CRISP-DM LIFECYCLE DOCUMENTATION FOR A RETAIL DATASET

CRISP-DM is a process model that describes common approaches used by data mining experts. It is an open standard and the most widely used analytics model. This document contains models that describes the lifecycle of a data science project that is intended to solve a retail business problem. This project is broken into 6 distinct phases starting from the inception of the project all the way to its final impact. The project runs on a CRISP-DM (Cross-Industry Standard Process for Data Mining) lifecycle. The key phases for the project (Figure 1): Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. In the documentation below, you will find detailed information on each of the phases, practical information (code implementation), important considerations, ethical and moral policies to be undertaken.

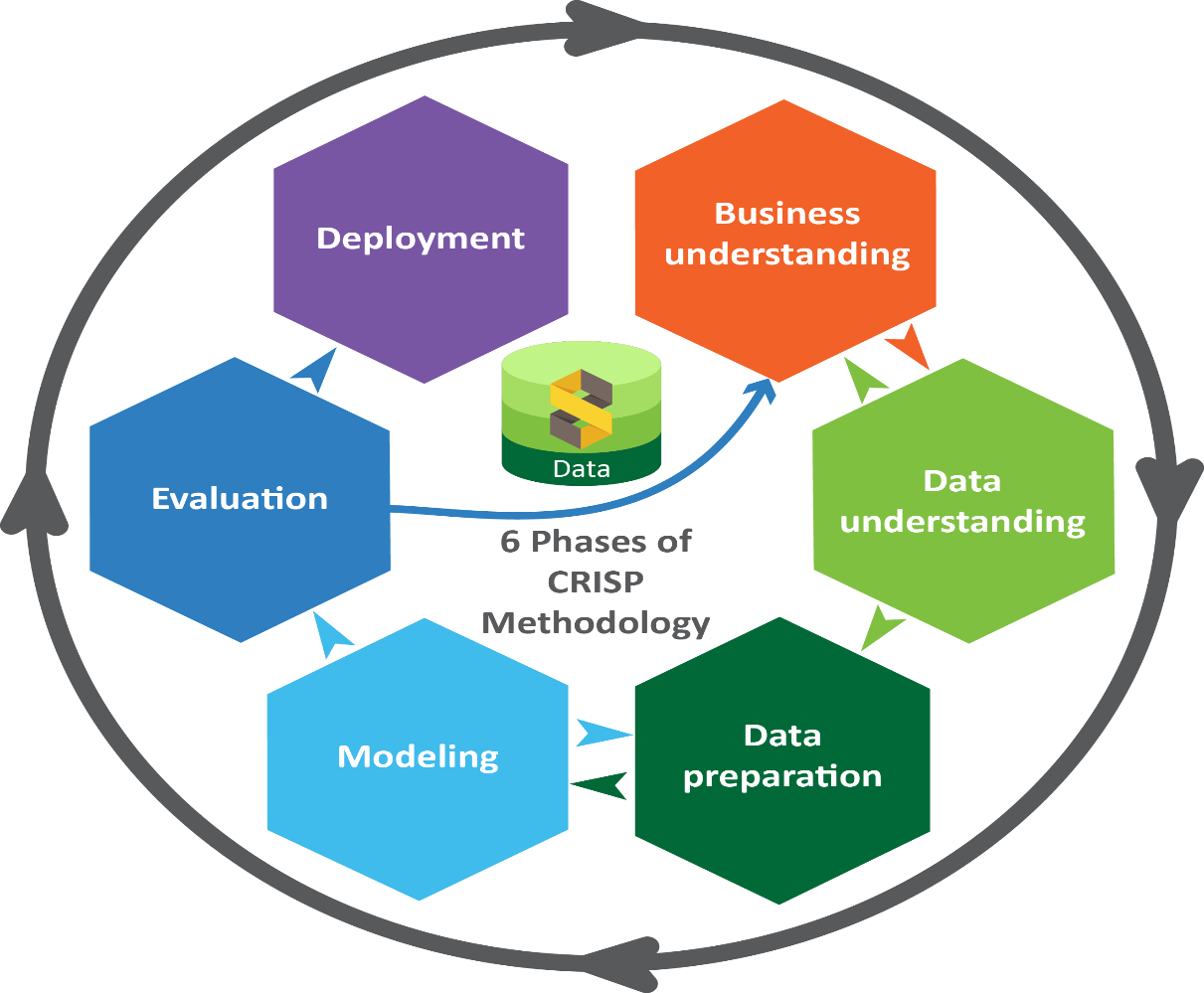


Figure 1: Phases of the CRISP-DM

**PHASE ONE: BUSINESS UNDERSTANDING**

Project Overview: Customer Sales Analysis Using Machine Learning and Tableau

Team B—consisting of Rebecca Marriott, Nouman Mehar, and Suneeta Vota—undertook a data science project focused on analysing customer purchasing behaviour. The core aim of this project was to leverage machine learning and visualization tools to derive actionable insights from customer sales data. By combining data modelling with Tableau dashboards, the team aimed to understand patterns in purchasing behaviour and enhance the personalization of customer experiences within Enterprise Resource Planning (ERP) systems.

**Project Purpose and Objectives**

The overarching purpose of the project was to utilize data science techniques within ERP systems to enrich and personalize customer engagement. The project targeted the analysis of customer purchase histories, behavioural traits, and demographic data to uncover insights that can lead to improved marketing effectiveness, personalized product recommendations, and heightened customer satisfaction.

A key objective of the project was to predict customer gender based on behavioural and transactional data. Accurate gender prediction could support more tailored marketing strategies and customer segmentation. To achieve this, the team employed machine learning techniques aimed at uncovering patterns and associations within a rich dataset of customer interactions.

Dataset Overview

The dataset consisted of 302,010 records and included 30 distinct features. The data was structured across three primary dimensions:

Customer Information: Customer ID, name, contact details, age, gender, income level, and market segment (Premium, Regular, or New).

Transaction History: Data on transaction IDs, last purchase dates, prior purchase amounts, customer feedback, and preferences in shipping and payment methods.

Product Attributes: Product category, brand, and type, including items like electronics, clothing, and groceries.

In addition, the dataset included geographic indicators (USA, UK, Canada, Australia, and Germany) and temporal attributes such as the year, month, date, and time of purchases, enabling deep trend analysis.

Data Preparation and Challenges

Data preprocessing, led by Nouman Mehar, was a critical phase of the project. The dataset exhibited common real-world challenges such as null values and duplicate entries. Addressing these was essential to improving the reliability and performance of the predictive models.

One of the major issues encountered was class imbalance in the gender label. The team employed SMOTE (Synthetic Minority Over-sampling Technique) to address this imbalance and improve model fairness. Careful evaluation using precision-recall curves was used to ensure the models weren’t simply overfitting to the dominant class.

**PHASE TWO: DATA UNDERSTANDING**

Acquiring, storing, linking, understanding data for analysis on a project often entails an involved and iterative process, requiring working closely with the team to work on the choice of the dataset to ensure that the dataset collected and chosen answers the business objectives. Consistency and reliability in the data format was essential to ensure protection of private and sensitive information. During this phase of the work, the team was required to apply skills working with and structuring data to get it into a storage format that is appropriate for linking it with the data sources. Each of the steps require active communication with the project’s stakeholders to understand the context in which the data were collected and structured to ensure data definitions actually describe the events they are supposed to reflect.

**Data Collection:**

**Data Source:**

The data was collected through teamwork. We used Kaggle, which is a platform for finding datasets, participate in data science competitions, collaborate with other data scientists etc.

The data was collected through teamwork. We used Kaggle, which is a platform for finding datasets, participate in data science competitions, collaborate with other data scientists etc.

The title of our dataset is: Retail Transaction Dataset

**The link to access this dataset:**

https://www.kaggle.com/datasets/prasad22/retail-transactions-dataset/data

Data Files: Retail Transactions Dataset.csv

**Data Quality Issues:**

* In our dataset we did not have any outliers or duplicates.
* Missing values in Promotion column.
* Inconsistent customer names.

**ABOUT THE DATASET**

This dataset was created to simulate a market basket dataset, providing insights into customer purchasing behaviour and store operations. The dataset facilitates market basket analysis, customer segmentation, and other retail analytics tasks.

Dataset Information (Attributes and Variables):

The columns provide information about the transactions, customers, products, and purchasing behaviour, making the dataset suitable for various analysis, including market basket analysis and customer segmentation. Here's a brief explanation of each column in the Dataset:

**PHASE THREE: DATA PREPARATION**

Data preparation, often referred to as “data wrangling” or “data munging,” is a vital step to ensure that the dataset is in optimal shape before modelling. This phase was critical in addressing the quality and usability of the retail dataset.

**Key Tasks:**

1. Handling Missing Values
   * **The Promotion column had over 333,943 missing values. The team considered several strategies:**
     + Removal: Dropping rows would significantly reduce the dataset size.
     + **Imputation:**
       - Mode Imputation for categorical values like promotion status.
       - Mean/Median Imputation for numeric columns (if applicable).
     + Final decision: Filled missing promotion values with “None”, treating them as a lack of promotion.
2. Dealing with Duplicates
   * The dataset was scanned using pandas. duplicated() function.
   * No duplicates were found and thus no rows were removed.
3. Correcting Inconsistencies
   * **Inconsistent entries in customer names were standardized using:**
     + Lowercasing
     + Stripping whitespace
     + Removing special characters where necessary
4. Encoding Categorical Variables
   * Gender (target): Encoded as binary (0 for Male, 1 for Female)
   * Market segment, Payment method, Product category: One-Hot Encoding
   * Geographic data (Country): Label encoded or One-Hot depending on model compatibility.
5. Feature Engineering
   * **Derived new features:**
     + Recency: Days since last purchase
     + Total Purchases: Aggregated count per customer
     + Customer Lifetime Value: Aggregated spend.
   * Temporal features (year, month, day) were extracted to identify seasonal patterns.
6. Normalization and Scaling
   * Applied Min-Max Scaling to numeric columns (e.g., age, income, amount spent) for algorithms sensitive to magnitude (KNN, Logistic Regression).

**PHASE FOUR: MODELLING**

Modelling

We tested the performance of numerous models for gender prediction. We evaluated each model using accuracy and F1-score, for both male and females – as class imbalance was a significant challenge in our dataset.

**Our target variable was ‘gender’ (categorical value), we tested the following models on our dataset:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1-Score (Male)** | **F1-Score (Female)** | **Notes** |
| **Logistic (raw)** | 62% | 0.00 | 0.77 | Biased to Female |
| **Logistic (balanced)** | 52% | 0.45 | 0.57 | Better class balance |
| **SMOTE + Logistic** | 53% | 0.44 | 0.59 | Slight improvement |
| **Random Forest** | 59% | 0.22 | 0.73 | Struggles with Males |
| **XGBoost** | 56% | 0.01 | 0.77 | Worst male recall |
| **KNN (k=3)** | 55% | 0.36 | 0.65 | Balanced but moderate |

As you can see, the Logistic Regression (raw) model was biased towards predicting females. For this reason, we tried the balancing the Logistic Regression. This resulted in the accuracy to drop by 10% to 52%, however, the performance was more balanced, with the F1-score for male increasing to 0.45 and for females decreasing slightly to 0.57. This trade-off shows that that the class weighing helped.

We tried to balance this further by using SMOTE, but this resulted in a very slight improvement of 1%.

The Random Forest model struggles with male predictions but had a decent accuracy of 59%.

XGBoost had a F1- score of just 0.01 for males.

The KNN (4=3) model had a moderately balanced performance for males and females and moderate accuracy of 55%.

Below we used a bar chart to compare just some of the models with

A graph of different colored bars

AI-generated content may be incorrect.

Visuals for Logistic Regression – Balanced

Precision-Recall Curve - Important when one class dominates.

A graph of a line graph

AI-generated content may be incorrect.

Value counts: Gender

**Male 187596**

**Female 114093**

This evaluates how well the model performs, especially when dealing with **imbalanced classes**.

**AP (Average Precision)**: The area under this curve, **AP = 0.69** in this case - the model maintains **moderate precision** across varying levels of recall.

* This is significantly better than random.
* It suggests your model has some predictive power, even if the ROC AUC was low (0.54).

ROC Curve - Especially useful for class imbalance — this shows performance trade-offs.

A graph with a line

AI-generated content may be incorrect.

This shows the model's ability to distinguish between classes (e.g., predicting gender).

**AUC Score = 0.54**

* **AUC (Area Under Curve)** ranges from 0 to 1:
* **0.5**: Random performance.
* **1.0**: Perfect classification.
* **< 0.6**: Very poor performance.
* **0.54**: This is **only slightly better than random**. The model **is not effectively distinguishing** between the classes.

Confusion Matrix - actual vs. predicted class counts

The model is **heavily biased toward predicting “Female (1)”**

It **correctly identifies most females**, but **struggles with males** — often misclassifying them as female

Out of **22,819 actual males**, only **3,485 were correctly predicted** → that’s **very poor performance for class 0 (Male**

The model is much better at identifying **Female (1)** than **Male (0)**

A blue squares with numbers and a number on them

AI-generated content may be incorrect.

Visuals for Random Forest

Feature Importance Plot - This will show which features are driving the model's predictions.

A graph with blue and white stripes

AI-generated content may be incorrect.

This ranks features based on their importance in the Random Forest model.

Higher values mean the feature contributes more to the model's predictions.

* The most important features are: Zipcode, Customer\_ID, Amount, Total\_Amount.
* We should consider dropping Customer\_ID to see what difference it makes to the model’s ability to predict gender correctly.

ROC Curve - Especially useful for class imbalance — this shows performance trade-offs.

This shows the model's ability to distinguish between classes (e.g., predicting gender).

**AUC Score = 0.54**

* **AUC (Area Under Curve)** ranges from 0 to 1:
* **0.5**: Random performance.
* **1.0**: Perfect classification.
* **< 0.6**: Very poor performance.
* **0.54**: This is **only slightly better than random**. The model **is not effectively distinguishing** between the classes.

A graph with a line

AI-generated content may be incorrect.

Precision-Recall Curve - Important when one class dominates.

A graph of a line

AI-generated content may be incorrect.

It evaluates how well your model performs, especially when dealing with **imbalanced classes**.

**AP (Average Precision)**: The area under this curve, **AP = 0.69** in this case - the model maintains **moderate precision** across varying levels of recall.

* This is **significantly better than random** (which usually gives a low AP, often near the positive class ratio).
* It suggests your model has some predictive power, even if the **ROC AUC was low (0.54)**.

Misclassification Analysis - bar chart of correct vs. incorrect predictions

A graph of a bar graph

AI-generated content may be incorrect.

**Correct Predictions**: Represented by the left bar, with a height around **36,000**. This indicates the number of samples where the Random Forest model predicted gender correctly

**Incorrect Predictions**: Represented by the right bar, with a height around **24,000**. This indicates the number of misclassified samples—i.e., where the predicted gender did not match the actual gender

The Random Forest model made **more correct predictions than incorrect ones**, suggesting it performs reasonably well.

Confusion Matrix - actual vs. predicted class counts – Random Forest

A diagram of a graph

AI-generated content may be incorrect.

The **confusion matrix** shows the performance of the **Random Forest** model on gender classification where:

* **0 = Male**
* **1 = Female**

**True Positives (TP)**: 32,393 — Female correctly predicted as Female

**True Negatives (TN)**: 3,485 — Male correctly predicted as Male

**False Positives (FP)**: 19,334 — Male incorrectly predicted as Female

**False Negatives (FN)**: 5,126 — Female incorrectly predicted as Male

High accuracy for Females: Recall = 86.3%

Poor performance for Males: Recall = 15.3%

Visuals for KNN k=3

Confusion Matrix - actual vs. predicted class counts – KNN k=3

A blue squares with white text

AI-generated content may be incorrect.

**Accuracy:** **55%**  
This means the model predicted the correct gender **55% of the time** on the test set.

This shows the model is **relatively balanced**, but its predictive power is **weak for both classes**, especially due to the high number of false positives/negatives.

Classification Report – Heatmap

A graph with blue squares

AI-generated content may be incorrect.

**Class 0 = Male**

* **Precision**: 0.39 (many false positives)
* **Recall**: 0.51 (barely half the actual males are correctly predicted)
* **F1-score**: 0.44 (low overall performance)

**Class 1 = Female**

* **Precision**: 0.63 (relatively better)
* **Recall**: 0.52
* **F1-score**: 0.57
* The model is **moderately biased towards predicting females.**

**Tableau**

CUSTOMER SPENDING ANALYSIS DASHBOARD

Insights by Gender, Income, Category, and Time

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Gender Distribution**

Pie chart shows an unequal distribution between Male and Female consumers.

* Balanced representation ensures fair analysis across genders.
* Important for marketing and segmentation strategies.

A screenshot of a computer

AI-generated content may be incorrect.

**Spending Over Time**

* Timeline from 2023 to 2024.
* Observation:
  + High spending peak in early 2023, followed by a dip and gradual increase into 2024.
  + Males have slightly higher total amounts towards 2024.
* Implications:

Possible seasonal effects or market events.

A screenshot of a computer

AI-generated content may be incorrect.

**Amount by City**

* Map visualisation shows spending concentration by city.
* Observation:
  + Major cities have higher spending (indicated by larger circles).
  + London and surrounding areas show a dominant share.
* **Product category preferences segmented by gender**
* Males outnumber females in all the listed product categories, however, if we consider the gender imbalance, then we can assume that there is no significant difference between male and female engagement.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Age vs Income Analysis**

* Scatter plot analysis with filters for Gender and Income Levels.
* Key insights:
  + High-income individuals mostly spend around the 150K mark.
  + Younger individuals (30–35 age group) contribute significantly to spending.

**Key Takeaways**

* Gender preferences vary significantly by product category.
* Spending over time shows a recovery trend in 2024.
* Income and age play vital roles in determining spending behaviour.
* Regional spending insights can support targeted campaigns.

# **PHASE FIVE EVALUATION OF KEY POINTS**

# Model Performance Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1-Male** | **F1-Female** | **Strengths** | **Weaknesses** |
| Logistic (Raw) | 62% | 0 | 0.77 | High precision for Females | Severely biased, fails to detect Males |
| Logistic (Balanced) | 52% | 0.45 | 0.57 | Better class parity | Accuracy trade-off |
| SMOTE + Logistic | 53% | 0.44 | 0.59 | Slight performance boost with SMOTE | Still underperforms for Male class |
| Random Forest | 59% | 0.22 | 0.73 | Decent overall accuracy | Weak recall for Males |
| XGBoost | 56% | 0.01 | 0.77 | Strong on dominant class | Extremely weak for minority class (Males) |
| KNN (k=3) | 55% | 0.36 | 0.65 | Balanced performance | Moderate across all metrics |

Visual Evaluations:  
- Precision-Recall Curve: AP = 0.69 (moderate). Models perform better than random chance but struggle with minority class.  
- ROC Curve: AUC = 0.54 for most models — only slightly better than guessing.  
- Confusion Matrix: Consistent male misclassification, e.g., only 3,485 out of 22,819 Males correctly predicted by Random Forest.  
  
Final Thoughts:  
- Model Bias: Persistent bias against Male predictions due to class imbalance.  
- Data Quality: Promotion column had major missingness affecting downstream model predictions.  
- Recommendations:  
 - Drop Customer\_ID as a feature to avoid data leakage.  
 - Explore deeper models or ensemble methods.  
 - Revisit feature selection and engineering to improve discriminatory power.

**PHASE SIX: DEPLOYMENT**

**Deployment Plan:**

* Integrate the recommendation system into the retail website/app.
* Use customer segments to tailor marketing campaigns.
* Implement sales prediction model to optimize inventory management.

**Monitoring and Maintenance:**

* Continuously monitor model performance and update models with new data.
* Gather feedback from users to improve recommendation system.
* Regularly retrain models to adapt to changing customer behaviour and market trends.

**6.3 Documentation and Reporting:**

Use and develop a dashboard.

