Predicting Customer Churn in Online Retail: A CRISP-DM Report

1. Business Problem

1.1 Introduction and Objective

Customer attrition erodes revenue and inflates marketing spend. Empirical studies indicate that acquiring a new buyer can cost five-to-twenty-five times more than keeping an existing one hbr.org. The Online Retail Customer Churn Dataset contains 5630 observations and twenty demographic-behavioural variables; roughly 17% of records are labelled as churned kaggle.com.

Our objective is to classify current customers into churners and non-churners using behavioral and transactional data so that targeted retention strategies can be applied to those identified as churn customers. We cast the task as binary classification, train Logistic Regression, Decision Tree, and Random Forest models, and judge them with F1-score metrics that are robust to class imbalance.

1.2 CRISP-DM Phase 1: Business Understanding

The business problem states straightforwardly: maximise net lifetime value by reducing voluntary churn. A churn event removes all future cashflows from a customer and forces the firm to invest in fresh acquisition. Our model becomes a decision instrument. Management sets an intervention threshold; when the predicted risk rises above that line, the CRM system triggers an e-mail incentive or personal contact. The threshold itself is chosen through a cost–benefit analysis that weighs the margin restored by a successful save against the expense of unnecessary incentives. This cost-sensitive framing ties model performance directly to profit and guards against the temptation to optimise statistical precision in isolation from economics.

1.3 CRISP-DM Phase 2: Data Understanding

The Kaggle file mixes static descriptors (age, marital status, region) with transactional signals (tenure, days-since-last-order, order count, cashback). Three structural facts:

1) Class imbalance—about 17 % positive labels—motivates F1 and AUC over naïve accuracy.
2) Right-skew in several continuous fields (*Warehouse-to-Home*, *OrderCount*) suggests log or quantile transforms for linear models.
3) Missing values cluster in optional survey items; median/mode imputation preserves sample size.

2. CRISP-DM Phase 3: Data Processing & Modeling

The process of preparing the customer churn dataset for model testing included initial data integrity checks, data type adjustments, and feature transformations, culminating in a cleaned and scaled dataset. See the steps below:

- **Initial Data Inspection**: The dataset was loaded and examined for the presence of missing values and duplicate rows.
- **Missing Value and Duplicate Handling:** It was confirmed that the dataset contained no missing values and no duplicate rows, thus no imputation or duplicate removal was required.
- **Data Type Conversion:** The boolean columns 'Email_Opt_In' and 'Target_Churn' were converted to integer data types (0 or 1).
- **Categorical Feature Encoding:** One-hot encoding was applied to the nominal categorical features, specifically 'Gender' and 'Promotion_Response', to transform them into a numerical representation. This process created new columns for each unique category within these features (e.g., 'Gender_Female', 'Gender_Male', 'Gender_Other', 'Promotion_Response_Ignored', etc.).
- **Numerical Feature Scaling:** Numerical features were scaled using two different methods:
- **Standardization (StandardScaler):** Transformed numerical features to have a mean of 0 and a standard deviation of 1.
- **Normalization (MinMaxScaler):** Scaled numerical features to a fixed range, typically between 0 and 1.

3. Model Building & Comparison

3.1 Logistic Regression

1) Description of the Model

Logistic Regression is a linear classification algorithm commonly used for binary outcomes. It models the probability of a binary target variable using a logistic function. While simple and interpretable, it assumes linear relationships between features and the log-odds of the outcome. In our case, Logistic Regression acts as a baseline model to compare against more complex classifiers like Random Forest. The advantage of choosing logistic regression lies in efficiency, low variance, and transparency of feature coefficients.

2) Model Building Process

1. Data Preprocessing: Non-informative columns such as Customer_ID were dropped. The label column Target_Churn was isolated as the response variable y.
2. Train-Test Split: The dataset was split 80-20 into training and testing sets using a fixed random_state for reproducibility.
3. Feature Scaling: All numerical features were standardized using StandardScaler to ensure fair weightage in the model.
4. Model Training: A LogisticRegression model was instantiated with max_iter=1000 to ensure convergence and trained on the scaled training data.
5. Evaluation: Predictions were made on the test set and evaluated using accuracy, precision, recall, F1 score, and RMSE.

6. Model Registration: The trained model was registered to jrjModelRegistry using registerAJrjModel() with appropriate metadata, transformer, and predictor functions.

3) Metrics Used & Results

The following classification metrics were used to assess the model's performance:

- Accuracy (0.465): Out of all predictions made, 46.5% were correct. This measures how often the model got it right overall.
- Recall (0.660): The model identified 66.0% of true churners, meaning among all customers who actually churned, 66.0% were successfully identified, making it relatively sensitive to churn cases.
- Precision (0.496): Among all customers the model predicted as churners, 49.6% really were. This indicates some false positives but reasonable targeting accuracy.
- F1 Score (0.567): An F1 score of 0.567 shows the model is fairly balanced — it catches many churners without raising too many false alarms.
- RMSE (0.535): The root mean square error reflects overall prediction error and was included for completeness.

4) **Insights & Business Implications**

- The model is strong at catching customers likely to churn (eg. high recall), but less precise, which means the company might reach out to some customers who weren't planning to leave.
- Many churners identified had low satisfaction scores, high return counts, or had not purchased for over 200 days. These numbers show early warning signs.
  This model can serve as a first filter to flag risky customers and should be paired with business rules.

3.2 Decision Tree

1) Description of the Model

Decision Tree is a non-parametric, hierarchical classification method that recursively splits feature space into homogeneous subgroups. In our churn context, a balanced-class DecisionTreeClassifier captures nonlinear interactions between customer attributes (e.g. recency, frequency, monetary) and churn risk.

2) Model Building Process

- Dropped Customer_ID and isolated Target_Churn as the label y.
- One-hot encoded flags (e.g. Gender_Female, Email_Opt_In)

2. Train/Validation/Test Split

- Stratified 80/20 split into train+validation (800 rows) and test (200 rows).
- Further 75/25 split of train+validation into training (600 rows) and validation (200 rows).

3. Baseline Training

- Fitted DecisionTreeClassifier(class_weight='balanced', random_state=42) on the 600-row training set.
- Evaluated on validation to establish baseline metrics.

4. Hyperparameter Tuning

- Ran 5-fold GridSearchCV over:

- *max_depth ∈ {None, 3, 5, 7} / min_samples_leaf ∈ {1, 5, 10} /min_samples_split ∈ {2, 5, 10}*
- Scored on precision, recall, and F1; refit the best F1 model.

5. Feature Selection
- Computed importances from the tuned tree and selected the six features above the mean importance:*Age,Total_Spend,Years_as_Customer,Num_of_Purchases,Average_Transaction_Amount,Last_Purchase_Days_Ago*

6. Final Training & Registration
- Retrained DecisionTreeClassifier(**best_params_, class_weight='balanced') on train+validation using only the six selected features.

3) Metrics Used & Results

- Accuracy (49.5 %): Overall fraction of correct labels.
- Recall (77.1 %): Of all true churners, 77.1 % are flagged. High recall ensures most at-risk customers receive retention outreach.
- Precision (51.3 %): Among those predicted to churn, 51.3 % actually do. Roughly half of "at-risk" alerts convert, implying some wasted outreach.
- F1 Score (61.6 %): Harmonic mean of precision and recall.

4) Insights & Business Implications

- Maximize churn coverage: With 77 % recall, the model flags most at-risk customers—valuable for proactive retention.
- Control outreach costs: 51 % precision means half of interventions go to loyal customers. To reduce waste:

1. Raise classification threshold for higher precision (accept some drop in recall).
2. Use low-cost channels (email/SMS) for broad alerts and reserve high-touch efforts for top-probability churners.

3.3 Random Forest

1) Description of Model

Random Forest is a robust, ensemble-based method that handles both numerical and categorical features well, avoids overfitting through bagging, and is highly interpretable through feature importance. It performs well on complex classification tasks, especially with moderate data size and mixed-type input features, making it a good choice for churn prediction.

2) Model Building Process

1. Data Preprocessing: Unnecessary columns like Customer_ID were dropped. The target column Target_Churn was separated as the label y.
2. Train-Test Split: The data was split into training and testing sets using an 80-20 ratio with a fixed random_state for reproducibility.
3. Model Training: A basic RandomForestClassifier was initialized with default settings and trained on the training data.
4. Evaluation: The model was evaluated on the test set using accuracy, recall, precision, and F1 score.
5. Optimize model: Choose the most important features (top 8) to try to optimize the model.

6. Model Registration: The model was registered using registerAJrjModel() with metadata.

3) Metrics Used & Results

We used four classification metrics to evaluate model performance and support business needs:

- Accuracy (0.540): The proportion of all correct predictions. The model correctly predicted churn 54% of the time.
- Recall (0.632): The ability to identify actual churners. Of all customers who actually churned, 66% were successfully identified by the model.
- Precision (0.558): The proportion of churn predictions that were correct. Among all customers predicted to churn, 55.8% actually did.
- F1 Score (0.593): The harmonic mean of precision and recall. This score balances both metrics, and a value of 0.593 indicates the model achieves a fair compromise between capturing churners and minimizing false alarms.

4. Insights and Ethical reflections

4.1 Logistic Regression

This foundational model, valued for its simplicity and efficiency, set the framework for our assessments. While it achieved a recall of 0.660 in identifying true churners, indicating decent sensitivity, its overall accuracy was lower at 0.465, with precision at 0.496 suggesting a notable rate of false positives. The F1 score of 0.567 highlighted a need for improved balance between precision and recall. Ethically, its linear nature makes it susceptible to amplifying existing biases in training data, potentially leading to unfair or ineffective targeting. The high rate of false positives could also lead to customer annoyance from misdirected retention efforts, and its simplicity might oversimplify complex churn drivers, raising concerns about the completeness of insights used for decision-making.

4.2 Random Forest

This model showed improved performance over Logistic Regression, achieving an accuracy of 0.540 and a higher F1 score of 0.592, while maintaining a strong recall of 0.632 for identifying churners. Its precision also improved to 0.558, indicating fewer false alarms compared to the baseline. This model's strength lies in handling diverse feature types and mitigating overfitting through bagging, providing valuable insights through feature importance. However, its "black box" nature can make direct interpretability challenging, raising ethical concerns about explaining specific churn predictions. The model can perpetuate data biases if not carefully managed, potentially leading to unfair outcomes or an over-reliance on automated predictions at the expense of nuanced human judgment.

4.3 Decision Tree

The Decision Tree captures most churners effectively (77% recall), but only half of its churn predictions are accurate (51% precision), implying some wasted marketing efforts. Key churn signals include low recent spend, infrequent purchases, shorter customer tenure, and long purchase gaps. Businesses should deploy broad, low-cost retention campaigns, reserving

expensive outreach for the highest-risk customers, and consider adjusting the classification threshold to balance recall and precision according to budget constraints.

5 Insights & Business Implications

Based on this analysis, the Tuned Decision Tree model emerges as the most promising model for churn prediction, primarily due to achieving the highest recall at 0.77. While the Random Forest also shows strong performance with a 0.540 accuracy and F1 score (0.592). The Logistic Regression serves as an interpretable baseline, but its lower accuracy (0.465) and higher false positive rate make it less effective. However, if we compare only the baseline model the random forest stand out since the slightly higher F1 score among all. Ethically, the Decision Tree's interpretability is a significant advantage, fostering transparency, though its propensity for overfitting necessitates careful management to ensure generalized and fair predictions. Regardless of the chosen model, the pervasive ethical concern across all three is their potential to perpetuate biases present in historical data, demanding vigilant monitoring and responsible deployment to ensure equitable outcomes in customer retention strategies.

6. Team Contribution Statement

Raoyi Li authors the present sections: Introduction and Objective, Business Understanding, Data Understanding, the Contribution Statement itself and performs full linguistic proof-reading of the final manuscript.

Grace Aiyedogbon constructs the unified preprocessing script, publishes descriptive statistics, and later synthesises insights, conclusions, and ethical analysis while curating the presentation deck.

LEE Yi-Chieh, Jenny TING, and **Keer HUANG** respectively implement the Decision Tree, Logistic Regression, and Random Forest models, record hyper-parameters, log metrics, create the comparison table, and maintain the project README in compliance with the jjModelRegistry standard.

7. References

Dataset: https://www.kaggle.com/datasets/hassaneskikri/online-retail-customer-churn-dataset

Comparison Base Model:

Metric	Accuracy	Recall	Precision	F1 Score
Logistic Model	46.50%	66.00%	49.60%	0.567
Decision Tree	56.00%	59.00%	58.00%	58.00%
Random Forest	54.00%	63.20%	55.80%	0.593