National College of Ireland

A systematic evaluation of regressions and loss functions for the prediction of monetary value in RFM analysis

MSc Research Project
Data Analytics

Shiva Prasad Aruva
Student ID: x22115188

School of Computing
National College of Ireland

Supervisor: Dr Giovani Estrada

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Shiva Prasad Aruva
Student ID:	x22115188
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr Giovani Estrada
Submission Due Date:	04/12/2023
Project Title:	A systematic evaluation of regressions and loss functions for the prediction of monetary value in RFM analysis
Word Count:	11190
Page Count:	28

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Shiva Prasad Aruva
Date:	4th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A systematic evaluation of regressions and loss functions for the prediction of monetary value in RFM analysis

Shiva Prasad Aruva
x22115188

Abstract

RFM, which stands for Recency, Frequency, and Monetary value, is one of the most important methods for market research. It is employed to rank and categorise customers into segments. The traditional ranking of each variable (R, F, and M) is a number from one to five, thus resulting in up to 5*5*5=125 potential customer segments. This research study looks into the reduction of those segments using k-Means and the usage of these clusters for the prediction of monetary value. We take the optimal number of customer segments as a feature for regression. The best loss function and boosting algorithm for the prediction of monetary value is presented. Overall, we show that Extra Trees regression with negative median absolute error loss is the best combination for the prediction of monetary value. By identifying significant trends and distinctive client segments based on their purchase behaviour, the study intends to support targeted marketing initiatives and individualized customer engagement. The suggested approach makes use of the RFM analysis to determine customer scores, then applies the elbow method using k-means clustering to obtain the optimal number of customer clusters, and then use those clusters as a novel feature for the prediction of monetary value. We will show how the monetary value predictions can offer e-commerce enterprises useful insights. Our initial exploration shows encouraging results for the prediction of monetary value using the described methodology.

1 Introduction

The RFM model, which stands for Recency, Frequency, and Monetary Value, is a data-driven technique used by businesses to analyse and categorize their customers based on their purchasing behaviour Liao et al. (2022). It has evolved over decades, adapting to changing business landscapes and technological advancements. This section provides an extensive historical account of the RFM model, its origins, development, and its present-day significance.

Moreover, the origins of the RFM model can be found in the middle of the 20th century, at which point direct marketing began to become more and more popular Yavari et al. (2013).Mail-order companies had a big problem in the beginning: sorting through long lists of potential clients to find actual ones. The people who were most likely to react to their marketing efforts needed to be identified in a methodical manner by marketers. Sorting manually by purchase date was one of the first techniques used. The idea was

simple: new offers were more likely to be accepted by recent customers Alizadeh Moghaddam et al. (2018).

This strategy developed into a systematic methodology that included both frequency and monetary value over time. In this case, the "R" in RFM stands for recency, signifying the recentness of a customer's purchase or interaction with a business. Recency in customer analysis started with the discovery that customers were more likely to respond to marketing campaigns if they had interacted with a business recently. To effectively tailor marketing efforts, businesses initially concentrated on classifying customers into segments based on their recent purchases. The term "frequency" in RFM refers to how often a customer has interacted with the company over a given period of time. It became clear that regular buyers were more devoted to the company and thus more valuable

Meaning of RFM

Before proceeding any further, it is important to formally introduce the concepts that will be used throughout the report.

- ✓ **Recency (R)**: Recency is a crucial factor in understanding customer behaviour. It signifies the time that has elapsed since a customer's last interaction or purchase. Customers who have engaged recently are often considered more relevant and engaged. For instance, a customer who made a purchase in the last week is viewed as more likely to make another purchase compared to a customer who last made a purchase several months ago. Analysing recency helps in identifying customers who might require re-engagement strategies or targeted promotions.

- ✓ **Frequency (F)**: Frequency refers to how often a customer engages with or purchases from the business within a specific period. Customers who make frequent purchases are often seen as loyal customers and are more likely to continue their engagement with the brand. Analyzing frequency helps in understanding customer loyalty and devising strategies to encourage repeat purchases. For instance, a customer who buys from an online store multiple times a month is considered a high-frequency customer.

- ✓ **Monetary Value (M)**: Monetary Value represents the total amount of money a customer has spent on purchases. Customers who spend more are generally more profitable for the business. Analysing monetary value helps businesses segment customers based on their spending capacity and tailor marketing efforts accordingly. For instance, high monetary value customers may be targeted with premium offers or exclusive deals. Understanding customer behaviour through the RFM model involves combining these three components to segment customers effectively. For example, a customer who has recently made frequent purchases with high monetary value is a highly valuable customer and should be retained and nurtured. On the other hand, a customer who hasn't engaged for a while, makes infrequent purchases, and has a low monetary value might be at risk of churning. Analysing these segments helps businesses develop targeted strategies, such as offering exclusive deals to retain valuable customers or re-engaging at-risk customers with special promotions According to Huang et al. (2020), in summary, the RFM model offers an organized method for analyzing consumer behavior and grouping them into relevant customer segments based on their transactional activities. These segments can

then be utilized to customize marketing campaigns, allocate resources optimally, and raise customer satisfaction levels, all of which will eventually improve business performance and foster sustainable growth.

Consequently, companies began to view frequency as an important component of customer segmentation. The "M" in RFM stands for monetary value, which is the total amount of money a customer has spent on purchases over a specific time period. According to Khajvand et al. (2011). knowing the monetarily valued transactions of a customer can provide valuable information about their purchasing power and overall revenue contribution. This feature quickly evolved into a crucial component of customer segmentation. The digital transformation of the RFM model occurred with the introduction of computers and sophisticated data analytics. Large volumes of consumer data could now be processed by businesses effectively, allowing for more accurate and sophisticated RFM analysis. Furthermore, the model was further improved by integrating machine learning and predictive analytics, which made it possible to predict future customer behavior and monetary value with accuracy. Furthermore, by utilizing sophisticated algorithms, artificial intelligence, and big data analytics, the RFM model Dogan et al. (2018) has continued to develop in recent years. It continues to be a vital tool for companies looking to boost revenue growth, improve customer engagement, and optimize marketing strategies. Increasing comprehension and anticipating consumer behavior is essential to modern business strategies. It gives companies the ability to break down and examine customer interactions, providing insightful data about customer involvement and possible financial contributions. The "Monetary Value" component of the RFM model is of particular interest. This element represents the cost incurred by a client, indicating their purchasing power and monetary contribution to the company. Precise estimation of this financial worth is crucial since it allows companies to customize their marketing strategies, maximize resource distribution, and enrich customer connections Christy et al. (2021).

This study initiates a methodical assessment of regression methodologies to precisely predict monetary value within the RFM framework. Regression, a foundational tool in predictive analysis, strives to establish a correlation between independent variables (like recency and frequency) and the dependent variable (monetary value). Our comprehensive assessment aims to pinpoint the most efficient regression methodologies for this precise prediction objective. RFM, which stands for Recency, Frequency, and Monetary Value, is a fundamental framework in the realm of customer-centric analytics and marketing.

The process entails assessing and comprehending consumer behavior based on three crucial factors: the frequency and recency of a customer's purchases, as well as their spending amount (monetary value). Haiying and Yu (2010) In order to improve customer relationships, optimize profits, and simplify marketing strategies, businesses can make well-informed decisions by using the RFM model to understand the dynamics of their customer base.

The RFM (Recency, Frequency, Monetary Value) model has completely changed how companies handle segmenting and customer analysis. It is a strategic framework that aids in the effective understanding and use of customer transaction data by organizations. The RFM model classifies customers into segments that can guide targeted marketing and operational strategies by analyzing the time since a customer's last purchase, how frequently they make purchases, and the monetary value of those transactions. Customer segmentation is one of the RFM model's main applications. Businesses can categorize

their consumer base into segments that reflect their purchasing patterns by allocating numerical scores based on factors like frequency, recency, and monetary value.

Marketing initiatives can be focused on specific customer segments, like "at-risk customers" or "high-value customers. For instance, in order to keep their loyalty, high-value clients might get special offers, and in order to keep at-risk customers from leaving, re-engagement campaigns could be directed towards them. Moreover, RFM analysis can motivate targeted marketing campaigns. Marketing messages that resonate with customers can be more effectively tailored when one is aware of their purchasing behavior. To increase the chance of conversion, a customer who regularly purchases a certain product, for example, might get tailored promotions about it. Furthermore, customer life-cycle management is facilitated by the RFM model. It aids companies in determining a customer's current position in their brand journey. A long-term, high-value customer may receive different nurturing than a new one. Tailoring strategies to each stage of the customer life cycle can maximize revenue generation and customer satisfaction. Furthermore, the RFM model is an effective instrument for forecasting future consumer behavior. Businesses can forecast potential future purchases from a customer by looking at past recency, frequency, and monetary value.Planning marketing strategies, managing inventories, and estimating demand are all made easier by this predictive power Sun and Meng (2022).

All things considered, the RFM model is a flexible and strong instrument that helps companies comprehend, classify, and interact with their clientele. It converts transactional data into meaningful insights that businesses can use to improve customer experiences, make data-driven decisions, and spur growth and profitability in a cutthroat industry.

The RFM (Recency, Frequency, Monetary Value) model is extremely significant when it comes to customer relationship management and marketing. It is a pillar for companies trying to understand and maximize their consumer interactions. Businesses can customize their strategies to increase customer satisfaction and retention while also optimizing sales and revenue by utilizing the RFM model. The RFM model, first and foremost, offers a methodical approach to customer segmentation. It enables companies to group their clientele according to the frequency, recentness, and monetary amount of their transactions. Hu et al. (2020).

Through this segmentation, companies can discern different customer segments, each requiring distinct marketing strategies. For example, a "high-value" customer may warrant personalized attention and exclusive offers to maintain their loyalty, while a "potential loyalist" might need targeted campaigns to encourage repeated purchases. Furthermore, the RFM model aids in resource allocation and optimization. By identifying high-value customers and understanding their purchasing patterns, businesses can direct their resources, such as marketing budgets and promotional activities, towards retaining and satisfying these valuable customers. This targeted approach ensures that resources are efficiently utilized to yield the highest possible return on investment. Predictive Analytics, powered by the RFM model, is another crucial aspect. It enables businesses to forecast future customer Behaviour based on historical purchasing data Jianping (2011). Comprehending a client's prospective future purchases enables proactive tactics, like tailored product suggestions or well-timed promotions. This proactive strategy improves client satisfaction and sales by increasing customer engagement. Furthermore, the customer

life-cycle benefits greatly from the RFM model. It helps to determine a customer's current position in their relationship with the brand. Various phases of the customer life cycle necessitate customized methods, and the RFM model assists in efficiently modifying tactics to fit each phase, fostering connections and promoting sustained allegiance Sheshasaayee and Logeshwari (2018)

Essentially, the RFM model is a strategic asset for modern businesses rather than just a tool. It gives businesses the ability to better understand the nuances of consumer behavior, maximize marketing initiatives, enhance customer satisfaction, and eventually increase profitability. Its importance stems from its capacity to convert data into useful insights, which makes it a vital instrument for companies aiming to prosper in the cutthroat marketplaces of today.

Data collection is the systematic process of gathering, measuring, and recording information to understand a specific phenomenon or answer research questions. It involves using various methods like surveys, interviews, observations, or existing records to collect relevant data. The collected data is then analysed to derive insights, identify patterns, and make informed decisions. Maintaining the quality and validity of the findings depends on ensuring the data is handled ethically, reliably, and accurately throughout the collection process Hu et al. (2020). Hu et al. (2020).Further Data analysis involves examining, cleaning, transforming, and interpreting data to extract valuable insights and inform decision-making. It encompasses methods to uncover patterns, trends, and relationships within the data, enabling a deeper understanding of the subject being studied. Through statistical techniques, visualization, and interpretation, data analysis provides a foundation for evidence-based decision-making in various fields, aiding in problem-solving, strategy development, and optimizing outcomes. Moreover RFM analysis is a technique used by businesses to segment and understand their customers based on three key dimensions: Recency (last purchase), Frequency (number of purchases), and Monetary Value (total spending). By assigning scores and categorizing customers into segments based on these dimensions, businesses gain insights into customer behavior and preferences Zong and Xing (2021).

It evaluates three key aspects: recency of the last purchase, frequency of purchases, and the monetary value of those purchases. Customers are segmented based on these factors to identify their engagement levels and potential value. The model helps businesses target specific customer segments with tailored strategies, improving marketing efficiency Wei et al. (2010). Recency signifies how recently a customer made a purchase, frequency measures the purchase repetition, and monetary value gauges the amount spent, collectively offering insights into customer preferences and profitability. Next, RFM data consists of three components: Recency, Frequency, and Monetary Value, essential in customer analytics. Recency indicates how recently a customer made a purchase, reflecting their engagement Cheng and Chen (2009).

Frequency signifies how often a customer makes purchases, indicating loyalty. Monetary Value represents the total spending by a customer, reflecting their financial contribution. Collectively, these components help categorize customers into segments based on their purchasing behavior. This segmentation allows businesses to tailor marketing and operational strategies to maximize customer engagement and revenue generation. RFM data is crucial for customer-centric decision-making and efficient resource allocation. Again, In a predictive modeling context, features refer to the input variables or attributes used to make predictions. These are the measurable characteristics that help

in understanding and predicting the target variable Shirole and Saraswati (2021). The target variable, on the other hand, is the variable being predicted or estimated based on the features. The goal of the model is to learn patterns and relationships within the features to accurately predict the target variable. Features provide the basis for prediction, while the target variable is what the model aims to predict or explain using these features.

Additionally A statistical method for comprehending and measuring the relationship between one or more independent variables and a dependent variable is a regression model. It seeks to identify the curve or line that best captures this relationship. Based on the supplied independent variables, the model forecasts the value of the dependent variable. It facilitates comprehension of the effects of modifications to the independent variables on the dependent variable. In many disciplines, including economics, finance, psychology, and more, regression models are used to predict, forecast, and comprehend the causal relationship between variables. The model's goal is to minimize the difference between the predicted values and the actual observed values of the dependent variable Chui-Yu Chiu and Kuo (2009).Moreover Monetary prediction involves forecasting or estimating future monetary values based on historical data, trends, and relevant factors. It utilizes statistical methods, machine learning algorithms, or financial models to analyze patterns and behaviors related to financial transactions, investments, or economic indicators Dogan et al. (2018). The goal is to provide insights into potential financial outcomes, helping individuals, businesses, or organizations make informed decisions regarding budgeting, investment strategies, resource allocation, or financial planning. Accurate monetary prediction is crucial for effective financial management, risk assessment, and achieving financial goals. It empowers stakeholders to navigate financial landscapes with more confidence and optimize their financial resources Dogan et al. (2018).

In the world of modern business, especially in the ever-changing domains of e-commerce and retail, there is a complex problem coupled with a significant opportunity. The identification of consumer bias and inclinations lies at the core of this problem, making it a crucial endeavor in the modern era given the abundance of data available to us. The Recency, Frequency, Monetary (RFM) analysis framework, which dates back to its inception, has become a fundamental concept for defining customer segments based on past purchasing patterns. Predicting Monetary Value, or "M" in RFM, becomes crucial in this framework. The veracious prediction of a customer's forthcoming financial engagements ineluctably underpins the capacity of enterprises to fine-tune marketing strategies, personalize customer experiences, and ultimately engender revenue escalation.

With methodical determination, this report sets out to investigate regression techniques and their applicability in the foresightful estimation of Monetary Value, a fundamental concern that permeates the retail and e-commerce industries. In fact, Monetary Value—which represents total financial patronage—bears a powerful testimony to the financial inclinations of a particular customer.

Our expedition herein is dedicated to the discernment of its intricacies, as deciphered through the prism of regression models. Among these techniques, Hist Gradient Boosting Machine (GBM) neural network emerges as a noteworthy player. It excels in capturing intricate patterns within data and has demonstrated remarkable performance in a diverse array of applications. In this report, we embark on a systematic evaluation of Hist GBM

neural network within the context of RFM analysis to ascertain its efficacy in predicting Monetary Value.

As we proceed, it is critical that we comprehend the profound effects that precise monetary value prediction can have on the e-commerce environment. Our goal in producing this report is to bring regression modeling and RFM analysis together to redefine customer engagement and revenue actualization in the context of e-commerce.

1.1 Research Gap

A number of original ideas can be explored, for instance the use of deep neural networks for the prediction of monetary value. Moreover, the prediction can be made more accurate by incorportating the optimal number of customer segments. Deep neural networks and other advanced regression techniques were used to automatically learn intricate patterns and representations within the data, especially in cases where traditional feature engineering methods may fall short. The inclusion of additional features is expected to improve the predictive power of your models as well. We included the optimal assignment of customer segments as an input feature for regressions. Machine learning algorithms, such as regression or gradient boosting, can leverage these features to discover non-linear relationships and interactions that might not be captured by RFM scores alone. This can lead to more accurate monetary value predictions.

1.2 Research question

This research project aims to compare unsupervised clustering algorithms for RFM segmentation and explore how machine learning algorithms can be effectively harnessed to comprehend diverse customer types and their behaviors within the e-commerce landscape.

Traditionally, market analysts use R and F to predict monetary value (M); however we could also use the optimal cluster assignment for the prediction of M. The central research questions are framed as follows:

1. How do you obtain and use the optimal number of customer segments in the prediction of monetary value?

2. What numerical algorithm give us the best prediction of monetary value?

1.3 Research objectives

In this thesis, the optimal cluster assignment, obtained through k-Means and elbow method, will be included as a feature for the prediction of monetary value using boosting methods. In order to address the research questions, we have to follow a number of steps:

1. Prepare a dataset to calculate RFM: RFM ranks customers in three categories, the R(Recency), F(Frequency) and M(Monetary value). Each category has a number of levels, typically 1 to 5, so each customer belongs to one of the 125 segments.

2. Find the optimal number of customer segments using k-Means: Market analysts normally focus on a few segments that can help optimize marketing strategies and

drive business growth. The specific segments they seek to identify may vary depending on the goals of the analysis and the nature of the business. Include this cluster assignment for the numerical prediction of monetary value.

3. Prepare algorithms to predict monetary value: several algorithms were employed, including regressions, boosting, and neural networks.

We choose to use R and F for clustering; however, other features could be incorporated, such as demographic data or behavioural data to enrich the clustering analysis.

2 Related Work

2.1 Diverse Customer Segmentation Approaches

As the title suggests, An Innovative Approach to Multi-criteria Market Segmentation presents a segmentation algorithm based on Multiple, competing goals López and Chavira (2019). It divides markets by employing pairwise assessments of numerous criteria to pinpoint unique customer segments. This method provides a thorough grasp of customer preferences, which proves advantageous for honing marketing strategies and elevating customer contentment. Furthermore, the paper showcases the practical use of this approach through a case study, underlining its adaptability and usefulness in intricate and ever-changing market environments. Another paper introduces a customer segmentation model that places a strong emphasis on the value contributed by various customer segments Cuadros and Domínguez (2014). It underscores the significance of comprehending customer value generation as a foundation for devising successful marketing strategies. Furthermore, the paper delves into how this model can assist businesses in customizing their marketing strategies to optimize customer value, ultimately leading to improved overall performance and higher customer satisfaction.

2.2 Customer Churn Prediction

In Machine learning algorithms have also been applied to forecast customer attrition in recent studies. For Example a recent paper de Lima Lemos et al. (2022) used transaction history and customer support contacts to predict customer attrition using a random forest method. The results showed that higher churn rates were associated with customer complaints and service delays. This suggests that resolving customer complaints can lower attrition and raise satisfaction.

2.3 Targeted Marketing and Purchase Intent Prediction

Another interesting paper Bulut (2014) utilizes latent topic models to forecast user conversions by uncovering the underlying themes within user queries and ad content. According to the research, this method performs better than conventional models at differentiating between ad clicks that convert well and those that don't, giving advertisers a more efficient way to allocate resources and optimize their campaigns. The study emphasizes how latent topic modeling can help advertisers make better decisions to improve their search advertising performance by improving predictions for ad click conversions

2.4 The Role of Conventional Statistical Approaches

A recent paper Rajula et al. (2020) examined the benefits and drawbacks of traditional machine learning and statistical techniques in the medical domain. It looks into how well they work in different areas of medicine, such as treatment, medication development, and diagnosis. The study helps determine which approach is best for a given medical application by highlighting the advantages and disadvantages of both through a comparative analysis. This study adds to our understanding of the rapidly changing landscape of machine learning in healthcare by offering insightful information about how it might be used to advance diagnosis, treatment plans, medication development, and medical decision-making. They studied customer attrition in a telecom company and contrasted the precision of traditional statistical models with machine learning algorithms. The results made clear how important it is to combine the two methods in order to comprehend customer behavior better.

2.5 Innovative Approaches: Deep Clustering and Tailored Pricing

This paper Yao et al. (2023) presented a novel baseline technique for deep clustering, called Multi-CC: A Fresh Benchmark for Enhanced Deep Clustering, with an emphasis on improving speed and performance. Known as Multi-CC, this method advances the field by providing faster and more accurate clustering, thus enhancing the potential of deep learning models in this situation. A novel technique for clustering time series data is introduced in another fascinating paper by Eid et al. (2021), titled "Novel Deep Clustering Technique and Indicator for Soft Partitioning of Time Series Data." This method uses a special soft partitioning indicator to improve the accuracy of time series data clustering. It has numerous applications in the fields of environmental monitoring, healthcare, and finance.

2.6 Churn Prediction Estimation

This study looks into the prediction of customer churn. A fundamental machine learning task, Malyar et al. (2020), is becoming more and more important as companies gather more and more customer data. Businesses can create customized pricing strategies to retain customers by using this data to develop models that predict customer churn. The study looks at churn prediction methods that are currently in use, presents a method to figure out churn periods, and chooses the best data labeling strategy for binary classification. In practice, the Prozorro dataset is used. Presumably, the dataset was obtained through a collaboration or agreement with the Prozorro system and is not available to the general public. However, the results use ensemble tree methods (Random Forest, XGBoost, LightGBM) and provide access to additional datasets for customer churn prediction, such as the Telco Customer Churn dataset on Kaggle Platform. [1] One of the many contemporary applications is customer churn prediction, which calls for a specialist in the field to have knowledge of mathematics, machine learning, and problem formulation.

[1] `https://www.kaggle.com/datasets/blastchar/telco-customer-churn`

2.7 Machine Learning Models for Customer Relationship Management

This paper examines Goel and Kalotra (2022).that Leveraging Machine Learning in Customer Relationship Management (CRM) that has the potential to revolutionize today's business landscape. This study investigates consumer perceptions and satisfaction levels regarding CRM practices in financial institutions, identifying opportunities for enhanced customer satisfaction. Recognizing customers as a vital business asset, firms allocate substantial resources to marketing for customer expansion. Various supervised and unsupervised machine learning techniques are employed to enhance the customer experience and business profitability. This paper reviews CRM literature, focusing on machine learning for customer identification, attraction, retention, and development. Computational statistics and machine learning closely align, facilitating data-driven CRM strategies that utilize tabulation and other methods to visualize data and relationships.

2.8 Predicting Customer Behavior in Support Channels Using Machine Learning

The reported paper has been explained about enhancing customer satisfaction through consistent support channels is crucial as customer expectations for prompt issue resolution have risen significantly Begović et al. (2023). This research focuses on applying data research and machine learning techniques to predict customer behavior in support channels. Using historical service data, classification algorithms are employed to predict customer patience levels during support interactions, aiming to optimize support center management

2.9 Prediction and Analysis of Customer Churn

The reported work explain about In the context of China's expanding automobile market, this study emphasizes the importance of using customer consumption data for tailored retention efforts Zhang et al. (2023). Analyzing an open dataset of auto dealer customer churn, it employs data analysis, visualization, and various sampling techniques to address sample imbalance. Machine learning models like XGBoost are used to predict and identify lost customers effectively, achieving high Recall (0.926) and AUC (0.842) values, aiding in customized retention measures for automotive enterprises.

2.10 Credit Risk Prediction

The paper had been explained about credit scoring is vital for loan institutions to assess customer repayment capability Li (2019). This study employs the XGBoost algorithm to distinguish between good and bad customers. Comparing it with logistic regression, the research demonstrates that XGBoost significantly outperforms it in identifying non-repaying customers, showcasing its superior performance.

2.11 Prediction and Analysis of Customer Churn of Automobile Dealers

In the rapidly growing Chinese automobile market, this research Zhang et al. (2023) underscores the importance of leveraging customer consumption data for tailored retention strategies. It analyzes an open dataset of auto dealer customer churn from a business intelligence perspective. Employing data analysis, visualization, and various sampling techniques to address sample imbalance, the study employs machine learning models like XGBoost and AdaBoost. The results highlight XGBoost as the optimal predictive model, achieving high Recall (0.926) and AUC (0.842) values. This aids in effectively identifying lost customers and enabling customized retention measures for automotive enterprises.

2.12 Predicting Customer Behavior in Support Channels

In this paper Begović et al. (2023), support channels offer a distinct opportunity to enhance customer satisfaction through consistent issue resolution. Recent surveys reveal heightened customer expectations for swift support services, contrasting with earlier, more patient attitudes. To meet these demands, support channels must deliver exceptional service, demonstrating respect for customer time and choices. This research focuses on applying data research techniques and machine learning to predict customer behavior in support channels, using historical service data and classification algorithms to forecast customer patience during service interactions.

2.13 Customer Churn Prediction in Bank Based on Different Machine Learning Models

This Reserch paper Li and Chen (2022) has been explained about amidst intense competition in the banking sector due to the rise of internet finance, customer retention has become a top priority. This paper utilizes an open dataset to perform descriptive statistical analysis on various features. Logistic, Random Forest, and Support Vector Machine (SVM) models are applied to predict customer churn, with evaluation metrics like AUC curves, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Squared Error (MSE). SVM emerges as the best-performing model, offering suggestions for banks based on feature importance and descriptive statistics to enhance customer retention.

2.14 Customer Behavior Prediction using Deep Learning Techniques

In this paper, Nisha and Singh (2023), enhancing e-commerce success requires a deeper understanding of customer behavior online. Previous research has mostly focused on purchase intent using sales rank as a proxy, but translating intent into action isn't guaranteed. This study analyzes over 50,000 web sessions to predict online shopper actions, highlighting platform engagement and customer characteristics as key predictors. Deep learning outperformed traditional supervised methods like Decision Trees, Support Vector Machines, Random Forests, and ANNs, providing valuable insights for platform development and contributing to e-commerce forecasting literature.

2.15 A Meta-learning based Stacked Regression Approach for Customer Lifetime Value Prediction

The Work in this paper Gadgil et al. (2023) focuses on Modern deep learning approaches, such as gradient boosting machines and extreme gradient boosting, are incorporated into this research to investigate customer lifetime value, which is evaluated using an online shopping dataset and a meta-learning on stacked regression model to illustrate its capacity. The techniques used in this work include the Time Series Regression Problem, which brings a set of basic models like Random Forest, XGBoost, etc., using a stacking-based approach. The input feature set was used to train each of these models. To create the final estimate, the predictions from these models are then combined with the original inputs into a meta-model, such as a linear regression." "A little lower RMSE and MAE were produced by the Stacked Regressor (proposed model), indicating a lower likelihood for company losses.

2.16 Customer Shopping Pattern Prediction: A RNN Approach

In this research work Salehinejad and Rahnamayan (2016) The application of deep learning, recurrent neural networks, recency frequency monetary modeling (RFM), recommender systems, and shopping patterns to predict customer behavior is addressed. A Recurrent neural networks (RNNs) were utilized in the study to support a customer shopping pattern model based on client loyalty number (CLN), recency, frequency, and monetary (RFM) variables. The experiment's findings revealed the effectiveness of RNNs in predicting customers' RFM values, and it was suggested that this concept could be used later on by recommender systems for managing reward schemes and special promotional offers.

2.17 Air Pollution Comparison RFM Model Using Machine Learning Approach

This Research Mohammad and Kashem (2022) Mainly focuses on Air-pollution Comparison RFM Model Using Machine learning With Specific Implementation of K means clustering and the elbow method Approach. This paper includes the air pollution data in Urban areas in clustering Values. The methods in this paper included the RFM With Air pollution clinical data and outcomes Data Pre-processing, Excuting RFM Values With Air Quality, Cluster declaration Improving efficiency with k-means applying the elbow method for obtaining K values with an optimal solution.

The form has not achieved well with the clusters of various volumes and different quantities. The Limitation gaps in this research can focus on developing different clustering algorithms with an air pollution prediction model.

3 Methodology

The methodological steps followed in the present research report are described below:

Goal and Application Understanding -is the first step in KDD, 1and prior knowledge of the subject matter is necessary. The decision regarding the use of the data mining-generated patterns and modified data is made here, and this is a crucial assumption that,

if not made appropriately, might result in inaccurate interpretations and unfavorable outcomes for the user.

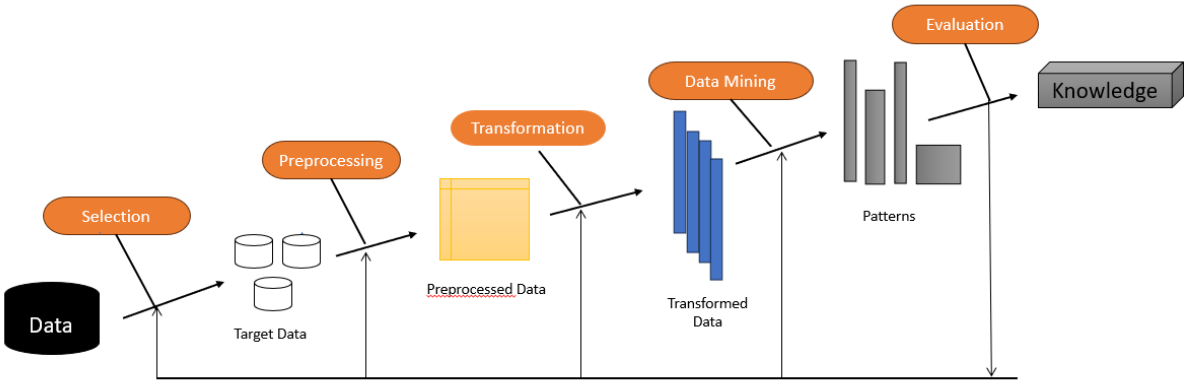

Figure 1: KDD process
Fayyad et al. (1996)

3.1 Knowledge Discovery in Databases (KDD)

- **Data Collection** – Relevant customer data was collected from e-commerce platforms or databases. The dataset used included the following attributes: invoice number, stock code, description, quantity, invoice date, unit price, customer ID, and country.

- **Data Integration** – Data from various sources of heterogeneous data is combined into a single source through the process of data integration. The acquired data is selected and grouped into relevant clusters based on the quantity, quality, and significance of its accessibility. These characteristics are essential for data mining because they form the foundation of data mining and influence the types of data models that are produced. Data integration is accomplished using data synchronization and migration technologies, as well as the ETL process (extract, load, and transform).

- **Data Cleaning** – Here Useless information is eliminated from the data collection through the process of data cleaning. Cleaning is performed by KDD when values are absent. Low-quality, redundant, and noisy data is removed from the dataset. This step is important to improve the validity and effectiveness of the information.

As a preprocessing step, a check for duplicates was performed. In particular, the data shape and unique values in different columns were checked. Nulls in the customer ID column were also checked. If an invoice number exists for a null customer ID where a customer ID is present, the customer ID nulls were filled. Since the customer IDs are missing, these orders were assumed to not have been made by the customers already present in the dataset (i.e., because those customers already have IDs).

These orders are also not wanted to be assigned to those customers because the insights drawn from the data would be altered.Instead of dropping the null CustomerID values which amount to 25 percent of the data, let's assign those rows a unique customer ID per order using InvoiceNo. This will act as a new customer for each unique order.It is checked if InvoiceNo has a unique mapping with Customer ID so that each InvoiceNo corresponding to Null CustomerID can be assigned as a New Customer. Since both values are equal, we know that all the different orders that didn't have a customer ID got assigned a unique NewID and no duplicates were created.The object type was converted to datetime for InvoiceDate and the first and last dates were checked.The cancellations column was added based on the definition that InvoiceNo starts with C. The UnitPrice variable was analyzed.

Data Transformation – Transform customer data into their RFM scores:

• Recency (R) - The number of days since each customer's last purchase is calculated. This can be calculated from the Invoice Date attribute. Customers with the lowest number of days will be assigned the highest R score.

• Frequency (F) - The total number of transactions or orders for each customer is counted from their records in the dataset. Customers with the highest transaction counts are assigned the highest F score.

• Monetary (M) - The total sales amount or revenue generated from each customer is calculated by summing the Quantity Unit Price for all invoices belonging to a customer. Customers with the highest revenues are assigned the highest M score.

- **Patterns**– This is the procedure's most widely recognized feature. The visualizing, increasing, and plotting of such patterns are done in a way that is especially helpful for the KDD process. To accomplish our research goals we generated various graphs. This stage of the process incorporates the techniques of grouping, clustering, and regression Here we Build scatter plots, heat maps, etc. to visually explore relationships between RFM attributes and clusters.

3.2 RFM Analysis

The gathered data was subjected to RFM analysis, which yielded RFM scores for every customer. Based on the relevant metrics, scores for recency, frequency, and money were assigned. This stage establishes a foundation for segmentation by quantifying consumer behavior. The day the analysis was performed was designated as the analysis date, and one year's worth of data from the chosen date was used to calculate recency. This day was taken as the next-to-last date in the data. These rows were removed as customerID even though columns were made to handle CustomerID Nulls (NewID) because the analysis would be distorted by these fictitious customer IDs, particularly with regard to frequency. For each customerID, the recency, frequency, and monetary value columns were determined by combining the remaining dataset.

The lowest frequency, monetary value, and highest recency value are allotted to customers with the lowest RFM scores, and vice versa. Using RFM scores, manual segments can be created in this way: a group of loyal customers with high frequency, a group of high spenders with high monetary values, and a group of lost customers with high recency. Good and Loyal Customers with High RFM values may receive rewards; they do not require significant discounts. To keep customers with high recency (as well as high frequency and monetary values) from leaving, aggressive discounting could be used.

3.3 Traditional K-means Clustering

The traditional K-means clustering algorithm was applied to the RFM scores. The elbow method was used to determine the optimal number of clusters (K). Then, the K-means algorithm was implemented to partition customers into distinct clusters based on their RFM scores, with a range of k from 2 to 40 values of K in K-means.

3.4 Elbow Method

The elbow method is a heuristic technique used to identify the optimal number of clusters (K) in a K-means clustering algorithm. It is used to evaluate the within-cluster sum of squared distances between data points and their assigned centroids to help determine a suitable value for K. The approach involves running the K-means algorithm for a range of K values, calculating the sum of squared distances (inertia), and plotting these values against the number of clusters.

We apply this in our K-means clustering by the following steps:

- **Choose a Range of K:** A range of values for K that needs to be explored is selected, typically from 1 to a certain maximum number of clusters, to begin the process.
- **Run K-means for Each K:** K-means clustering is performed for each K in the chosen range, and the sum of squared distances (inertia) between data points and their corresponding centroids is computed.
- **Plot the Elbow Curve:** Create a plot displaying the inertia against the number of clusters (K).
- **Identify the Elbow Point:** The "elbow point," or the point at which inertia starts to decrease more slowly, is located by analyzing the plot. A compromise between reducing inertia and avoiding an unduly complicated model is represented by this point.
- **Select the Optimal K** The optimal number of clusters (K) is identified by the value at the elbow point. This K value is the most suitable choice for your K-means

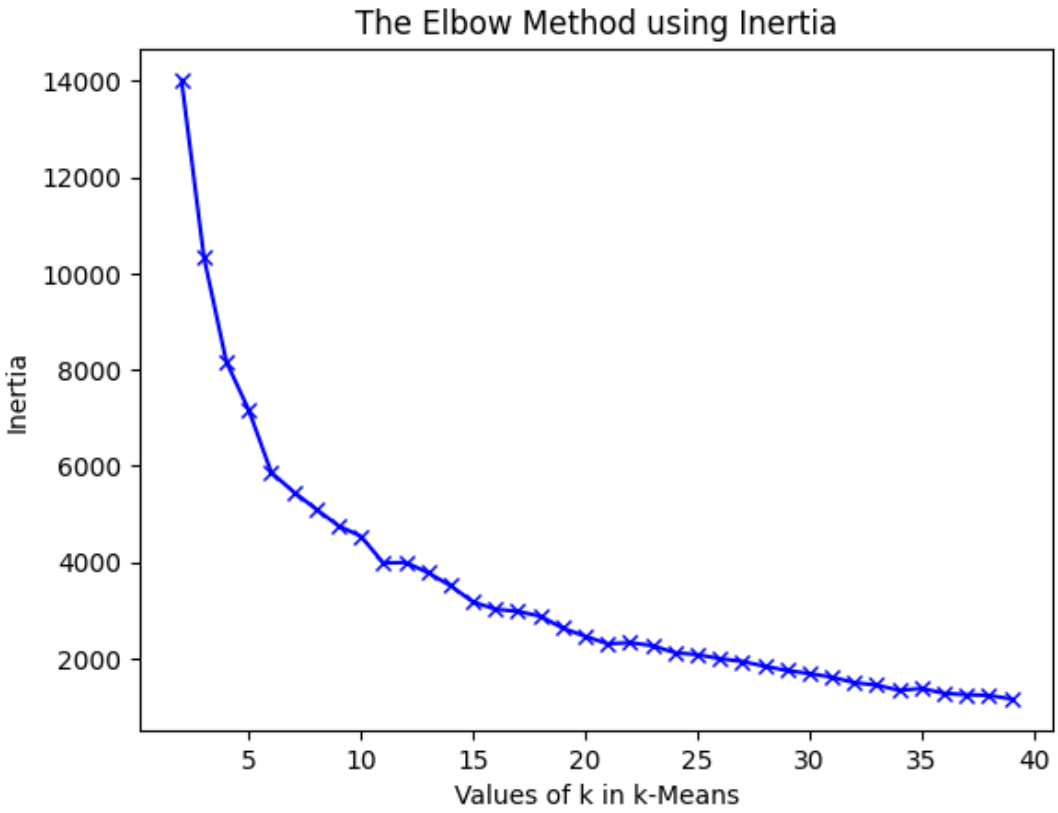

Figure 2: Elbow Method

clustering model based on the elbow method analysis. By using the elbow method, an informed decision about the appropriate number of clusters for your dataset can be made, facilitating more accurate and meaningful cluster analysis. To find an optimal number of clusters, the Elbow method was used. In this method, errors are plotted against K (cluster value) to identify an optimal number of clusters.

3.5 Evaluation of K-means Clustering

The quality of the K-means clustering results was evaluated by calculating metrics such as the silhouette coefficient, intra-cluster distance, and inter-cluster distance. These metrics assessed the cohesion within clusters and the separation between clusters, providing insights into the effectiveness of the clustering approach.

3.6 Neural Networks (NN) Training

Neural networks (NNs) are a type of unsupervised neural network that are used for feature learning, dimensionality reduction, and collaborative filtering. NNs belong to the family of generative stochastic artificial neural networks and can be used as building blocks for more complex neural network architectures, including deep learning models.

In Figure 3 a breakdown of the neural network architecture is given. The components are described below:

1. **Visible Layer:**

• Number of Units: The number of visible units is determined by the number of visible variables, which is equal to the number of selected features. In this case, the number of visible units depends on the length of the selected features list, and there appear to be three selected features (Recency, Frequency, and Monetary).

2. **Hidden Layer:**

• Number of Units: The number of hidden units is set to 100, as specified by the number of hidden variables. We ran it with 50 epochs.

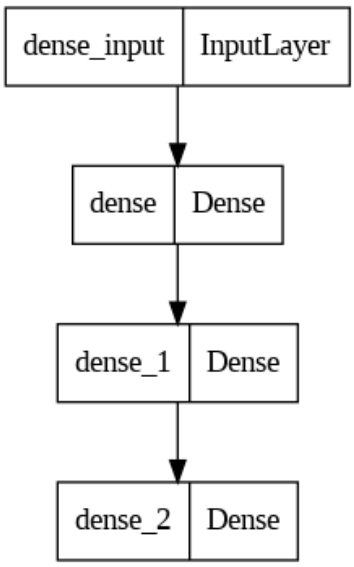

Figure 3: Hand crafted neural network

During training, NNs have their weights adjusted to understand the underlying patterns in the data. An energy-based model is used by NNs to represent the likelihood of different configurations of visible and hidden nodes. They utilize an algorithm called Contrastive Divergence, which adjusts the weights and biases of the NN to minimize the difference between the input data and the Neural Network's reconstructions.

After training on unlabeled data, meaningful features can be extracted by neural networks (NNs). These learned features can serve as the foundation for neural networks, enhancing their ability to understand complex data. By initializing the weights of neural networks with those learned by NNs, a head start is provided for deep learning models, aiding them in their quest to uncover intricate patterns in vast datasets.

3.7 Enhanced K-means Clustering

The K-means clustering algorithm was applied to the extracted latent features obtained from the Neural Networks (NN). The optimal number of clusters (K) was determined using the elbow method or silhouette analysis. Customers were partitioned into clusters based on the enhanced feature representations derived from the Neural Networks (NN).

3.8 Evaluation of Enhanced K-means Clustering

The quality of the enhanced K-means clustering results was evaluated by calculating metrics such as the silhouette coefficient, intra-cluster distance, and inter-cluster distance. These metrics were then compared with those obtained from the traditional K-means clustering to assess the improvement achieved by incorporating Neural Networks (NN)-based feature learning.

4 Design Specification

The machine learning system applied in this project, focusing on client segmentation within the e-commerce domain, is outlined in the design specification, with essential components, techniques, and expected outcomes. A detailed overview of the chosen methods, algorithms, and performance evaluation metrics for the project is provided in this section. Neural Networks (NN), a type of neural network that allows for feature learning and uncovering intricate patterns in customer behavior, were used. This method enhanced the cluster formation from the RFM(recency, frequency, monetary) vectors and facilitated the identification of latent variables and relationships within the data, enhancing the granularity of client segmentation. Hist Gradient Boosting Regressor with and without the use of Neural Networks (NN) was further used to predict the monetary value, verifying that the use of Neural Networks resulted in a significant reduction in the error of the model.

5 Implementation

The implementation phase entails the meticulous execution of the machine learning model development process, and each step is crucial to ensure the model's functionality, deployability, and effectiveness in real-world scenarios. An in-depth overview of the tools employed, data selection, exploratory data analysis, data cleaning, and the programming language used is provided in this section.

Tools Used: The implementation process leverages the following tools:

- **Google Colab**: A powerful environment for executing Python code, running Jupyter notebooks, and utilizing machine learning libraries is seamlessly provided by Google Colab. The advantage of cloud-based computing resources is offered, ensuring efficient execution of resource-intensive tasks.

- **Python Programming Language**: Python has been chosen as the programming language for this project. Its rich ecosystem of libraries for data manipulation, visualization, and machine learning makes it a comprehensive toolkit for efficient project

development.

- **Data Selection**: Relevant customer data was collected from e-commerce platforms or databases. The dataset used includes the following attributes: Invoice number, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, and Country.

- **Exploratory Data Analysis**: Exploratory Data Analysis (EDA) is performed to visualize and understand data patterns. With the aid of Python libraries such as Pandas and Matplotlib, prominent variables are plotted against degrees of injury and injury counts. Trends, correlations, and important features that influence the machine learning model are revealed by the visualizations.

Data Cleaning: Reliable modeling outcomes are ensured by ensuring that the data is clean. This phase encompasses several tasks:

- Handling Null Values: Null values are identified and handled by using Pandas functions.
- Removing Duplicates: Duplicates are identified and removed to ensure consistency of the data.
- Handling Missing Values: Columns with significant missing data are addressed.
- Feature Selection: Features are selected after Neural Networks (NN) are used.

5.1 Evaluation Technique

The effectiveness of our Customer Segmentation methodology is assessed through a set of performance evaluation metrics. The optimal number of cluster assignments was determined using the elbow method. As previously described, the Elbow Method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters. The WCSS represents the sum of squared distances between data points and their corresponding cluster centroids. As the number of clusters increases, the WCSS tends to decrease since each data point can be closer to its cluster centroid. However, adding more clusters beyond a certain point doesn't significantly reduce the WCSS. The "elbow point" on the graph, where the rate of decrease slows down, is considered an indication of the optimal number of clusters. This method helps strike a balance between minimizing intra-cluster distances and avoiding excessive cluster fragmentation.

As for the regression analysis, a number of metrics were used:

- **Neg Mean Absolute Error**: The mean absolute error is defined as the average difference between the model's output (predictions) and observations (actual values), ignoring the sign of these differences to prevent cancellations between positive and negative numbers. The predicted MAE would likely be much smaller than the

actual differences between the model and the data if the sign had not been ignored. [2]

- **Neg Mean Squared Error** : The Mean Squared Error (MSE), perhaps the most basic and widely used loss function, is frequently covered in beginning machine learning classes. To obtain the MSE, the difference between the ground truth and the model's predictions is squared and averaged over the entire Dataset. The MSE will never be negative because the errors are always squared. [3]

- **Neg Median Absolute Error :** Absolute error is used in machine learning to describe the amount of difference between an observation's true value and its predicted, and it is one of the most commonly used loss functions for regression problems. The size of mistakes for the entire group is determined by the MAE, which takes the average of absolute errors for a set of observations and forecasts. MAE is also known as the L1 loss function, and it serves as an easy-to-understand quantifiable measurement of errors for regression problems. [4]

6 Evaluation

We have four experiments or use cases for the prediction of monetary value. Results here will focus on the absolute number of the prediction error. A zero error will be associated to a perfect regression, with a zero error in the prediction. We used a number of loss functions, which also calculate the sign of the error, indicating that a negative number is an underestimate of the predicted value. The first three use cases are boosting techniques with the following loss functions:

1. Negative Mean Absolute Error
2. Negative Mean Squared Error
3. Negative Median Absolute Error

For a more comprehensive evaluation, a neural network was implemented using Keras[5]. This is the fourth use case. It is important to note that the results shown below are median values taken from a 10-fold cross-validation run.

6.1 Experiment 1 / Negative Mean Absolute Error

The negative mean absolute error (negMAE) metric is used to evaluate the performance of a regression model. Results from this loss function appear in Table 1. MAE measures the average absolute errors between the actual and predicted values, and is calculated by taking the average of the absolute differences between the actual and predicted values. The negative sign in negMAE indicates that lower values are better, meaning that a lower negMAE suggests a better fit of the model.

[2] https://insidelearningmachines.com/mean_absolute_error/
[3] https://towardsdatascience.com/understanding-the-3-most-common-loss-functions-for-machine-learni
[5] https://keras.io/

Different regression models were applied to predict monetary values, and negMAE was used as the evaluation metric. The lowest error was obtained with the Hist Gradient Boosting Regressor model, and when enhanced features (extracted using a neural network) were used, the minimum error was reduced to -1139.197221480816.

Table 1: Prediction of monetary value using negative mean absolute error as loss function

Algorithm	**Neg Mean Absolute Error**
Linear Regression	-1788.7525735153486
Gradient Boosting Regression	-1150.6701817716903
Random Forest Regression	-1152.7328102525303
Ada Boost Regression	-1278.0704683271101
Extra Tree Regression	-1161.1913684962356
Hist Gradient Boosting Regression	-1139.197221480816

6.2 Experiment 2/ Negative Mean Squared Error

Negative Mean Squared Error (NegMSE) is another metric used to evaluate the performance of regression models. Results of this loss function appear in Table 2. Like Mean Squared Error (MSE), NegMSE measures the average of the squares of the errors between the predicted values and the actual values. However, in NegMSE, the values are negated, meaning that the errors are squared and then negated before averaging.Different regression models were applied to predict monetary values, and NegMSE was used as the evaluation metric. The lowest error was obtained with the Hist Gradient Boosting Regressor model, and when enhanced features (extracted using a neural network) were used, the minimum error was reduced to -29528770.740563903.

Table 2: Prediction of monetary value using negative mean squared error as loss function

Algorithm	**Neg Mean Squared Error**
Linear Regression	-31929208.60168594
Gradient Boosting Regression	-29522432.91974823
Random Forest Regression	-29588080.64847672
Ada Boost Regression	-29924778.552389223
Extra Tree Regression	-29542264.68516937
Hist Gradient Boosting Regression	-29528770.740563903

6.3 Experiment3 / Neg Median Absolute Error :

Negative Median Absolute Error (NegMedAE) is a metric used to evaluate the performance of regression models. Results of this loss function appear in Table 3. It is based on the Median Absolute Error (MedAE or MAE), but the values are negated. MedAE measures the median of the absolute errors between the predicted

values and the actual values. By negating the MedAE, the sign of the errors is flipped, turning them into negative values.

Different regression models were applied to predict monetary values, and NegMedAE was used as the evaluation metric. The lowest error was obtained with the model, and when enhanced features (extracted using a neural network) were used, the minimum error was reduced to -95.26689064856708.

Table 3: Prediction of monetary value using negative median absolute error as loss function

Algorithm	**Neg Median Absolute Error**
Linear Regression	-1312.248557475692
Gradient Boosting Regression	-98.1816030868504
Random Forest Regression	-95.96193378569068
Ada Boost Regression	-230.34802524631726
Extra Tree Regression	-95.26689064856708
Hist Gradient Boosting Regression	-95.78768151981473

6.4 Experiment 4/ Neural network

For the final comparison we created a neural network, as described in Figure 3. It uses mean squared error as loss function, and the default Adam optimizer.

Table 4: Prediction of monetary value using mean squared error

Algorithm	**Mean Squared Error**
Neural network	-253.07057922363282

6.5 Discussions

The conducted project was employed to extract insights from e-commerce customer data using a comprehensive methodology. The traditional K-means clustering approach was enhanced through the integration of Neural Networks (NN) for feature learning. This discussion section delves into the key findings, implications, and limitations of the project.

From Table 1 we can see that Histogram Gradient Boosting Regression outperformed other boosting algorithms. As expected, the worst performing algorithm was the linear regression.

Table 2 shows Gradient Boosting Regression outperform other boosting algorithms. Linear regression was the worst one. The best result equals to -5433.45, which is not as good as Histogram Gradient Boosting Regression from Table 1. It means the Negative Mean Squared Error loss is not good for the prediction of monetary value.

As for Table 3, the best estimation came from Extra Tree Regressions. Linear regression is still the worst estimation across all use cases.

Results from the neural network, Table 4, shows the mean square error estimation is comparable to Ada Boost Regression (Table 3).

It can be seen from the results that negative median absolute error is the best loss function to predict monetary value in RFM. Most results from this loss functions are very competitive. The best one, Extra Tree Regression, is not much better than the next three ones from Table 3. In other words, by using this loss function and a range of boosting algorithms we can get a very good estimation of the monetary value.

Other results worth discussing are as follows.

Implications The traditional K-means clustering can be used by businesses to gain insights into basic customer segments, which can be utilized for basic targeting and segmentation strategies. However, the enhanced K-means clustering using Neural Networks (NN)-derived features provides the most value. This approach uncovers hidden patterns and behaviors, allowing businesses to tailor marketing efforts more effectively. The identified customer segments can be leveraged to design personalized marketing campaigns, recommend products based on individual preferences, and optimize pricing strategies. Additionally, the enhanced clusters provide a deeper understanding of customer lifetime value and purchasing patterns, facilitating better inventory management and revenue forecasting.

Limitations: The methodology offers promising results, but there are certain limitations that should be considered. The success of the approach is heavily dependent on the quality and quantity of the data collected. Incomplete or noisy data can lead to inaccurate clustering outcomes. Additionally, the effectiveness of the Neural Networks (NN) training process relies on parameter tuning and the chosen architecture, which may require expertise in machine learning. Another limitation is the assumption that K-means clustering is the most suitable algorithm for customer segmentation. Depending on the data, other clustering algorithms such as hierarchical clustering or DBSCAN may yield different results. Additionally, the project focuses on behavioral data and does not incorporate external factors such as demographics, which could enrich the segmentation insights.

7 Conclusion and Future Work

The main findings of the current research are two fold: the usage of cluster segments as a feature for regression, and the study of best loss function for the estimation of monetary value. The best loss function turned to be negative median absolute error which enables a number of boosting algorithms to accurately predict monetary value. While the best algorithm was Extra Trees, other boosting algorithms, such as Gradient Boosting, Random Forest, and Histogram Gradient Boosting were also very good.

The Valuable insights for e-commerce businesses are provided by the customer segmentation results obtained from the enhanced K-means clustering with Neural Networks (NN). The need for targeted marketing strategies and personalized engagement to maximize the value of Premium Customers is suggested by the identification of this segment as distinct. Retention strategies and initiatives to enhance loyalty and repeat purchases can benefit the Regular Customers segment. These insights enable businesses to tailor their marketing efforts, product offerings, and customer experiences to specific customer segments, leading to improved customer satisfaction and increased revenue.

The results highlight the effectiveness of enhanced K-means clustering with Neural Networks (NN)-based feature learning for achieving more accurate and meaningful customer segmentation in the e-commerce domain. Neural Networks (NN) incorporation provides deeper insights into customer behavior and enhances the clustering results. The higher silhouette score obtained from the enhanced approach indicates improved separation and distinctiveness among the customer segments, further enhancing the segmentation granularity.

It is important to note that the optimal number of clusters and the silhouette scores may be varied depending on the dataset and specific business requirements. Therefore, businesses should consider these factors and analyze further the segments obtained to derive actionable insights.

The value of incorporating Neural Networks (NN)-based feature learning into the K-means clustering process for customer segmentation in e-commerce has been demonstrated by the findings. The use of Hist Gradient Boosting Regressor offers more precise segmentation and results in lower error values when compared to using either KNN alone or KNN in conjunction with Neural Networks (NN) for the task at hand. Businesses can achieve more accurate and nuanced customer segmentation using the enhanced approach, which provides valuable insights for targeted marketing strategies, personalized customer engagement, and revenue optimization.

Future Directions: To further enhance the methodology, external data sources such as demographic information, online behavior, and social media interactions could be integrated. This could lead to more comprehensive and accurate customer segments. Additionally, more advanced unsupervised learning techniques beyond Neural Networks (NNs), such as variational autoencoders, could be explored to potentially uncover even deeper insights.Furthermore, the methodology could be validated across diverse datasets and industries to ascertain its generalizability. Sensitivity analysis and parameter optimization could be conducted on larger datasets to ensure stable results.

In conclusion, a multi-faceted approach to customer segmentation is presented in this project, integrating traditional K-means clustering with boosting algorithms for the prediction of monetary value. We have shown that boosting algorithms with negative median absolute error loss function outperform other regressions. In doing

so, this approach offers businesses the opportunity to have intricate customer behavior patterns uncovered, thereby enabling marketing strategies to be more effective and revenue to be optimized. While certain limitations exist, the potential benefits of the approach make it a valuable tool in the realm of e-commerce analytics.

References

Alizadeh Moghaddam, S. H., Mokhtarzade, M. and Moghaddam, S. A. A. (2018). Optimization of rfm's structure based on pso algorithm and figure condition analysis, *IEEE Geoscience and Remote Sensing Letters* **15**(8): 1179–1183.

Begović, M., Avdagić-Golub, E., Memić, B. and Kosovac, A. (2023). Predicting customer behavior in support channels using machine learning, *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, pp. 1275–1280.

Bulut, A. (2014). Topicmachine: Conversion prediction in search advertising using latent topic models, *IEEE Transactions on Knowledge and Data Engineering* **26**(11): 2846–2858.

Cheng, C.-H. and Chen, Y.-S. (2009). Classifying the segmentation of customer value via rfm model and rs theory, *Expert Systems with Applications* **36**(3, Part 1): 4176–4184.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417408002091*

Christy, A. J., Umamakeswari, A., Priyatharsini, L. and Neyaa, A. (2021). Rfm ranking – an effective approach to customer segmentation, *Journal of King Saud University - Computer and Information Sciences* **33**(10): 1251–1257.
URL: *https://www.sciencedirect.com/science/article/pii/S1319157818304178*

Chui-Yu Chiu, Zhi-Ping Lin, P.-C. C. and Kuo, I.-T. (2009). Applying rfm model to evaluate the e-loyalty for information-based website, *International Journal of Electronic Business Management* **7**: 278–285.
URL: *https://web.archive.org/web/20130124013516id/http* : *//ijebm.ie.nthu.edu.tw* : $80/IJEBM_Web/IJEBM_static/Paper - V7_N4/A06.pdf$

Cuadros, A. J. and Domínguez, V. E. (2014). Customer segmentation model based on value generation for marketing strategies formulation, *Estudios Gerenciales* **30**(130): 25–30.
URL: *https://www.sciencedirect.com/science/article/pii/S012359231400045X*

de Lima Lemos, R. A., Silva, T. C. and Tabak, B. M. (2022). Propension to customer churn in a financial institution: a machine learning approach, *Neural Computing and Applications* **34**(14): 11751–11768.
URL: *https://doi.org/10.1007/s00521-022-07067-x*

Dogan, O., Ayçin, E. and Bulut, Z. (2018). Customer segmentation by using rfm model and clustering methods: A case study in retail industry, *International Journal of Contemporary Economics and Administrative Sciences* **8**: 1–19.

Eid, A., Clerc, G., Mansouri, B. and Roux, S. (2021). A novel deep clustering method and indicator for time series soft partitioning, *Energies* **14**(17).
URL: *https://www.mdpi.com/1996-1073/14/17/5530*

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI Mag.* **17**: 37–54.
URL: *https://api.semanticscholar.org/CorpusID:61287995*

Gadgil, K., Gill, S. S. and Abdelmoniem, A. M. (2023). A meta-learning based stacked regression approach for customer lifetime value prediction, *Journal of Economy and Technology* .
URL: *https://www.sciencedirect.com/science/article/pii/S2949948823000045*

Goel, R. and Kalotra, A. (2022). Machine learning models for customer relationship management to improve satisfaction rate in banking sector, *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pp. 172–175.

Haiying, M. and Yu, G. (2010). Customer segmentation study of college students based on the rfm, *2010 International Conference on E-Business and E-Government*, pp. 3860–3863.

Hu, X., Shi, Z., Yang, Y. and Chen, L. (2020). Classification method of internet catering customer based on improved rfm model and cluster analysis, *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pp. 28–31.

Huang, Y., Zhang, M. and He, Y. (2020). Research on improved rfm customer segmentation model based on k-means algorithm, *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*, pp. 24–27.

Jianping, W. (2011). Research on vip customer classification rule based on rfm model, *MSIE 2011*, pp. 336–338.

Khajvand, M., Zolfaghar, K., Ashoori, S. and Alizadeh, S. (2011). Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study, *Procedia Computer Science* **3**: 57–63. World Conference on Information Technology.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050910003868*

Li, X. and Chen, Z. (2022). Customer churn prediction in bank based on different machine learning models, *2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*, pp. 274–279.

Li, Y. (2019). Credit risk prediction based on machine learning methods, *2019 14th International Conference on Computer Science Education (ICCSE)*, pp. 1011–1013.

Liao, J., Jantan, A., Ruan, Y. and Zhou, C. (2022). Multi-behavior rfm model based on improved som neural network algorithm for customer segmentation, *IEEE Access* **10**: 122501–122512.

López, J. C. L. and Chavira, D. A. G. (2019). Multicriteria market segmentation: An outranking approach, *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pp. 36–42.

Malyar, M., Mykola Robotyshyn, M. and Sharkadi, M. (2020). Churn prediction estimation based on machine learning methods, *2020 IEEE 2nd International Conference on System Analysis Intelligent Computing (SAIC)*, pp. 1–4.

Mohammad, J. and Kashem, M. A. (2022). Air pollution comparison rfm model using machine learning approach, *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pp. 1–5.

Nisha and Singh, A. S. (2023). Customer behavior prediction using deep learning techniques for online purchasing, *2023 2nd International Conference for Innovation in Technology (INOCON)*, pp. 1–7.

Rajula, H. S. R., Giuseppe, V., Manchia, M., Antonucci, N. and Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment, *Medicina Journal* **56**.

Salehinejad, H. and Rahnamayan, S. (2016). Customer shopping pattern prediction: A recurrent neural network approach, *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–6.

Sheshasaayee, A. and Logeshwari, L. (2018). Implementation of rfm analysis using support vector machine model, *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, pp. 760–763.

Shirole, Rahul, S. L. and Saraswati, J. (2021). Customer segmentation using rfm model and k-means clustering, *International Journal of Scientific Research in Science and Technology* **8**: 591–597.
URL: *https://ijsrst.com/paper/8152.pdf*

Sun, J. and Meng, Z. (2022). Research on customer value identification of video-on-demand services based on rfm improved model, *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pp. 464–467.

Wei, J.-T., Lin, S.-Y. and Wu, H.-H. (2010). A review of the application of rfm model, *African Journal of Business Management December Special Review* **4**: 4199–4206.
URL: *https://www.researchgate.net/publication/228399859*

Yao, Y., Yang, Y., Zhou, L., Guo, X. and Wang, G. (2023). Multi-cc: A new baseline for faster and better deep clustering, *Electronics* **12**(20).
URL: *https://www.mdpi.com/2079-9292/12/20/4204*

Yavari, S., Valadan Zoej, M. J., Mohammadzadeh, A. and Mokhtarzade, M. (2013). Particle swarm optimization of rfm for georeferencing of satellite images, *IEEE Geoscience and Remote Sensing Letters* **10**(1): 135–139.

Zhang, D., Zhang, C. and Zheng, C. (2023). Prediction and analysis of customer churn of automobile dealers based on bi, *2023 IEEE 6th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, Vol. 6, pp. 368–372.

Zong, Y. and Xing, H. (2021). Customer stratification theory and value evaluation—analysis based on improved rfm model, *Journal of Intelligent & Fuzzy Systems* **40**: 4155–4167. 3.
URL: *https://doi.org/10.3233/JIFS-200737*