**Diagnostic Analytics of Sales and Customer Insights**

**Sakshi Sandeep Salvi**

**Business Analytics, Conestoga College, Institute of Technology and Advanced Learning**

**INFO8146: Diagnostic Analysis**

**Maria Aiello**

**Jun 15, 2025**

# **Table of Contents**

Executive Summary ...................................................................................................................... 3

Azure Tools Configuration and Usage ............................................................................................ 4

- Microsoft Azure Setup

- Data Selection

- Pipeline and Cleaning Setup

Actor Interaction Diagram & Roles ............................................................................................... 7

Azure Synapse Analytics Insights .................................................................................................. 9

- SQL Insight 01

- SQL Insight 02

Power BI Visualizations and Insights ............................................................................................ 11

- Insights based on the Year

- Insights based on Regionality, Seasonality, and Retention Strategy

Diagnostic Analytics Framework ................................................................................................. 19

- Observations, Hypotheses, Root Cause

Actionable Recommendations ...................................................................................................... 21

Application of Critical Thinking .................................................................................................. 22

Conclusion ................................................................................................................................. 23

**Executive Summary**

This report presents a diagnostic analytics assessment conducted using Microsoft Azure and Power BI. The primary objective was to analyze a rich customer and sales dataset to uncover patterns in customer behaviour, identify trends in purchasing activity, and assess risks related to customer churn.

The whole data preparation process involved ingestion, cleaning, and transformation using a variety of Azure technologies, such as Azure Data Factory to move the data, Azure Blob Storage to store the data securely, and Azure Synapse Analytics for structured querying and aggregation. Power BI was then used to generate interactive dashboards and graphs that display key performance indicators (KPIs), including average churn probability, total lifetime value (LTV), customer count, and average purchase value.

The analysis made available some important insights. The customers acquired via loyalty programs are always high in lifetime value, and their purchasing behaviour is also quite good as compared to any other retention programme. Nevertheless, the churn probability was evenly distributed throughout the dataset, which indicated that there may be a problem with the predictive model or the inputs of the data. The further investigation revealed a significant difference between performance in various regions and seasons, which indicated the evident possibility of improvement and segmentation.

Three targeted recommendations close the report these are enhance the churn prediction logics to better match the reality of customer retention behaviours, increase the loyalty-based programs to exploit high-value segments, and regionalize and seasonalize the marketing campaigns. These evidence-based observations are intended to help business teams, particularly marketing and operations, make informed decisions that would maximize customer engagement and long-term value.

**Azure Tools Configuration and Usage**

1. **Microsoft Azure Setup**

A screenshot of a computer

Description automatically generated

* Created a **Microsoft Azure account** with access to free-tier and core services via the **Azure for Students** subscription offered by the college.
* Browsed through the **Azure Portal** to familiarize myself with the interface to perform further configurations.
* Then I created the **resource groups, R1\_SS, where I arranged all elements of the analytical workflow**.
* After creating the resource group, I deployed the **Azure Data Factory by naming it R1SS** to prepare it for pipeline creation.
* Next, I created the **Azure Synapse Workspace** and linked it with the storage container where I later uploaded the dataset that I had selected via Kaggle.
* Then I created and linked a **storage account** (Azure Blob Storage) by enabling the **Azure Data Lake Storage Gen2** for dataset upload.
* Then I explored **Azure documentation and portal tutorials** to guide initial setup decisions.

1. **Data Selection**

Dataset Name: Sales and Customer Insights Dataset

Dataset URL: <https://www.kaggle.com/datasets/imranalishahh/sales-and-customer-insights>

* I chose the dataset based on various demographic aspects mentioned in the requirement criteria provided by the professor.
* Some main aspects were the number of years, customer behaviour, sales performance, and churn-related attributes for high-quality insights.
* The dataset was selected based on some potential business questions like:
  + Will the data give trends on a yearly basis?
  + How do regions and seasons influence the sales performance?
  + How can retention strategies impact the trend analysis?
  + Will the dataset provide any kind of profit-based analysis?
* Checked whether the dataset was in CSV, not too large enough to work in the free version, supported diagnostics analytics, and compatibility with the Azure Data Lake for performing Azure Data Analytics in the Synapse.

1. **Azure Pipeline and Cleaning Setup**

* I enabled the **Azure Data Lake Storage Gen2** to store and work on the raw dataset using **Azure Blob Storage**.
* Then I uploaded the raw dataset into the Azure Blob Container: container2 to perform data profiling.
* Then the Pipeline 01 was created to start the data cleaning process using **Azure Data Factory**.
* Data was imported into the pipeline using the file path in the source dataset tab.

1. **Dataflow 01**

Created the dataflow 01 to perform key data preparation steps:

* Derive Column: This was used to extract the date parts as months and years.
* Aggregate: This was used to perform various aggregations and produce values like total lifetime value, average order value, average churn probability, and customer count.
* Sort: This was used to organize the dataset.
* Sink: This was used to transfer the clean data to a new location.

A screenshot of a computer

Description automatically generated

1. **Dataflow 02**

Created the dataflow 02 to perform to handle missing values, duplicate records, and anomalies.

A screenshot of a computer

Description automatically generated

**Actor Interaction Diagram and Roles**

1. **Actors Identification**

|  |  |
| --- | --- |
| Actor | Role Description |
| User End (Customer) | Source of data generation through retail transactions and purchase. |
| Data Engineer | Manages the data ingestion, pipelines, storage, and dataflows. |
| Azure (Platform Actor) | Handles the security of the platform and storage for Synapse and Data Lake. |
| Azure Data Factory | Executes the daily ingestions from the source systems which contains the data. |
| Azure Data Lake | It stores the raw and updated data. |
| Data Analyst | Cleans, Analyzes, and performs queries for SQL and Power BI using Synapse. |
| Azure Synapse Analytics | Synapse helps to clean the data and gives responses to various queries for high quality Power BI dashboards. |
| Power BI | It’s a platform that helps to generate high quality dashboards with real-time data. |
| Business Stakeholder | Retail Managers, who review the dashboards for insights and better decision making. |

1. **Interaction Summary**

* **Customer (End User) → Azure Data Factory:** Indicates the source of the daily sales data generated by the customer that is to be used for analysis.
* **Azure Data Factory → Azure Data Lake:** Raw data is stored and ingested.
* **Data Engineer → Azure Platform + Azure Data Lake + Azure Data Factory:** The data engineer is one of the main actors among all because they are in charge of configuration and managing the pipelines.
* **Azure Data Lake → Azure Synapse Analytics:** Data is cleaned here and prepared for analysis.
* **Data Analyst → Azure Synapse Analytics:** Here, some queries are run to check the metrics and the dataset for visualization is refined.
* **Azure Synapse → Power BI:** It provides a refined dataset to Power BI for dashboard creation.
* **Power BI → Stakeholders:** It provides insights through dashboards to support business decision-making.
* **Azure (Platform Actor) →** Is responsible for data security, storage, data access control, and configuration of the data.

**A diagram of a software company

Description automatically generated**

**Azure Synapse Analytics Insights**

The **Azure Synapse Analytics** workspace was created, and SQL queries were performed. Below are some images and insights for the SQL queries performed.

1. **SQL Insight 01**

A computer screen shot of a computer

Description automatically generated

**Quantitative Insights:**

* The sales gap between the highest and the lowest region is nearly 7,400, which has a variance of less than 3%.

**Qualitative Insights:**

* These numbers indicate a balanced market performance by all the regions, flagging a well-distributed global sales strategy.
* South Africa ranks first with the highest total sales, indicating a better sales strategy or promotional events leading to a boost in sales.
* Since all the regions are closely grouped, a small adjustment like targeted sales or loyalty programs can help in increasing the sales of the lower-ranked regions, making shifts in the ranks, and increasing the ROI.

1. **SQL Insight 02**

A screenshot of a computer

Description automatically generated

**Quantitative Insights:**

* The total lifetime value between electronics and sports-related product categories is nearly 38,362, which has a variance of less than 3.1%.

**Qualitative Insights:**

* Electronics contribute the highest to the total value, suggesting that the product may have higher value or that customer engagement is higher in this category.
* Clothing follows closely with the electronics, indicating that customers are equally loyal towards this category and can produce chances of an increase in the ROI via this category.
* There is a tight clustering of all the categories, flagging that the store has a well-balanced product strategy.

**Power BI Visualization and Insights**

The cleaned dataset was imported from Azure Synapse Analytics, and numerous interactive visualizations were created using Power BI.

**Insights based on the Year**

1. **Power BI – Lifetime Value Trends by Launch Year**

**A graph showing the sales growth

Description automatically generated with medium confidence**

**Insights:**

* The Loyalty Program has consistently maintained a stable lifetime value between 0.53M and 0.54M, indicating good customer retention.
* In 2021, the email campaign strategy had the greatest lifetime value (0.60M), but by 2022, it had dropped to 0.47M. Flagging that this strategy may be a short-term solution, but didn’t work in the long run.
* Discount strategy values increased somewhat over time, rising from 0.55M in 2020 to 0.56M in 2022. This shows a consistent growth in gaining customers, or maybe the order frequency increased.
* 2023 displays 0 for all the strategies, most likely because of inadequate data or new strategies might have launched which had no lifetime value yet. This can be kept for prediction or future real-time analysis.

1. **Power BI – Average Order Value by Launch Year and Seasons**

**A graph of blue bars

Description automatically generated**

**Insights:**

* Fall typically produces the highest average order value, particularly in 2021, when it exceeds all other seasons. This may suggest that fall promotions and customer behaviour result in increased expenditure per transaction.
* Summer and Spring seasons remain competitive, resulting in continuous performance and consumer involvement throughout the year. This reflects effective retention and product demand.
* Winter may underperform because of the weather conditions or holidays, or weaker spending cycles. Bundled offerings or sales can help improve performance.
* The data for 2023 is incomplete, with only Spring showing a minimal value. This should be considered with caution as it can be used for prediction purposes.

1. **Power BI – Churn Probability by Peak Year and Region**

**A blue rectangular object with white text

Description automatically generated**

**Insights:**

* North America has the greatest churn probability, indicating a risk of customer disengagement. To address this, retention methods, including tailored promotions and loyalty programs, should be prioritized.
* South America and Europe are slightly close, with steady but vulnerable probabilities. Both regions might benefit from proactive engagement programs to prevent an increase in churn.
* Asia has the lowest churn, indicating higher retention or customer satisfaction. A successful strategy in Asia might cause trouble in the higher-risk regions.
* A regional spread indicates that the churn behaviour varies widely, emphasizing the necessity for specialized retention methods rather than a common mutual strategy for all.

1. **Power BI – Order Value Trend by Retention Strategy over Launch Date**

**A graph showing a line graph

Description automatically generated with medium confidence**

**Insights:**

* Email campaigns have a high vulnerability, with average order values peaking at 1,800 in early 2021. While effective for short-term pushes like seasonal or promotional, they lack stability over time.
* Loyalty Program maintains a stable trend, with fewer sudden spikes or drops. It flags that there is a strong customer retention or consistent average order value performance over time.
* Discount Strategy varies considerably, with peaks around 1,500 and dips below 500. It suggests that when utilized correctly, it can increase the order value, although its influence may differ depending on timing or customer behaviour.
* By 2023, all strategies show a considerable fall, possibly due to limited data from new customers or due to the introduction of new strategies.

**Insights based on Regionality, Seasonality, and Retention Strategy**

1. **Power BI – Lifetime Value by Region**

**A blue rectangular bars with white text

Description automatically generated**

**Insights:**

* Europe has the highest total lifetime value at 1.31M, showing strong customer retention and high spending behaviour. This is likely due to effective retention measures or a mature customer base.
* North America follows closely with at 1.26M, indicating a well-established and profitable market. This could be boosted further with loyalty or targeted campaign efforts.
* Asia has a neutral performance of 1.19M, below the top two regions. A valuable market with potential is needed for future expansion.
* South America has the lowest total lifetime value at 1.13M but retains a solid overall presence. Promotional campaigns might benefit from greater targeting or engagement of customers.

1. **Power BI – Average Order Value by Retention Strategy and Seasons**

**A graph showing the sales of marketing

Description automatically generated with medium confidence**

**Insights:**

* Winter has the greatest average order value (just over $10,000) in the Loyalty Program, indicating significant customer engagement and spending throughout the colder months for loyalty retention.
* Spring shows a consistent performance, particularly in Email Campaign and Discount Strategies. This reflects the effective mid-year promotions or seasonal purchasing habits.
* Summer underperforms in Discount Strategy, with values dropping below 8K. This flags weaker reactivity to discounts during the summer, maybe due to off-season product alignment or customer disinterest.
* The fall season is the most balanced, with consistently high order values across all strategies. The fall is a great season for maximizing order value across all retention strategies.

1. **Power BI – Lifetime Value vs Churn Probability by Region**

**A graph of blue dots

Description automatically generated**

**Insights:**

* The scattered plot displays a dense distribution of churn probability between 0.2 and 0.9 in all the regions. Flagging that most customers are in the mid-to-high churn risk category, with a low representation of the outliers.
* Most customers have a lifetime value of less than 20K, with a few exceeding 40K. These customers with high value occur in all the countries but are rare.
* There is no linear relationship between turnover and lifetime value. Flagging that customers with high lifetime value may have high churn risk, indicating that value alone may not ensure retention.
* All regions have an evenly distributed market, indicating that the turnover habits are global and not region-specific.

1. **Power BI – Churn Risk Category by Strategy**

**A blue squares with white text

Description automatically generated**

**Insights:**

* The Email Campaign has the highest proportion of high-risk customers (42.86%). This method may not be beneficial for long-term engagement or loyalty.
* Loyalty Program has the lowest high-risk group (37.84%) but contributes most to the low-risk customers (42.94%). Loyalty programs are most effective in reducing churn and increasing the retention of the customers.
* The Discount Strategy is mostly evenly spread, but still compromises 34.74% high-risk consumers and the highest percentage of moderate-risk customers (24.77%). It may attract a large customer base, but may not be effective in maintaining the customer base after transformation.

**Diagnostic Analytics Framework**

1. **Framework overview**

The diagnostic analytics framework combines both SQL queries and Power BI dashboard insights to study the customer behaviour, regional trends, seasonal trends, and retention strategies. It also provides patterns of past data to justify the observed change in critical performance measures, particularly those involving a decline in sales.

1. **Key Observations**

* There is substantial region, season and retention-strategy variance in Lifetime Value (LTV) and order behaviour.
* Email Campaigns generate short-term bursts but lead to high churn and reduce lifetime value.
* Churn risk is evenly distributed globally, indicating that no regional factor is involved in the decline of sales.
* Loyalty programs decrease churn, but enhance customer value, particularly during colder months.
* 2023 sales and retention data appear to be less or nearly missing, flagging a decline or incomplete data provided that there might be some changes in the retention strategy or sales strategy.

1. **Hypothesis Behind Declining Sales Trends**

* An Email Campaign Strategy can be overused or ineffective.
* High churn probability risk and decreasing lifetime values indicate that this strategy is not maintaining long-term relationships with the customers.
* Lower Sales in Asia and South America can be due to weaker customer outreach or weaker loyalty programs.
* These regions flag high churn and lower lifetime value, making them highly susceptible.
* Decline in the Average Order Value during Summer and Winter is an indicator of poor seasonal promotions.
* This flags that the seasonal promotions are not impactful or influential enough to increase customer engagement.
* Inconsistent use of Discounts without any follow-up loyalty programs.
* A lot of customers obtained via discounts leave soon unless they are moved to a loyalty model.
* The 2023 data gap may reflect late onboarding or ineffective campaigns.
* There is almost no data about 2023, leading to no lifetime value or order value logs.

1. **Root Cause Analysis**

* Lack of consolidation in the Retention Strategy implementation – relying on one-time promotions leads to high churn probability risk.
* Seasonal and regional categorization is not used at its best – most campaigns appear to be generic or global rather than customized according to the geography.
* Customer acquisition is not strongly supported by stable value creation – e.g. discounts or sales offers without any follow-up loyalty programs.
* Data modelling for churn may need adjustments – constant or extreme value of churn probability can be risky for the sales and indicate one of the critical reasons for the decline in sales.

**Actionable Recommendations**

1. **Empower loyalty programs in all regions.**

* Increase the inclusion of loyalty-based incentives in places where lifetime value is minimal, particularly in Asia and South America.
* Automate the requirements of post-purchase enrolment in loyalty to maintain the customers that are acquired during the promotion.

1. **Restrict sending one-time email campaigns and using follow-up campaigns.**

* Limit the use of email campaigns with short-term usage resulting in high churn.
* Implement scheduled follow-up messages to create repeated interaction.

1. **Regional Campaigns: Construct region-wise based on performance trend.**

* Customize campaigns based on the diagnosis done and work on sales strategy as per regions, and not globally.
* Initiate campaigns that are geared towards the behaviour of customers in the regions and their peak season demands.

1. **Improve data quality and the churn probability model.**

* Repeat churn prediction logic procedures and check whether high uniform rates of churn imply overfitting or missing time.
* Confirm what has been inputted in the models and retrain with new information about their behavioural patterns.

1. **Rebuild Summer and Winter Campaign Strategies**

* When the orders are low at such seasons, this implies that there should be a new creative, a bundle of deals, or time-limited loyalty boosters.

**Application of Critical Thinking in the Analysis**

1. **Spotted Churn Probability Modelling Issue –**

* Spotted incompatible data in the churn probability field, like uniformity of values, and challenged the validity of the models without assuming it is accurate.
* Recurred into the suggestion to reengage the churn model and to reinforce its logic and inputs.

1. **Compared Seasonal Order Value Trends –**

* Identified seasonal trends in the value of orders over the years to see whether there were periodical declines or whether they were random.
* Exposed uniform underperformance in the Summer and the Winter, providing recommendations for restructuring seasonal campaigns.

1. **Examined Retention Strategy Impact –**

* Examined how the retention strategy affected more than shallow numbers, such as the number of customers.
* Learned that Email Campaigns, when active, disproportionately contributed to the churn and did not correspond to the belief that more is better.

1. **Recognized gaps in 2023 sales data –**

* Identified holes in the 2023 performance data and did not come up with hasty conclusions.
* This more cautious reading caused more informed planning and filtered dashboard visuals.

1. **Balanced short-term and long-term plans and initiatives –**

* Healthy short-term and long-term viewpoints by separating metrics that drive spikes in order value and those that drive the development of customer lifetime value.
* Reported the idea of mixing discount strategies with loyalty call-backs.

**Conclusion**

In this report, sales and customer insights will be investigated with Microsoft Azure and Power BI, utilizing diagnostic analytics to identify the patterns of customer behaviour and performance during the seasons, as well as the churn risk. The workflow consisted of ingestion and transformation of the data using Azure Data Factory, Blob Storage, and Synapse Analytics and visualization of the data and creation of insights in Power BI.

Some of the important findings feature spotlights on: Loyalty Programs would create high lifetime value and reduce churn. To clients with high churn rates, Email Campaigns are related to high short-term success, but less to long-term success. The order values and the sales performance vary according to the season, with the best value of the performances in Fall and the lowest in both Summer and Winter. All of the regions have a high risk of churn, with 2023 data only covering some of the parameters, so there is no possible way to track the recent trends.

The critical thinking was used to question the accuracy of churn probability, regional and seasonal underperformance, and differentiate between short-term tactics and long-term thinking on value strategies. The report has five practical recommendations, such as the refinement of retention strategies, enhancement of the churn model and conduct of region and season-based campaigns to help realize the reversal of declining sales patterns and achieve optimal customer value.