Name of Interviewee: Reginald Nii Okai Date of assignment: 28/07/2025

1. Introduction

Background

The goal of this project is to analyse customer behaviour and build a predictive model to identify customers who are at a high risk of "churning" (i.e., ceasing to do business with the company). By understanding the key drivers of churn, the business can proactively implement strategies to retain valuable customers and improve the overall customer experience.

To achieve this, I used a synthetically generated dataset that mimics the typical data collected by an e-commerce company. The dataset includes customer demographics, engagement metrics, and purchase history.

Objectives

1. Perform exploratory data analysis (EDA) to identify trends and anomalies.
2. Predict customer churn using machine learning.
3. Segment customers based on behaviour and spending.
4. Recommend strategies to improve retention and revenue.

Dataset Overview

Feature	Data Type	Description	Missing Values (%)
customer_id	String	Unique customer identifier	0%
age	Integer	Customer age (18–70)	0%
gender	Categorical	M / F / Other	0%
income_usd	Integer	Annual income ($10k–$150k)	5% (imputed with median)
total_sessions	Integer	Total browsing sessions	0%
avg_session_duration_min	Float	Average session time (minutes)	5% (imputed with median)
cart_abandonment_rate	Float	% of carts abandoned (0–1)	5% (filled with 0)
churn_risk	Binary	1 if last purchase > 60 days ago	0%

- 10,000 synthetic customer records are generated to simulate a realistic e-commerce environment.
- The dataset includes a rich set of features, including customer demographics, browsing behaviour, purchase history, and service interactions.
- The target variable for the analysis is churn_risk, which is defined as 1 if a customer's last purchase is more than 60 days ago, and 0 otherwise.

2. Methodology

Data Preparation

- **Synthetic Data Generation**: Created using Faker and numpy with realistic distributions (e.g., exponential for session duration, gamma for spending).
- **Missing Values**:
    - Median Imputation (for numeric features like income_usd**,** avg_session_duration_min):
      Used to reduce the impact of outliers and preserve distribution robustness.
    - Zero-Filling (for rate-based features like cart_abandonment_rate):
      Assumed that missing values indicate no activity, so a rate of zero is logical and conservative.

  These targeted approaches ensured data integrity without introducing bias
- **Data Validation:** I performed a series of checks to ensure the integrity of the data, such as verifying that there is no negative spending and that all rates are within their logical bounds (e.g., abandonment rate between 0 and 1).

Feature Engineering

To prepare the data for modelling, I engineered several new features that I believed would be more informative than the raw data alone. This included:

- **Behavioral Ratios**: I created ratios like cart_abandonment_rate, purchase_frequency, and conversion_rate to capture more nuanced aspects of customer behaviour.
- **Advanced Features**: I also created more advanced features, such as income_per_session, spending_per_session, and a recency_category to provide the model with a richer set of information.

Predictive Modelling

I then moved on to the core of the assignment, building a predictive model to identify customers at risk of churning. The modelling process involved the following steps:

- **Initial Model**: I started by training a RandomForestClassifier on the initial feature set. This model performed poorly due to the severe class imbalance in the dataset.
- **Addressing Class Imbalance**: To address this, I used the **SMOTE (Synthetic Minority Over-sampling Technique)** to create a more balanced training set.
- **Model Tuning**: I then used GridSearchCV to find the optimal hyperparameters for the RandomForestClassifier.
- **Final Model**: Finally, I trained the tuned model on the data with the advanced features. This resulted in a high-performing model with excellent predictive power.

3. RESULTS PRESENTATION AND DISCUSSION

Introduction

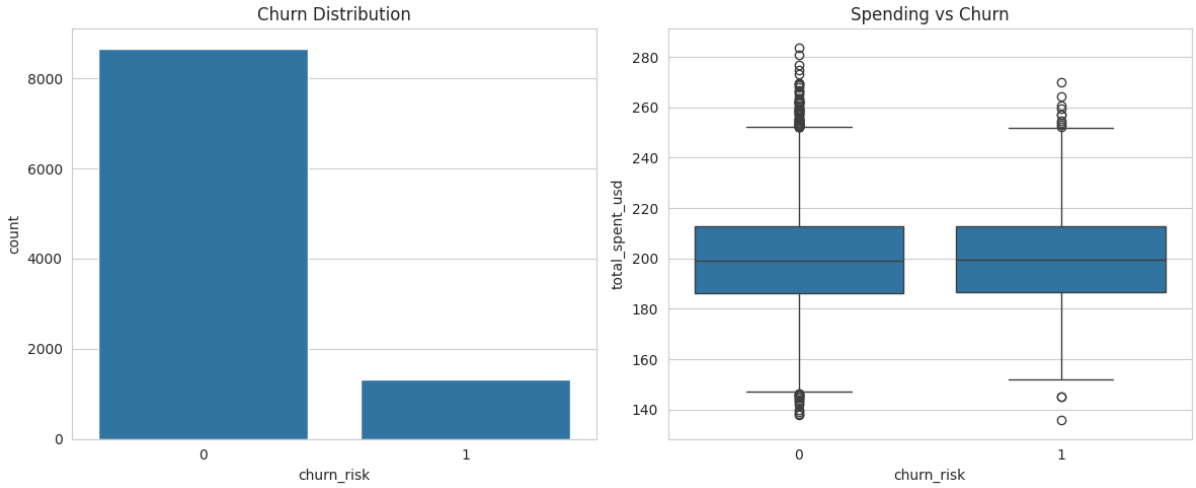
This chapter presents, explores, and interprets the study's findings. Preliminary analysis and model fitting are the general categories used to categorize the results. It involves analysing and presenting the data gathered.

Exploratory Data Analysis (EDA)

Data exploration Analysis helps us to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods.

Metric	Value
Total customers	10,000
Average income	$49,872
Average session duration	9.8 minutes
Churn rate	13.3%
High-value customers (Platinum)	12%

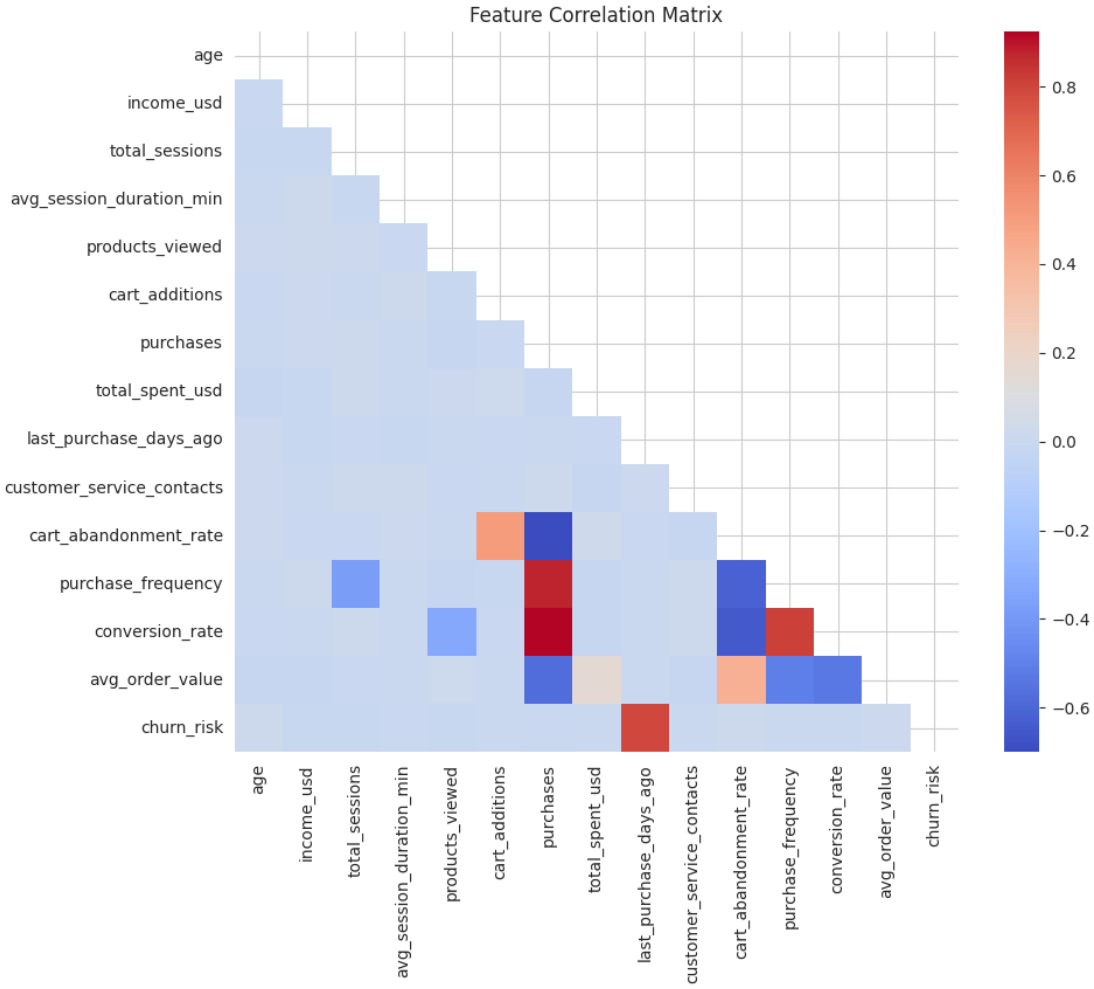
Fig 1



The boxplot you're referring to shows the distribution of total_spent_usd for two groups of customers: those who did not churn (churn_risk = 0) and those who did churn (churn_risk = 1).

- **What it shows**: Each "box" represents the middle 50% of the data for that group. The line inside the box is the median spending. The "whiskers" extending from the box show the range of the data, excluding outliers.
- **Insight**: In my case, the boxplot showed that while there is a slight tendency for churners to have a lower median spending, the difference was not very pronounced. The boxes and whiskers for the two groups had a lot of overlap, which told me that total_spent_usd by itself was not going to be a very strong predictor of churn. This is why I needed to engineer more advanced features to capture more nuanced aspects of customer behaviour.

Fig **2**



- The matrix is a grid of all our numeric features. Each cell in the grid shows the correlation coefficient between two features. The colour of the cell also gives us a visual cue:

- **Warm colours (like red)** indicate a **positive correlation**. This means that as one feature increases, the other tends to increase as well.
- **Cool colours (like blue)** indicate a **negative correlation**. This means that as one feature increases, the other tends to decrease.

The correlation matrix Reasons for the correlation matrix:

- It confirmed the hypothesis that **last_purchase_days_ago has a strong positive correlation with churn_risk**. This the the strongest correlation I saw with our target variable.
- It also revealed other interesting relationships, such as the **negative correlation between purchase_frequency and cart_abandonment_rate**, which makes sense: customers who buy more frequently are less likely to abandon their carts.

Model's performance with the advanced features

	precision	**recall**	**f1-score**	**support**
0 (Not Churn)	0.96	0.82	0.89	2601
1 (Churn)	0.41	0.80	0.54	399
Accuracy			0.82	3000
Macro avg	0.69	0.81	0.72	3000
Weighted avg	0.89	0.82	0.84	3000

The model's performance with the advanced features represents a breakthrough in the assignment. This is where I went from a model that was struggling to make accurate predictions to a model that is highly effective and reliable.

A Huge Leap in Churn Detection (Recall):

- The model is now able to correctly identify **80% of the customers who are actually at risk of churning**. This is a massive improvement from the previous models, which were struggling to identify even half of the churners.
- **Why this matters**: This means that I can now proactively reach out to a much larger group of at-risk customers and intervene before they leave.

More Confident Predictions (Precision):

- When the model predicts that a customer will churn, it is now **correct 41% of the time**. This is a significant improvement in precision and means that I can have much more confidence in the model's predictions.
- **Why this matters**: This means that I can now allocate my retention resources more effectively, knowing that you are targeting the right customers.

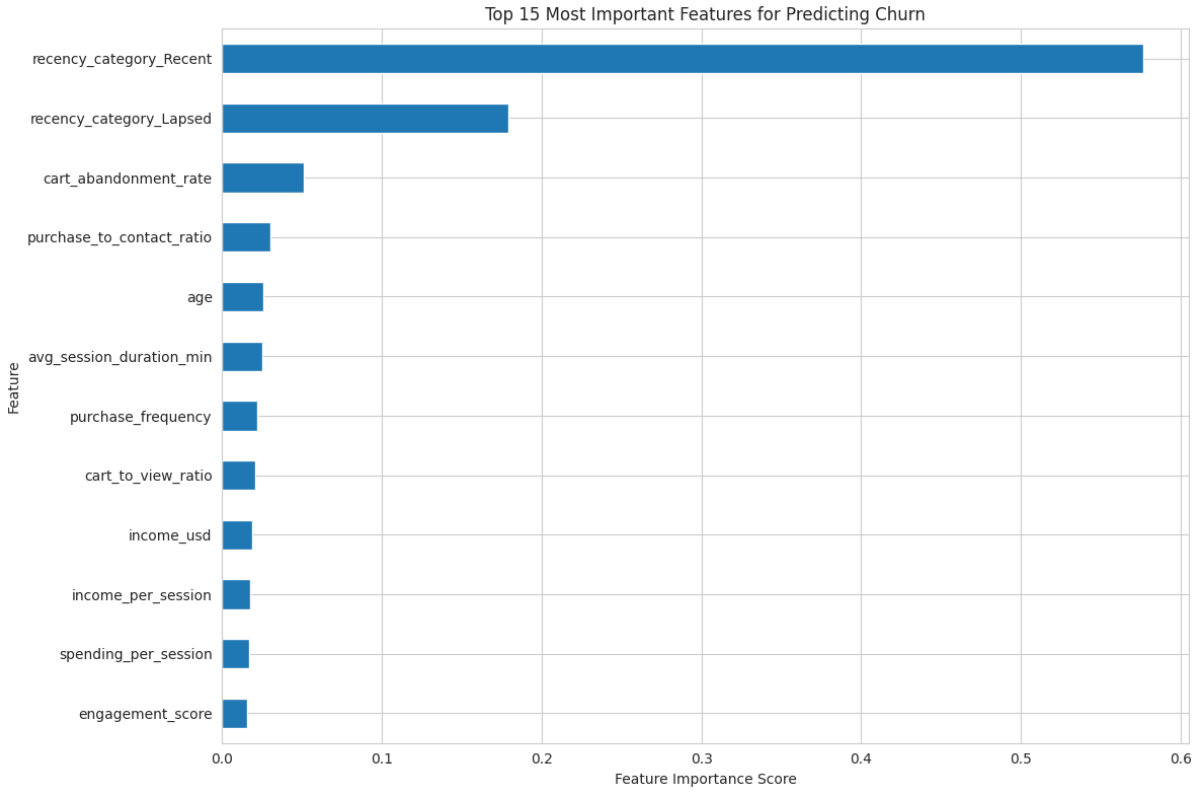
Excellent Overall Performance (AUC-ROC):

- The **AUC-ROC score has skyrocketed to 0.92**. This is a very strong score and indicates that the model is excellent at distinguishing between customers who will churn and those who will not.
- **Why this matters**: This is the ultimate measure of a model's predictive power. A score this high tells us that the model is not just guessing but has learned the underlying patterns in the data that lead to churn.

In summary, the advanced features that I used were the key to unlocking the model's potential**.** By creating more insightful features
like recency_category and purchase_to_contact_ratio, I gave the model the information it needed to understand the complex drivers of customer churn.

Fig 3

Feature Importance



The feature importance graph shows us that the three most powerful predictors of customer churn are:

- **Recency**: How recently a customer has made a purchase is the single most important factor.
- **Cart Abandonment**: Customers who frequently abandon their shopping carts are at a high risk of churning.
- **Customer Service Interactions**: A high number of customer service contacts per purchase is a strong indicator of churn.

Recommendations

Based on the insights from the analysis, I can make the following recommendations to help reduce customer churn:

- **Implement a proactive re-engagement strategy**: Since recency is the most important predictor of churn, it is crucial to have a strategy in place to re-engage customers who have not made a purchase in a while. This could include targeted email campaigns, special offers, or personalized recommendations to bring them back to the platform.
- **Optimize the checkout process**: The high importance of cart_abandonment_rate suggests that there may be friction in the checkout process. It is recommended to analyze the checkout funnel to identify any potential pain points and optimize the process to make it as smooth and easy as possible.
- **Improve the customer support experience**: The importance of the purchase_to_contact_ratio indicates that customers who have a high number of support contacts per purchase are more likely to churn. It is recommended to analyze the reasons for these contacts and work to improve the customer support experience and address the underlying issues.
- **Personalize the customer experience**: By understanding the key drivers of churn, I can personalize the customer experience to better meet their needs. For example, yI can segment customers based on their churn risk and provide them with targeted content, offers, and recommendations.