# 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 time for getting goods from the warehouse to the transport hub, from the transport hub to the next transport hub etc.   
        
      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.

Selecting First ML solution - because the data points for individual steps are not present.

**Benefits of Proposed Solution**

**Process improvements -**

Predicting delivery dates more accurately will not impact any process in the e-commerce setup.

Reduction in Inventory costs.

**Monetary benefits -**

The monetary benefits of accurate delivery can be divided into two streams

* 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)
  3. Average order value - 100 Brazilian real (Assumed)
  4. Revenue increased by = 0.1 \* 272.44 \* 100 = 2724.4 per day.
* Inventory management   
    
  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)
  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 DS approach**

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.

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)

**Limitation**

1. RMSE might be high
2. High complexity model - non linearity , thus non linear model might be needed - is the team capable or not.
3. Think of edge cases - data for special cases is not available

**Define Appropriate Success Metrics**

1. Early delivery rates
2. Late delivery rates
3. Churn rate after estimated dates is shown

**Prioritising Use Case**

Refer to the use case prioritisation framework linked [here](https://docs.google.com/spreadsheets/d/1EGoVubzdetsv8YhrKeenSq_uaBSubtsgZm2WydL1U1I/edit#gid=704154787).

| Delivery Date Prediction | Raw Scores | Normalized Scores |
| --- | --- | --- |
| **Feasibility Score** | 4.1 | 0.82 |
| **Complexity Rating** | 16.5 | 0.825 |
| **Strategic Value** | 142 | 0.7 |
| **Business Value (Enter in numericals the business value generated)** | 3746.05 | 0.97 |
| Final | | 3.31 |

**Cite references**

1. Similar delivery date estimation application form example on amazon blogs. <https://aws.amazon.com/blogs/industries/how-to-predict-shipments-time-of-delivery-with-cloud-based-machine-learning-models/>
2. Paper suggesting various solutions for delivery date production <https://arxiv.org/pdf/2105.00315.pdf>

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 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:**

**ML Solution**

Supervised sentiment analysis can be performed using the review\_score to help label the data fast. Unsupervised learning can be performed to estimate the sentiments of these reviews and hence the popularity of a product. For unsupervised, the reviews have to be translated to English.

1. Utilize machine learning techniques, such as supervised learning for sentiment analysis and topic modeling, to analyze customer reviews in Portuguese. Translate a subset of reviews into English for training and use the trained model to predict sentiment and extract topics from the remaining Portuguese reviews.
   1. Utilize NLP techniques to analyze and extract sentiment from customer reviews written in Portuguese. This can help OLIST gain insights into customer experiences and identify areas of improvement based on the sentiment expressed.
   2. Analyze the comments left on e-commerce orders using **olist\_order\_reviews\_dataset**. Understand the data and number of comments available and missing reviews. Create a function to search for some regular expression RegEx and remove or identify sites,hyperlinks, line breaks,dates, numbers, special characters, and whitespaces.With considerations that the comments are in portuguese. Apply some advanced text transformations like stopwords removal, stemming and the TF-IDF matrix process. Implement Feature extraction using the BoW produces reduced and simplified representation of an entire document, ignoring aspects like grammar, word appearance order, and semantic relations between words and phrases. Or use frequency-inverse document frequency (TF-IDF). Followed by a classical machine learning (ML) classifier, such as Logistic Regression, Support Vector Machines, Gradient Boosting Decision trees, or Random Forests, in sentiment classification tasks. As positive or negative.
   3. Implement topic modeling algorithms to identify common topics or themes within customer reviews. This can provide valuable information about the main concerns, preferences, and feedback from customers, helping OLIST address specific issues.
   4. Use a combination of language translation and sentiment analysis to convert Portuguese customer reviews into English (or another language with better NLP support) and then perform sentiment analysis using existing sophisticated NLP algorithms.
2. LSTM is useful due to its memory control for data. Negation is crucial for accurate sentiment analysis. For example, "great" and "not great" have different meanings. LSTM can learn this distinction and predict negation. It can also infer grammar rules from text. However, LSTMs may struggle with long sentences, forgetting distant words, and processing word by word.
3. Transformer models This model differentially weights the significance of each part of the data to identify the context that confers meaning to each word.
4. Deep learning approach results in the most accurate sentiment analysis. No need for manual work to define classification features with the benefit of considering the impact of word order. To increase the accuracy and efficiency of sentiment analysis.

**Non ML solution**

Manually analyze customer reviews, create a sentiment lexicon with positive and negative words, assign sentiment scores, and classify reviews as positive, negative, or neutral based on scores. Extract insights by analyzing sentiment patterns and trends.

This manual approach is suitable for small teams with limited resources and ML expertise, but it is time-consuming and less scalable than ML solutions.

* The best machine learning solution for OLIST's sentiment analysis use case would be (NLP) model (1) for Sentiment Analysis." This approach directly aligns with the problem statement and business goals, enabling the Chief Marketing Officer to gain valuable insights into customer experiences, identify improvement areas, and make data-driven decisions to enhance customer satisfaction and improve OLIST's offerings. Additionally, this solution has the potential to scale efficiently, making it feasible for a small data science team with limited resources to implement effectively.

**Benefits of Proposed Solution**

**Process improvements -**

* By understanding customer sentiment and identifying areas of improvement, OLIST can enhance its product offering and improve customer experience.
* Marketing and Customer Support teams to incorporate sentiment analysis results into their decision-making processes, product improvements, and targeted marketing strategies.

**Monetary benefits -**

* Increased customer satisfaction can lead to higher customer retention, reduced churn, and improved marketing ROI, resulting in higher revenues and profits.
* Calculating the impact of improved customer satisfaction on revenue growth, repeat purchases, and customer retention.
* 27% of customers had negative comments based on the rule we made thus if we identified there pain points we can avoid there churn and increase profits lets almost 11000 customers had negative reviews we could retain those as improving deliveries or products listed diversifying sellers and avoiding scam lets assume 70% of the 11,000 customer with negative comments will churn thus and cost of retaining one customer is 10 BRL while acquisition of new customer is 100 BRL thus the company will save 90\*11000 =990000 BRL throughout this period. Or 1294/day. Avoiding churn of 14 persons per day with 70% probability to buy a minimum product with 150 BRL on average = 1470 RBL / day

Ex: if cost of acquisition of customer assumed is 100 BRL and you successfully retained a customer for 10 BRL then you save costs by retaining customers rather than loosing and performing new acquisition.

* Recommending products that customers like which would lead to more sales and revenue.

**Summarise the Solution**

* Data Collection: customer review\_comment \_message and other data from OLIST's e-commerce platform, ensuring it includes relevant attributes such as order ID, customer unique ID, review text, and delivery feedback.
* Data Preprocessing: Preprocess the text data by removing special characters, converting text to lowercase, and tokenizing the sentences into individual words.
* Use customer reviews from the **olist\_order\_reviews\_dataset**. Carry out EDA and see the most frequent words.
* Analyze available and missing reviews, use regular expressions to remove or identify unwanted elements like sites, hyperlinks, line breaks, dates, numbers, special characters, and whitespaces in Portuguese comments. Apply text transformations like stopwords removal, stemming, and TF-IDF matrix process. Implement feature extraction using Bag-of-Words (BoW) or TF-IDF to simplify the document representation, disregarding grammar, word order, and semantic relations.
* Language Translation (Optional): As the customer reviews are in Portuguese, translate the text into English or another language with better NLP support to leverage existing sophisticated NLP algorithms.
* Utilize NLP techniques such as bag-of-words, word embeddings, or pre-trained language models like BERT to analyze the sentiment of each review. This process will classify reviews as positive, negative, or neutral based on the sentiments expressed.
* Customer Segmentation: After sentiment analysis and topic modeling, perform customer segmentation based on the sentiments and topics to identify groups of customers with similar feedback and experiences.
* Visualizations and Insights: Visualize the results, sentiment distribution, and identified topics through graphs and charts to make it easier for stakeholders to understand and act upon the insights.
* Proof of Concept (POC): Develop a POC on a subset of the data to demonstrate the effectiveness and potential benefits of the sentiment analysis approach.
* Continuous Improvement: Implement an iterative approach to continuously improve the sentiment analysis model based on feedback and evolving customer preferences.

**Limitation**

* Translation accuracy and loss of context.
* Language-specific nuances may be missed.
* Limited training data.
* Potential loss of information and intent.
* Domain-specific challenges may arise.

**Prioritising Use Case**

The spreadsheet is available for details about prioritising.

Table Summary:

| Sentiment Analysis | Raw Scores | Normalized Scores |
| --- | --- | --- |
| **Feasibility Score** | 3.5 | 0.7 |
| **Complexity Rating** | 14.5 | 0.725 |
| **Strategic Value** | 14 | 0.7 |
| **Business Value (Enter in numericals the business value generated)** | 2764.11 | 0.71 |
| Final | | 2.84 |

**Appropriate Success Metrics**:

* Sentiment Accuracy: Measure the accuracy of the sentiment analysis model in correctly classifying customer reviews as positive, negative, or neutral.
* Customer satisfaction scores, sentiment score improvement, or the number of actionable insights generated, to evaluate the effectiveness of the sentiment analysis model.
* Operational Efficiency: Evaluate the efficiency of the sentiment analysis solution in automating the analysis of customer reviews and extracting insights. Measure the time and resources saved compared to manual review processes, indicating improved operational efficiency.
* Actionable Insights: Determine the number of actionable insights derived from sentiment analysis that have been implemented and resulted in positive outcomes. This metric reflects the solution's ability to drive meaningful changes and improvements in the business based on customer feedback.

**Cite references**

1. Sentiment analysis usage and steps to implement.

<https://aws.amazon.com/what-is/sentiment-analysis/>

1. Amazon Product Review

<https://www.cs.jhu.edu/~mdredze/datasets/sentiment/>

1. Women’s clothes ecommerce reviews.

https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Customer Churn

**Problem statement:**

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.

**Proposed Solutions:**

* ML Solution
  1. Implement a logistic regression model to predict customer churn based on various customer attributes such as purchase history, engagement metrics, and demographic information. Identify significant factors influencing churn and prioritize retention strategies accordingly. LogisticRegression, KNeighborsClassifier, SVC, DecisionTreeClassifier, RandomForestClassifier, LGBMClassifier
  2. Build a random forest model to capture complex interactions and patterns in customer data, enabling accurate prediction of churn. Leverage features such as customer behavior, product preferences, and customer support interactions to identify at-risk customers and tailor retention initiatives.
  3. Utilize gradient boosting algorithms like XGBoost or LightGBM to develop a powerful predictive model for customer churn. Incorporate a wide range of customer data, including transactional history, browsing patterns, and customer feedback, to identify factors contributing to churn and recommend personalized retention actions.
* Non ML solution
  1. Segment customers based on their behavior, preferences, and buying patterns. Use these segments to understand different churn risk profiles and develop targeted retention strategies. Personalize communication, offers, and incentives to meet the specific needs of each segment.
  2. Regularly collect customer feedback through surveys to measure satisfaction and identify pain points. Analyze the survey results to uncover areas for improvement and take proactive measures to address customer concerns, improving overall customer experience and reducing churn.
  3. Implement loyalty programs to incentivize repeat purchases and reward customer loyalty. Offer exclusive benefits, discounts, and personalized offers to loyal customers, encouraging them to stay engaged with the platform and reducing the likelihood of churn.
  4. Strengthen customer support services by providing timely and personalized assistance. Invest in training support staff to address customer queries and concerns effectively, ensuring a positive experience and building long-term customer

We can select A and C in ML solutions. The choice between machine learning and non-machine learning solutions depends on factors such as available data, resources, expertise, and the desired level of automation. Machine learning solutions provide advanced predictive capabilities but may require more data and technical expertise. Non-machine learning solutions focus on customer-centric approaches and can be implemented with fewer resources.

**Benefits of Proposed Solution**

**Process improvements-**

* Implementing a machine learning solution automates the churn prediction process, saving time and effort compared to manual analysis. This allows the sales and marketing team to focus on executing effective retention strategies and enhancing the overall customer experience while increasing revenues by focusing only on customers with high probability to churn thus reducing marketing expenses.

**Monetary benefits -**

1. By accurately identifying at-risk customers, OLIST can proactively implement targeted retention strategies, reducing churn rates and improving customer loyalty.
2. Reducing churn improves profitability by increasing customer lifetime value. Acquiring new customers can be expensive, while retaining existing customers can be more cost-effective. For example, offering discounts and free delivery to attract new customers might require a significant investment ex: 70% discount, but offering a smaller discount (10)% to an existing customer can still bring them back to the platform.
   1. Existing customer revenue with low recency 273 BRL thus saving 100 customers per month = 100\* 273 = 27300 BRL of revenue might be lost if no retention is made.
   2. Cost of acquisition is 50 BRL/customer assumed thus 273-50 =223 per customer savings
      1. Cost of retention of 1 customer = 10 BRL assumed thus saving 80% than money spent on acquisition 50-10 = 40 BRL Savings= 40\*100 customer per month assumed = 4000 BRL/ Month savings
      2. Cost of sending 10 BRL discount to customers not churning if done on randomly on 50% of customers 96000\*0.5 = 48000 customers \* 20 =9,600,000 RBL on 760 days thus 1263/ day RBL for unoptimized retention while optimized retention will allow you to send to only 10% (assumed) thus 10\*9600 = 96000 96000/760=126BRL/day Savings = 1263-126 =1137 BRL/day
3. Leveraging advanced analytics and machine learning to address customer churn positions OLIST as an innovative and customer-centric e-commerce company, differentiating it from competitors and attracting new customers.

**Limitations**:

* Data quality and availability: The effectiveness of the machine learning solution heavily relies on the quality and availability of customer data. If the data is incomplete, inconsistent, or biased, it may impact the accuracy of the churn prediction model. The customer data is complete along with product and payment dataset that helped derive new features.
* Technical expertise and resources: Developing and deploying machine learning models require a certain level of technical expertise and resources. The team needs to have the skills and knowledge to preprocess the data, tune the model parameters, and deploy it into the production environment is the team capable or not.
* Interpretability: Gradient Boosting models are considered as black-box models, meaning they lack interpretability compared to simpler models like logistic regression. Understanding the specific factors driving churn predictions may be challenging. Interpretability is really important and using a logistic regression model might be more convenient based on skills and resources available.
* CV F1 Score, Test Accuracy, Test Precision, Test Recall, Test F1 values shall be monitored

**Summarise the Solution**

Gradient Boosting models can capture complex interactions and patterns in OList customer data, enabling accurate identification of at-risk customers and predicting churn with high precision

1. Use relevant customer data, including purchase history, engagement metrics as, review score,demographic information (location), city, customer\_id, product category name, customer\_id and customer\_unique\_id..
2. Clean and preprocess the data, handling missing values, outliers, and ensuring data quality.
3. Extract meaningful features from the data that capture customer behavior, preferences, and interaction patterns frequency of purchase and assume that a customer churns if he didn't purchase 6 months after his first order. Segment customers on RFM after extracting features as frequency, resenancy and monetary and extract the data to be used from model training.
4. Train and optimize a Gradient Boosting model using the prepared dataset to predict customer churn probabilities.
5. Evaluate the model's performance using appropriate metrics as accuracy and confusion matrix, and fine-tune the model if necessary.
6. Deploy the model into the production environment, integrate it with the existing systems, and continuously monitor its performance, updating the model periodically as needed.

**Define Appropriate Success Metrics**

1. Churn Prediction Accuracy: Metrics such as accuracy, precision, recall, and F1 score. A higher accuracy indicates a more reliable model for identifying at-risk customers.
2. Churn Reduction Rate: This metric reflects the effectiveness of the implemented strategies in reducing churn and improving customer loyalty.
3. Customer Lifetime Value (CLV) Improvement: Monitor changes in CLV for customers who were identified as at-risk but successfully retained. An increase in CLV indicates that the implemented retention initiatives have positively impacted customer behavior and long-term value.

**Prioritising Use Case**

The spreadsheet is available for details about prioritising.

Table Summary:

| Customer Churn | Raw Scores | Normalized Scores |
| --- | --- | --- |
| **Feasibility Score** | 3.7 | 0.74 |
| **Complexity Rating** | 16.83 | 0.8415 |
| **Strategic Value** | 18 | 0.9 |
| **Business Value (Enter in numericals the business value generated)** | 3867 | 1 |
| Final | | 3.48 |

**Cite references**

**NA**

# Customer Acquisition Cost Optimisation

**Problem statement:**

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.

**Proposed Solutions:**

ML Solution

1. Customer Segmentation: Use machine learning algorithms to segment customers based on their behavior, preferences, and purchase history creating an RFM framework or K-means clustering. This can help identify high-value customers who are more likely to have a higher lifetime value and optimize marketing efforts towards them. OLIST can develop targeted marketing strategies and allocate resources more effectively to optimize customer acquisition cost.
2. Predictive Modeling: Build machine learning models, such as logistic regression or random forest, to predict the likelihood of customer conversion or purchase. By analyzing various customer attributes and historical data, this model can identify potential customers with a higher probability of conversion. OLIST can then focus its marketing efforts on these high-potential customers, reducing acquisition costs and improving conversion rates.
3. Lifetime Value Prediction: Develop a predictive model to estimate the lifetime value of customers acquired through different marketing campaigns. By understanding the potential revenue generated from each customer, the marketing team can focus on campaigns that yield higher returns on investment.
4. Campaign Effectiveness Analysis: Use machine learning to analyze the effectiveness of various marketing campaigns. This can involve tracking customer responses, conversion rates, and ROI for different campaigns to optimize future marketing strategies.

Non ML solution

1. A/B Testing: Implement A/B testing for different marketing campaigns to compare their performance and effectiveness. This approach allows the marketing team to experiment with different strategies and identify the most successful ones without extensive ML implementation.
2. Customer Surveys and Feedback: Conduct surveys and collect feedback from customers to understand their satisfaction levels, preferences, and reasons for choosing or leaving Olist. This can provide valuable insights for optimizing marketing strategies without relying on complex ML models.

Customer Segmentation (1) segmenting customers based on their behavior and purchase patterns, Olist can identify high-value customers and focus marketing efforts on acquiring similar customers. This approach allows the marketing team to optimize resources and target campaigns to specific customer segments that have a higher likelihood of generating a higher lifetime value.

.

**Benefits of Proposed Solution:**

1. Reduced Marketing Costs: By targeting marketing efforts towards high-value customer segments, OLIST can allocate marketing budgets more efficiently, leading to reduced overall marketing costs.
2. Increased Revenue: By optimizing marketing strategies towards high-value customers, OLIST can expect to generate more revenue from these customers, thereby increasing overall revenue.
   1. Ex: If marketing is focused on high value customer spending 2500 (Assumed) in 1 year compared to low value customers spending 200 per year we can infer that the ROI on marketing would be 90% higher when implementing focused marketing as well as saving marketing costs on false customers or customers not willing to use Olist.
3. Personalized and targeted marketing efforts can lead to a better customer experience, increasing customer satisfaction and loyalty.
4. With a more efficient and targeted marketing approach, OLIST can enhance the efficiency of its marketing campaigns and overall customer acquisition process.

**Summarise the Solution**

1. Combine and merge relevant data as marketing qualified lead dataset with all marketing details and customer transactions, campaign details, customer behavior, demographics, and customer lifetime value. Clean and preprocess the data, handling missing values, outliers, and inconsistencies.
2. Feature Engineering: Create additional features such as customer recency, frequency, and monetary (RFM) scores, customer loyalty indicators, and campaign response metrics. These features will help in segmenting customers effectively.
3. Customer Segmentation Model: Utilize unsupervised machine learning algorithms such as K-means clustering or hierarchical clustering to segment customers into distinct groups based on their RFM scores, purchase patterns, and response to marketing campaigns.
4. Lifetime Value Prediction: Build predictive models to estimate the lifetime value of customers in each segment. This will help in identifying high-value customers who are likely to contribute more to revenue in the long term.
5. Marketing Strategy Optimization: Utilize customer segments and lifetime value predictions to optimize marketing strategies. Focus marketing efforts on high-value customer segments, tailor campaigns to target specific customer groups, and design personalized offers to improve customer acquisition efficiency.
6. Proof of Concept: Create a proof of concept using a subset of the data and evaluate the segmentation model's performance and its impact on customer acquisition cost optimization.

**Limitations**:

1. The success of customer segmentation heavily relies on the availability of rich and diverse customer data which is missing with alot of null values for marketing leads. If the data is limited or of poor quality, it may lead to less accurate segmentation results.
2. Implementing machine learning algorithms for customer segmentation requires skilled data scientists or analysts who are proficient in machine learning techniques and data manipulation.
3. Resource Constraints: As the data science team at OLIST is small, implementing customer segmentation may require additional resources and time, which could be a limitation in terms of feasibility.
4. Overfitting: Inaccurate segmentation models may lead to overfitting, where the model performs well on the training data but fails to generalize well on unseen data.
5. Privacy Concerns: Utilizing customer data for segmentation purposes may raise privacy concerns and require careful handling and compliance with data protection regulations.
6. Implementation Cost: Integrating the segmentation model into OLIST's existing systems and processes may involve additional costs and require changes to the current marketing practices.
7. Continuous Model Updates: Customer behavior and preferences change over time, necessitating regular updates to the segmentation model to keep it relevant and effective.
8. Marketing Strategy Adaptation: The success of customer segmentation relies on the marketing team's ability to adapt and tailor marketing strategies based on the identified customer segments. If the marketing team is not receptive to such changes, the benefits may be limited.

**Prioritising Use Case**

The spreadsheet is available for details about prioritising.

Table Summary:

| **Customer Acquisition Cost Optimisation** | Raw Scores | Normalized Scores |
| --- | --- | --- |
| **Feasibility Score** | 3.63 | 0.726 |
| **Complexity Rating** | 14 | 0.7 |
| **Strategic Value** | 10 | 0.5 |
| **Business Value (Enter in numericals the business value generated)** | 3003 | 0.78 |
| Final | | 2.7 |

**Appropriate Success Metrics**:

1. Customer Acquisition Cost (CAC): Calculate the cost incurred to acquire a new customer through marketing campaigns. Lower CAC indicates cost optimization.
2. Return on Investment (ROI): Evaluate the revenue generated from the marketing campaigns compared to the cost invested. Positive ROI signifies successful campaigns.
3. Customer Retention Rate: Monitor the percentage of customers who continue to make purchases from OLIST after the initial acquisition. Higher retention rate indicates effective customer targeting and satisfaction.
4. Conversion Rate: Measure the percentage of potential customers who become actual customers after being exposed to marketing campaigns. Higher conversion rate signifies better campaign effectiveness.
5. Customer Churn Rate: Monitor the rate at which customers stop making purchases or leave OLIST after the initial acquisition. Lower churn rate indicates better customer retention.
6. Average Order Value (AOV): Calculate the average value of orders placed by customers. Higher AOV implies more successful targeting of high-value customers.
7. Customer Satisfaction Score (CSAT): Gather feedback from customers to assess their satisfaction level with OLIST's services and products. Higher CSAT indicates improved customer experience.
8. Marketing Channel Performance: Evaluate the performance of different marketing channels or leads (e.g., email, social media, ads) in terms of customer acquisition and revenue generation.

**Cite references**

NA

# Fraud Detection

**Problem statement:**

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 organisation against such actions.

**Proposed Solutions:**

ML Solution

1. Implement anomaly detection algorithms, such as Isolation Forest or One-Class SVM, to identify unusual patterns or outliers in transaction data. These algorithms can detect fraudulent activities that deviate significantly from normal customer behavior.
   1. High accuracy in identifying unusual patterns that may indicate fraudulent behavior but requires labeled fraud data for training, which does not exist as well as continuous monitoring to avoid generating false positives and negatives.
2. Train supervised machine learning or Implement a random forest classifier to detect fraudulent transactions based on various features such as transaction amount, customer behavior, purchase history, order\_delivered\_customer\_date,payment details. order\_purchase\_timestamp, order\_id, customer\_id and geolocation,payment\_type reviews or order delivery, order\_value . Train the model using historical data labeled as fraudulent or non-fraudulent, allowing it to learn patterns and make predictions on new transactions.
   1. Ability to leverage historical fraud data to make accurate predictions. Random forest models can handle high-dimensional data with complex relationships effectively, they are robust to outliers and noise in the data, model can provide feature importance, helping to identify the most influential factors contributing to fraud.
   2. Requires labeled training data, which may be time-consuming and expensive to collect since no labeled fraud and non-fraud exist but could be derived. The data would be biased and the model may struggle with imbalanced datasets, where fraudulent transactions are rare compared to legitimate ones.

Non ML solution

1. Manual Review and Verification: Assign a dedicated team of fraud analysts to manually review and verify transactions flagged as potentially fraudulent based on predefined rules and suspicious indicators. The team can analyze transaction details, customer behavior, and conduct further investigations to identify fraudulent activities.
   1. Human expertise in identifying subtle fraud patterns adapt quickly to new fraud techniques.
   2. Time-consuming, resource-intensive and with limited scalability, especially for a large volume of transactions. As well as human error and bias,
2. Rule-Based Systems: Establish a set of rules and thresholds to flag transactions that deviate from predefined criteria, such as unusually large purchases or transactions from high-risk regions. These rules can be based on industry best practices or OLIST's specific fraud patterns.
   1. Simple and straightforward implementation and easy to understand and adjust rules based on evolving fraud patterns.
   2. Limited ability to detect new or unknown fraud patterns thus, generating a high number of false positives or false negatives, requires resources for continuous monitoring and updating of rules.
3. Collaboration and Industry Networks: Share fraud-related information and collaborate with other e-commerce companies, financial institutions, or industry networks to leverage collective knowledge and resources in detecting and preventing fraud.
   1. Access to shared fraud databases and insights.Enhanced fraud detection capabilities through collective intelligence.
   2. Dependence on the willingness of other organizations to share information.Data privacy and security concerns in sharing sensitive information.

The best fraud detection solution for OLIST combines machine learning (2 random forest classifier) with (1) manual review. This improves accuracy and efficiency by utilizing the strengths of both methods. The classifier handles complex patterns and provides insights into important features, while manual review brings human expertise in detecting sophisticated fraud. Although the classifier requires labeled data and updates, and manual review is resource-intensive, the combination optimizes fraud detection, reduces false positives, and benefits from automated algorithms and human intelligence.

**Benefits of Proposed Solution**

**Process Improvements Business:**

1. Automation and Efficiency: Machine learning solutions streamline fraud detection by automating the analysis of large volumes of transactions, reducing manual effort and increasing efficiency.
2. Timely Fraud Detection: Implementing machine learning algorithms and continuous monitoring enables proactive identification of fraudulent activities, allowing OLIST to take prompt action and to focus on other concerns.
3. Expertise and Adaptability: Leveraging manual review and collaboration with industry networks ensures human expertise in identifying sophisticated fraud patterns and adapting to emerging threats.

**Monetary Benefits for Business**:

1. Reduced Financial Losses: By accurately detecting and preventing fraudulent transactions, OLIST can minimize financial losses associated with fraud which are reported as huge.
   1. Assuming 800 orders are fraudulent each 2 years 0.8% thus an average transaction amount of 600BRL thus 800\* 600 = 480,,000 BRL total company loss over almost 760 days. = 631 BRL/day
2. Enhanced Reputation and Customer Trust: Effective fraud detection measures can improve OLIST's reputation and customer trust, leading to increased customer retention and acquisition.
3. Cost Savings: Detecting fraud early can minimize the costs associated with chargebacks, investigation, and recovery.

**Summarise the Solution**

**ML Solution:**

1. Extract and derive features to identify fraudulent or non-fraudulent actions and label the data.
2. Train the model on data labeled as fraudulent or non-fraudulent from step 1.
3. Use various features such as transaction amount, customer behavior, and purchase history.

Benefits: Ability to handle high-dimensional data, robustness to outliers and noise, identification of important fraud-contributing features.

Non-machine learning solution:

1. Manual review and verification by a dedicated team of fraud analysts.

Benefits: Detection of subtle fraud patterns, quick adaptation to new fraud techniques, uncovering sophisticated fraud schemes missed by machine learning.

Combine machine learning with manual review:

To leverage strengths of both methods for enhanced fraud detection accuracy and efficiency.

Be aware of limitations and assumptions to ensure effective implementation and management of the fraud detection system.

**Limitations**:

1. Requires a significant amount of labeled training data, frequent retraining and updating for emerging fraud patterns.
2. Scalability and continuous monitoring.
3. Time-consuming and resource-intensive, scalability challenges, potential for human error and bias, continuous training and knowledge updates required.
4. Requires continuous training and knowledge updates to stay current with evolving fraud tactics.

**Appropriate Success Metrics**:

1. Fraud Detection Accuracy: Measure the percentage of fraudulent transactions correctly identified by the system to assess the effectiveness of the solution.
2. False Positive Rate: Evaluate the proportion of legitimate transactions incorrectly flagged as fraudulent, which should be minimized to avoid inconveniencing customers.
3. Detection Time: Measure the time taken to detect fraudulent transactions from the moment they occur to mitigate financial losses.

**Prioritising Use Case**

The spreadsheet is available for details about prioritising.

Table Summary:

| Fraud Detection | Raw Scores | Normalized Scores |
| --- | --- | --- |
| **Feasibility Score** | 3.9 | 0.726 |
| **Complexity Rating** | 9.83 | 0.4915 |
| **Strategic Value** | 16 | 0.8 |
| **Business Value (Enter in numericals the business value generated)** | 631 | 0.16 |
| Final | | 2.23 |

**Cite references**

R. Jhangiani, D. Bein and A. Verma, "Machine Learning Pipeline for Fraud Detection and Prevention in E-Commerce Transactions," 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0135-0140, 2019.

https://thepaypers.com/online-fraud-prevention/brazil/7

# Price Optimisation

**Problem statement:**

Pricing is one of the most important pieces of business for an e-commerce organsiation. 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 optimisation algorithm so as to maximise the sales and revenue.

**Proposed Solutions:**

ML Solution

1. Implement a machine learning algorithm, such as regression or neural networks, to predict optimal pricing based on factors like location from olist\_geolocation\_dataset, customer attitudes, and competitor pricing. Use historical data and customer segmentation to train the model and generate price recommendations.

Non ML solution

1. Perform market research and competitor analysis to manually determine optimal pricing strategies based on factors like customer preferences, market demand, and cost considerations. Utilize pricing models and business expertise to make informed pricing decisions.

The chosen solution approach combines the use of machine learning and non-machine learning techniques to achieve optimal price optimization. This allows OLIST to leverage data-driven insights and market expertise to maximize sales and revenue.

**Benefits of Proposed Solution**

**Process Improvement:**

1. Automated Pricing: Machine learning algorithms can automate the price optimization process, saving time and effort compared to manual analysis by the sales team.
2. Data-Driven Decisions: Both solutions allow for data-driven decision-making, leveraging analytics to inform pricing strategies to help sales and marketing departments.

**Monetary Benefit:**

1. Increased Sales: By optimizing prices, OLIST can drive more sales and generate higher revenue by finding the sweet spot that balances customer demand and profit margins.
2. Customer Satisfaction: Offering competitive and fair prices can enhance customer satisfaction and loyalty.

**Summarise the Solution**

* Utilize machine learning algorithms to predict optimal prices based on various factors as location from olist\_geolocation\_dataset, customer attitudes, and competitor pricing. This approach automates the pricing process and leverages historical data and customer segmentation.
* Non ML solutions conduct manual market research and competitor analysis to determine optimal pricing strategies. This approach relies on business expertise and market understanding.

**Limitations**:

* The need for extensive data, its quality and bias and the assumption that past trends will continue to hold in the future.
* Data quality and availability, Assumptions about customer attitudes, Complexity of pricing factors as market dynamics, Generalization to new scenarios, Interpretability.
* Need for continuous monitoring and updates

**Prioritising Use Case**

The spreadsheet is available for details about prioritising.

Table Summary:

| Price Optimisation | Raw Scores | Normalized Scores |
| --- | --- | --- |
| **Feasibility Score** | 3.27 | 0.654 |
| **Complexity Rating** | 12.5 | 0.625 |
| **Strategic Value** | 18 | 0.9 |
| **Business Value (Enter in numericals the business value generated)** | 1800 | 0.47 |
| Final | | 2.64 |

**Appropriate Success Metrics**:

* Revenue Growth: Measure the increase in overall revenue generated by implementing the price optimization algorithm or manual pricing strategies.
* Sales Volume: Track the change in the number of products sold to assess the impact of price optimization on sales.
* Profit Margin: Monitor the improvement in profit margins by optimizing prices to find the balance between maximizing sales and maintaining profitability.
* Customer Satisfaction: Gather feedback from customers to gauge their satisfaction with the pricing strategies implemented. This can be measured through surveys, reviews, or customer retention rates.
* Market Share: Assess the growth in market share as a result of effective price optimization, indicating the ability to attract and retain customers.
* Competitor Comparison: Compare the pricing strategies and market position against competitors to evaluate the effectiveness of the price optimization efforts.

**Cite references**

NA