# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | *Enhancing Customer Satisfaction and Reducing Churn at Jade Inc. through Comprehensive Data Analysis* |
| **Requestor** | *Mr. Bilal* |
| **Date of Request** | *2024-07-02* |
| **Target Quarter for Delivery** | *2* |
| **Epic Link(s)** | <https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction> |
| **Business Impact** | *Reducing Churn caused due to Customer Dissatisfaction* |

About the Company

Jade Inc. is a leading e-commerce company that prides itself on delivering a seamless online shopping experience to its diverse customer base. Despite its success, the company is facing challenges in maintaining high levels of customer satisfaction, a critical factor for fostering customer loyalty and sustaining business growth. Our current analysis aims to identify key factors influencing customer satisfaction using a comprehensive dataset, which includes variables such as customer tenure, preferred login devices, satisfaction scores, and purchase behaviors. By examining these factors, we will propose actionable insights to enhance service quality and customer experience, ensuring Jade Inc. continues to thrive in the competitive e-commerce landscape.

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

**The specific ask:**

Analyze the dataset of **Jade Inc.** to identify key factors influencing customer satisfaction and propose actionable insights to enhance service quality and customer experience.

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

**Research 1: Factors influencing satisfaction and loyalty in online shopping: an integrated model**

This research examines how external factors like technology acceptance and website service quality, alongside internal factors like specific holdup cost, influence customer satisfaction and loyalty. Findings indicate that customer e-satisfaction directly impacts e-loyalty, while technology acceptance and website service quality positively influence both e-satisfaction and e-loyalty. However, specific holdup costs impact e-loyalty directly but not e-satisfaction. This highlights the importance of enhancing website functionality and service quality to foster customer loyalty​ (Lin & Sun, 2009).

**Research 2: Click to smile: Mastering customer satisfaction in ecommerce**

In the competitive e-commerce landscape, customer satisfaction (CSAT) is critical for business growth. Studies show that satisfied customers drive repeat purchases and act as brand ambassadors through positive reviews and recommendations. Research by Salesforce reveals that 91% of customers repurchase from companies after a positive service experience. Additionally, personalized experiences significantly boost satisfaction and revenue, with McKinsey & Company reporting a 40% revenue increase from personalized customer interactions. Efficient website user experience, product presentation, smooth checkout processes, and strong post-purchase communication are essential to maintaining high satisfaction levels (DHL, 2024).

**Key Competition**

Key competitors in the e-commerce market include Amazon, eBay, and Walmart. These companies offer seamless user experiences, diverse product selections, and robust customer service. Amazon, for example, dominates with approximately 38% of the U.S. e-commerce market share and is known for its Prime membership benefits, personalized recommendations, and fast shipping. eBay focuses on a vast marketplace with auctions and direct sales, while Walmart leverages its extensive physical presence to enhance its online offerings with services like same-day delivery and in-store pickup. These companies emphasize customer satisfaction through innovative solutions, extensive product ranges, and exceptional service

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

**Project Gains**

* **Revenue Gains:** Improved customer satisfaction can lead to increased repeat purchases and customer lifetime value.
* **Quality Improvements:** Enhancing service quality based on satisfaction factors can lead to higher customer retention.
* **Cost and Time Savings:** Reducing churn can lower acquisition costs and increase operational efficiency.

**Implications of Inaction:**

* Decline in customer retention and satisfaction.
* Increased operational costs due to higher churn rates.
* Loss of competitive edge in the market.

## *Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.*

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

**Key Questions:**

* What are the primary factors influencing customer satisfaction?
* How do different demographic variables affect satisfaction?
* What operational changes can improve customer satisfaction scores?

**Hypotheses:**

* Tenure with the company positively influences satisfaction.
* Higher engagement on the app leads to higher satisfaction scores.
* Complaints negatively impact satisfaction.

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

**Assumptions:**

* The data is accurate and up-to-date.
* Customer satisfaction scores are a reliable measure of overall satisfaction.
* All relevant factors affecting satisfaction are included in the dataset.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |
|  | **Key Performance Indicators (KPIs):**   * Increase in average customer satisfaction score. * Reduction in the number of customer complaints. * Improvement in repeat purchase rates. |
|  |  |

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

**Type of Analysis:** *logistic regression, linear regression, Chi-square test*

*The initial approach will be to use a decision tree to determine which dealer level variables (size, region, segmentation...) are most significant related to a dealer’s likelihood to churn. We will also use other techniques to verify our findings.*

The type of analysis used in this project would be:

* 1. **Exploratory Data Analysis (EDA)**: I'll start with EDA to get a solid understanding of the dataset. This will involve creating visualizations like histograms, box plots, and correlation matrices to spot trends and outliers.
  2. **Predictive Modeling**: Techniques such as Random Forests or Gradient Boosting Machines will be used to predict customer churn and satisfaction levels.
  3. **Linear Regression**: This method can analyze the relationship between continuous variables, such as how the amount of time spent on the app, or the number of devices registered affects customer satisfaction scores. It helps in understanding the linear relationship between these factors.
  4. **Logistic Regression:** To create a forecast that is more reliable and accurate, a Random Forest ensemble learning technique generates several decision trees during training. Predictions are made by averaging the findings (in regression) or selecting the majority vote (in classification), with each tree in the forest being trained on a random subset of the data. This methodology improves robustness and lessens the overfitting issue that is frequently linked to individual decision trees.
  5. **Random Forest:** To create a forecast that is more reliable and accurate, a Random Forest ensemble learning technique generates several decision trees during training. Predictions are made by averaging the findings (in regression) or selecting the majority vote (in classification), with each tree in the forest being trained on a random subset of the data. This methodology improves robustness and lessens the overfitting issue that is frequently linked to individual decision trees.
  6. **Decision Trees:** For both classification and regression tasks, a supervised learning approach called a decision tree is employed. It is a decision tree model that shows potential outcomes, resource costs, and utility of various choices
  7. **Support Vector Machine:** One popular supervised learning approach for classification tasks is the Support Vector Machine (SVM), which can also be used for regression. The main goal of SVM is to locate the hyperplane in a high-dimensional space that best divides the data points of various classes. The objective is to produce the strongest possible classification borders by maximizing the margin between the closest points of the two classes, or support vectors.
  8. **Gradient Boosting Machine:** Gradient Boosting is an ensemble technique wherein models are built one after the other, with each model fixing the mistakes of its predecessor. Both regression and classification problems frequently make use of it. The model is constructed iteratively, with decision trees usually serving as weak learners, and each new model is trained to forecast the residuals (errors) of the preceding models. The method is iterated until the model achieves the required degree of accuracy, with the aim of minimizing a certain loss function.
  9. **K-Nearest Neighbours:** For applications involving regression and classification, the K-Nearest Neighbors (KNN) algorithm is an easy-to-understand technique. The majority class of a data point's closest neighbors is used to classify it. Within KNN, the "K" stands for the number of nearest neighbors taken into account. The majority class label of the K nearest points is assigned after the algorithm calculates the distance between the test data point and all training data points.
  10. **Neural Network:** An algorithmic system called a Neural Network is made up of layers of connected "neurons" that resemble the structure of the human brain in order to identify patterns. Complex tasks like natural language processing, image and audio recognition, and other high-dimensional data challenges are where it excels. Every neuron takes in information, processes it, and then sends the result to the layer above. This process gradually changes the data into a format that may be used to anticipate the desired variable.

**Methodology:** *Key questions from ‘Analytics objective’ will be tackled in ascending order as outlined in ‘5.0 Timelines and deliverable section’.*

*We will start by identifying all dealers that were active in the first quarter of 2018. We will then define the response variable to be a 1 if they are still active, and 0 otherwise. We will build a decision tree based on this sample, and observe which variables are the most important in determining whether these dealers are still active. We can then repeat this analysis using a sample based on the dealers that were active in the second quarter of 2018. The idea is to check if the same variables are being identified as the most important drivers of churn, or if the importance of variables change as we get closer to the present day.*

**Methodology:**

The methodology stated below follows a chronological order

1. **Data Preparation:** Clean and preprocess data to handle missing values and outliers.
2. **Exploratory Data Analysis (EDA):** Identify patterns and relationships in the data.
3. **Modeling:** Build predictive models using linear regression and decision trees. Using clustering to segment customers in bins based on their segments.
4. **Validation:** Validate models using cross-validation techniques.
5. **Insights & Recommendations:** Derive actionable insights and propose recommendations.

**Output:** *The output will be a set of insights, rules and strategic recommendations that will help us to evaluate dealers based on likelihood to churn and positioning of sales-match.*

Using the methodology given above, we can get key insights into the drivers of customer satisfaction. Additionally, it’ll also help in providing strategic recommendations to improve service quality.

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection:** Audience are the executives while the population are the customers of the company

**Observation window:** Data from the year 2022-23

**Inclusions:** All customers with a satisfaction score

**Exclusions:** IDs of customers

**Data Sources:** Internal data of the company regarding customer transactions and customer feedback surveys.

**Audience Level:** Upper Management.

**Variable Selection:  Target Variable:** Churn

** Predictors:**

* Tenure
* PreferredLoginDevice
* CityTier
* WarehouseToHome
* PreferredPaymentMode
* Gender
* HourSpendOnApp
* NumberOfDeviceRegistered
* PreferedOrderCat
* MaritalStatus
* Complain
* OrderAmountHikeFromlastYear
* CouponUsed
* OrderCount
* DaySinceLastOrder
* CashbackAmount

**Derived Variables:**

* **None**

**Assumptions and data limitations:**

* Potential bias in self-reported satisfaction scores.
* Data is complete

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| Risk | Likelihood (based on historical data) | Delay (based on historical data) | Impact |
| *Churn rate being inflated by counting multiple contracts from the same rooftop as individual observations.* | *Low* |  | *Once analysis begins, we can quantify the inflation. However, this approach allows us to compare how the same dealer performed across different contracts and find useful patterns.* |
| *Data Accuracy and Completeness* | *Low* |  | *There is no certainty about the data as it can be biased and incomplete.* |
| *Changes in customer behavior over time.* | *High* |  | *There is no certainty if the customer satisfaction score is relevant in the current market scenario as it changes with the time.* |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | Kick-off / Formal Request | Initiate the project with a formal request and kick-off meeting | Define project goals, deliverables, and timelines | 2 | *2024.06.08* |
| 2. | Assessment / Triage | Initial assessment of the dataset and project requirements | Evaluate data quality, identify initial issues, and prioritize tasks | 3 | *2024.06.18* |
| 3. | Prioritization | Prioritize tasks based on the assessment and project goals | Create a priority list of tasks to address data issues and analysis | 2 | *2024.06.20* |
| 4. | Data Exploration & Analysis   * Issues with duplicates * Issues with Spend data | Explore and analyze the dataset | Handle duplicates, assess spend data issues, derive variables | 10 | *2024.07.22* |
| 5. | Story Board 1 | |  | | --- | |  |   Create the first storyboard outlining initial findings and plans | Present initial insights and proposed analysis framework | 3 | *2024.07.25* |
| 6. | QA Output | Quality assurance review of the analysis output | Ensure accuracy and reliability of the analysis results | 2 | *2024.07.26* |
| 7. | Internal team Presentation | Present findings to the internal team | Share progress and gather feedback | 2 | *2024.07.29* |
| 8. | Go/No Go | Decision point for proceeding with the next phase | Evaluate if the project is ready to move forward | 1 | *2024.08.05* |
| 9. | Story Board 2 | Create a refined storyboard with updated insights and recommendations | Present refined insights and actionable recommendations | 3 | *2024.08.15* |
| 10. | Pilot | Implement a pilot based on the analysis findings | Test the proposed solutions on a small scale | 7 | *2024.08.16* |
| 11. | Delivery & sign-off | Final delivery of the project and obtain sign-off | Present final report, insights, and action plans | 2 | *2024.08.17* |

|  |
| --- |
|  |