Data-Driven Optimization of Newspaper Distribution and Customer Retention Strategies in the Unorganized B2C Sector

Mid-Term Submission *for the BDM Capstone Project*

Submitted by

Name: **K Yuvaraj**

Roll Number: **22F3002987**

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Content

1. Executive Summary and Title

S. K. News Paper Mart is an unorganized B2C newspaper distribution service located in Mogappair, Chennai, operated single-handedly by Mr. Kumar for **over 23 years**. While he caters to around **80 households**, the business faces several key operational challenges. **Externally**, he faced the lack of route optimization and **Internally**, he faced difficulty in tracking customer churn, and inefficiencies in demand forecasting. This project titled "**Data-Driven Optimization of Newspaper Distribution and Customer Retention Strategies in the Unorganized B2C Sector**" aims to address the rising fuel costs and customer attrition due to digital migration and service gaps.

Data was manually collected and digitized for **three months** (**March to May 2025**), resulting in three structured datasets: **Customer Master Data**, **Daily Delivery Log**, and **Route & Fuel Log**. Key metadata includes customer preferences, delivery dates, fuel usage, and cost metrics. Descriptive statistics revealed that a majority of customers reside in a few key zones, missed deliveries are minimal but concentrated, and weekend demand is higher for certain newspapers. This provided strong groundwork for deeper analysis.

For midterm analysis, key methodologies applied include **descriptive analytics**, **time series forecasting using Prophet**, **geospatial delivery mapping**, and **classification modeling with Random Forest**. Initial results revealed stable newspaper demand with mild weekend peaks, visually clustered delivery zones useful for future route optimization, and key churn indicators such as missed deliveries and shorter subscription lengths. These insights are helping transform an intuition-driven business into a data-informed operation.

.

2. Proof of Originality

1. Interview with Business Owner – Transcript Included **:** Interaction Video Folder
2. Authorization Letter from the Owner **:** Letter From Owner
3. Small Video Clips for proof **:** Clips of Mr.Kumar Distributing Newspapers
4. Photo Proofs
    - Mr. Kumar organizing his papers : Starting Point at Thirumangalam
    - Mr. Kumar with his vehicle and Newspapers : Activa 3G
    - Handwritten Notes and Accounts : Images

3. Metadata

The data required for this project was initially obtained in handwritten form and partly based on memory as narrated by the owner, Mr. S. Kumar. After multiple interactions and consolidation steps, I was able to digitize and organize the information into three clean and structured datasets. These datasets form the foundation of the analysis aimed at solving the challenges faced in the newspaper distribution process. The three consolidated datasets are:

1. **Customer_Master_Data** : sheets1
2. **Daily_Delivery_Log** : sheets2
3. **Route_and_Fuel_Log :** sheets3

- **Data Collection Duration** : 3 months
- **Data Collection Dates** : March 1, 2025 to May 31, 2025

3.1 Customer Master Data

This is the core reference dataset that captures customer-specific details, such as newspaper/magazine preferences, delivery days, address, latitude and longitude.

Dataset Dimensions : 58 rows x 12 columns

Columns :

1. **Customer_ID** : Unique ID assigned to each customer.
2. **Customer_Name** : Name of the Customer.
3. **Address** : Delivery address of the customer.
4. **Area_Zone** : Area of the customer.
5. **Delivery_Days** : Specifies the days on which delivery is made.
6. **Start_Date** : Year in which delivery started.
7. **End_Date** : Year the customer stopped (if applicable).
8. **Status** : Indicates if the customer is Active or Inactive.
9. **Preferred_Newspaper** : Newspapers subscribed by the customer.
10. **Magazines_Subscribed** : Magazines (if any) subscribed by the customer.
11. **Latitude** : Latitude of the Area.
12. **Longitude** : Longitude of the Area.

3.2 Daily Delivery Log

This dataset records daily deliveries made to each customer and tracks whether newspapers and magazines were successfully delivered on each day.

Dataset Dimensions : 5336 rows x 6 columns

Columns :

1. **Date :** Date of delivery.
2. **Customer_ID :** Unique ID from master data.
3. **Newspaper_Delivered :** Newspapers delivered on that day.
4. **Magazines_Delivered :** Magazines delivered (if any).
5. **Delivered? :** Whether delivery was completed.
6. **Notes :** Additional information (e.g., reason for missed delivery).

3.3 Route and Fuel Log

This dataset helps estimate operational costs associated with daily deliveries. Mr. Kumar uses a two-wheeler (Active 3G) for newspaper distribution. The petrol price was considered ₹104/litre for cost calculations.

Dataset Dimensions : 92 rows x 6 columns

Columns :

1. **Date :** Date of delivery.
2. **Total_Customers_Served :** Total number of customers served on that day.
3. **Estimated_Route_Distance (km) :** Approximate distance traveled during delivery.
3. **Fuel_Used_Liters :** Estimated fuel used based on distance and mileage.
4. **Fuel_Cost(Rs) :** Total cost incurred for fuel.
5. **Delivery_Notes :** Any noteworthy observations (e.g., holiday, rain).

4. Descriptive Statistics

In this section, I performed a detailed descriptive statistical analysis on the three structured datasets: Customer Master Data, Daily Delivery Log, and Route and Fuel Log. The objective was to extract baseline insights into customer demographics, delivery performance, and operational efficiency.

4.1 Overview:

Total number of customers : 58

Total delivery records : 5336

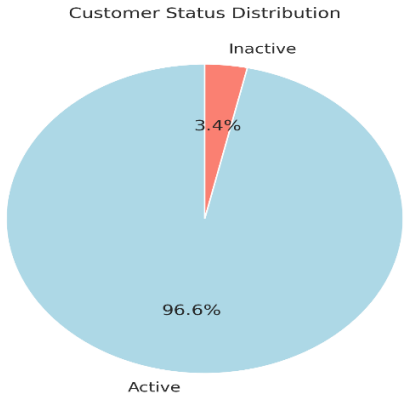
Total route distance covered : 4659 km

Total fuel used : 116.48 litres

Total fuel cost : Rs.12113

Average customers served per day: 55.16

4.2 Statistics in Customer Master Data



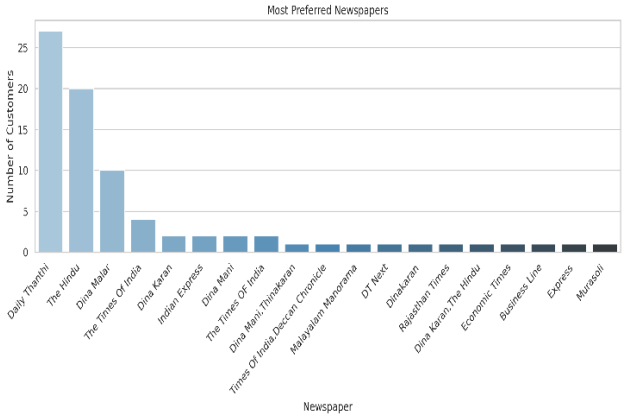


Figure 1 Customer Status Distribution **Figure 2 Most Preferred Newspapers**

1. A majority of the customers (96.6%) are marked as Active, indicating a stable and consistent customer base. A smaller portion (3.4%) are Inactive, which is due to address shifts, subscription cancellations, or transition to digital platforms. This segmentation helps in identifying retention opportunities and assessing churn impact on daily delivery volume (Refer Figure 1)
2. Daily Thanthi and The Hindu appear as the most frequently subscribed newspapers as shown in Figure 2, indicating strong demand for both English and Tamil dailies. A few customers prefer multiple newspapers, suggesting opportunities for bundled distribution and cross-promotions.

4.3 Statistics in Daily Delivery Log

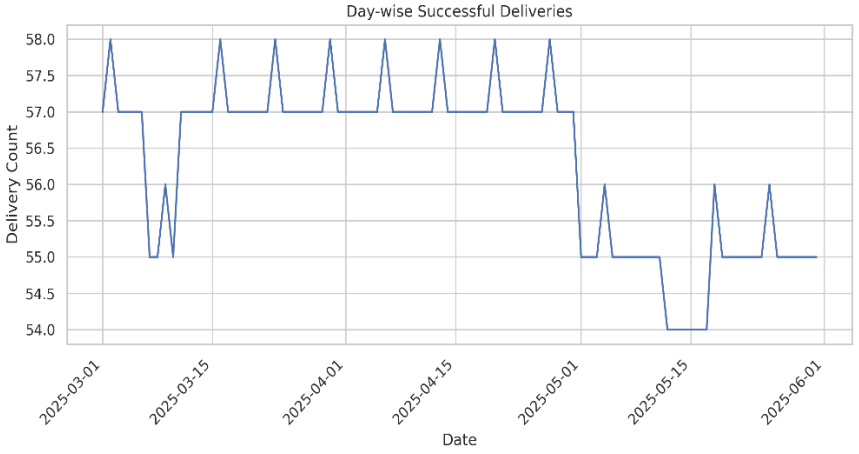
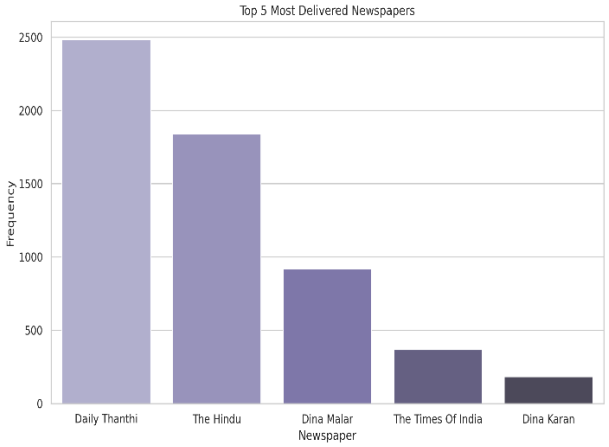


Figure 3 Day-wise Successful Deliveries

1. Delivery volumes remain mostly stable, with minor dips on weekends or public holidays, indicating predictable demand. This visualization helps anticipate lower volume days and plan workload accordingly (Refer Figure 3).



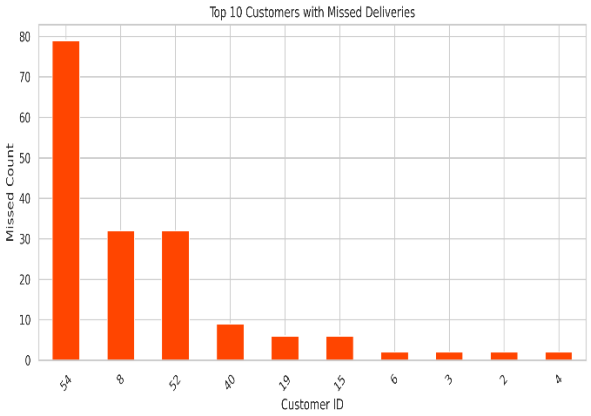


Figure 4 Top 5 Most Delivered Newspapers **Figure 5 Customers with Missed Deliveries**

2. The most frequently delivered newspapers include Daily Thanthi (Refer Figure 4), The Hindu and Dina Malar aligning with customer preferences from the master data. These high-frequency items should be prioritized in inventory planning to avoid understocking.
3. A small group of customers accounts for a majority of missed deliveries, suggesting delivery inconsistencies. These customers may be at risk of churn, and require targeted follow-up or confirmation of details. This insight helps prioritize customer retention strategies and operational checks (Refer Figure 5) .

4.4 Statistics in Route and Fuel Log

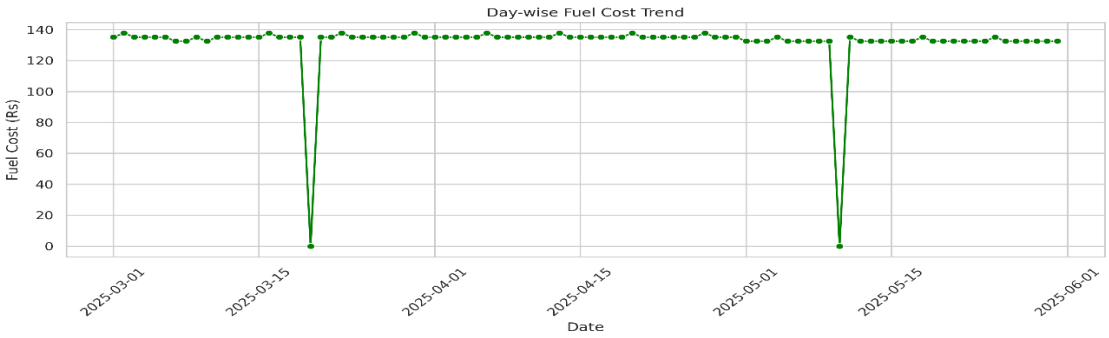


Figure 6 Day-wise Fuel Cost Trend

1. Fuel costs show relatively stable trends with occasional spikes that may align with longer routes or higher customer volume days. Tracking daily fuel cost is essential for cost optimization and sustainability in the long term (Refer Figure 6) .

5. Explanation of Analysis Procedure

This section details the methodological framework adopted for transforming raw, semi-structured data into actionable insights. The process involved systematic data consolidation, descriptive analysis, and feature derivation, followed by preparation for advanced analytical techniques such as forecasting, clustering, and classification. Each method was chosen based on its contextual

relevance, interpretability, and alignment with the project's operational objectives in an unorganized service setting.

5.1 Data Consolidation

The three raw datasets are consolidated and the below is the information on that.

1. In Customer Master Data ,The raw dataset contained 58 customer records with significant missing data in critical columns like Area_Zone (53% missing) and Magazines_Subscribed (93% missing). In the cleaned version, missing Area_Zone values were filled manually based on address patterns to enable area-based clustering later. The latitudes and longitudes of the area is added in the sheets for analysis purpose.
2. In Daily Delivery Log, it lacked semantic consistency in the Notes column. The cleaned version introduced structured defaults by marking empty Notes fields with a placeholder '-' .
3. In Route and Fuel Log, the cleaned version was added with two new columns using formulas:
   Fuel_Used_Liters = Estimated Distance / Vehicle Mileage
   Fuel_Cost = Fuel Used × Price per Liter (₹104)

Link to the raw data : Raw Data

Justification:

These enhancements enabled better analysis of operational efficiency and cost forecasting. It brings the balance in the data which is essential for the greater accuracy.

5.2 Descriptive Statistical Analysis

Descriptive statistics were used as the first analytical method to explore and summarize the core structure of the data. This included calculating counts, proportions, and averages, as well as generating visualizations such as pie charts, bar graphs, and time series plots.

Justification:

The business operates in an unstructured, manual environment with limited digital history. Descriptive summaries are ideal for highlighting immediate insights without requiring complex inputs or assumptions. Visual tools also help the business owner better understand his operations and validate the findings.

5.3 Visual Exploratory Data Analysis (EDA)

Visual EDA techniques such as line charts, grouped bar plots, and pie charts were applied using Python libraries (Matplotlib and Seaborn). This allowed effective communication of insights from the delivery logs and operational metrics.

Justification:

Tabular summaries are not intuitive for manual operators. Visualizations communicate trends such as fuel cost patterns or missed delivery clusters more clearly than raw numbers. EDA also supports later decisions like customer segmentation or route optimization, making it a bridge between data cleaning and model building.

5.4 Derived Field Computation (Feature Engineering)

Certain fields in the dataset such as Fuel Used (Liters) and Fuel Cost (Rs) were not present in the raw files. These were computed using known relationships:

Fuel Used = Distance ÷ Mileage

Fuel Cost = Fuel Used × Petrol Rate

Justification:

These derived fields are lightweight but powerful. Instead of implementing GPS tracking or vehicle telemetry (which is impractical for the business), simple formulas based on business inputs (e.g., mileage, fuel rate) provide actionable insights. They also enable cost per customer and cost per kilometer comparisons — critical KPIs for route and fuel planning.

5.5 Time Series Forecasting (Newspaper Demand Prediction)

To predict future newspaper demand, I plan to apply time series forecasting techniques on the historical daily delivery data. The goal is to identify trends, seasonality, and fluctuations, especially across weekends, public holidays, and low-demand weekdays. The primary model selected for this purpose is Facebook Prophet, a robust time series forecasting tool well-suited for business-oriented time series data.

Justification:

Newspaper delivery data is sequential and strongly time-dependent, making time series forecasting the natural analytical choice. Weekend effects, holiday dips, and changing customer behavior influence demand in a non-linear way. Prophet is chosen over traditional models like SARIMA because it handles seasonality natively (e.g., weekly patterns) without requiring manual differencing or stationarity transformations and also easily incorporates holidays and outliers, which is crucial in newspaper delivery.

5.6 Route Optimization Using TSP (Travelling Salesman Problem)

To optimize daily delivery routes and minimize fuel usage, this project adopts the Travelling Salesman Problem (TSP) framework. By using the GPS coordinates (latitude and longitude) of all

customer locations, the goal is to calculate the most efficient route that starts with paper collecting point and ends at Mr. Kumar's home, while visiting each delivery point exactly once.

Justification:

TSP delivers a step-by-step delivery sequence that minimizes total travel distance and operational cost. This makes it ideal for small-scale, fixed-loop deliveries like newspaper distribution. Unlike clustering, which provides rough zones, TSP results in an actual delivery path, which Mr. Kumar can follow directly.

5.7 Classification (Customer Churn Prediction)

The goal is to build a classification model to identify which customers are at risk of churn (i.e., likely to become inactive). A customer is labeled as "churned" if they are marked as Inactive in the master sheet or have missed a threshold number of deliveries (e.g., more than 5) during the 3-month period. Random forest is used for this prediction.

Justification:

Random forest is selected because it is an ensemble learning and it is best where the relationships in the dataset is non-linear. Random forest is better than Decision tree because it reduces overfitting because here the active customers rate or way more than churned customers.

6. Results and Findings

6.1 Newspaper Demand Prediction

As part of the demand analysis, daily newspaper delivery counts were extracted from the Daily Delivery Log dataset for each newspaper title. Using these counts, time series forecasting was implemented using the Facebook Prophet model for multiple newspapers.

The model outputs include:

1. A **blue trend line** representing predicted newspaper demand over time
2. A **light blue** confidence interval band showing the uncertainty range
3. **Black dots** indicating actual delivery data used to train the model

Link to the all newspaper forecast pngs : Forecast Pngs

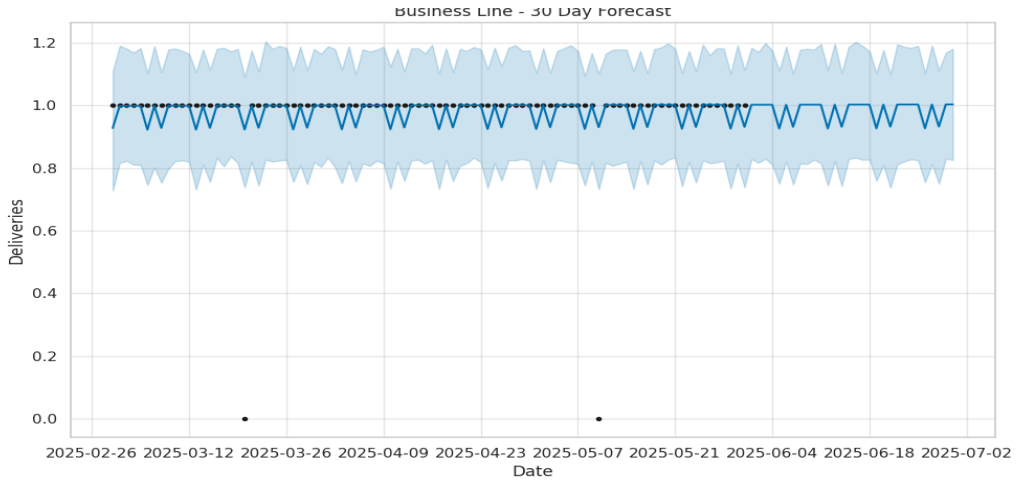


Figure 7 Business Line Forecast

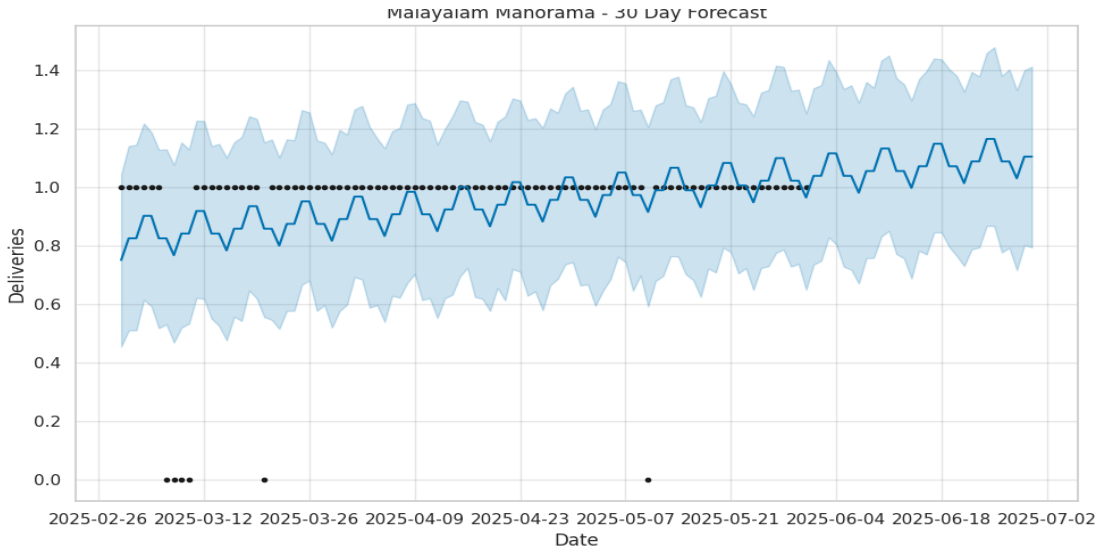


Figure 8 Malayalam Manorama Forecast

Observations :

Business Line is **low variability in demand** can be seen in Figure 7 and it does not need aggressive demand planning whereas Malayalam Manorama is **dynamic and inconsistent behavior** can be seen in Figure 8 which needs better stock planning.

6.2 Travelling Salesman Problem

As part of the delivery cost optimization, customer locations were geocoded using latitude and longitude based on their delivery addresses. These coordinates were then plotted on a folium map to visually represent delivery points, with each location marked by a numbered orange circle indicating how many customers share that spot (Refer Figure 9).

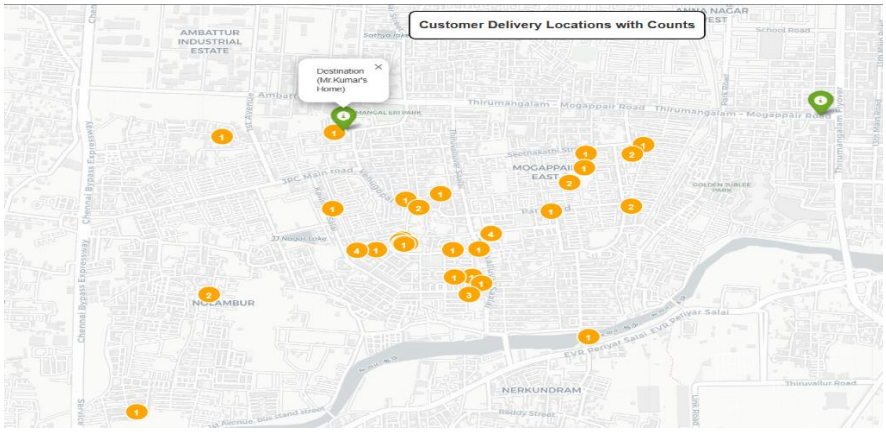


Figure 9 Customer Delivery Locations

In addition to that, a reference point (Mr. Kumar's starting location) and a destination point were marked in green to serve as fixed anchors for future path calculations. TSP will be done in the future to find out the shortest path so that the fuel cost can be reduced and can also save time.

6.3 Customer Churn Prediction

To identify customers at risk of churn (i.e., likely to become inactive), a binary classification setup was established using delivery and subscription-based features. The churn label was generated by marking a customer as "churned" if they were listed as Inactive in the Customer Master Data or had missed more than 5 deliveries in the 3-month period. For midterm analysis, a Random Forest Classifier was trained using the following engineered features: Total_Deliveries, Missed_Deliveries, Subscription_Length, Preferred_Newspapers_Count, Magazine_Subscribed .

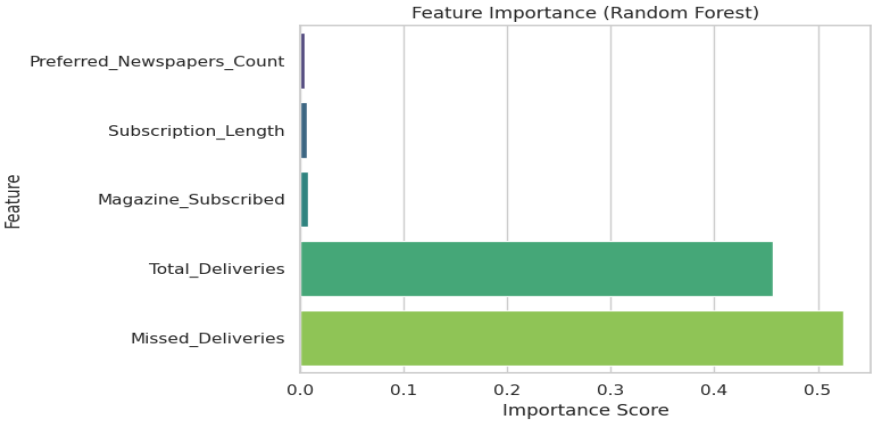


Figure 10 Feature Importance

Observations :

Missed Deliveries was the most influential feature, indicating that customers who missed a higher number of deliveries are significantly more likely to churn (Refer Figure 10).