may compromise model accuracy
2. Edge cases - data for special cases is not available

**Success Metrics**

1. Reduction in customer churn or churn rate
2. Increase in revenue
3. Improved customer experience

## Prioritising Use Case

Refer to the use case prioritisation framework placed at the excel sheet.

## Adoption Examples from Similar Domain

1. Customer Churn prediction in ECommerce Sector

<https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12319&context=theses>

1. E-Commerce Customer Churn Prediction

<https://www.analyticsvidhya.com/blog/2022/06/e-commerce-customer-churn-prediction/>

1. Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees

<https://www.sciencedirect.com/science/article/pii/S1877050922001703>

# Customer Acquisition Cost Optimisation

## Problem statement:

1. 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.
2. 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.

## Proposed Solutions:

1. ML Solution
   1. Artificial intelligence can help in each step of customer acquisition and customer retention, from boosting a company’s search engine ranking so that a prospect finds a business’ website to answering questions and making the ordering process more effortless. While businesses make significant investments in lead generation, they may be neglecting the customer acquisition and customer retention strategy of offer optimization. Artificial intelligence can help optimize offers by making them more personalized. AI can power bots to interact with potential customers and answer their questions.
   2. Email marketing can be personalized to each prospect through artificial intelligence that catches the attention of new leads and targets the unique interests of prospects. Once a prospect becomes a customer, AI can continue to boost engagement and repeat sales. Analysis of customer engagement online can help companies determine the most effective marketing strategies, most popular products, and best production schedules to meet demand. AI customer acquisition marketing can make an entire company run more efficiently.
   3. To boost customer acquisition and CX, companies worldwide use Big data to create prospect-centric marketing campaigns and improve conversion rates. Marketing teams harness the power of artificial intelligence, its robust capabilities along with machine learning to build customer loyalty, increase brand recognition, and tailor strategies that generate better results.
   4. **Lower your customer acquisition costs by utilizing marketing automation in conjunction with customer relationship management (CRM) tools. Marketing automation already reduces costs because fewer staff members need to focus on menial tasks.**
   5. **A company can think of ways to collect contact details for targeted marketing down the line. A communications software company may publish gated content and write topics such as “**[**call dropping**](https://www.dialpad.com/blog/dropped-calls/)**”.**
2. Non ML solution
   1. Companies can lower their customer acquisition costs by **increasing their repeat customer rate, purchase frequencies, and average order values**. Utilize strategies like customer feedback loops, add loyalty programs, and add customer education programs. On top of this, keep a watchful eye on your churn rates. There are different ways to reduce Customer Acquisition costs:-
      1. Prioritize Appropriate Audiences
      2. Retarget Customers
      3. Improve Customer Retention
      4. Try Affiliate Programs
      5. Create Content and Assess the Effectiveness
      6. A/B Test and Optimize Your Pages
      7. Improve the Sales Funnel

## Benefits of Proposed Solution

### Process improvements –

1. Predicting customer acquisition cost optimisation more accurately will not impact any existing process in the e-commerce setup.
2. However, there will be new processes which will be added to take action on the insights provided by the ML solution to improve customer experience & to minimize customer acquisition costs.
3. Improved profits to the company.

### Monetary benefits –

1. The monetary benefits of predicting customer acquisition cost optimisation which will result in reduction of the customer acquisition costs can be calculated as follows:-
   1. Reducing customer acquisition costs will increase in the revenue of the company.   
      Based on customer segmentation,we found that 30% customers are not adding any lifetime value so we are neglecting them and discounts are offered to the rest of the 70% customers. On implementing this,we found that:-
      1. Increase in number of new customers after marketing expenditure - 8% (Assumed\*)
      2. Number of customers daily - 272.44   
         (total number of orders in 1 year = 99442 / number of days in a year 365)
      3. Average order value - 100 Brazilian real (Assumed)

Revenue increased by = 0.08 \* 272.44 \* 100 = 2179.52 per day.

* 1. Inventory management. Predicting the customer acquisition cost optimisation correctly will reduce the pressure on the supply chain. The through out rate of deliveries of the entire system increases but predicting the cost optimisation correctly. The warehouses will have to hold lesser inventory.
     1. Reduction of time items stay in warehouse = 40% (assumed)
     2. In the same time frame warehouse will be able to hold more items = 40% (assumed)
     3. Warehouse cost per item delivered will reduce = 25% (assumed)
     4. Initial warehouse cost per item delivered = 5 BR (assumed)
     5. Total savings per day = 272.44 \* 0.75 \* 5 BR = 1021.65 BR

## Summarise the Solution

1. Customer data analytics give an overall picture of what a business’s customers look and act like, which can help with targeted customer acquisition efforts. By leveraging data science models, companies can strategize cross-channel marketing to target prospective customers and maximize return on campaigns. Apart from this, if we are looking to increase customer reach, data analytics can help us by tapping preferences and purchase patterns such as demographics, affinity, in-market, remarketing, etc.
2. We can also automate and optimize data feed updates, create add groups to target specific products, manage campaigns like thematic search, retargeting strategy, and implementation, ad extensions more effectively using data analytics for acquiring new customers. Aligning bidding strategy with KPIs, improving quality score on Google, increasing online visibility by enhancing local search results and optimizing user experience, optimizing marketing strategies and improving the effectiveness of advertising by selecting the suitable attribution model are some other ways in which data analytics can help us with customer acquisition.
3. To improve personalization. Customer personalization must become part of a company’s business strategy. For personalization to be effective, it requires a systemic and sustained effort on part of the marketing team.
4. AI*-*powered *content:* Machines have started writing content; not just some incoherent stuff but the text that seems it is written by a human. There are tools out there, for example, that can rank your content, ad copy, or headline, even simplify them.
5. Automated email marketing*:* The deployment of machine learning in marketing can help marketing teams hyper-personalize their campaigns. There are a variety of automated email marketing programs available that are powered by machine learning.
6. Social Listening: Machine learning-powered social listening tools can help in following brand mentions and hashtags across many social media channels. These can be used by digital marketers to understand how audiences react to specific content products and help create content that finds an echo with their audience.

## Project Roadmap

1. Goals and initiatives – The project will be able to achieve the company’s short-term goals of increasing the efficiency of services, increasing revenue, and improving customer experience.
2. Releases and Deliverables – this project will have the following releases and deliverables: -
   1. Proof of Concept (PoC).
      1. Timeline – one month
      2. Deliverables – process setup and preliminary design, review, mock-up and trials.
   2. Beta release – Beta trials in staging area
      1. Timeline – three months
      2. Deliverables – Testing of project in staging area, vulnerability assessment , performance testing, stress testing, bug removal
   3. Final release – In production area
      1. Timeline – one month
      2. Deliverables – deployment to production area, performance monitoring, drift monitoring

## Limitations

1. A preliminary study of the project finds all necessary information and facilities needed adequately in place for its execution. No limitations are envisaged at this stage.

## Prioritising Use Case

Refer to the use case prioritisation framework placed at the excel sheet.

## Adoption Examples from Similar Domain

1. Customer Acquisition Cost: How to Calculate, Reduce & Improve It

<https://neilpatel.com/blog/customer-acquisition-cost/>

1. Using Machine Learning to Detect Customer Acquisition Opportunities and Evaluating the Required Organizational Prerequisites

<https://www.diva-portal.org/smash/get/diva2:1366207/FULLTEXT01.pdf>

# Fraud Detection

## Problem statement:

1. 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.

## Proposed Solutions:

1. ML solutions

* Intention here is to identify customer and merchant frauds.
* More labelled data needs to be captured such as sold product details including billing and shipping, IP addresses, Merchant profiles, payment details.
* A small team to act on suspected fraudulent transactions should be set up to work with Customer and Merchant

In the absence of labelled data, below is the proposed approach.

* 1. **Neural Networks (Autoencoders)**
     1. An autoencoder is a neural network that learns by encoding its input data to a lower or higher dimension in a hidden layer and tries to decode back to the original input. This process is called reconstruction. The hidden layers are often used to enforce the model to prioritize what properties are the most useful for reconstructing the output. Autoencoders are not designed to reconstruct perfectly since it would defeat the purpose of finding patterns in the data.
     2. These can adapt to the change in the behavior of normal transactions and identify patterns of fraud transactions. The process of the neural networks is extremely fast and can make decisions in real time.
     3. Use unsupervised learning in order to get a rough estimation of whether or not a transaction is fraudulent. Based on this estimation and a threshold, a new labelled dataset can be created in order to train with supervised learning
     4. Data imbalance needs to be handled.
  2. **Random Forest** 
     1. Once we get a rough estimation of whether or not a transaction is fraudulent, we can use random forest to predict on the labelled data.
     2. Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data. Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data and that they may over-fit data sets that are particularly noisy.

1. Non - ML solutions
   1. **SME Support**
      1. Large pool of SMEs to monitor and identify suspected fraudulent transaction.
   2. **Limits on daily spending.**
      1. Limit on daily spending set by the customer.
   3. **Manually derive insights from data**
      1. **Customer behavior**

Flag off irregular behaviors based on past data (Back to back orders, High value orders)

* + 1. **Geolocation**

Check transactions where there is a mismatch between customer address and order shipping address

* + 1. **Reviews**

Customer reviews could potentially identify certain merchant frauds

Selecting the ML solution - An automatic process will be in place for identifying fraudulent transactions.

## Benefits of Proposed Solution

### Process improvements –

Making each customer’s experience as satisfying as possible during the fraudulent cases.

Acting on suspected fraudulent cases will trigger actions on rogue merchants and customers.

Detecting fraud more accurately will not impact any existing process in the e-commerce setup.

### Monetary benefits -

A new process of identifying and handling fraud transactions can result not just in better customer experience but also an increase in revenue and active customers.

Many of the successful handling of fraud transactions can be **effectively marketed** thereby increasing the brand value, being the go-to marketplace for both customers and merchants and lesser customer churn/inactive customers.

With the assumption number of fraudulent transactions acted upon quickly with new model and process introduced, we anticipate a 5% increase in new active customers and 10% increase in revenue from active customer bases. This doesn’t take way in risk on brand and image of not acting on fraudulent transactions.

* 1. Reduction in the number of people who drop out - 10% (Assumed)
  2. Increase in customer base – 5%(Assumed)
  3. Number of customers daily - 272.44   
     (total number of orders in 1 year = 99442 / number of days in a year 365)
  4. Average order value - 100 Brazilian real (Assumed)
  5. Revenue increased by new active customer bases = 0.05 \* 272.44 \* 100 = 1.362.2 BR per day.
  6. Revenue increased by new active customer bases = 0.1 \* 272.44 \* 100 = 2,724.4 BR per day.

## Summarize the DS approach

Due to the absence of labelled data, the recommended approach is that of a combination of unsupervised and supervised techniques.

An autoencoder will be used to generate labels which will in turn be used by a classification model to detect fraud in real time.

## Project Roadmap

1. Goals and initiatives – The project will be able to achieve the company’s short-term goals of increasing the efficiency of services and increasing revenue.
2. Releases and Deliverables – this project will have the following releases and deliverables: -
   1. Proof of Concept (PoC).
      1. Timeline – one month
      2. Deliverables – process setup and preliminary design, review, mock-up and trials.
   2. Beta release – Beta trials in staging area
      1. Timeline – three months
      2. Deliverables – Testing of project in staging area, vulnerability assessment , performance testing, stress testing, bug removal
   3. Final release – In production area
      1. Timeline – one month
      2. Deliverables – deployment to production area, performance monitoring, drift monitoring

## Limitation

1. Labelled data not available and many other relevant features are not available in the data set.
2. High complexity model – The team may not have the required skillset.

## Success Metrics

1. Customer satisfaction
2. Increase in revenue by ~10%
3. Increase active customer base by ~5%
4. Achieve a high negative predictive value (low number of false negatives as compared to true negatives) rather than high accuracy.

## Prioritising Use Case

Refer to the use case prioritisation framework placed at the excel sheet.

## Adoption Examples from Similar Domain

1. Fraud Detection using Machine Learning and Deep Learning

<https://www.researchgate.net/publication/339411416_Fraud_Detection_using_Machine_Learning_and_Deep_Learning>

1. Fraud Detection using Machine Learning

<https://cs229.stanford.edu/proj2018/report/261.pdf>

# Price Optimisation

## Problem statement:

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.

## Proposed Solutions:

Any proposed solution has to serve following goals

* The best initial price to set products in order to generate the most revenue and profit
* The best overall price to keep your products at , for e.g. to keep below competition price
* The best price to discount your products to based on people’s willingness to buy
* To get the best price for stock clearance.

We will need following features to build a good price optimization model

1. willingness to pay (data available)

* segment customer on their spending
* how frequently they are buying

2. Reviews (data available)

* % of positive or negative reviews

3. Local demand (data available, can be derived on geo location)

4. Global demand (data available, can be derived on geo location)

5. Events e.g. festive seasons, Black Friday, (data available, can be derived on date of purchase)

6. Business operating costs (some cost available such as delivery cost)

7. Competitors prices (not available)

8. Seasonal weather information (data available, can be derived on date of purchase)

9. Warehouse stock data (not available)

10. Product category : Basics, seasonal and occasional

We can see that we about 60 % of the features are either available or can be derived from available data. So based this we propose following ML solution.

1. ML solutions -

ML solution would be a combination of following supervised and un-supervised learning.

1. Build a segmentation model which will segment data on factor such as seasonality, geo location, demands, events such as festivals , customer profile. For this we can leverage traditional un-supervised ML algorithms such
   * K-means clustering
   * KNN (k-nearest neighbors)
   * Hierarchal clustering
   * DB scan
2. Then for each segment, we will build predictive models using Linear Regression to predict the price point that should be set. This will allow a dynamic pricing model to be set which will allow the company to be responsive to changing dynamic market condition and thus gaining advantage over the competition. This model will also allow the company to offer discount to customer, which in turn drive more sales and revenue.
3. Non - ML solutions
   1. One of the solution is manually fix prices by looking at the competitors website etc.
   2. Another solution could be rely on the old historical prices and the sales response to adjust current prices and do this manually.
   3. And a third option is to offer discount to a clearance sale in case goods are unsold.

## Benefits of Proposed Solution

### Process improvements -

Currently process of setting price is mainly rule of thumb as opposed to based on actual data of sales, cost, market condition etc.

Since the machine learning model would be predicting prices in response to the market condition based on real data, it makes the process of setting prices more dynamic, responsive and automatic. This in turn would increase sales, revenue, reduce cost by getting rid of surplus stock.

Improved profits to the company.

### Monetary benefits -

The monetary benefits of price optimsation can be divided into following streams

* Price optimisation will increase in the revenue and profits of the company because of higher prices  
  1. Average increase in price of the product because of dynamic pricing - 10% (Assumed)

(for e.g. during festival, certain items can priced higher due to more demand,

Also, certain regions can afford more higher pricing)

* 1. Number of customers daily - 272.44   
     (total number of orders in 1 year = 99442 / number of days in a year 365)
  2. Average order value - 100 Brazilian real (Assumed)
  3. Revenue increased by = 0.10 \* 272.44 \* 100 = 2724.4 BR per day.
* More sales will occur as we will be able to offer lower prices than competition.   
  1. Additional sales as result of lower prices as compared to competition – 5 % (Assumed)
  2. Number of customers daily - 272.44   
     (total number of orders in 1 year = 99442 / number of days in a year 365)
  3. Average order value - 100 Brazilian real (Assumed)

Revenue increased by = 0.05 \* 272.44 \* 100 = 1362.2 BR per day.

* Increased sales as result of discount   
  1. Additional sales as result of discounts – 5 % (Assumed)
  2. Number of customers daily - 272.44   
     (total number of orders in 1 year = 99442 / number of days in a year 365)

Average order value - 100 Brazilian real (Assumed)

Revenue increased by = 0.05 \* 272.44 \* 100 = 1362.2 BR per day.

Total increase in sale per day = 5,448.8 BR per day.

* Cost saving a result of stock clearance  
  1. Loss to due surplus unsold stock (in previous years) : BR 60,000 (assumed)
  2. Stock sold as result of stock clearance as prices suggested by ML model : BR 20,000 (assumed)

Cost saving per day = BR 20,000 / 365 = BR 54.80

## Summarise the DS approach

The DS approach we have proposed is a combination of segmentation (un-supervised learning) and regression (supervised learning). By segmenting the data on various factors such geographies, seasons, weathers or local / global conditions we are able to take account the nuances of that segments. Then harnessing the predictive power of regression in specific segment we are able to get accurate price point prediction within that segment. This will also allow company respond more quickly to any changes in that segment.

## Project Roadmap

1. Goals and initiatives – The project will be able to achieve the company’s short-term goals of increasing the efficiency of services, increasing revenue, and improving customer experience.
2. Releases and Deliverables – this project will have the following releases and deliverables: -
   1. Proof of Concept (PoC).
      1. Timeline – one month
      2. Deliverables – process setup and preliminary design, review, mock-up and trials.
   2. Beta release – Beta trials in staging area
      1. Timeline – three months
      2. Deliverables – Testing of project in staging area, vulnerability assessment , performance testing, stress testing, bug removal
   3. Final release – In production area
      1. Timeline – one month
      2. Deliverables – deployment to production area, performance monitoring, drift monitoring

## Limitation

1. Pricing optimization depends on the astute application of meticulously gathered data. It will never succeed if it is based on speculation. To prevent these issues, make sure monitoring of price and input data is required.
2. Another issue that frequently arises is pricing optimization that fails because of data errors and inconsistencies.
3. Price optimisation model will not be able to take account the value of the product in mind of the buyer. For e.g. an apple product has a higher value in mind of the buyers as opposed to other brands. Taking of value can result in higher profits, but a data driven can not take account of value.
4. The price optimization model may suggest heavy discounting which can lead to erosion of profits.

## Success Metrics

1. Increase in revenue
2. Increase in overall sales
3. Increase in profits (topline)
4. Decrease in maintenance cost as a result of clearance of surplus stock.

## Prioritising Use Case

Refer to the use case prioritisation framework placed at the excel sheet.

## Adoption Examples from Similar Domain

1. Machine Learning E-Commerce: Optimizing Prices with ML

<https://prisync.com/blog/machine-learning-e-commerce/>

1. Pricing Optimization Strategies for E-commerce Businesses

<https://www.mailmunch.com/blog/4-pricing-optimization-strategies-e-commerce-businesses>

1. Machine Learning Helps Companies Optimize Prices

<https://www.electrifai.net/blog/machine-learning-helps-companies-optimize-prices#:~:text=What%20is%20pricing%20optimization%3F,achieved%20between%20value%20and%20profit>

# Appendix

The study was done using references from online resources

1. Online Fashion Commerce: Modelling Customer Promise Date

<https://aws.amazon.com/blogs/industries/how-to-predict-shipments-time-of-delivery-with-cloud-based-machine-learning-models/>

1. Paper suggesting various solutions for delivery date production <https://arxiv.org/pdf/2105.00315.pdf>
2. Article detailing use of sentiment analysis on towardsdatascience.com

<https://towardsdatascience.com/five-practical-use-cases-of-customer-sentiment-analysis-for-nps-a3167ac2caaa>

1. Analytics Vidhya website use case on sentiment analysis

<https://www.analyticsvidhya.com/blog/2022/02/sentiment-analysis-with-nlp-deep-learning/>

1. Real world examples of implementation of sentiment analysis

<https://www.repustate.com/blog/sentiment-analysis-real-world-examples/>

1. According to study at e-com sites, generally 60% of the data is not-reliable. <https://www.ngpf.org/blog/question-of-the-day/qod-what-percent-of-reviews-posted-on-popular-e-commerce-sites-are-fake/>
2. Study at Preventing Customer Churn article

<https://www.totango.com/customer-churn/preventing-customer-churn>

1. Paper suggesting various solutions for customer churn

<https://paperswithcode.com/paper/customer-churn-prediction-in-telecom-using>

1. Malaysian telecommunications company customer churn case study <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051585/>
2. 8 Simple Ways to Reduce Customer Acquisition Costs

<https://www.datafeedwatch.com/blog/reduce-customer-acquisition-cost>

1. How to Utilize AI for Customer Acquisition

<https://www.infopulse.com/blog/ai-for-customer-acquisition>

1. HOW AI CAN CUT YOUR CUSTOMER ACQUISITION COSTS

<https://www.analyticsinsight.net/how-ai-can-cut-your-customer-acquisition-costs/>

1. Examples of customer acquisition implementation by industries

<https://www.zembula.com/blog/5-best-examples-customer-acquisition/>

1. E-commerce Fraud Detection and Prevention: The In-depth Guide

<https://spd.group/machine-learning/e-commerce-fraud-detection/>

1. Machine Learning E-Commerce: Optimizing Prices with ML

<https://prisync.com/blog/machine-learning-e-commerce/>

1. Pricing Optimization Strategies for E-commerce Businesses

<https://www.mailmunch.com/blog/4-pricing-optimization-strategies-e-commerce-businesses>