Yinan Wang

10/22/2024

Project Introduction

**Literature Review**

Customer segmentation is a common method in marketing and data analysis. It is used to classify customers according to factors such as their behavior, purchase frequency and consumption amount, and then develop more accurate marketing strategies. Tsai and Chang (2011) used the RFM model to study customer clustering, and they explained and confirmed the importance of weight allocation for different industry characteristics in clustering. Their research shows how marketing strategies can be improved by analyzing customer loyalty and profitability. This approach has great application potential in the aviation industry.

Traditional customer clustering methods such as K-means clustering can only handle smaller data sets due to computational power limitations. As big data technologies become more advanced, tools such as Apache Spark enable the analysis of large-scale data sets. It enables more real-time, scalable customer segmentation. Zaharia et al. (2016) illustrate Apache Spark's ability to handle large-scale distributed data, making it particularly suitable when dealing with the vast customer data of airlines. However, despite the increasing application of big data technology in multiple fields, research on the application of Spark in aviation customer segments is still relatively scarce.

Although there has been much progress in customer segmentation research in the aviation industry, few studies have applied modern big data tools such as PySpark to large-scale, real-time customer segmentation analysis on the Databricks platform. This project will fill this research gap by leveraging PySpark's ability to process large-scale data to perform cluster analysis for diverse customer booking behaviors and preferences.

**Citation**

https://www.ice.nchu.edu.tw/Pic/Writings/2779\_Group%20RFM%20analysis%20as%20a%20novel%20framework%20to%20discover%20better%20customer%20consumption%20behavior%20.pdf

<https://www.researchgate.net/publication/310613994_Apache_spark_A_unified_engine_for_big_data_processing>

**Research Questions:**

1. How to effectively group customers based on booking behavior?
2. How can airlines use customer segmentation to optimize service?

The significance of the research is to use PySpark to do big data processing, help airlines identify high-value customer groups, and finally develop more accurate marketing strategies for them. This not only improves customer satisfaction, but also optimizes operational efficiency, which has a wide range of practical applications.

**Descriptive statistics**

**EDA**

Calculating average passengers per booking:

A screen shot of a computer code

Description automatically generated

Distribution of bookings via Internet versus other channels:

A screenshot of a computer

Description automatically generated

Ratio of round-trip to one-way bookings:

A screenshot of a computer

Description automatically generated

Peak times and days for flight departures:

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated

Wants\_extra\_baggage, Wants\_preferred\_seat, Wants\_in\_flight\_meals: Percentage of customers opting for extra services:

A graph with different colored bars

Description automatically generated

Customer Service Preferences(Bar Plot)

A screen shot of a graph

Description automatically generated

Flight Departure Patterns (Heatmap)

A blue and yellow squares with numbers

Description automatically generated

Sales Channel Usage

A pie chart with numbers and a black background

Description automatically generated

Customer Segmentation (PCA Visualization)

A blue dots in a white background

Description automatically generated with medium confidence