DATA-DRIVEN MARKETING STRATEGIES REPORT

Marketing Analytics Project

Farabi Issa (922689)

Wani Godwill (925318)

Introduction and Business Context

Objective of the Project

The objective of this project is to develop quantitative, data-driven strategies to:

- Understand customer behavior and value
- Identify customers at risk of churn
- Interpret customer satisfaction signals
- Design actionable and economically justified marketing campaigns

To achieve this goal, the project integrates Exploratory data analysis, RFM customer segmentation, Supervised churn prediction, Sentiment analysis on customer reviews and a fully specified data-driven marketing campaign

Business Questions Addressed

The analysis is driven by the following core business questions:

Customer Value

- Who are the most valuable customers?
- How is revenue distributed across the customer base?

Customer Behavior

- How frequently do customers purchase?
- How recent is their activity?

Customer Risk

- Which customers are likely to churn in the near future?
- Can churn be predicted before it happens?

Customer Satisfaction

- How do customers perceive the service?
- Does dissatisfaction anticipate churn?

Marketing Actionability

- Which customers should be targeted?
- When should campaigns be triggered?
- What incentives are economically justified?

Data Overview

The project uses a relational dataset describing the full customer lifecycle; Users & Profiles: demographic and platform attributes, Orders & Order Details: transactional behavior and revenue, Reviews &Labeled Reviews: textual feedback and sentiment. This structure allows the integration of behavioral, demographic, economic, and emotional signals.

- **Extract Transform Load (ETL) Pipeline**

Data Acquisition

Downloaded all raw datasets from Google Drive using automatedscripts and organized them consistently under data_input/

Structured Loading

Imported each dataset with strict dtype rules, controlled NA parsing, and unified schema definitions for IDs, timestamps, booleans, and text fields.

- **Cleaning & Normalization**

Standardized textual fields (whitespace cleanup, lowercasing, noise removal). Normalized categorical values and parsed list-like attributes for consistent downstream use.

- **Key Integrity Validation (PK + FK)**

Validated primary keys by detecting null or duplicate IDs. Ensured referential integrity by checking that all foreign-key fields (e.g., user.uid, order.uid) existed in their parent tables, removing inconsistent rows.

- **Business-Rule Filtering**

Applied domain constraints such as valid status categories (item.status, card.status) and removed logically invalid records (e.g., negative amounts, impossible timestamps).

- **Persisting Cleaned Data**

Saved all cleaned and validated dataframes as PKL files to support fast, reproducible loading for modeling and analysis tasks.

Data Analysis:

Business Context & Key Insights

Objective

- To understand who the customers are, where revenue comes from, and how customers behave, in order to support segmentation, churn prediction, and data-driven marketing strategies.

Data Preparation

- Integrated **o**rders, order details, users, and profiles into a unified dataset (customer_orders_df).
- Standardized revenue (converted from cents to euros).
- Ensured consistent customer identifiers for reliable aggregation and analysis.

Insights from Data analysis

Geographic Insights

- Revenue is highly concentrated geographically.
- Milan is the dominant market, generating by far the highest total revenue.
- Other key cities (Bergamo, Brescia, Monza) contribute significantly but at much smaller scale.
- Customer volume and revenue are strongly aligned geographically, confirming urban market dependence.

Business implication:

Focus marketing investment and partnerships on top cities, while testing growth strategies in secondary markets.

Customer Loyalty & Purchase Behavior

- 58.1% of customers are one-time buyers, while 41.9% are repeat customers.
- Some smaller cities show higher orders per customer, indicating stronger local loyalty despite lower volume.

Business implication:

There is a clear opportunity for retention campaigns, especially targeting first-time buyers to increase repeat purchases.

Insights from Data analysis

Demographic Profile

- Average customer age: ~28 years (young user base).
- Gender distribution: ~59% male, ~41% female.
- Customer base is predominantly Italian-speaking, with English as the main secondary language.

Business implication:

Marketing communication should prioritize Italian language, mobile-friendly channels, and products appealing to young adults.

Product Performance

- Skipass products dominate revenue, representing the core business.
- Bundles, insurance, rentals, and experiences provide meaningful but secondary contributions.

Business implication:

- Skipass should remain the anchor product.
- Cross-selling opportunities exist for rentals, insurance, and experiences.

Channel & Source Analysis

- Website generates the majority of revenue, followed by the mobile app.
- Most users register via the website, with the app as a growing channel.

Business implication:
Optimize website conversion and use the app for retention and personalized offers.

•

Temporal Trends

- Order volume shows strong seasonality, with clear peaks aligned to the winter season.
- Confirms demand is highly time-dependent.

Business implication:
Campaigns should be seasonally timed, with churn prevention before peak periods.

RFM Analysis
Purpose of the RFM Analysis

The objective of the RFM analysis is to segment customers according to their economic value and engagement behavior, in order to support data-driven marketing decisions.

Business questions this analysis aims to answer:

. Which customers are the most valuable for the business?

- Which customers are actively engaged, and which are becoming inactive?
- How can customers be grouped into actionable segments for targeted strategies?

RFM Analysis

Data Preparation and KPI Construction

Data sources

The analysis uses cleaned and serialized (`.pkl`) datasets, including

-orders,

Order_details,

Users,

Profiles

cards

Building the Order-Level KPI Table

To perform RFM analysis, transactions must be aggregated at the **order level**.

- Filtering fulfilled items. Only items with status `"fulfilled"` are kept, ensuring that revenue reflects completed purchases.
- Date parsing. Order creation dates (`createdAt`) are converted to datetime format to enable temporal analysis.
- Joining orders and order details.
- Revenue computation
  Item revenue is computed and aggregated to obtain total revenue per order.
- Order_kpi table: user_uid, order_uid, order_date, order_revenue, source, tenant

Construction of RFM Variables

Definition of RFM Metrics & Interpretation of Raw RFM Values

- Using the `order_kpi` table, RFM metrics are computed at the level:
- **Recency (R)**:
  Number of days since the customer's most recent purchase, computed relative to a snapshot date. Customers with low recency values are recently active
- **Frequency (F)**:
  Number of distinct orders placed by the customer. High frequency indicates repeat purchasing behavior.
- **Monetary (M)**:
  Total revenue generated by the customer across all orders. High monetary values identify economically valuable customers

RFM Scoring Methodology

- **Quantile-Based Scoring**
- Each RFM variable is transformed into a score from 1 to 5 using quantile thresholds.
- Score **5** represents the most desirable behavior
- Score **1** represents the least desirable behavior

Special handling is applied to Recency:

- Lower recency (more recent purchase) - higher score

This approach ensures Balanced segment sizes, Robustness to outliers and comparability across dimensions.

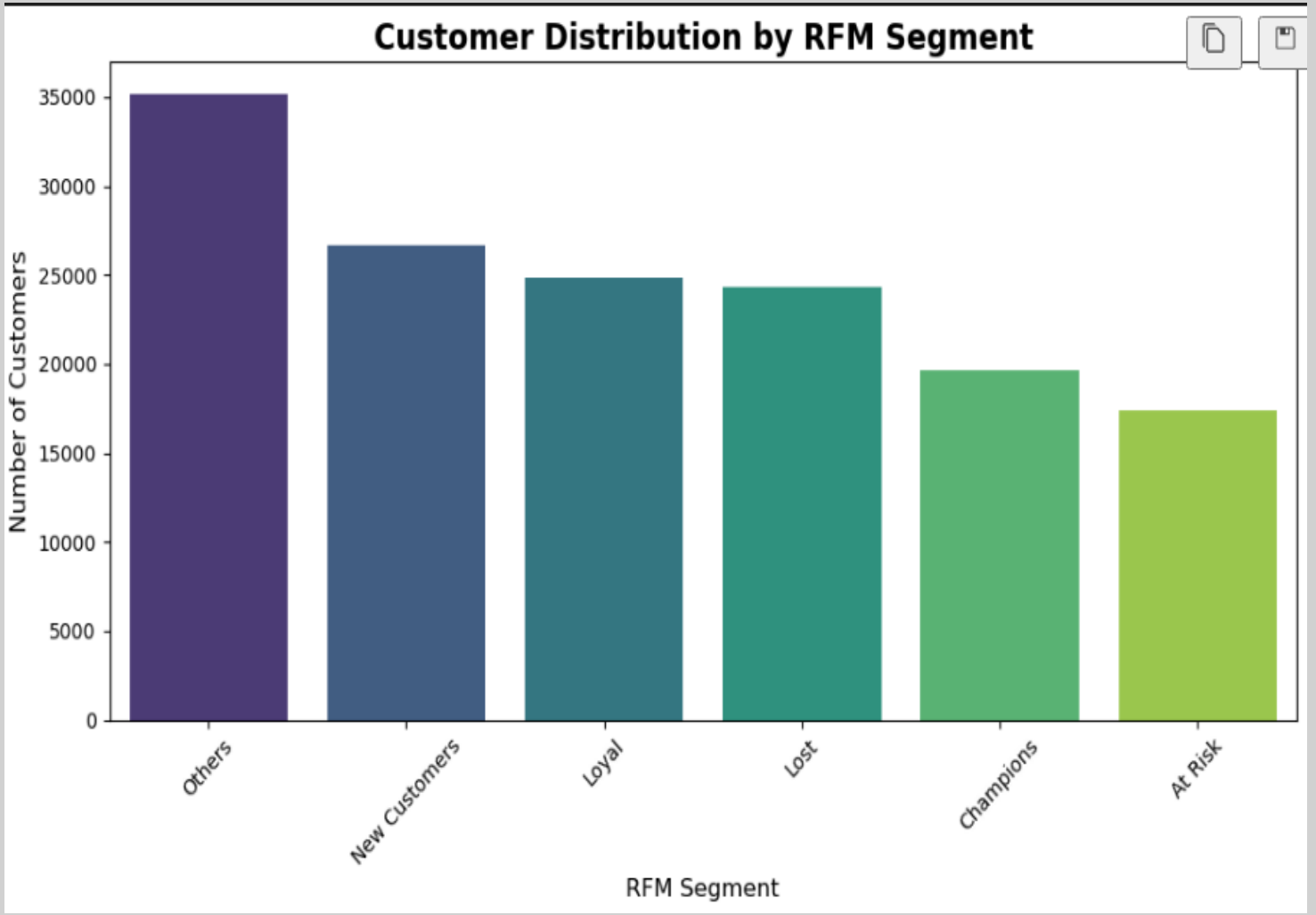
- **RFM Score Composition**

For each customer: R_score, F_score, and M_score are computed and Scores are concatenated into a three-digit **RFM code** (e.g. 434, 113) providing a compact behavioral fingerprint for each customer.

Customer Segmentation Logic

Segment Definition

- Customers are assigned to business-interpretable segments based on their RFM scores
- **Champions**: high R, F, and M
- **Loyal**: frequent and recent customers
- **New Customers**: recent but low frequency
- **At Risk**: previously frequent but not recent
- **Lost**: low recency, frequency, and monetary value
- **Others**: mixed or intermediate behaviors



Churn Prediction Model

The objective of this analysis is to:

- Predict which customers are likely to churn
- Estimate churn propensity at the individual customer level

Using the Order-KPI table built previously in RFM analysis, (One row per order, Customer identifier(user_uid), Order date, Order revenue and Contextual attributes (source, tenant)

Enrichment with User and Profile dataset.

Behavioral features are enriched with:

- Acquisition channel (source), Language, City, Gender (sex) and Customer level

This allows the model to learn structural and demographic churn patterns, such as:

- Channel-specific churn risk
- Geographic or demographic differences

Churn Definition Logic

A time-based churn definition is adopted to avoid subjective assumptions:

- Snapshot date:The most recent order date in the dataset (2025-02-15).
- Churn period:90 days of inactivity.
- Cut-off date (T0)**:** Snapshot date minus 90 days 2024-11-17.
- A customer is labeled as: **Churned (1)** if no order occurs after T0 **Active (0)** otherwise

Churn Distribution Analysis

The first churn distribution shows:

- 79,106 churned customers
- 69,016 active customers

This reveals:

- A slightly imbalanced dataset and a significant churn problem, with more than half the customer base inactive over the defined window

Churn analysis

Feature Engineering (Leakage-Free)

Avoiding Data Leakage

- features are computed using only data before T0. This prevents future information from leaking into the model and artificially inflating performance. This design choice ensures Realistic evaluation and Trustworthy deployment in production environments

Behavioral Features

- Frequency (freq_total): number of orders before T0
- Monetary value (monetary_total): total revenue before T0
- Tenure (tenure_days): time between first order and T0
- Average order value: monetary / frequency
- Recent activity (orders_last_30d): orders in the last 30 days before T0

Feature Selection and Design Matrix Construction

Feature Types

- Numerical features (frequency, monetary value, tenure, recency intensity)
- Categorical features (source, language, sex, level, city)

Preprocessing Strategy

- To ensure model compatibility and robustness:
- Missing numerical values are imputed using the median
- Numerical variables are standardized
- Categorical variables are one-hot encoded
- Missing categorical values are treated as "Missing"
- The resulting design matrix contains: 101,748 rows and 400 encoded features

Model Choice and Benchmarking

Algorithms Selected

- **Logistic Regression:** Interpretable, Linear baseline and Useful for understanding directional effects
- **Random Forest:** Non-linear, Captures interactions and Robust to feature scaling and noise.

Evaluation Metrics

•ROC-AUC
Measures ranking quality across all thresholds (primary metric)

•Precision & Recall
Particularly important for churned customers

•F1-score
Balances precision and recall

•Accuracy
Reported but not relied upon due to class imbalance

Model Performance and Interpretation

Logistic Regression Results

- ROC-AUC: 0.793
- High recall for churned customers (0.96)
- Lower recall for active customers (0.29)

Interpretation:

- Strong ability to identify churners
- Acceptable trade-off for retention campaigns, where missing churners is more costly than contacting some active users

Random Forest Results

- ROC-AUC: 0.7569
- Balanced but weaker than logistic regression

Interpretation:

- Captures non-linearities
- Slightly inferior ranking performance in this context

Model Selection

The **Logistic Regression model** is preferred for: Higher ROC-AUC, Greater interpretability and Stability in large-scale deployment

Sentiment Analysis

Customer reviews provide a complementary and essential perspective, capturing emotions, satisfaction, and dissatisfaction, which often precede behavioral changes such as churn

The objective of this sentiment analysis

- Automatically classify customer reviews into sentiment categories
- Scale sentiment evaluation to hundreds of thousands of reviews
- Integrate emotional signals into customer analytics and marketing decisions

Data Context and Preparation

Data Sources (wo datasets are used)

- **Unlabelled reviews** (reviews.csv): Containing over 105,000 customer reviews, each linked to a user and written in free text.
- **Labelled reviews** (reviews_labelled.csv): A large supervised learning dataset with 462,744 reviews annotated as: positive, neutral and negative.

These datasets are loaded, validated, cleaned, and serialized into .pkl files for reproducibility

Data Cleaning and Validation

Before modeling, extensive text preprocessing is applied

- Lowercasing all text
- Removing URLs and HTML tags
- Removing non-informative characters
- Normalizing whitespace and punctuation

Sentiment Modeling Approach

Problem Framing

Sentiment analysis is framed as a multi-class supervised classification problem with three target classes: Negative, Neutral and Positive

Feature Representation: TF-IDF

The textual data is transformed using TF-IDF (Term Frequency – Inverse Document Frequency):

Why TF-IDF?

- Efficient for large text corpora, Interpretable and Proven baseline for sentiment analysis

Algorithm Choice

A Logistic Regression classifier is trained with: High iteration limit (max_iter=1500) and Class balancing to handle uneven sentiment distribution

Rationale:

- Strong baseline for text classification, Scales well to large datasets and produces stable and interpretable results.

Model Evaluation and Results

Performance Metrics

- The model is evaluated using: Precision, Recall, F1-score and Overall accuracy

Model Evaluation and Results

Classification Results

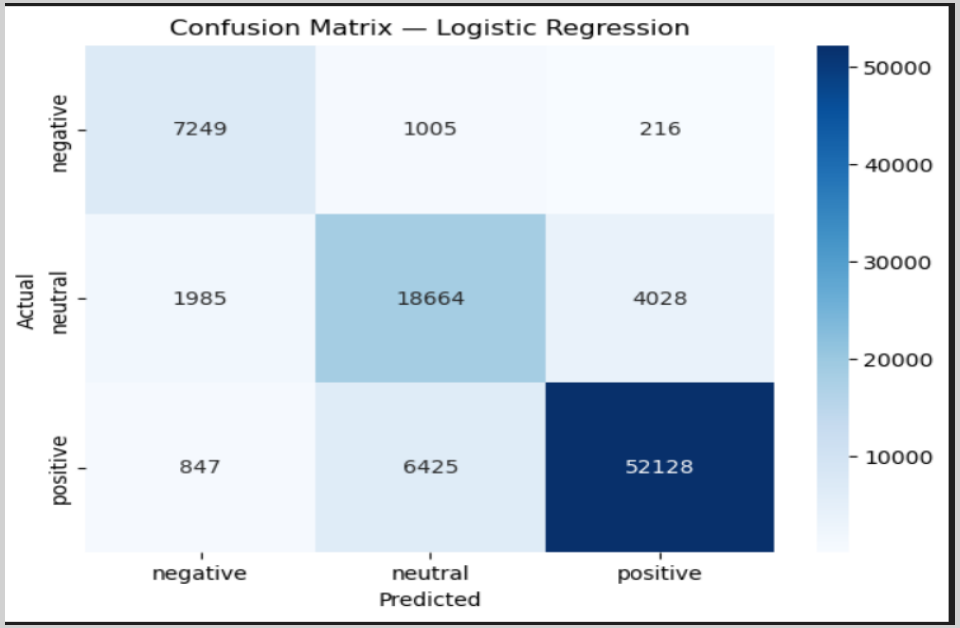
Sentiment	Precision	Recall	F1-score
Negative	0.72	0.86	0.78
Neutral	0.72	0.76	0.74
Positive	0.92	0.88	0.90

Overall accuracy: **84%**

Interpretation:

- The model achieves 84% accuracy, showing strong overall performance. It is effective at identifying negative sentiment (high recall),crucial for detecting dissatisfied customers early.
- The model performs well on positive sentiment (high precision and recall), allowing identification of satisfied and loyal customers.
- Neutral sentiment is harder to classify due to its ambiguous nature but still reaches acceptable performance.

Confusion Matrix Interpretation



- The confusion matrix highlights; Strong performance on positive sentiment, Some confusion between neutral and negative reviews and Very limited misclassification of strongly positive reviews
- Neutral sentiment represents a semantic middle ground and is inherently harder to classify, while strong emotions are easier to detect.

Business Interpretation of Sentiment Results

Negative Sentiment

- Indicates dissatisfaction
- Often precedes churn
- Requires immediate attention

Neutral Sentiment

- Indicates weak engagement
- Represents a potential early churn signal

Positive Sentiment

- Associated with loyalty
- Opportunity for reinforcement and advocacy

DATA-DRIVEN MARKETING CAMPAIGN
Customer Retention Strategy

Customer Retention Strategy

Customer Retention Strategy

Campaign Objective

- The objective of this marketing campaign is to retain high-value customers who are at risk of churning, using predictive analytics to deliver timely and personalized interventions.
  The campaign leverages three analytical pillars:
- **Customer value segmentation (RFM Model);** identifies which customers are the most valuable.
- **Churn probability estimation (Random Forest Model);** identifies which customers are at highest risk of leaving.
- **Sentiment analysis on reviews;** provides insight into customers' emotional tone and potential dissatisfaction patterns

When the Campaign Is Triggered

Triggering Logic

- The campaign is triggered when the churn model predicts that a customer has a probability **≥ 0.65** of churning:

CHURN_PROB_THRESHOLD = 0.65

- This makes the campaign **reactive but data-driven**: customers become eligible as soon as their churn probability exceeds the threshold

Periodicity

- The campaign is **launched every 30 days**. This frequency aligns with the “orders_last_30d” behavioral feature used in the churn model. It ensures that the model continuously receives new data and re-evaluates churn risk.

Why 30 days?

- Monthly cadence allows the company to respond before customers completely disengage.
- Matches the business’s natural seasonal purchasing cycles

Target Audience

Target Selection Conditions

rom the campaign pipeline (selection block) we see that customers must satisfy two conditions:

- **Condition 1 — High Churn Probability (**scored["churn_proba"] >= prob_threshold)
- **Condition 2 — High Monetary Value (**scored["monetary_total"] >= monetary_cutoff)
- Where the cutoff is the **5**0th percentile of total spending: value_percentile = 0.5
- Thus, only the top 50% of customers in monetary value are considered for retention.

Size of the Target Audience

Number of campaign targets: 34,862

These are customers who:

- Spend above the median total monetary value
- Have ≥ 0.65 probability of churning
- Have meaningful behavioral patterns showing a risk of disengagement
- Are strategically important to retain

Call to Action

Campaign CTA (Call to Action**):**

- Make your next booking with Snowit within the next 14 days and receive an exclusive reward tailored to your profile.

Why 14 days?

- Short enough to create urgency
- Long enough for customers to act
- This CTA pushes the “next order,” which resets churn risk and increases predicted lifetime value

Marketing Promotion Offered

Personalized Incentive Structure

Based on the customer profile, the incentive may include:

- **A. Discount on next purchase (5–15%)**

Ideal for price-sensitive customers

Highly effective for medium-value segments

- **B. Loyalty points multiplier**

Encourages long-term engagement

Suitable for heavy-spending loyal customers

C. Sentiment-Aware Messaging

Using sentiment analysis (e.g in our sentiment_analysis), negative reviewers get:

- Apology messaging
- Customer care follow-up
- Additional reassurance (service quality guarantees)
- Positive or neutral reviewers receive:
- Standard promotional offers
- Loyalty-focused messaging

Cost–Benefit Analysis

Costs

- **Promotional incentives**
  - Discounts reduce margin
  - Loyalty points represent deferred cost
- **Operational execution**
  - Email/SMS cost per user
  - Customer support for sentiment-flagged cases

Benefits.

Retention of high-value customers
Each saved customer yields strong long-term profit.

Reduction in churn rate
Predictive targeting reduces wasted marketing effort.

Revenue uplift
Retained customers often purchase multiple times.

Costs Related to Model Prediction Errors

False Positives (FP)

Model predicts churn, but customer would not have churned.

Cost:

- Unnecessary discount or promotion expenditure
- Lost margin

Benefit:

- Possible increase in goodwill or purchases

False Negatives (FN)

Customer churns, but model predicted they would not.

Cost:

- Loss of a potentially high-value customer.
- Revenue permanently lost.

This is the most expensive error, particularly because the campaign targets high-value users.

Mitigation:

- Use probability threshold tuning
- Increase monitoring frequency
- Incorporate sentiment signals to improve accuracy

- Thank you !