**DATA 690**

**Machine Learning Data Science Project**

**Jean Paul Melendez**

**Jon McKeeby**

**Customer Insights through Transactional Machine Learning**

**June 3rd, 2025**

**Executive Summary**

This project focuses on applying machine learning techniques to predict customer purchasing behavior using a transactional dataset from a UK-based online retailer. The dataset, comprising over 500,000 records, includes information such as invoice numbers, product descriptions, quantities, unit prices, customer IDs, and transaction dates. The primary goal is to extract actionable insights from this data and develop predictive models to support data-driven decision-making in an e-commerce environment.

The strategy begins with extensive data preprocessing to manage missing values, remove duplicates, and correct anomalies. Next, exploratory data analysis (EDA) is performed to identify key trends, seasonality, and patterns in customer purchasing behavior. Feature engineering is used to convert raw variables into meaningful features, followed by clustering algorithms such as K-Means to segment customers based on transactional traits. Supervised learning models, including Random Forest and Logistic Regression, are trained to predict future purchases, customer churn, and frequency of transactions. Model evaluation will rely on standard classification metrics such as accuracy, precision, recall, and F1-score to assess performance and reliability.

The results are expected to yield valuable insights, including the identification of loyal customer segments, high-revenue products, and repeat purchasing tendencies. These findings can inform targeted marketing campaigns, inventory planning, and customer relationship management strategies. By leveraging machine learning, this project demonstrates how businesses can transform raw retail data into predictive intelligence that supports competitive advantage and strategic growth. The outcome reinforces the importance of data science in optimizing modern e-commerce operations and enhancing long-term customer experience through effective predictive analytics and segmentation.

**Table of Contents**

[Project Scope 3](#_Toc1971401152)

[Problem Description 4](#_Toc295479660)

[Project Importance 4](#_Toc309016351)

[Review Existing Research in Area 4](#_Toc26485380)

[Review Critical Success Factors (CSFs) 4](#_Toc1089682474)

[Review Key Performance Indicators (KPIs) 4](#_Toc621793074)

[Dataset Description 4](#_Toc1012618880)

[Data Analytics Tools 5](#_Toc1034140899)

[Project Milestones – Lessons Learned 5](#_Toc561754698)

[Data Selection 6](#_Toc25766198)

[Data Summary 6](#_Toc264919807)

[Data Definition/Data Profile 6](#_Toc751195011)

[Descriptive Statistics 6](#_Toc43271108)

[Data Visualizations 6](#_Toc1698881189)

[Data Visualization Definitions 7](#_Toc2030392255)

[Data Visualization 1 7](#_Toc114070683)

[Data Visualization 2 7](#_Toc2130639856)

[Data Visualization 3 7](#_Toc651651397)

[Summary 7](#_Toc814329542)

[Predictive Models 7](#_Toc841049789)

[Predictive Modeling Definitions 8](#_Toc964245446)

[Predictive Model 1 8](#_Toc774809834)

[Predictive Model 2 8](#_Toc659246463)

[Predictive Model 3 8](#_Toc1316207860)

[Review of Machine Learning Models 8](#_Toc1749511472)

[Final Report 8](#_Toc201099021)

[Findings 9](#_Toc1723296491)

[Review of Ethical, Privacy, and Security Aspects for Your Selected Project 9](#_Toc287563454)

[Review of Success or Completion 9](#_Toc1793924783)

[Recommendations for Future Analyses 9](#_Toc845534454)

[References 9](#_Toc1694641193)

# Project Scope

**Unit 2 Assignment**

# Problem Description

In today’s competitive e-commerce landscape, businesses face mounting pressure to understand customer behavior in order to drive sales, improve inventory management, and deliver personalized experiences. Despite having access to vast amounts of transactional data, many organizations lack the analytical infrastructure or expertise to effectively leverage this data for strategic decision-making. This project addresses this challenge by examining historical transaction records from a UK-based online retail store to identify patterns in customer purchases and build predictive models capable of forecasting future buying behavior. The ultimate goal is to move from reactive sales analysis to proactive, data-driven marketing and operational planning.

The dataset used in this project contains over 500,000 rows of transactions, offering detailed insight into product purchases, quantities, customer IDs, invoice dates, and pricing information. However, the raw nature of the data also introduces issues such as missing values, duplicates, and anomalies that must be addressed before any modeling can be performed. By applying data preprocessing techniques, exploratory analysis, and machine learning algorithms, this project aims to uncover behavioral trends and predict outcomes such as customer churn, high-value segments, or likely repeat purchases. The research problem that I am analyzing is how machine learning can be used to predict customer purchasing behavior and segment customers based on transactional data from an online retail platform.

# Project Importance

I selected this project because of its direct relevance to real-world business challenges faced by modern e-commerce platforms. With the explosion of online retail activity, companies now collect vast amounts of customer transaction data but often lack the expertise to translate that data into actionable insights. As someone interested in the intersection of data science and business strategy, I saw an opportunity to apply machine learning techniques to gain a deeper understanding of consumer behavior. This project also allows me to apply skills acquired throughout my program to a realistic business scenario, reinforcing both my technical capabilities and strategic thinking.

This project is important because it demonstrates how predictive analytics can help businesses move from reactive to proactive decision-making. Online retailers face intense competition and rising customer expectations, which makes it critical to predict purchasing behavior, anticipate customer needs, and personalize marketing efforts. According to Wuennenberg et al. (2023), predictive analytics is key to identifying behavioral patterns and determining key performance indicators that can improve operational efficiency and strategic planning in logistics and retail environments. By implementing predictive models on the Online Retail dataset, this project illustrates how data can be harnessed not just for reporting, but for forecasting and optimization, which are essential capabilities for digital commerce success.

The primary beneficiaries of this project are e-commerce businesses, marketing analysts, data science teams, and business intelligence professionals. Retailers can use insights from predictive models to enhance customer segmentation, tailor product recommendations, reduce churn, and improve inventory decisions. At the same time, the project provides valuable learning experience for data analysts seeking to apply machine learning in practical settings. It bridges the gap between raw data and business strategy, demonstrating the impact of data science on customer-centric innovation and decision-making.

# Review Existing Research in Area

The use of predictive analytics in retail has gained increasing attention in both academic and business communities. Research shows that applying machine learning to customer transaction data can significantly improve decision-making in areas like customer segmentation, inventory control, and marketing strategy. Wuennenberg et al. (2023) presented a methodology for determining key performance indicators (KPIs) in internal logistics using a data-driven approach. Their work highlights the importance of data preprocessing and model interpretability when applying predictive analytics in operational settings. These insights are directly applicable to this project, which involves cleaning and preparing transactional data to build models that classify and forecast purchasing behavior in online retail.

Additionally, Daradkeh (2019) emphasized the critical success factors (CSFs) necessary for implementing enterprise-level analytics and visualization systems. His findings revealed that top management support, data quality, flexible infrastructure, and user-centric implementation are vital for success. This aligns with the goals of this project, which depend on the quality of the Online Retail dataset, the effectiveness of data visualization tools, and the interpretability of machine learning models. Other studies have also explored classification algorithms and clustering techniques for customer segmentation, indicating that combining unsupervised learning for grouping customers with supervised models for predicting behavior can offer powerful business intelligence. Together, these studies provide a strong foundation for leveraging machine learning in retail and validate the relevance and feasibility of this project.

# Review Critical Success Factors (CSFs)

Critical Success Factors (CSFs) refer to the essential conditions, elements, or activities that must be present or effectively managed to achieve the desired goals of a project. According to Daradkeh (2019), successful enterprise data analytics and visualization ecosystems rely on several interrelated CSFs, including top management support, data quality and integrity, appropriate team skills, scalable infrastructure, and iterative implementation strategies. These factors serve as guiding principles that help ensure the success of complex analytics projects by aligning strategic goals with the technical and organizational capabilities necessary for effective execution.

To ensure the success of this project, three CSFs stand out. First, data quality and preprocessing are critical. The raw transactional dataset contains noise, missing values, and duplicates, which must be addressed through rigorous cleaning and transformation processes. Second, technical proficiency with machine learning tools such as Scikit-learn, pandas, and visualization libraries is essential to building and interpreting predictive models. These tools will be used for clustering, classification, and evaluation of model performance. Lastly, the project requires strong analytical and interpretive skills to draw meaningful insights from model outputs and translate them into business recommendations. These skills must be supported by a clear understanding of the data science life cycle, which ensures an organized, ethical, and reproducible research process. With the right tools, a structured approach, and technical competency, the project is positioned for successful outcomes.

# Review Key Performance Indicators (KPIs)

Key Performance Indicators (KPIs) are measurable values that organizations use to evaluate the effectiveness of a process, project, or system in achieving predefined goals. In the context of data science and machine learning, KPIs are essential for assessing the quality and impact of predictive models and determining whether they are delivering value to stakeholders. According to Abbasian et al. (2023), KPIs serve as “foundation metrics” that guide performance evaluation and ensure model outputs align with business objectives. They help quantify not only technical success, such as model accuracy, but also broader organizational outcomes like increased efficiency, profitability, or user satisfaction. Therefore, selecting appropriate KPIs is crucial for interpreting the success of this project beyond technical performance alone.

Three specific KPIs will be used to evaluate this online retail analytics project. First, a project completion KPI will be measured by adherence to the milestone timeline defined in the capstone project schedule, including successful delivery of the data preprocessing, modeling, and reporting phases. Second, a model performance KPI will focus on the technical quality of the predictive models, evaluated through metrics like accuracy, precision, recall, and F1-score. These will ensure the models are not only functioning but are also effective at capturing customer behavior patterns. Third, a business impact KPI will assess how well the insights generated from the models support organizational goals, such as improved customer segmentation, targeted marketing, or inventory optimization. Together, these KPIs provide a balanced approach for evaluating both technical success and practical business value.

# Dataset Description

The dataset used for this project is the "Online Retail" dataset, a transactional dataset sourced from a UK-based e-commerce store. It contains over 500,000 rows of transactions made between December 2010 and December 2011. Each transaction includes fields such as Invoice Number, Stock Code, Item Description, Quantity, Invoice Date, Unit Price, Customer ID, and Country. This dataset provides rich insights into customer purchasing behavior, item popularity, and pricing, which makes it an ideal foundation for predictive modeling in a retail setting. Despite some limitations such as missing customer IDs in many rows, it offers sufficient complexity and volume to support a robust machine learning project.

This project qualifies under Option 2 for dataset complexity, as it contains over 100,000 records and more than 15 variables, including numeric and categorical attributes that can be engineered into predictive features. The dataset will be used to train and test machine learning models aimed at predicting purchasing patterns and segmenting customers based on transaction behavior. It is publicly available from the UCI Machine Learning Repository and is frequently used in academic and commercial data science applications. Proper preprocessing, including handling missing values, duplicate records, and date formatting, will be essential before model development. The use of this dataset aligns with best practices for exploratory data analysis and predictive modeling within the e-commerce domain.

# Data Analytics Tools

For this project, I will use Python as the primary programming language due to its versatility, large open-source ecosystem, and strong support for data science and machine learning workflows. Key Python libraries such as pandas and NumPy will be used for data preprocessing and manipulation, including handling missing values, aggregating customer-level features, and transforming categorical variables. For data visualization, Matplotlib and Seaborn will help uncover patterns, trends, and anomalies through bar charts, heatmaps, and time series plots. These tools will support the exploratory data analysis (EDA) phase and enable effective communication of insights to both technical and non-technical audiences.

To develop and evaluate machine learning models, I will rely on Scikit-learn, a widely used Python library that offers a comprehensive suite of supervised and unsupervised learning algorithms. For customer segmentation, I will implement K-Means clustering, while classification models such as Logistic Regression and Random Forest will be used to predict purchasing behavior. I will use Google Colab or Jupyter Notebook as my development environment for ease of code execution, documentation, and collaboration. These tools align with the data science life cycle outlined by Stodden (2020), which emphasizes reproducibility, transparency, and toolchain integration throughout the analytics pipeline. This toolset allows for end-to-end execution from data ingestion to predictive modeling and visualization, ensuring a streamlined and effective analytics workflow.

# Project Milestones – Lessons Learned

|  |  |  |
| --- | --- | --- |
| **Unit 2 Assignment**  **Project Scope** | For Unit 2, I completed the Project Scope section by defining the problem statement, outlining its importance, reviewing existing research, identifying critical success factors (CSFs), establishing key performance indicators (KPIs), and providing a detailed description of the dataset and tools I plan to use. I also ensured that scholarly sources were integrated to support my rationale and used APA 7 formatting for all references. | This assignment helped me develop a clearer understanding of how to structure a data science project from the ground up. I learned how to connect technical elements like machine learning tools and data sources with business goals and outcomes. I also gained insight into the importance of aligning project objectives with measurable success criteria (KPIs and CSFs). Writing the scope forced me to think critically about feasibility, stakeholder impact, and the real-world value of predictive modeling. |
| **Unit 4 Assignment**  **Data Selection and Analysis** | For this assignment, I analyzed the Online Retail Dataset by summarizing its key characteristics, including the number of rows and columns, timeframe, and variable definitions. I documented the origin of the dataset and evaluated its importance for customer behavior analysis in e-commerce. I created a data definition table outlining each variable’s data type, null values, outliers, and quality concerns. I also defined descriptive statistics using a scholarly source and described how variables like Quantity and UnitPrice can be evaluated using measures such as mean, standard deviation, and quantiles. | This assignment helped me deepen my understanding of how to critically assess a dataset before moving into visualizations or modeling. I learned how to identify potential data quality issues and how descriptive statistics can provide immediate insight into trends and anomalies. I also gained experience organizing a data profile in a way that supports future modeling steps. This exercise emphasized the importance of data integrity and reinforced how foundational data review is to any successful machine learning project. |
| **Unit 5 Assignment Presentation** | *Describe the tasks that you completed.* | *Reflect on what you learned as part of this assignment.* |
| **Unit 5 Assignment**  **Visualization** | *Describe the tasks that you completed.* | *Reflect on what you learned as part of this assignment.* |
| **Unit 7 Assignment**  **Predictive Modeling** | *Describe the tasks that you completed.* | *Reflect on what you learned as part of this assignment.* |
| **Unit 8 Assignment Final Report**  **Findings** | *Describe the tasks that you completed.* | *Reflect on what you learned as part of this assignment.* |

# Data Selection

**Unit 4 Assignment**

### Data Summary

The dataset used in this project is the Online Retail Dataset, which contains over 541,909 rows and 8 columns representing customer transaction data from a UK-based online retailer. The dataset spans a one-year period, from December 1, 2010, to December 9, 2011, and includes features such as InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. Each record corresponds to a specific product purchase by a customer, providing valuable insight into consumer purchasing patterns, including the volume of purchases, timing, and pricing. Despite some limitations like missing CustomerID values and negative Quantity entries indicating returns, the dataset offers a robust foundation for predictive analytics.

This dataset is critical for modeling real-world retail behavior and is widely cited in academic and applied machine learning studies. It supports various analytical approaches including customer segmentation, sales forecasting, and churn prediction. Originally collected by a UK-based e-commerce company and made available through the UCI Machine Learning Repository, the dataset provides a standardized and ethically accessible source for data science experimentation. In this project, the data is divided into a training set and a validation set, where the training portion is used to build predictive models and the validation set is used to evaluate their performance. This approach supports fair model assessment and helps prevent overfitting by testing on unseen data.

### Data Definition/Data Profile

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Field Name** | **Definition** | **Data Type** | **Outliers** | **Nulls** | **Potential Quality Issues** |
| InvoiceNo | Unique identifier for each transaction | Categorical | Returns marked with "C" | None | Canceled invoices must be filtered |
| StockCode | Unique product/item code | Categorical | Non-standard codes | None | Some codes represent adjustments or fees |
| Description | Description of the product | Text | Not applicable | Some (~1%) | Some missing or vague product descriptions |
| Quantity | Number of items purchased per transaction | Integer | Negative values | None | Negative quantities indicate product returns |
| InvoiceDate | Date and time of the transaction | DateTime | None | None | Must be parsed for time-based analysis |
| UnitPrice | Price per item in pounds | Float | Zeros or extremely high values | None | Zero prices may indicate promotions or data errors |
| CustomerID | Unique identifier for each customer | Categorical | Not applicable | High (~25%) | Many records missing CustomerID; affects personalization |

### Descriptive Statistics

Descriptive statistics are summary measures that quantitatively describe the main features of a dataset. They are used to present key characteristics of variables such as central tendency (mean, median, mode), dispersion (range, variance, standard deviation), and distribution shape (skewness, kurtosis). These statistics provide a foundation for data understanding before applying more advanced modeling techniques. According to Santucci (2023), descriptive statistics are essential tools in research for summarizing, organizing, and interpreting data in a meaningful way.

In the Online Retail dataset, several fields are practical for descriptive statistical analysis. For example, the Quantity variable can be examined using the mean, which indicates the average number of items purchased per transaction, and the standard deviation, which reveals the spread of purchasing behavior. The UnitPrice field is also valuable for computing measures like median price, interquartile range, and identifying outliers such as unusually high or zero-priced items. InvoiceDate allows for time-based grouping to analyze trends in purchase frequency by month or season. Although CustomerID is categorical, it becomes useful when aggregated, for instance, calculating the average purchase value per customer or the number of transactions per customer, which supports behavioral segmentation.

# Data Visualizations

**Unit 5 Assignment**

*Present the results of the three (3)* ***data visualizations*** *you have created.*

### Data Visualization Definitions

*In at least 2 well-written paragraphs:*

*Identify and define the visualization techniques you will use.*

*Provide at least one (1) citation and reference from a scholarly journal defining each visualization technique that you use.*

### Data Visualization 1

*Present the visualization.*

*In at least 1 well-written paragraph, describe the visualization.*

*In at least 2 paragraphs, review the insights of the visualization.*

### Data Visualization 2

*Present the visualization.*

*In at least one (1) well-written paragraph, describe the visualization.*

*In at least two (2) paragraphs, review the insights of the visualization.*

### Data Visualization 3

*Present the visualization.*

*In at least one (1) well-written paragraph, describe the visualization.*

*In at least two (2) paragraphs, review the insights of the visualization.*

### Summary

*In at least one (1) well-written paragraph, review the findings from the two (2) visualizations. Citations and references are welcomed.*

# Predictive Models

**Unit 7 Assignment**

*Present the results of three (3)* ***predictive models*** *created. Provide comparisons of the models and appropriate conclusions.*

### Predictive Modeling Definitions

*In at least two (2) well-written paragraphs:*

*Identify and define the machine learning techniques you will use.*

*Provide at least one (1) citation and reference from a scholarly journal defining each predictive technique that you use.*

### Predictive Model 1

*Present Predictive Model 1.*

*In at least two (2) well-written paragraphs, describe the predictive model and its findings.*

### Predictive Model 2

*Present Predictive Model 2.*

*In at least two (2) well-written paragraphs, describe the predictive model and its findings.*

### Predictive Model 3

*Present Predictive Model 3.*

*In at least two (2) well-written paragraphs, describe the predictive model and its findings.*

### Review of Machine Learning Models

*In at least three (3) well-written paragraphs, summarize your findings.*

*Review the predictive models.*

*Identify if either machine learning model is the champion model.*

# Final Report

**Unit 8 Assignment**

### Findings

*In at least six (6) well-written paragraphs, describe the findings of the project.*

*Visuals and tables are welcomed.*

### Review of Ethical Aspects for Your Selected Project

*In at least three (3) well-written paragraphs, evaluate your execution of the project.*

### Review of Success or Completion

*In at least two (2) well-written paragraphs, evaluate your execution of the project.*

### Review of KPIs

*In at least three (3) well-written paragraphs, evaluate your project in relation to the KPIs.*

### Review of Lessons Learned

*In at least three (3) well-written paragraphs identify three Lessons Learned.  One paragraph per lesson learned.*

### Recommendations for Future Analysis

*In at least two (2) well-written paragraphs, identify opportunities for future analysis.*

# References

Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Abad, Z. S. H., Thieme, A., Sriram, R., Yang, Z., Wang, Y., Lin, B., Gevaert, O., Li, L.-J., Jain, R., & Rahmani, A. M. (2023). *Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI* [Working paper]. arXiv. <https://doi.org/10.48550/arXiv.2309.12444>

Chen, D., & Lundberg, S. (2011). *Online Retail Dataset*. UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/Online+Retail>

Daradkeh, M. (2019). Critical success factors of enterprise data analytics and visualization ecosystem: An interview study. *International Journal of Information Technology Project Management, 10*(3), 34–55. <https://doi.org/10.4018/IJITPM.2019070103>

Santucci, A. C. (2023). Data description in research. *Salem Press Encyclopedia of Health*. <https://search-ebscohost-com.ezproxy.umgc.edu/login.aspx?direct=true&db=ers&AN=93871877&site=eds-live&scope=site>

Wuennenberg, M., Muehlbauer, K., Fottner, J., & Meissner, S. (2023). Towards predictive analytics in internal logistics – An approach for the data-driven determination of key performance indicators. *CIRP Journal of Manufacturing Science and Technology, 44*, 116–125. <https://doi.org/10.1016/j.cirpj.2023.05.005>