Hotels choice prediction in security crisis times in Burkina Faso using exogenous factors, machine learning and Multi-Criteria Optimisation

Diarra Mamadou [1], Sanou Issiaka [2], Séré Abdoulaye [3]
{ diarra.md21@gmail.com [1], issiakadossama7473@gmail.com [2], abdoulaye.sere@recifaso.org [3] }

Ecole Polytechnique de Ouagadougou (EPO), Ouagadougou, Burkina Faso [1],
Université Nazi Boni (UNB), Bobo-Dioulasso, Burkina Faso [2,3]

Abstract. Customer satisfaction and retention are two priorities for the hotel industry. To achieve these objectives, hotels need to offer quality services. However, due to the security situation in Burkina Faso, external factors have become crucial in selecting accommodation sites. So, predictive analysis of data related to these exogenous factors is therefore becoming imperative for decision support.

Our decision for this predictive analysis was to compare some classification algorithms and multi-criteria optimization.

The simulation of different classification algorithms and multi-criteria optimization has shown that exogenous factors have a significant impact on customer choice. The results achieved an average accuracy of 80%.

Keywords: Hotel Market, External Environmental Factors, Prediction, Machine Learning, Multi-Criteria Optimisation.

1 Introduction

Tourism continues to flourish in Burkina Faso, despite the difficult security situation. The organisation of major international events such as Semaine Nationale de la Culture (SNC), Salon International de l'Artisanat de Ouagadougou (SIAO), Tour du Faso, Festival Panafricain du Cinéma et de la Télévision de Ouagadougou (FESPACO), Ouagadougou international tourism and hotel trade fair (SITHO) and many others are a perfect illustration of this. These events are a focal point for the populations of Burkina Faso and other countries, creating a need for comfort and security.

In an era of digital revolution, increasingly data-driven, the benefits of machine learning are clear to see. In an ever-changing business environment, where every detail counts in the decision-making process, machine learning algorithms can provide a strategic advantage that can make all the difference.

JRI 2023, December 18-20, Ouagadougou, Burkina Faso

DOI 10.4108/eai.18-12-2023.2348129

In this way, the prediction of certain endogenous factors such as comfort, room type and overnight stay, and exogenous factors such as airport and train stations, security services and hospitals and pharmacies, relating to hotel sites will help to influence customer choice. To predict this choice, we are carrying out a comparative study between Machine Learning, which refers to a set of inductive processes whose objective is to learn from data, and Multi-Criteria Optimisation, an approach that enables the best possible solution to be found by optimising a set of criteria and constraints.

This work is organised into three main parts. First, we present the state of the art, then we develop our contribution through the prediction methods chosen, and finally the evaluation and results.

2 Related Works

2.1 Hotel market and external environmental factors

The hotel industry includes all activities related to traditional hotels and large hotel groups, with the provision of short-stay accommodation for business travellers (and even other groups), as well as accommodation facilities for longer or shorter stays. External environmental factors represent characteristics or trends in the company's environment over which it has no control.

In this study, the factors concerned were safety, health, accessibility and distance from the hosting site.

2.2 Machine Learning, multi-criteria optimisation

2.2.1 Machine Learning Concepts

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves [1].

It can be categorized as supervised and unsupervised learning. Supervised learning is the main type of ML, where the algorithm learns from historical data that has both input and output values. It is widely used because it mostly needs less data for learning and resulting model eventually generalizes well. The core requirement of supervised learning based applications is proper labeling of the complete dataset with no missing class label and the quality of those labels also makes adifference. Amongst several techniques, support vector machines, Decision trees, Neural networks, Deep learning, Boosting, Classifiers, K-nearest neighbor, Genetic algorithm and classifier ensemble likeNeural Network, Logistic Regression, Random Forest, Naive Bayes, Support Vexteur Machine, Decision Trees and K-Nearest Neighbor [2, 3].

The main objective of machine learning is to analyse data in order to identify its structure, particularly when this structure is unknown in advance. It also aims to use computer programmes and systems capable of adapting and changing when confronted with new data. It is therefore based on computer algorithms, which are processes that enable a mathematical problem to be solved in a limited number of steps. Finally, machine learning techniques are used to build a mathematical model using training, validation and testing sets.

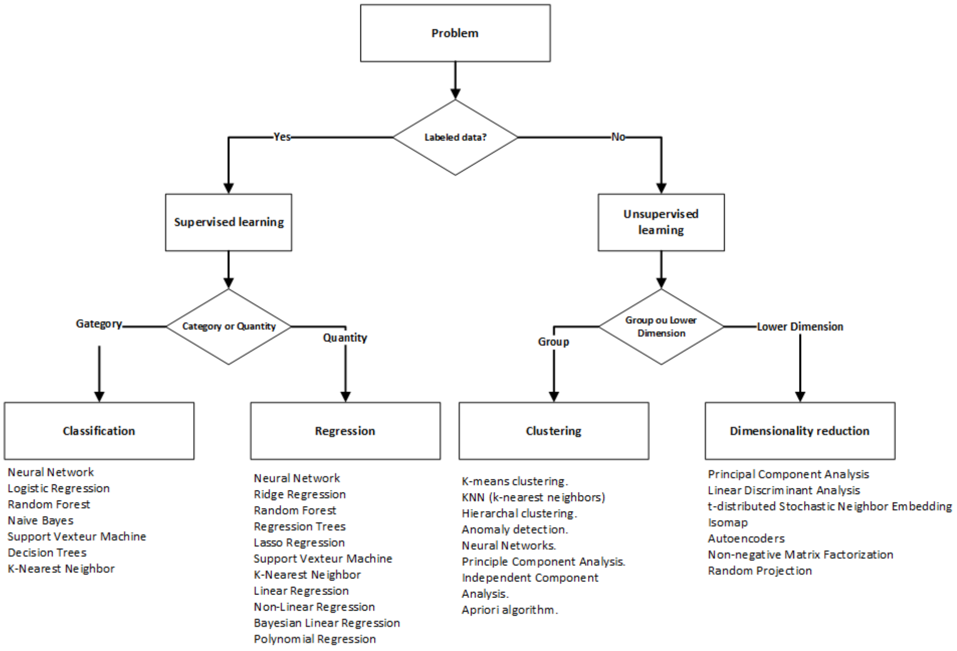


Fig. 1. Machine Learning.

2.2.2 Machine Learning Process

Without explicit programming, computers may learn from any incoming data through the machine learning process and improve their performance accordingly. It is a branch of artificial intelligence (AI) that deals with creating algorithms that let computers evaluate information, spot trends, and improve forecasts or choices. Fundamentally, a machine learning system makes use of data to continuously improve performance through iterations.

To carry out this comparative study of the performance between machine learning and multi-criteria optimisation in predicting the choice of a hotel establishment, we proceed as follows (Figure 3):

- **Data acquisition** is when you take samples from an environment or event and generate data. This data can then be converted by a computer into a graph or session. Data acquisition, also known as DAQ, usually involves collecting waveforms and signals and processing those signals to produce your desired information [4].

- **Data cleaning** is a critical aspect of machine learning, as a clean dataset cangenerate better results even with a less complex algorithm, compared to a complex algorithm processing a dataset with errors. A model's accuracy andreliability depend on the quality of the data used for training. Machine learning models are not intelligent and can only learn from the data presented to them. Therefore, inaccurate or incomplete data can lead to incorrect model features and flawed classification results [5].
- In machine learning, model deployment is the process of integrating a machine learning model into an existing production environment where it can take in an input and return an output [4].

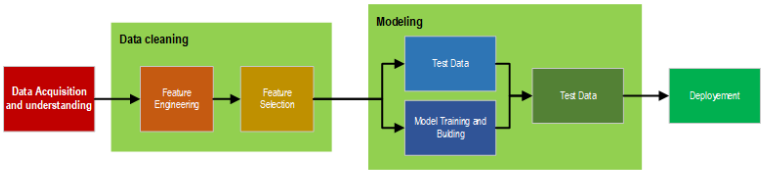


Fig. 2. Machine learning process.

- Model deployment is the process of putting machine learning models into production. This makes the model's predictions available to users, developers or systems, so they can make business decisions based on data, interact with their application (like recognize a face in an image) and so on.
- Model validation is the process by which model outputs are (systematically) compared to independent real-world observations to judge the quantitative and qualitative correspondence with reality.

2.2.3 Machine Learning Classification

Classification is a natural language processing task used in machine learning to assign labels to data items [6]. With an input training dataset, classification algorithms can identify categories and classify subsequent data accordingly. So they essentially identify and recognize patterns in the training data and use the findings to find similar patterns in future data.

Most classification models are supervised machine learning problems, since the input data has class labels. However, unlike unsupervised learning, supervised learning models use labeled input variables [7].

2.3 Multi-criteria Optimization

Multi-critera optimization deals with conflicting objectives and provides a mathematical framework to arrive at an optimal design state that accommodates various criteria. The process of multi-criteria optimization involves transforming the problem into a single criterion optimization problem, but the resulting solution is subjective to user parameter settings. Multi-criteria decision making

(MCDM) helps decision makers choose the most desirable alternative by considering multiple competing or conflicting criteria. MCDM methods include the Analytical Hierarchy Process (AHP) and the Elimination and Choice Translating Reality (ELECTRE) method, among others [8, 9] (Figure 3).

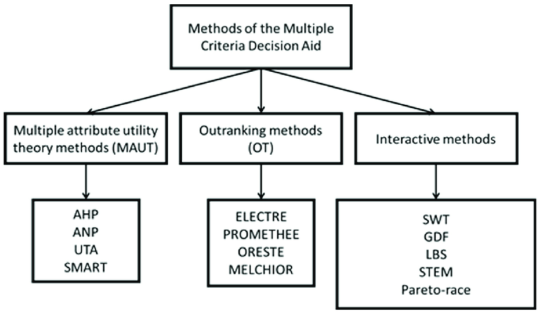


Fig. 3. Classification of multi-criteria decision aid (MCDA) methods according to P. Vincke. [9]

In this study, the multi-criteria optimisation method used is the Pareto chart, also known as ABC analysis. It is a method used to rank phenomena in order of importance or value in a given context. It is an effective decision-making tool in areas such as procurement, staff appraisal, sales, stock management and supplier relationship management.

The Pareto chart takes the form of a histogram ranking the causes of a problem in descending order, and has gained popularity from the many phenomena observed that obey the 20/80 law (20% of causes produce 80% of effects, and all you have to do is work on this 20% to strongly influence the phenomenon).

3 Our Contributions

3.1 Methodologies

To carry out this comparative study between the performance of machine learning and multi-criteria optimization in predicting the choice of a hotel establishment, we proceeded as follows:

3.1.1 Data acquisition

Our data come from the study: ”Orientation marché et performance commerciale des établissements hôteliers des villes de Ouagadougou et Bobo-Dioulasso : le rôle modérateur de l’environnement externe”. The initial data collection included 47 criteria and 500 records.

However, this study focused on 8 external criteria and 8 criteria relating to the hotel standing, which influence customer choice (Table 1). In addition, we found that the data collection covered 13 regional capitals in Burkina Faso, which justifies the number of records.

Table 1: Selection of determining criteria

Hotel standing	External criteria
Note	Location
Category	Airport
Bed places	Public service
Room price	Bus station
Restaurant	Road station
Number of beds	Pharmacy
Bed place	Security
Swimming pool	Healthcare

3.1.2 Data cleaning

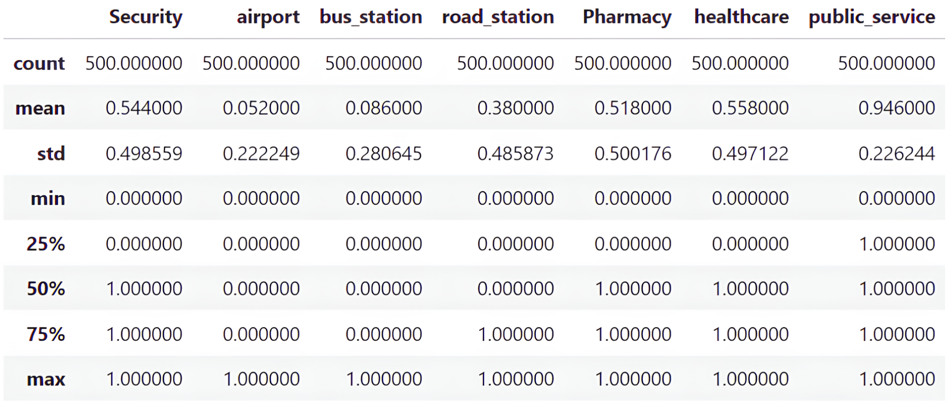
It consisted in extracting the criteria that were the subject of this study, namely 13 criteria with categorical values related to comfort and external environmental factors such as safety, health, admiring and public services and accessibility to the site.

The integration of huge amounts of variable data requires various strategies and resources to homogenize them. To ensure that all data is of the same quality, it must also be cleaned and filtered before converting and integrating it. To achieve this, we performed data mining before correcting the missing, erroneous and/or outlier data using the median method. Then we tagged the variables. And finally, we made a standardization of the data to improve their efficiency and so that they correspond to a predefined and constrained set of values without however distorting them. For numerical data, we used the min-max method or the normalization z-score, the one-hot encoding to transform categorical data into binary variables.

3.1.3 Create training and test datasets

For the simulation, we use a PC running Windows 10, with Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz 2.40 GHz with 16 Go of RAM. The 500 records were divided into training and test data with a proportion of 80 / 20 or 400 records for training and 100 for testing. The target values are the standing of the hotel, and external factors like Security, airport, bus station, road station, Pharmacy, healthcare and public service (Figure 4).

To determine the most decisive criteria in the choice of a hotel, we used the correlation matrix (Figure 5) and principal component analysis (PCA). Taking into account a correlation of over 50%, we found that 7 of the 16 criteria were highly correlated in terms of hotel choice. These highly correlated criteria are made up of 42.86% endogenous criteria and 57.14% exogenous criteria (table 2).



	Security	airport	bus_station	road_station	Pharmacy	healthcare	public_service
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000
mean	0.544000	0.052000	0.086000	0.380000	0.518000	0.558000	0.946000
std	0.498559	0.222249	0.280645	0.485873	0.500176	0.497122	0.226244
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
50%	1.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
75%	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Fig. 4. Dataset describe

Table 2: Selection of determining criteria

Features	Weight
nb_beds	-0.762993
bedplaces	0.746361
room_price	-0.714898
road_station	0.702439
Pharmacy	0.922100
Security	0.980008
healthcare	0.943320

3.1.4 Deploy ML model.

Once the data preparation process was complete, we extracted a different dataset for each target variable, before dividing them into a training set and a validation set in an 80/20 proportion. We then carried out a comparative study of the 5 classification models, Random Forest Regressor (RFR), Decision Tree Regressor (DTR), Support Vector Machines (SVM), K Neighbors Regressor(KNN),Naive Bayes Classifier(NBC) in Figure 1 with 6 metrics: MAE MSE RMSE R2 RMSLE MAPE [10].

The various simulations and the optimization of the hyperparameters gave us the scores summarized in the table below (Table 3 and Figure 6).

Analysis of results To evaluate the different models, we focused on 6 metrics, with particular reference to RMSE and R2-score. A descriptive approach to the various metrics reveals the following scores:

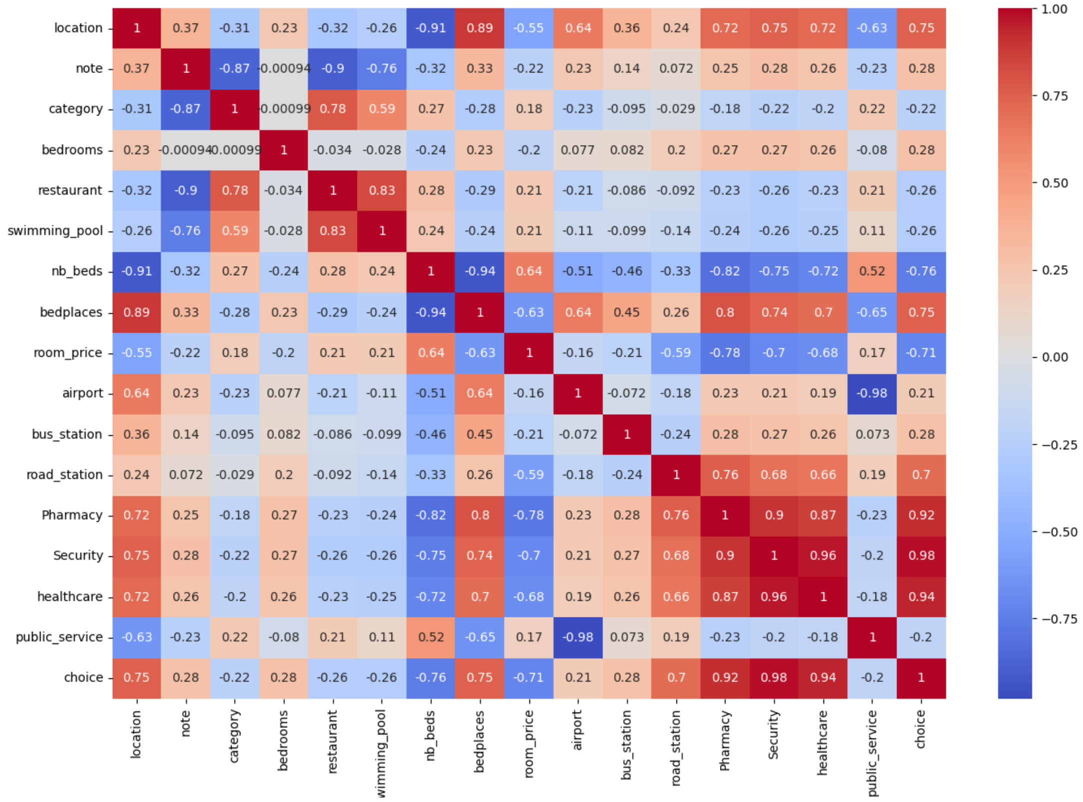


Fig. 5. Correlation matrix for dataset

- MAE, des scores qui varient de 0.1168 de 0.0480, avec la meilleure valeur pour le KNN et la pire pour le SVM.
- MSE, des scores qui varient de 0.0213 de 0.1444, avec la meilleure valeur pour le RFR et la pire pour le NBC.
- RMSE, des scores qui varient de 0.0213 à 0.3706, avec la meilleure valeur pour le RFR et la pire pour le NBC.
- R2, des scores qui varient de 0.4107 à 0.9122, avec la meilleure valeur pour le RFR et la pire pour le NBC.

Table 3: Comparison of different classification models

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
RFR	0.0467	**0.0213**	**0.1431**	**0.9122**	**0.0976**	0.0451
DTR	**0.044**9	0.0372	0.1883	0.8464	0.1298	**0.0334**
SVM	0.0480	0.0480	0.2113	0.8026	0.1465	0.0483
KNN	0.1168	0.0731	0.2625	0.7015	0.1842	0.1068
NBC	0.2579	0.1444	0.3706	0.4107	0.2167	0.2718

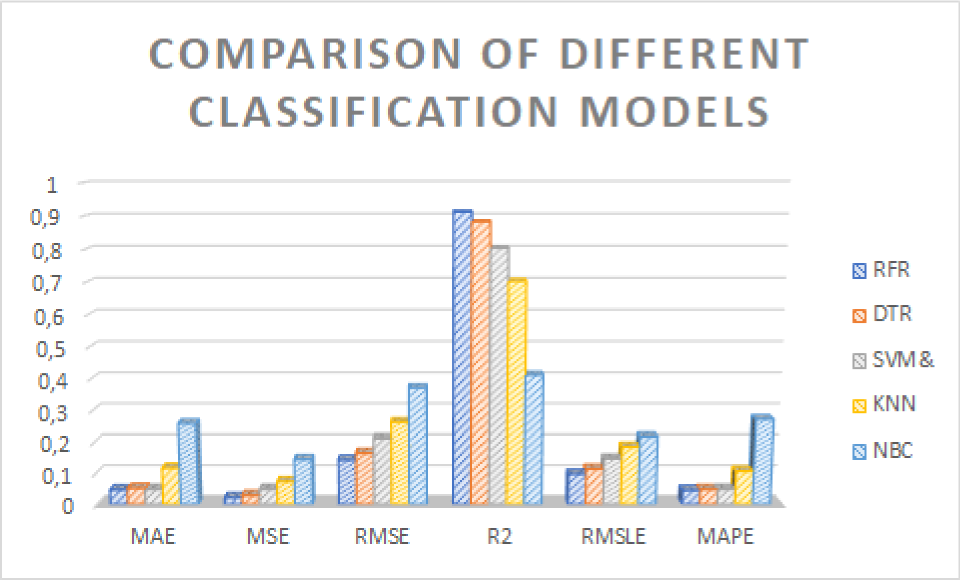


Fig. 6. Model comparison chart

- RMSLE, des scores qui varient de 0.0976 à 0.2167, avec la meilleure valeur pour le RFR et la pire pour le NBC.
- MAPE, des scores qui varient de 0.0334 à 0.2718, avec la meilleure valeur pour le DTR et la pire pour le NBC.

From the above, we can confirm that RFR, gives the best of 4 of the 6 metrics that were the subject of our study. Also, concerning the main metrics of this study, namely RMSE and R2-score, RFR comes first with a score of 0.0213 for RMSE and 0.9122 for R2-score.

3.1.5 Applying multi-criteria optimization

Implementation of the Pareto principle, also known as the 80/20 rule, based on 13 categorical criteria and 500 collected hits. An application of the correlation matrix and PCA enabled us to retain the 8 criteria with a correlation greater than 50 (Table 2).

To apply the 20/80 rule, we weighted the coefficients of the various criteria. This weighting is summarized in table 4

Table 4: Multicriteria optimisation: optimality of Pareto sense

Features	Weight
location	75.0492
nb_beds	76.2993
bedplaces	74.6361
room_price	71.4898
road_station	70.2439
Pharmacy	92.2100
Security	98.0008
healthcare	94.3320

Table 5: Multicriteria optimisation: optimality of Pareto sense

N°	***Features***	***Count***	***count cumsum***	***cumpercentage***
6	Security	98	98	0.150769
7	healthcare	94	192	0.295385
5	Pharmacy	92	284	0.436923
1	nb_beds	76	360	0.553846
0	location	75	435	0.669231
2	bedplaces	74	509	0.783077
3	room_price	71	580	0.892308
4	road_station	70	650	1.000000

Application of the Pareto principle to the various determining criteria (Table 5) shows that only 3 of the 8 factors have an influence on hotel choice. These are the highest level of security at 98, followed by health centers at 94 and pharmacies at 92. Also, his interpretation of Pareto's law allows us to assert, according to the decision curve (Figures 7,8), that his 3 criteria account for 80% of hotel choice decisions.

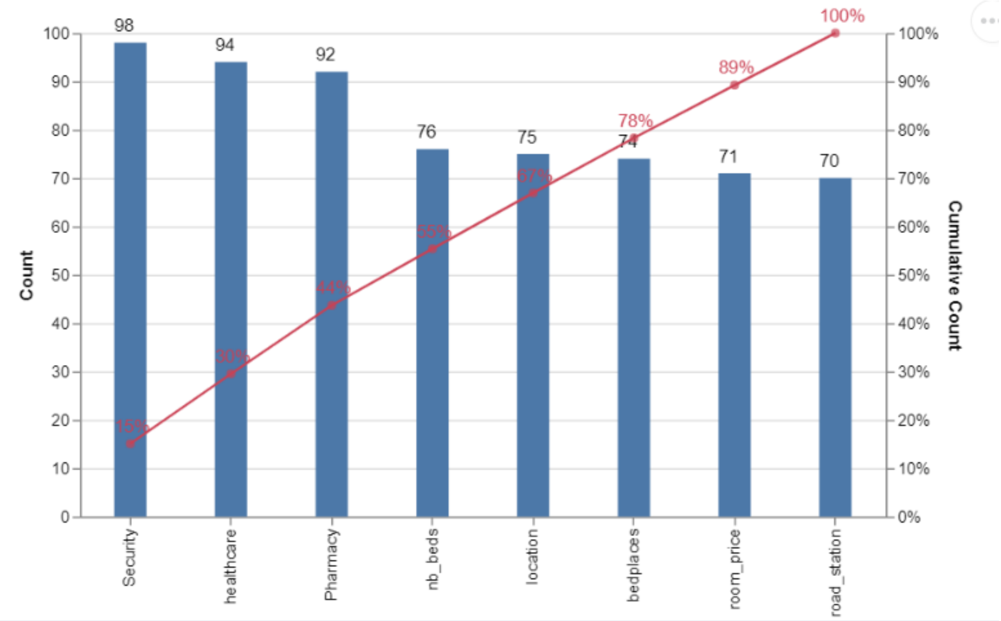


Fig. 7. Pareto chart of customer priorities

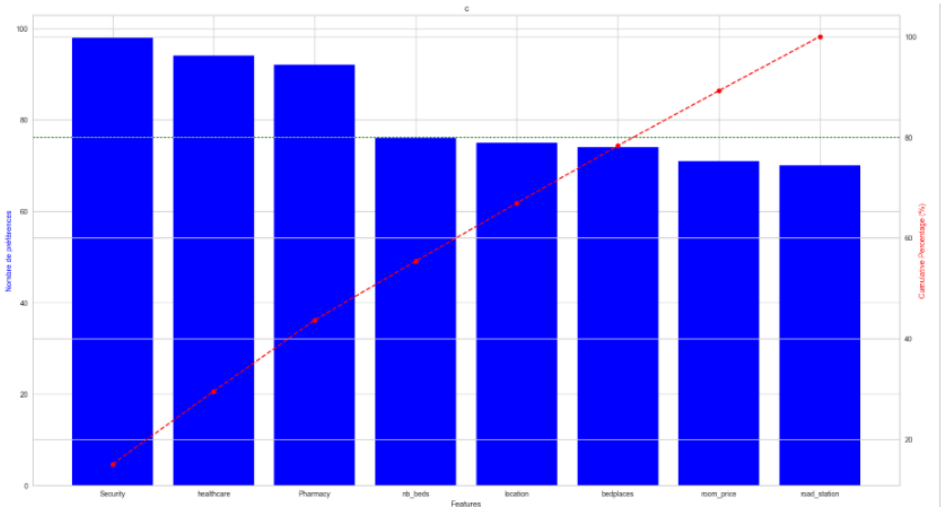


Fig. 8. Pareto chart of customer priorities

4 Conclusion and Outlooks

Our study highlights the link between market orientation and company performance, with more than 90% prediction of customer choice.

The comparative study also shows that decision trees are the method that gives the best prediction of the customer's choice. Indeed, it gets the lowest RMSE (0.0213) and the highest R2-Score (0.9122).

Another contribution of this study was the attempt to predict the choice of hotel establishments using multi-criteria optimization. Also, this approach allowed us to determine the criteria influencing the choices of customers. Thus, on the 8 criteria having a correlation greater than 50, only 3, safety, health centers and pharmacies are decisive in the choice of customers.

In short, this study proves that machine learning can predict customer choice and multi-criteria optimization determine the criteria influencing said choice.

In perspective, it would be interesting to resume a similar study in pheasants a comparative study between machine prediction before and after applying Pareto's law.

References

[1] Meena S. Applications and Limitations of Machine Learning Process.

[2] Kaddoura S, Popescu DE, Hemanth JD. A systematic review on machine learning models for online learning and examination systems. PeerJ Computer Science. 2022;8:e986.

[3] Nedal M, Kozarev K, Arsenov N, Zhang P. Forecasting solar energetic proton integral fluxes with bi-directional long short-term memory neural networks. Journal of Space Weather and Space Climate. 2023;13:26.

[4] Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (mlops): Overview, definition, and architecture. IEEE access. 2023.

[5] Singh H, Dogra A, Modi P. Data Cleaning in Eye Image Dataset. 2023.

[6] Sharifani K, Amini M, Akbari Y, Aghajanzadeh Godarzi J. Operating machine learning across natural language processing techniques for improvement of fabricated news model. International Journal of Science and Information System Research. 2022;12(9):20-44.

[7] Tatsat H, Puri S, Lookabaugh B. Machine Learning and Data Science Blueprints for Finance. O'Reilly media; 2020.

[8] Odu G, Charles-Owaba O. Review of multi-criteria optimization methods–theory and applications. IOSR Journal of Engineering. 2013;3(10):01-14.

[9] Nosal Hoy K, Solecka K, Szarata A. The application of the multiple criteria decision aid to assess transport policy measures focusing on innovation. Sustainability. 2019;11(5):1472.

[10] Gama F, Magistretti S. Artificial intelligence in innovation management: A review of innovation capabilities and a taxonomy of AI applications. Journal of Product Innovation Management. 2023.