BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2025, Volume 16, Issue 3, pages: 86-101
Submitted: April 15th, 2025 | *Accepted for publication:* July 21st, 2025

Customer Behaviour Analysis Through Clustering and Classification for Segmentation and Forecasting in Marketing Campaigns

Alexandar Danailov
Faculty of Mathematics and Informatics, University of Veliko Turnovo, St. Cyril and St. Methodius Veliko Turnovo, Bulgaria.
D1601@sd.uni-vt.bg

Abstract: *In this paper, the possibilities of extracting useful insights through data analysis using machine learning approaches are explored on a dataset focused on customer relationship management. Exploratory data analysis is applied to customers and their interactions, such as purchases, responses to marketing campaigns, and complaints. The detailed examination aims to assess the potential of using results from past marketing campaigns to predict customer reactions to future ones, thereby improving marketing effectiveness. To this end, a classification model is built to predict customer responses to the latest campaign. Evaluation metrics are calculated to assess the classification performance across different sets of selected features and classifiers, and the findings are summarized and discussed. Generalized Linear Model (GLM) and H2O Deep Learning (DL) models stood out as the best performers in the study.*

Keywords: *customer behaviour; marketing campaign; clustering; classification; machine learning; customer segmentation; predictive model; data analysis.*

How to cite: Danailov, A. (2025). Customer behaviour analysis through clustering and classification for segmentation and forecasting in marketing campaigns. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *16*(3), 86-101. https://doi.org/10.70594/brain/16.3/8

1. Introduction

The dynamic development of technologies and the competitive nature of the market pose challenges to companies in their efforts to attract and retain customers. Direct marketing campaigns remain a key tool for implementing personalised offers and generating revenue. For a campaign to be successful, it must be targeted to the right customers – those who are most likely to respond positively to the offer. In this regard, predictive models play an important role in optimising marketing strategies. They allow companies to identify target segments of the customer base using historical data and modern machine learning algorithms. Through precise targeting, companies can increase the success rate of campaigns, minimise the cost of non-converting customers, and maximise their profits. In the context of this study, the goal is to build a predictive model that will support the planning and implementation of a future direct marketing campaign to offer a new product. Available customer information serves as the basis for developing a model that allows us to predict which customers are most likely to purchase.

The study is based on real CRM (Customer Relationship Management) data from iFood, a leading Brazilian food delivery platform. Segmenting customers and predicting their response allows platforms like iFood to identify target groups with a higher propensity to purchase, personalise marketing messages, and optimise campaign spending. This is crucial in an environment where order frequency, activity time, and preferences for specific food categories or merchants can be highly variable. Therefore, the predictive models built in this study have direct application in increasing marketing effectiveness in the food delivery industry.

Two main approaches are used in the study to develop the model: clustering and classification. Clustering, as an unsupervised learning method, is used to group customers according to their behavioural and demographic characteristics. This helps to identify similar profiles among the customer base. Classification, in turn, is used to predict the likelihood of a customer making a purchase based on existing characteristics and the results of the pilot campaign.

The introduction of such techniques not only improves the effectiveness of campaigns but also provides valuable insights into customer behaviour and preferences. This article explores the methodology, data analysis, and results of the application of clustering and classification in the context of direct marketing.

The original contribution of the study lies in the integration of real CRM data from a pilot campaign with a clustering and classification approach, as well as in the comparative analysis of eight different machine learning models. Such a comprehensive framework is rarely applied in practical conditions, especially on real and unbalanced data, which makes the present work a valuable contribution to applied marketing analysis.

This article seeks to answer the following research questions:

1. Can historical CRM data be used to predict customer response in a marketing campaign?
2. Which machine learning algorithms show the highest efficiency in classifying the response?
3. To what extent does feature selection improve the performance of models?
4. Can clustering contribute to a better understanding of customer segments and behaviour?

Although the study focused on a food delivery platform, the approaches used and the results obtained are fully applicable to other sectors with intensive customer engagement, such as e-commerce, telecommunications, and personalised online services, where behavioural and demographic data play a key role in targeting and prediction.

2. Related Work

In recent years, interest in the use of predictive models in marketing has grown significantly (Gomes & Meisen, 2023). The development of classification and clustering algorithms based on customer data underpins research aimed at optimising marketing campaigns.

One of the main approaches in literature is the use of classification algorithms, such as logistic regression, decision trees, and ensemble learning methods such as Random Forest and

Gradient Boosting. A study by Neslin et al. (2006) shows that logistic regression is an effective method for predicting customer response to direct campaigns, especially when behavioural and demographic data are used.

Clustering also occupies a central place in research. Abednego, Nugraheni & Salsabina (2023), Godcares et al. (2023) and Namvar, Ghazanfari & Naderpour (2017) demonstrate how clustering techniques, such as K-Means and DBSCAN, can be used to segment customer bases for personalising offers. This allows companies to focus their efforts on homogeneous groups of customers with a high probability of purchase.

Increasing attention has been paid to combining these two approaches – clustering and classification. Wedel & Kamakura (2000) emphasise the importance of segmenting customers before applying classification models, which improves the accuracy and relevance of predictions.

Some modern approaches emphasise the importance of transforming input data to improve the prediction accuracy of models. Sana et al. (2022) propose a model for predicting customer churn in the telecommunications sector that combines data transformation and feature selection techniques. The results show that careful data preparation before training significantly improves prediction accuracy, especially when using ensemble models and logistic regression.

Other studies emphasise the use of more sophisticated models, including neural networks and deep learning. Onasanya & Okonkwo (2023) use advanced models that facilitate the decision-making process by providing more accurate predictions of future customer behaviour.

The current study draws inspiration from these approaches and combines them to build a comprehensive forecasting model that includes both clustering and classification. Furthermore, the current work contributes by using real data from a pilot campaign, providing practical insights into the effectiveness of these methods in the context of direct marketing.

More recent studies have also focused on more advanced architectures for customer behaviour prediction based on deep neural networks with a self-attention mechanism. For example, Wu (2023) presents a model for predicting customer churn in the telecommunications industry using deep learning and carefully selected behavioural characteristics.

Similar approaches that combine data transformation with machine learning have been successfully applied in the telecommunications industry. Studies such as (Herkert & Mullenbach, 2021) propose bimodal architectures that combine textual and numerical information, allowing for better traceability of the factors influencing the prediction. Such a perspective is particularly useful in marketing environments where transparency in decision-making has a direct impact on customer trust.

This paper presents a combined approach to analysing customer data, combining exploratory analysis, clustering, and classification for direct marketing purposes. A real-world dataset of customer interactions is used, with a special focus on predicting response to marketing campaigns. The research also includes a feature selection step to increase forecast accuracy with a limited number of variables. The paper's contribution lies in the application of an integrated methodology tailored to real business conditions, as well as in the comparative analysis of the effectiveness of different machine learning algorithms in the context of marketing forecasting.

3. Research Methodology

The main objective of the study is to develop a predictive model for optimising customer selection in the context of direct marketing campaigns, maximising profit when offering a new product.

3.1. Dataset Characteristics and Preprocessing

The analysis is conducted using a dataset obtained from the (iFood Business Data Analytics Test Repository) (https://github.com/nailson/ifood-data-business-analyst-test). iFood, as a leading food delivery platform in Brazil, operating in over a thousand cities, provided this dataset as part of its analytical study. The data represents a real-world database of customers from the food retail sector.

The dataset contains 2,240 customer records from a pilot marketing campaign, structured to represent typical customer interactions and behaviour in the context of a food retail business. The data collection reflects a controlled marketing experiment in which:

- Customers are randomly selected from the database.
- Contact is made by phone regarding a new product offering.
- Responses and purchases are tracked over the following months.

Initial data preprocessing includes:

- Age categorisation into meaningful groups.
- Income level discretisation.
- Feature engineering for purchase patterns.
- Binary coding of categorical variables.
- Standardisation of numeric features.

The dataset contains 39 features covering various aspects of customer demographics, behaviour, and campaign responses. The columns in the table are as follows:

- Age: The customer's age.
- Divorced, Married, Single, Relationship, Widowed: The customer's marital status, represented by binary indicators.
- Child, Teenager: Number of children at home, and number of teenagers.
- Higher, Primary, Secondary, Master, Doctorate: Educational level of the customer, represented by binary indicators.
- Income: The annual income of the customer.
- Days since last order: The number of days since the customer's last order.
- Wine, Fruit, Meat, Fish, Sugar, Other: The number of products purchased from different categories as a monetary value for the last 2 years.
- Purchase amount: The total amount of the customer's purchases (for the last 2 years).
- Discount purchases: The number of purchases made with a discount (for the last 2 years).
- Online purchases: The number of purchases made online (for the last 2 years).
- Catalogue purchases: The number of purchases made through a catalogue (for the last 2 years).
- Store purchases: The number of purchases made in a store (for the last 2 years).
- Monthly Website Visits: The number of monthly website visits by the customer.
- Accepted Offer 1, Accepted Offer 2, Accepted Offer 3, Accepted Offer 4, Accepted Offer 5: Binary indicators of the accepted offers by the customer.
- Accepted Offers: The total number of accepted offers.
- Complaints: The number of complaints submitted by the customer.
- Response: The customer's reaction to the last marketing campaign. This is the attribute that we use as a label in the classification stage.

3.2. Exploratory Data Analysis

Box plots are used for primary analysis of the distribution of key characteristics (Fig. 1). They allow visualisation of medians, interquartile ranges and outliers, helping to identify potential anomalies and differences between groups.

• **Age**: The distribution is concentrated - 50% of customers are between 42 and 61 years old, with no extreme values. This facilitates categorisation by age range when building models.

• **Income**: The main group has an income between 35,000 and 67,000 MUR. Several customers with very high incomes are observed, indicating the need for normalisation and the possibility of using income as a discriminating characteristic in segmentation.

• **Amount of purchases**: The data shows strong asymmetry, with several customers spending significantly more than the others. This draws attention to the existence of a VIP segment that should be taken into account when clustering.

• **Monthly Site Visits**: Most customers visit the site between 3 and 7 times per month, but a small number have significantly higher activity. This is an indicator of strong digital engagement and can be used as a predictive feature.

• **Purchase Frequency**: Half of the customers make between 6 and 18 purchases over 2 years, with no significant outliers. This is a robust indicator for inclusion in behavioural activity patterns.

Summary: The exploratory analysis revealed a robust central tendency across most variables, with a few notable exceptions (e.g., purchases and visits), which suggests the presence of subgroups with higher business value. These insights aid feature selection and inform clustering and classification processes.

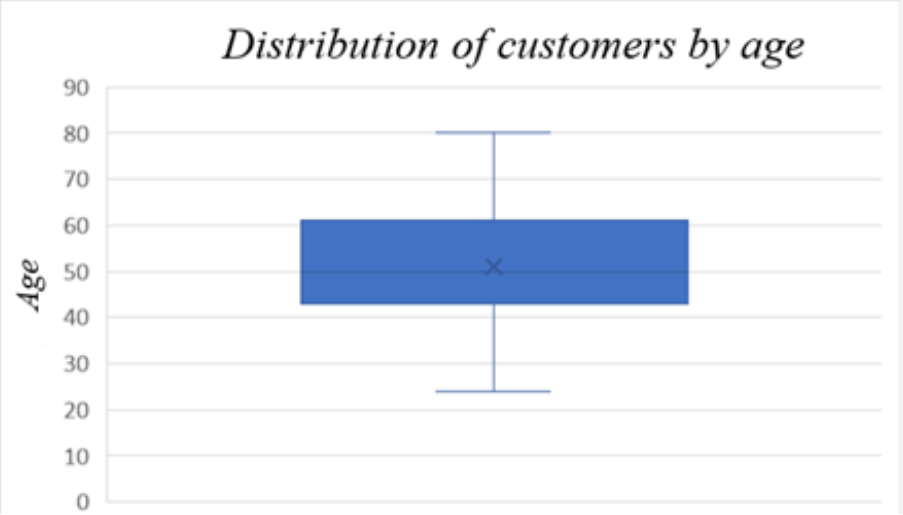

Figure 1. Distribution of customers by age

Heatmap for the distribution of clusters (Figure 2)

Figure 2 presents a heatmap showing the distribution of key characteristics – income and offer acceptance – between three clusters formed using K-Means (k=3).

Cluster 2 stands out with high income and active offer acceptance (especially offers 3 and 4), making it a strong target for marketing campaigns.

Cluster 1 includes customers with low income and low engagement, while

Cluster 0 represents moderately active customers with average income.

The colour scale ranges from red (below average) to green (above average), demonstrating the strength of the connections between the clusters and the characteristics under consideration.

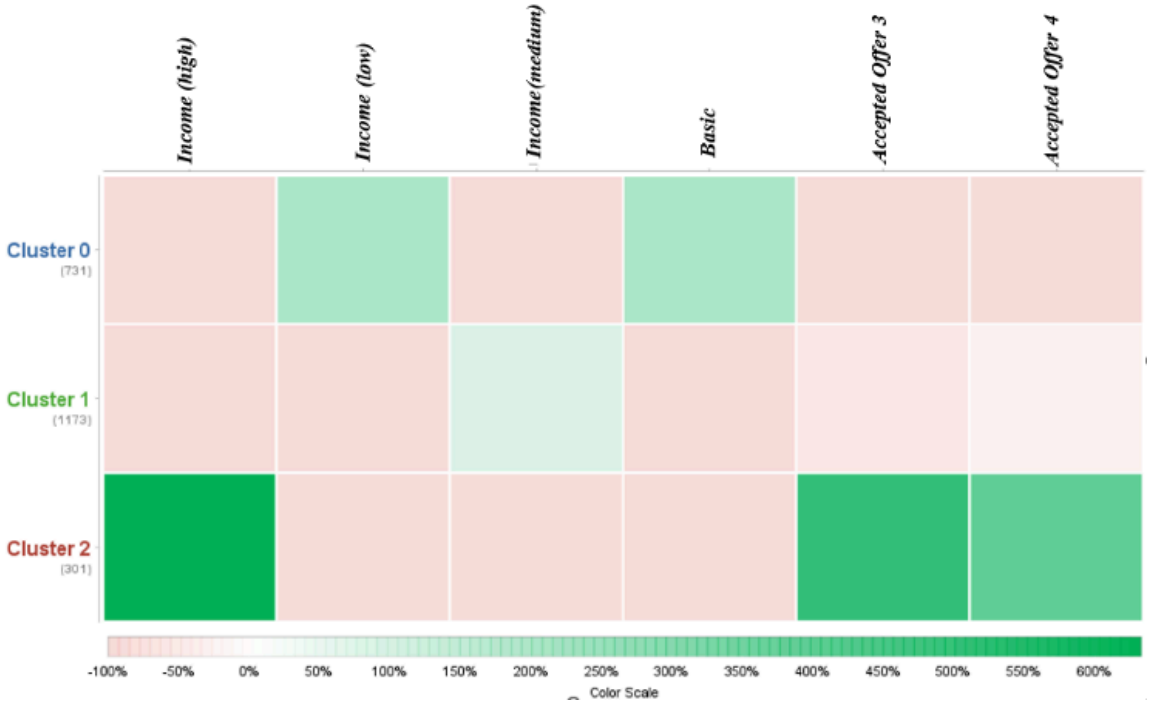

Figure 2. Heatmap for the distribution of characteristics

Analysis of the optimal number of clusters with the elbow method:

In order to determine the optimal number of clusters when applying the K-Means algorithm, the Elbow Method is used, which examines the relationship between the number of clusters (k) and the average distance between the objects and the corresponding centroids. For the analysis, two measures are used – Euclidean Distance and Cosine Similarity.

The results show a clear trend of decreasing the average value of the centroid distances with increasing the number of clusters (Fig. 3). At k=2, the average distance is approximately 3.22. As k increases to 6, a significant decrease is observed in values around 2.57 (for Euclidean Distance) and 2.54 (for Cosine Similarity). After this point, the decrease becomes less pronounced, at k=10, the values are 2.25 for both Euclidean Distance and Cosine Similarity, indicating that adding new clusters leads to minimal improvement in cluster cohesion. The elbow point is at 3 – 4 clusters, as after these values, the pace of decrease in internal distance slows down significantly.

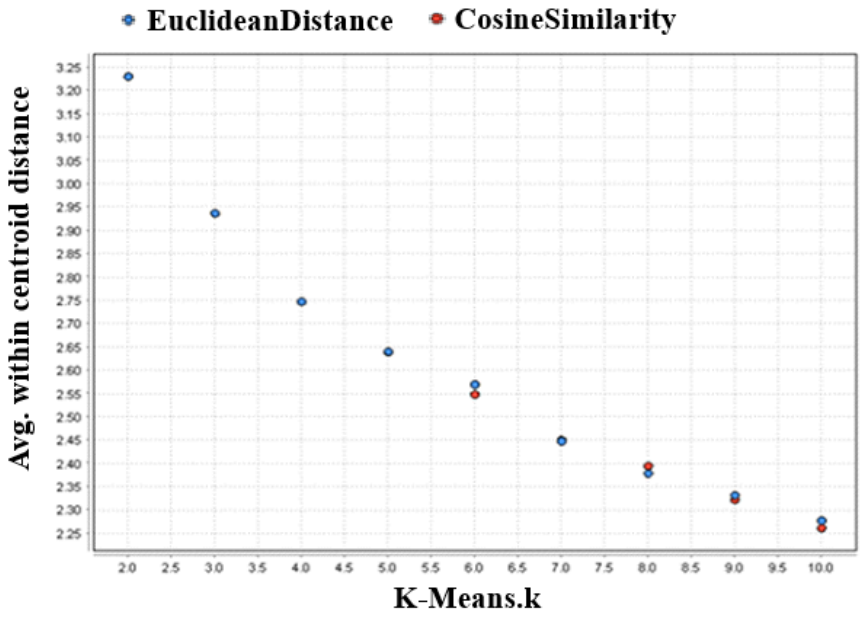

Figure 3. Clustering assessment

Table 1. Clustering Assessment

#	Cluster	Avg. within centroid distance
1	Cluster 0	0.058
2	Cluster 1	0.070
3	Cluster 2	0.086
4	Average	0.071

Table 1 presents the average distances from objects in each cluster to the centroid of that cluster for three clusters and the overall average for all clusters under k-means.

The data from the table can be analysed as follows:

Cluster 0 (0.058): the lowest average distance to the centroid. This indicates that the objects in this cluster are closer to each other and to the centroid, suggesting a better-defined cluster.

Cluster 1 (0.070): The average distance is larger compared to Cluster_0, but the cluster still looks relatively compact.

Cluster 2 (0.086): the highest average distance to the centroid. This may mean that the objects are more dispersed, and the cluster is less well defined than the others.

Average (0.071): The overall mean provides information about the average distance for all clusters together.

Therefore, Cluster 0 is the most compact, while Cluster 2 is the most dispersed. The differences between the mean values are not large, which indicates that the clustering is relatively balanced.

Analysis of clustering quality using the Davies-Bouldin Index:

For an objective assessment of the clustering quality, the Davies-Bouldin Index (DBI) is used in this study, which measures the ratio between the internal compactness and the inter-cluster difference. Low index values indicate well-separated and compact clusters, which is considered desirable in the analysis of structured or weakly structured data.

The results show that the DBSCAN algorithm demonstrates the best performance with a DBI value of 0.923, which indicates the formation of well-separated and internally homogeneous clusters. This is consistent with the characteristic strengths of the algorithm, related to its ability to identify clusters of arbitrary shape and to successfully deal with noise and isolated values in the data. However, it should be borne in mind that at high dimensionality (above 15–20 dimensions), DBSCAN becomes inefficient and leads to poorly defined clusters or too many small and unreliable clusters, as in our case. On the other hand, K-Means and X-Means show significantly higher DBI values (3.062 and 3.062, respectively), which indicates the presence of less separated clusters and potential overlap between them. The difference between these two algorithms is minimal, suggesting that the automatic determination of the number of clusters in X-Means does not lead to a significant improvement over the fixed number of clusters in K-Means.

Additionally, the Gaussian Mixture Model (GMM) achieves a DBI value of 2.606, which places it between DBSCAN and K-Means/X-Means in terms of segmentation quality. This shows that GMM manages to form more compact clusters than K-Means, thanks to its ability to model clusters using probability distributions, but it still lags behind DBSCAN when processing more chaotic or noisy structures.

The Gaussian Mixture Model (GMM) finds three balanced clusters: cluster 0 (731 examples), cluster 1 (743 examples), and cluster 2 (731 examples).

Cluster 0: Moderately active customers with high average income

The segment is characterised by a predominance of the age groups 41-50 years (30%) and 51-60 years (23%). Secondary education (44%) and master's degree (35%) dominate. Over 80% of customers have an average income, with 20% having a high-income status. Marital status includes mostly married (38%) and committed (26%) individuals. Responsiveness to marketing campaigns is moderate (16.7%), and the strongest activity is observed in in-store purchases (54%), followed by internet and catalogue purchases (12-18%).

Cluster 1: The most active and profitable segment

With a similar demographic structure to cluster 0, this segment is distinguished by the highest share of secondary education (45%). Approximately 82% have an average income. The key distinction is the highest responsiveness (19.7%) and the largest average purchase amount (55%). Significant activity is observed in the wine, fruit and meat categories (above 20% average), with internet and catalogue purchases ranging between 13-19%.

Cluster 2: Passive Customers with Limited Activity

The segment is characterised by a younger age structure (31-50 years old constitutes 60%) and 100% low income. It demonstrates the lowest responsiveness (9.6%) and minimal participation in promotional campaigns. User activity is extremely limited, with a purchase amount of only 2.9% and poor performance in all product categories.

3.3. Data classification algorithms

Various classification algorithms have been used to predict customer response to marketing campaigns, including k-NN (K-Nearest Neighbours), decision tree, naive Bayes model, generalised linear model (Nykodym et al., 2024), deep learning (Candel & Parmar, 2015), Gradient Boosted Trees, Bayesian Boosting (NB), and AdaBoost.

The models are trained on labelled data and validated for accuracy. The selected classification algorithms represent different classes of machine learning approaches; their main characteristics are summarised below.

Generalised Linear Model (GLM) is an extension of linear regression that allows modelling relationships between variables when the dependent variable is not continuous (e.g., categorical). It is preferred for small to medium-sized data sets where interpretability and simplicity are sought. An alternative approach that shows promising results is gradient boosting (GBM). According to Proksch (2020), GBM can outperform GLM in accuracy for certain tasks, such as credit scoring and customer behaviour prediction, although it requires a more complex setup and can be more sensitive to re-tuning.

Deep Learning (DL) on H2O is based on a multilayer neural network, which can contain many hidden layers consisting of neurons with activation functions.

k-NN is a simple and intuitive classification and regression algorithm that predicts the value of a given point based on its k nearest neighbours in the feature space. It works directly with the data without prior training (a "memory-based" algorithm). Requires the choice of a distance metric (e.g. Euclidean distance). Sensitive to the choice of the number of neighbours (k) and the scaling of the data.

Decision Tree is a structure that divides data into groups using a series of logical rules based on features. It is easy to interpret, as the decisions can be visualised. Susceptible to overfitting if techniques such as pruning are not applied. Supports both categorical and numeric data.

Bayesian Classifier is a classification algorithm based on Bayes' theorem. It calculates the probability that an object belongs to a particular class using prior probabilities and conditional probabilities.

Gradient Boosted Trees (GBM) is an ensemble method that sequentially builds multiple weak models (usually small decision trees), with each new tree learning to correct the errors of the previous ones. The algorithm uses gradient descent to minimise losses, making it particularly accurate even with complex and nonlinear dependencies in the data.

Bayesian Boosting combines the classic naive Bayes classifier with a boosting technique that improves accuracy by training on more difficult-to-classify examples. This combination preserves the speed and simplicity of Naive Bayes while simultaneously increasing its efficiency.

AdaBoost (Adaptive Boosting) is a classic boosting algorithm that iteratively builds an ensemble of weak models (most often decision stumps), increasing the weight of misclassified instances at each step.

The variety of classification algorithms selected provides the opportunity to identify the best solution for the specifics of the data.

Hyperparameter tuning and optimisation

To improve the predictive performance of the classification models, hyperparameter tuning is performed. To identify the optimal set of parameters for each model, a network search combined with 10-fold cross-validation is used. The evaluation is based on the F1-score due to the unbalanced nature of the response variable.

4. Experiments

The experiments are performed by executing custom-made processes with Altair RapidMiner (https://altair.com/altair-rapidminer).

4.1. Data preprocessing and feature selection

Data preprocessing

The first stage of working with a dataset is preprocessing. It is expressed in cleaning missing or incorrect values. The specific data set is provided for use, already cleaned, which favours direct passage to the next stage.

The following transformations are performed:

- Creation of a new field: "Purchases total" - a new calculated variable is created as the sum of purchases in a store, online and catalogue. This field serves as a quantitative indicator of engagement and is useful for behavioural analysis (for example, in RFM models).
- Exclusion of variables: Z_CostContact, Z_Revenue and MntGoldProds due to lack of explanatory value or low correlation with the target variable.
- Categorisation: Age is categorised into intervals (20-30, 30-40, etc.) to facilitate interpretation.
- Discretisation: Income is grouped into categories "low", "medium", "high", and automatic discretisation by binning is applied.

Data normalization

To facilitate analysis, data is transformed into a certain range, most often between 0 and 1. Normalisation uses the minimum and maximum of a given variable for the transformation. Normalisation is appropriate in a situation where the data has different scales, and the algorithm is sensitive to distances. In the dataset considered, normalisation has been performed for all attributes so that they vary in the range (0, 1).

Feature selection

In the context of data analysis, feature selection approaches are usually used to identify the most important features (variables) that affect the model or analysis.

The selection of features for clustering and classification is guided by their relevance, determined using a calculated weight. In this study, feature selection according to Information gain is used, and measures are calculated to assess the quality of classification for different numbers of selected features (Figures 4 and 5).

4.2. Metrics for Evaluating Models

The performance of the models is evaluated using four metrics: Accuracy, F1-measure, AUC (Area Under the ROC Curve) and AUPRC (Area Under the Precision-Recall Curve). . The results are visualised and compared in Table 2 and Figure 7.

General observations:

- **GLM** and **H2O Deep Learning** stand out as the most efficient and stable models across all metrics.
- **GBT, AdaBoost** and **Bayesian Boosting** also show very good performance, especially with a moderate number of features.

- **k-NN** and **Naive Bayes** perform significantly worse and do not benefit from adding more features.
- **DL** shows the highest F1-measure (75.8%), while **GLM** achieves the highest accuracy (~88.4%) and AUC (~88.1%).

Accuracy:

Accuracy shows the proportion of correctly classified cases. Although sensitive to unbalanced data (as in our case), it is used for comparative purposes along with the F1-measure.

Figure 4 illustrates the accuracy of the different models with an increasing number of selected features.

The main conclusions are:

• **GLM** achieves the highest accuracy (~88.4%) and demonstrates stability with different numbers of features.

• **H2O Deep Learning** has a similar accuracy (~87.6%), with smooth improvement, making it suitable for more complex **dependencies**.

• **GBT**, **AdaBoost** and **Bayesian Boosting** show consistently high accuracy (~86–87%), with a good balance between efficiency and adaptability.

• **Decision Tree** performs stably (~86.8%), but with less precision than the leading models.

• **k-NN** shows moderate performance (~85%), sensitive to feature selection.

• **Naive Bayes** has the lowest accuracy (~78.5%) and is not recommended.

Summary: GLM and DL stand out as the most accurate and reliable. Ensemble models also perform well, while NB and k-NN lag behind.

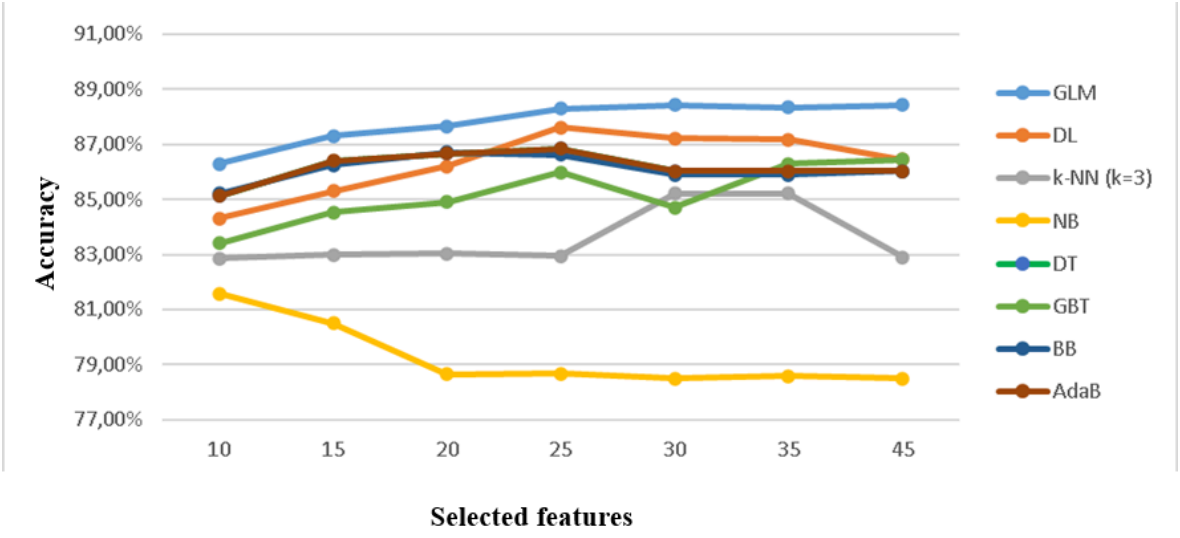

Figure 4. Assessment of the quality of classification models according to the accuracy of the model for different numbers of selected features

F1-measure (F1 Score):

The F1-measure is suitable for unbalanced data, as it considers both precision and recall. It is important for minority classes.

Figure 5 presents the changes in the F1-measure for different numbers of selected features. Main conclusions:

• **H2O Deep Learning** achieves the highest efficiency (up to 75.8%) and stability after 25 features, making it the most suitable model.

• **GLM** also performs strongly (up to 73.0%), with smooth improvement up to 30 features.

• **GBT**, **AdaBoost** and **Bayesian Boosting** show good stability (~70%), but the results stabilise or drop at higher complexity.

• **Decision Tree** is stable at moderate complexity (max. 70.4%), but becomes more sensitive with additional features.

• **k-NN** and **Naive Bayes** perform poorly (~58% and ~64%) and do not improve the result when adding new features.

The following summary can be made: DL is the best model according to the F1-measure. GLM and GBT are also reliable. Ensemble models perform stably up to a certain threshold, while k-NN and NB do not adapt well.

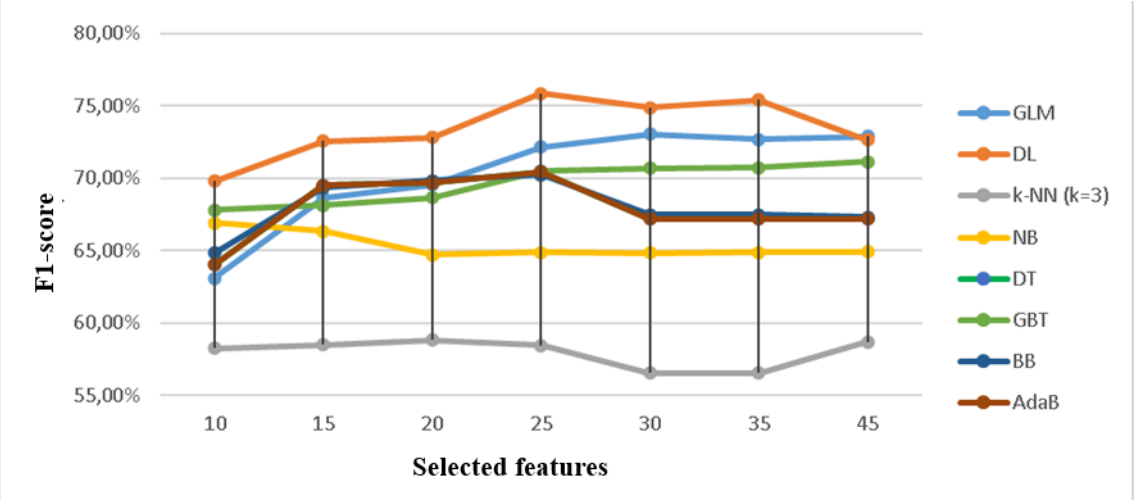

Figure 5: Evaluation of the quality of classification models according to the F1-measure for different numbers of selected features

Justification for the selection of 35 characteristics

Based on the empirical comparison of the performance with different numbers of input features for five classification models (GLM, DL, k-NN, NB, DT), it is found that the value 35 provides an optimal balance between classification quality and robustness to re-tuning. With H2O Deep Learning, almost maximum efficiency is achieved (75.37%), while with other models, the values stabilise or deteriorate with more than 35 features. This avoids both under-training (with too few attributes) and overtraining (with an excessive number). The results allow a reasonable limitation of the input space to 35 features, without significant loss of accuracy, which is why the remaining experiments in this study are conducted with them.

The evaluation of the models is done through two types of validation – cross-validation (CV) and Bootstrap estimation (Boot).

Figure 6 presents a comparative analysis of machine learning models based on their performance metrics. The classifiers used are those specified in section 3.3.

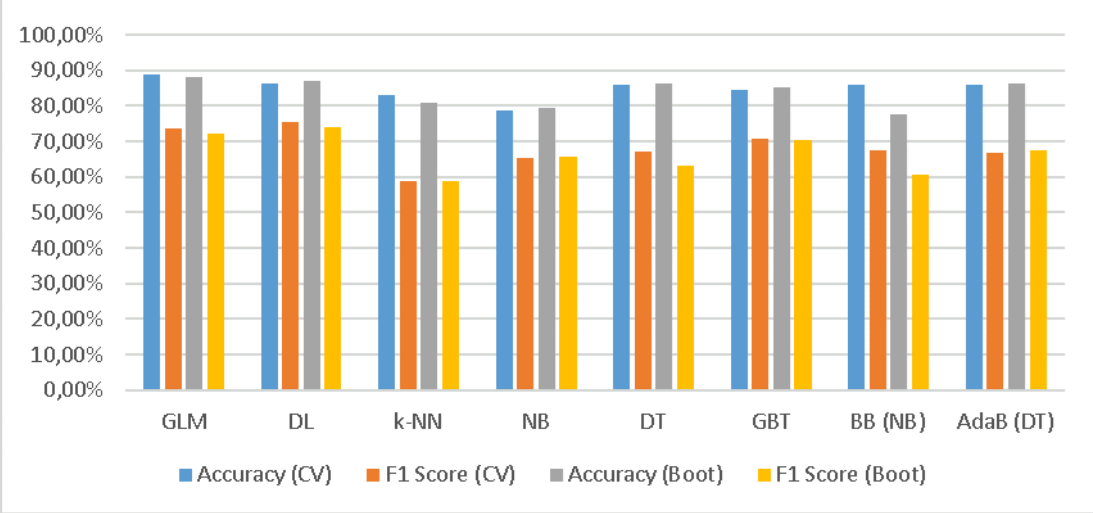

Figure 6. Accuracy and F1-score results for different classifiers and 35 selected features

A comparative analysis of eight machine learning models demonstrates significant differences in performance between the two validation methods. The results show that GLM achieves the highest accuracy of 88.71% under Cross-Validation and 88.08% under Bootstrap validation.

Comparison by validation methods

The analysis of the results shows mixed trends when comparing the two validation methods. Some models (DL, GBT, AdaB) show higher accuracy under Bootstrap validation, while others (GLM, k-NN, BB) perform better under Cross-Validation. The largest difference is observed for Bayesian Boosting (BB) and Naive Bayes (NB), where the accuracy drops from 85.90% (CV) to 77.59% (Bootstrap), suggesting instability of this model.

Performance by models

Best results: GLM demonstrates the highest performance with an accuracy of 88.71% (CV) and 88.08% (Bootstrap), followed by DL with an accuracy of 86.44% (CV) and 87.12% (Bootstrap). Interestingly, DL shows a higher F1-score of 75.27% in Cross-Validation, which is the highest value among all models.

Average performance: Decision Tree shows stable performance with accuracy around 86%, while GBT and ensemble methods (BB, AdaB) demonstrate moderate performance in the range of 84-86% for accuracy.

Lower performance: k-NN and NB show the lowest performance with accuracy below 84% and significantly lower F1-score. k-NN especially stands out with a very low F1-score of 58.86-58.95%, suggesting difficulties in recognising the positive class.

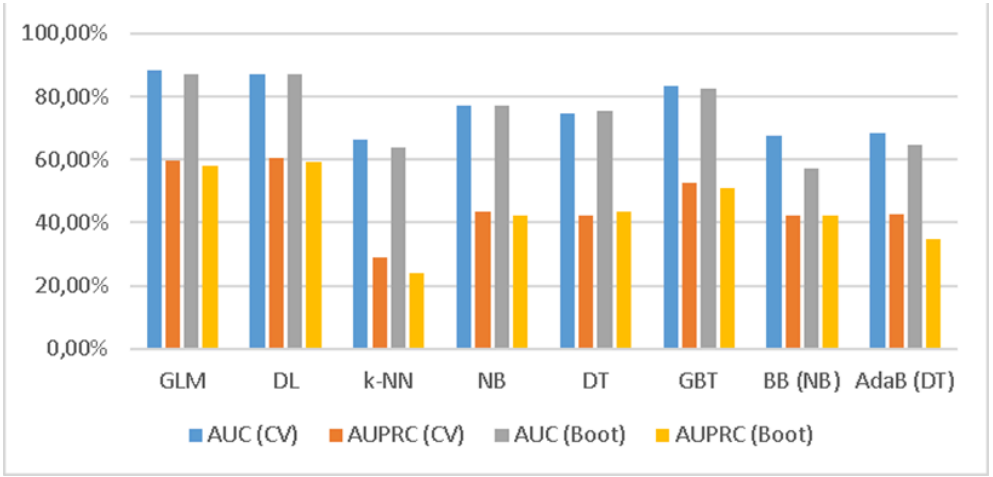

Figure 7. Comparison of the discriminatory ability of models using AUC and AUPRC metrics

The analysis of eight machine learning models using AUC and AUPRC metrics confirms and extends previous observations (Figure 7). GLM demonstrated the highest discriminatory ability with an AUC of 88.10% (CV) and 87.10% (Bootstrap), closely followed by DL with 87.20% (CV) and 86.90% (Bootstrap).

Analysis of AUC results

Excellent performance (AUC > 85%): GLM and DL show excellent class separation ability with AUC values above 85% for both validation methods. This high discrimination ability confirms their suitability for the task.

Good performance (AUC 80-85%): GBT achieves good results with an AUC of 83.40% (CV), positioning itself as the third-best model in terms of discrimination ability.

Problematic performance (AUC < 70%): k-NN, BB (NB) and AdaB (DT) show significantly lower performance, with k-NN standing out with a particularly low AUC of 66.40% (CV), close to random classification.

Analysis of AUPRC results

The AUPRC metric, which is particularly important in case of class imbalance, reveals additional problems:

Critical low values: All models show AUPRC below 61%, with the highest value being 60.50% for DL (CV). This confirms the presence of significant class imbalance in the data.

Dramatic drop in k-NN: An AUPRC of 28.80% (CV) and 24.00% (Bootstrap) in k-NN shows a complete inability of the model to cope with the imbalanced data.

Instability of ensemble methods: AdaB shows instability with AUPRC dropping from 42.50% (CV) to 34.70% (Bootstrap), suggesting overfitting.

4.3. Results and discussion

Analysis of the results obtained from the different algorithms

The results obtained show significant differences in the performance of the algorithms in terms of accuracy and F1-measure scores. The main differences can be attributed to the different approaches to training, optimisation and assumptions that go with each algorithm. These results are presented in detail in the previous chapter through graphs, tables and data analyses.

Impact of Feature Selection on Performance

Feature selection has a critical impact on the performance of algorithms. The results show that choosing an optimal set of features not only increases accuracy but also reduces complexity and processing time.

Computational complexity analysis

H2O Deep Learning models involve hierarchical feature extraction and parallelised stochastic gradient descent, which require significant computational resources and training time, especially for large datasets.

GLM (generalised linear models) use maximum likelihood estimation through iterative weighted least squares methods. The computational complexity of GLM is $O(Np^2 + p^3)$, where N is the number of observations (rows) and P is the number of predictors (columns).

Comparison: Simpler methods, such as GLM or decision trees, are less computationally intensive and faster to train, making them suitable when resources are limited or fast results are needed. The choice of algorithm depends on the specific problem, the size and complexity of the data, and the available computational resources.

Discussion of the applicability of the results for future campaigns

The obtained results provide a basis for applications in campaigns in areas such as marketing, security, industrial processes, etc.

ANOVA analysis and comparative evaluation of models

To check whether the differences between the average accuracies of the models for the different classifiers are statistically significant, an ANOVA test is applied. The result of the test is as follows:

Since the p-value is significantly smaller than the significance threshold ($p < 0.05$), there are statistically significant differences between at least two of the models considered.

Pairwise post-hoc t-tests

To establish which specific models these differences are observed in, paired t-tests are performed. The results are presented in Table 2, as follows:

- (-) indicates that there is no statistical significance in the difference between the average accuracies of the compared models.
- (*) indicates statistical significance at the 0.1 level.
- (**) indicates statistical significance at the 0.05 level.
- (***) indicates statistical significance at the 0.01 level.

Table 2: Results of post hoc t-tests for comparison of accuracy between models

#	Comparison	p-value
1	GL vs DL	0.095 (*)
2	GL vs NB	0.001 (***)
3	GL vs KNN	0.001 (***)
4	GL vs DT	0.072 (*)
5	DL vs NB	0.002 (***)
6	DL vs KNN	0.005 (***)
7	DL vs DT	0.442 (-)
8	NB vs KNN	0.007 (***)
9	NB vs DT	0.005 (***)
10	KNN vs DT	0.073 (*)

Discussion of results

The results of the t-tests show that the Naive Bayes (NB) model has a significantly lower efficiency compared to all others ($p < 0.01$), which positions it as the weakest model in the study.

On the other hand, the GL, DL and DT models do not show statistically significant differences among themselves, which defines them as comparably efficient. KNN also shows acceptable efficiency, but with a slight statistical difference compared to GL and DL.

4.4. Application of results in a real environment

Based on the obtained results, integration into real-time CRM processes is possible. The models can be used for automated calculation of response probability in real time, with the results used for targeting offers, personalised communication and optimisation of marketing campaigns. Segmentation through clustering (e.g. identification of VIP customers) supports dynamic management of customer groups. Integration can be implemented via REST API, with forecast results visualised in BI systems to support operational decisions.

5. Conclusion

In this article, a comparative analysis of different algorithms is performed, and the influence of feature selection on their performance is investigated. The main conclusions emphasise the importance of careful feature selection and an adaptive approach to algorithms, which significantly improves their accuracy and efficiency.

Additionally, the hybrid approach that combines exploratory analysis, clustering, and classification, coupled with large-scale comparative model evaluation (including GLM, DL, GBT, AdaBoost, and others), represents an original contribution that builds on existing academic and practical research.

The study identified GLM and DL as the most suitable models for predicting customer behaviour based on their performance on all key metrics.

Limitations of the study

The main limitations of the study include limitations in the dataset; the analysis is focused on a specific domain, which may limit the applicability of the results to other domains. Future research could consider a broader range of data and algorithms, as well as explore the impact of different parameters in more depth.

Directions for future research

Several directions for future research are suggested, including extending the analysis to new domains such as healthcare, the financial sector, and e-commerce. Investigating the impact of hybrid feature selection methods and exploring the possibilities for process automation. Furthermore, the integration of the algorithms into real-world applications will allow a better understanding of their advantages and limitations.

Scalability and real-world implementation for larger datasets. Trade-off between accuracy and computational resources: The results of this study show that H2O Deep Learning models offer the highest F1-score (~80%) but require significantly more computational resources. When scaled to larger datasets, the GLM model represents a good balance between accuracy (~70% F1-score) and computational complexity, making it a preferred choice for organisations with limited IT resources.

Acknowledgements

This publication is carried out with the financial support of project No: FSD-31-328-24/23.04.2025, titled "Research, Analysis, and Promotion of Mobile Technologies and Software Applications in Support of Students with Special Needs", funded under the regulation for conditions and procedures for evaluation, planning, allocation, and spending of state budget funds for research or artistic activity inherent to public higher education institutions by St. Cyril and St. Methodius University of Veliko Tarnovo.

References

Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management, 21*(3), 1–44. https://doi.org/10.1007/s10257-023-00640-4

Abednego, L., Nugraheni, C. E., & Salsabina, A. (2023). Customer segmentation: Transformation from data to marketing strategy. *IAIC International Conference Series, 4*(1), 139–152. https://doi.org/10.34306/conferenceseries.v4i1.645

Boaz, N. (2020). *iFood dataset* [Data set]. GitHub. https://github.com/nailson/ifood-data-business-analyst-test/blob/master/ifood_df.csv

Candel, V., & Parmar, D. (2015). *Deep learning with H2O* [White paper/manual]. H2O.ai, Inc.

Godcares, N., Sirsath, A., Bongale, A., Kadam, P., Jayawal, R., & Patil, S. (2023). Exploring customer segmentation in the context of market analysis. In *Proceedings of the IEEE 11th Region 10 Humanitarian Technology Conference* (pp. 444–449). https://doi.org/10.1109/R10-HTC57504.2023.10461815

Herkert, D., & Mullenbach, T. (2021). Customer churn prediction with text and interpretability. *Towards Data Science.* https://towardsdatascience.com/customer-churn-prediction-with-text-and-interpretability-bd3d57af34b1

Namvar, A., Ghazanfari, M., & Naderpour, M. (2017). A customer segmentation framework for targeted marketing in telecommunication. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 1–7). https://ieeexplore.ieee.org/document/8258803

Neslin, S. A., Grewal, D., Leghorn, R., Shankar, V., & Teerling, M. L. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Nykodym, T., Kraljevic, T., Wang, A., Wong, W., & Fryda, T. (2024). *Generalized linear modeling with H2O*. H2O.ai, Inc. https://h2o.ai/resources/booklet/generalized-linear-modeling-with-h2o/

Onasanya, A., & Okonkwo, R. (2023). Predictive analytics for customer behaviour: Developing a predictive model that analyzes customer data to forecast future buying trends and preferences, enabling small businesses to tailor their marketing strategies. *Unpublished manuscript.* https://www.researchgate.net/publication/378176015

Proksch, M. (2020, February 18). *From GLM to GBM – Part 1*. H2O.ai Blog. https://www.h2o.ai/blog/from-glm-to-gbm-part-1/

Sana, J. K., Abedin, M. Z., Rahman, M. S., & Rahman, M. S. (2022). A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. *PLOS ONE, 17*(12), Article e0278095. https://doi.org/10.1371/journal.pone.0278095

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations.* Springer. https://doi.org/10.1007/978-1-4615-4651-1

Wu, S. (2023). Customer churn prediction in the telecommunication industry. *Advances in Economics, Management and Political Sciences, 4*, 41–50. https://doi.org/10.54254/2754-1169/4/20221017