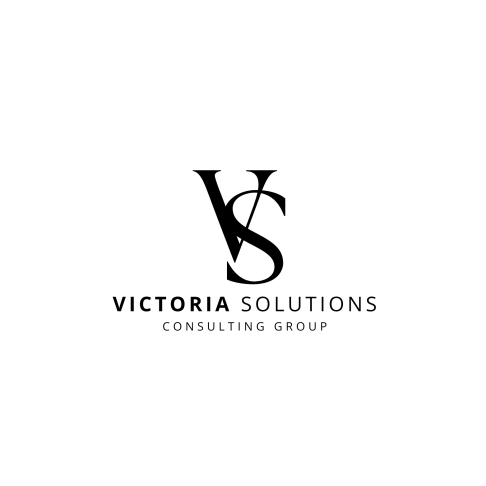
|  |  |
| --- | --- |
| Name | Bertrand Aime BANN SEMTSOU |
| Contact Number | 07476522259 |
| Project Title (Example – Week1, Week2, Week3) | WEEK\_4\_SALES\_&\_CUSTOMER\_BEHAVIOUR\_ANALYSIS |



**Project Guidelines and Rules**

1. **Formatting and Submission**
   * **Format:** Use a readable font (e.g., Arial/Times New Roman), size 12, 1.5 line spacing.
   * **Title:** Include Week and Title (Example - Week 1: TravelEase Case Study.)
   * **File Format:** Submit as PDF or Word file to contact@victoriasolutions.co.uk
   * **Page Limit:** 4–5 pages, including the title and references.
2. **Answer Requirements**
   * **Word Count:** Each answer should be 100–150 words; total 800–1,200 words.
   * **Clarity:** Write concise, structured answers with key points.
   * **Tone:** Use formal, professional language.
3. **Content Rules**
   * Answer all questions thoroughly, referencing case study concepts.
   * Use examples where possible (e.g., risk assessment techniques).
   * Break complex answers into bullet points or lists.
4. **Plagiarism Policy**
   * Submit original work; no copy-pasting.
   * Cite external material in a consistent format (e.g., APA, MLA).
5. **Evaluation Criteria**
   * **Understanding:** Clear grasp of business analysis principles.
   * **Application:** Effective use of concepts like cost-benefit analysis and Agile/Waterfall.
   * **Clarity:** Logical, well-structured responses.
   * **Creativity:** Innovative problem-solving and examples.
   * **Completeness:** Answer all questions within the word limit.
   * Refer to lecture notes and recommended readings.
   * Contact the instructor or peers for clarifications before the deadline.

**START YOUR PROJECT FROM HERE:**

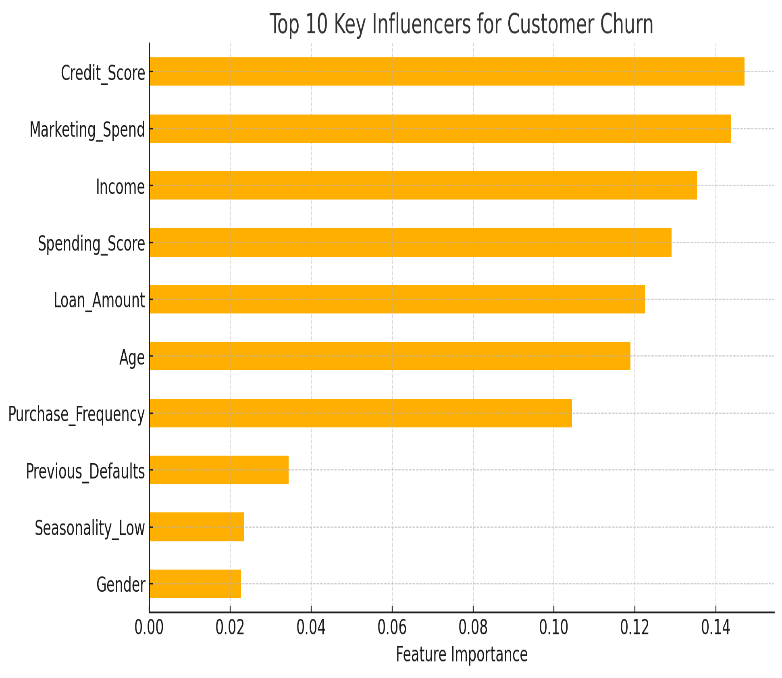
**Introduction**

Sales and customer behavior are critical key performance indicators for any company, as they directly inform customer insights and segmentation, sales forecasting and demand planning, marketing optimization, and customer retention strategies. To uncover actionable insights from our dataset, *raw\_sales\_dataset*, this comprehensive report integrates exploratory data analysis, anomaly detection, predictive modeling, clustering, and time-series forecasting.

The primary objective is to understand the key drivers of customer churn and sales performance, identify unusual patterns, and support data-driven strategic decision-making. The analysis was conducted using **Python** for advanced modeling and analytics, **Excel** for initial data cleaning and preprocessing, and **ChatGPT** to generate clear, insightful narrative explanations and summaries throughout the report.

**Data Cleaning**

The data cleaning process was conducted in two key phases:

* **Initial Cleaning in Excel:**  
  Using Excel's query editor, missing values in the columns **"Income"**, **"Credit\_Score"**, and **"Loan\_Amount"** were addressed by replacing them with the **mean** of their respective columns. This ensured that data integrity was maintained without introducing significant bias. Additionally, a thorough check for duplicates was performed, and **no duplicate records** were found in the dataset.
* **Outlier Removal Using ChatGPT:**  
  In the second phase, **ChatGPT** was utilized to assist in identifying and removing outliers from the dataset. This step enhanced the quality and reliability of the data by eliminating extreme values that could skew analysis. As a result of this cleaning, the dataset was refined to a total of **405 records**.

**Exploratory Data Analysis**

1. **Sales Insights**

The first graph highlights the top 10 key influencers for sales, revealing that **Loan\_Amount**, **Marketing\_Spend**, and **Income** are the most significant drivers, followed closely by **Credit\_Score**, **Spending\_Score**, and **Age**. This suggests that financially capable customers, supported by targeted marketing efforts, contribute heavily to sales performance. Behavioral factors like **Purchase\_Frequency** also play an important role, while demographic and contextual variables such as **Previous\_Defaults**, **Gender**, and **Seasonality** offer additional insight for refining customer segmentation.

A graph of blue lines

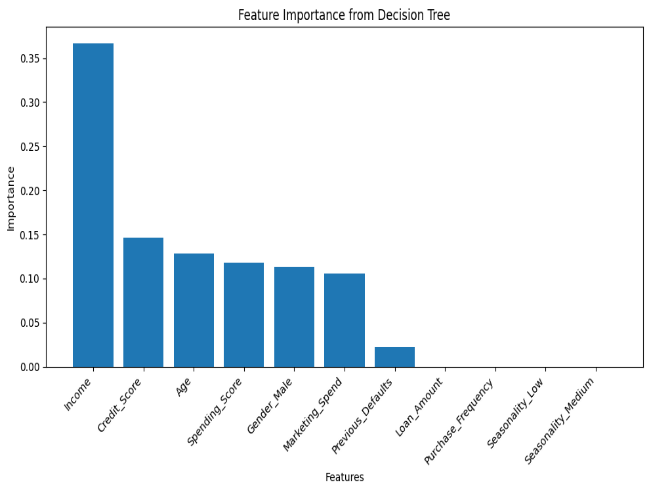
AI-generated content may be incorrect. The second graph, which visualizes sales data with anomaly detection, identifies both high and low sales across the customer base. These anomalies may indicate **high-value opportunities, potential fraud, or signs of customer disengagement and churn**. For the business, these findings underscore the importance of targeting high-income, financially stable customers, optimizing marketing spend, and continuously monitoring irregular sales patterns to enhance decision-making, improve customer retention, and drive revenue growth.

The SARIMAX model results [Ref: python\_Script, line code11], based on scaled inputs, show that all selected features have a statistically significant impact on sales. **Loan\_Amount** (coefficient = **0.217**, *p* < 0.001) and **Marketing\_Spend** (coefficient = **0.214**, *p* < 0.001) are the strongest predictors, indicating that increases in either variable lead to the largest relative boosts in sales. This suggests that offering larger loans and increasing marketing investments are highly effective strategies for driving revenue. **Credit\_Score** (coefficient = **0.159**, *p* = 0.004), **Spending\_Score** (coefficient = **0.147**, *p* = 0.002), and **Purchase\_Frequency** (coefficient = **0.132**, *p* = 0.004) also show meaningful and statistically significant positive effects, highlighting the importance of targeting financially stable, high-spending, and loyal customers. **Age** (coefficient = **0.101**, *p* = 0.043) has a smaller but still positive influence, suggesting older customers may be more financially stable or consistent purchasers. Model diagnostics support these findings, with low residual variance (σ² = **0.0952**), no autocorrelation (Ljung-Box *p* = **0.28**), and stable variance across errors (heteroskedasticity *p* = **0.70**).

**✅ Strategic Recommendations:**

* **Expand loan offerings:**  
  Increase access to financing options for qualified customers, as higher loan amounts are strongly linked to higher sales.
* **Increase and optimize marketing spend:**  
  Allocate more resources to marketing, particularly toward channels and customer segments that demonstrate high responsiveness.
* **Target financially strong customers:**  
  Use credit score data to identify and prioritize customers who are more likely to convert and spend more.
* **Leverage spending behaviour:**  
  Focus marketing and promotions on customers with high spending scores and past purchase volumes.
* **Invest in loyalty and retention programs:**  
  Reward frequent buyers with incentives or exclusive offers to reinforce repeat purchases and long-term engagement.
* **Segment customers by age and behaviour:**  
  Create personalized messaging or product bundles tailored to older, consistent buyers who exhibit strong engagement patterns.

**📊 2. Feature Importance for Churn Prediction**

The decision tree feature importance plot reveals that **Income** is the most critical factor in predicting customer churn, contributing approximately **36%** of the total model decision weight. Other notable features include **Credit Score**, **Age**, and **Spending Score**, all contributing meaningfully. Variables like **Loan Amount**, **Purchase Frequency**, and **Seasonality** have negligible importance, suggesting they are not strong predictors in this model.

**🌳 3. Decision Tree Structure & Churn Logic**

A diagram of a diagram

AI-generated content may be incorrect.The decision tree visual shows how key features interact to predict churn:

* Customers with **low income (≤ 0.185 scaled)** and **low age or spending score** are classified as **No Churn**, indicating that younger, lower-income segments may still be engaged if their spending behaviour is healthy.
* Customers with **high income** but **low marketing engagement or poor credit** tend to be classified as **churn risks**.
* **Gender**, **Spending Score**, and **Credit Score** influence several terminal nodes, highlighting their nuanced roles in retention and churn likelihood. These splits provide **interpretable business rules** that can be applied to segment customers by risk level and drive targeted retention actions.

**🎯 4. Classification Performance**

The decision tree classifier achieved an **overall accuracy of 63.9%** on the test set. However, the **recall for churners (class 1) is only 19%**, and **precision is 24%**, indicating the model is weak at **identifying true churn cases**. This is a key limitation: while the model performs reasonably overall, its **practical use for churn prevention is limited** unless improved.

**🧠 5. Customer Segmentation with K-Means**

The K-Means clustering visualizations (with and without PCA) show two clearly defined customer clusters, based on **Credit\_Score** and **Spending\_Score**:

* One cluster (e.g., high-spending, high-credit) likely represents **high-value, low-risk customers**.
* The other cluster may represent **low-spending or financially at-risk customers**.

This segmentation is highly actionable: it allows the business to develop **differentiated marketing and retention strategies**, such as loyalty rewards for high-value customers or re-engagement offers for at-risk segments.

**✅Recommendations**

* **Enhance churn prediction** by introducing more behavioral, transactional, and time-based features (e.g., time since last purchase, account age).
* **Target retention efforts** on customers identified in churn-prone branches of the decision tree (e.g., high-income but disengaged or low-credit individuals).
* **Use the K-Means clusters** for personalized campaign design—retain high-value users and uplift lower-value ones with targeted promotions.
* **Monitor and retrain** the model periodically to adapt to evolving customer behavior and market conditions.

**Conclusion**

This report has provided a comprehensive analysis of customer behavior and sales performance using a combination of data cleaning, exploratory analysis, predictive modeling, clustering, and time series forecasting. Key insights identified **Loan\_Amount**, **Marketing\_Spend**, and **Income** as the most influential drivers of sales, while **Income**, **Credit\_Score**, and **Age** were the strongest predictors of churn. The SARIMAX model demonstrated that targeted financial and marketing strategies can significantly boost sales, while the decision tree and clustering models offered clear segmentation for churn prevention and customer engagement.

Despite promising findings, the classification model's limited ability to detect churners highlights the need for enhanced feature engineering and more sophisticated algorithms. Future efforts should focus on incorporating behavioural and time-based indicators, as well as regular model retraining to reflect evolving customer patterns.

Overall, the analysis delivers actionable recommendations that can inform both strategic and tactical decision-making—improving marketing efficiency, guiding credit strategy, strengthening customer retention, and ultimately driving sustainable business growth.