MACHINE LEARNING ALGORITHMS FOR SALES PREDICTION IN STORES - AN APPLIED RESEARCH

Bogdan TIGANOAIA[1], Ionut-Petrisor ANGHEL[2]

Data Science and Machine Learning are trending topics in our current times and gaining knowledge. With algorithms and specific models, we can produce accurate results for analyzing large volumes of data. The objective of this article is to test several machine learning algorithms to establish the best method to predict sales in a retail store. We will also analyze and explore the recorded transactions to be able to determine representative information such as the customer category that purchase the most products and spend the most as possible depending on age, gender, stability in the city, occupation etc.

Keywords: Black Friday, Supervised learning, Tree-based methods, Ensemble methods, RMSE, ROC_AUC.

1. Introduction

Machine Learning algorithms are classified into three representative classes according to the learning model [1]:

- *Supervised learning:* models are built through a training process, models that adjust their variables to map the inputs to the corresponding output (regression or classification algorithms);
- *Unsupervised learning:* models are built based on the structures present in the input data (there is no target result, thence the algorithms will group the data set into different groups);
- *Reinforcement learning:* models are built through a trial-and-error process to achieve goals (the most interesting category having diverse applications such as chemical reaction optimization, AlphaGo, intelligent traffic light systems, personalized recommendations, etc.).

One of the most popular applications of Machine Learning in various fields is to predict customer's behaviour. The current context of this paper is that many stores offer during Black Friday highly promoted products at cheap prices, and this season is part of the winter holiday shopping period, therefore it's always the busiest shopping season of the year. So the main problem for a retail store in this season is to choose product price to earn more profit and with help of machine learning

[1] Faculty of Automatic and Computers, The National University of Science and Technology POLITEHNICA of Bucharest, Romania, e-mail: bogdantiganoaia@gmail.com

[2] Faculty of Automatic and Computers, The National University of Science and Technology POLITEHNICA of Bucharest, Romania, ionut.anghel2906@stud.acs.upb.ro

algorithms we can build a predictive model to help us to make decisions on the growth of our store based on the historical sales data.

The Black Friday is crucial season for the economy and some quantifiable statistics according to the National Retail Federation (NRF) are that 84.2 million people shopped in stores on Black Friday in 2019 and customers spent an average 361.90 dollars on holiday items over the five-day period. Furthermore, the biggest spenders were 25 to 34 year olds at 440.46 dollars, closely followed by those 35-44 at 439.72 dollars.[3]

Problem statement

In this problem we want to predict the total amount of money spent by a customer in a store during Black Friday. A file with transaction data exposing various characteristics of the buyers (age, geographic area, marital status, gender, occupation, etc.) is offered for analysis.

Motivation

Considering the study carried out by Bluecore in the year 2021 which illustrated the fact that more than half of the purchases came from the buyers who benefit from the Black Friday offers for the first time [2] it is desired to:

- Determining highly sought-after types of products during the busiest time of the year;
- Finding group of buyers who spend the most;
- Establishing the optimal marketing strategy by creating a personalized offer for customers depending on their purchasing behavior towards products from various categories.

Defining problem type

We have a target variable, namely the total price of purchased products by a customer so the problem we have to solve is part of the supervised learning category. Also, the label is continuous, therefore we will use regression-type algorithms.

Now, once the problem has been defined and its type has been identified, we will mention the classic methods that we can choose to solving it with their advantages and disadvantages. The following sections focus on:

- ***the data set*** – identifying missing values, determining the distribution of buyers according to gender, geographic region, sales distribution, the correlation matrix and adding new attributes in order to improve the accuracy;
- ***proposed methods*** – we apply a variety of standard algorithms from the sklearn, lightgbm, xgboost and catboost libraries and present the benefits offered by each algorithm;

[3] https://nrf.com/media-center/press-releases/thanksgiving-draws-nearly-190-million-shoppers

- ***results*** – evaluation of the performance of the applied methods according to the most important factors: R2, RMSE, ROC_AUC score;
- ***conclusions and improvements*** that could be made in the future in order to be able to establish an optimal marketing strategy according to the most requested products during Black Friday.

The methodology of the research implies the following steps:

1. An analysis of the literature on the research issue – the state of the art in order to show what other relevant research has been conducted. The topic is not new, so there are other existing approaches and solutions, but our results are better than the existing ones.
2. The description of the methods applied – see chapter 4 – proposed methods.
3. The use of these methods – see also chapter 4.
4. Discussion and findings – this section discusses the results and it is made a comparison with the results obtained in the reference source mentioned.

2. Available methods

The supervised machine learning technique involves training a model to predict a target variable based on a set of known features. Through this training process, the features of the known labels are adapted to obtain a general function that can predict the target value for new entries (unseen data), function defined in the form below where x is the feature vector and y is the variable to predict.

$$y = f(x) \tag{1}$$

$$x = [x_1, x_2, x_3, \dots] \tag{2}$$

The purpose of training the model is to obtain by applying a machine learning algorithm a reasonably accurate y-value for the x-values by reference to all cases in the training data set. Many supervised machine learning algorithms have been studied, which is why we can choose from a multitude of approaches to train a model for a given problem. These algorithms are divided into two classes, namely:

- ***Regression algorithms*** – a numerical value is predicted, for instance the purchase price for a certain product;
- ***Classification algorithms*** – the class or category to which a tag belongs is predicted, the y value in equation (1) being a probability vector between 0 and 1 that indicates the chance that said tag belongs to a class.

Except these classical methods, using advanced machine learning algorithms with deep neural networks we can improve the performance of the models. Deep Learning is a specialized subset of Machine Learning with both supervised and unsupervised learning that describes algorithms with brain like logical structure of algorithms, called artificial neural networks. Now, Deep Learning has attained so much usability with the emergence of cloud computing infrastructure and high-performance GPUs the features are extracted automatically and the algorithm learns from its own errors. [3]

2.1 Relevant methods

Determining the purchase price of a product is a regression problem, hence we can opt for one of the following methods used in practice:

- *Linear methods* – an affine function can be defined between the features of the vector x and the label y. In other words, x and y are linearly related by the function below:

$$f(x) = w^T x + b; w, x \in \mathbb{R}^d, b \in \mathbb{R} \tag{3}$$

An advantage of these methods is their simplicity and reduced training time.
Some examples in this category are: Linear Regression, Ridge, Lasso, etc.

- *Tree-based methods* – decision trees are built to make a prediction and are some of the best and most widely used methods due to their adaptability, stability, high accuracy, and ease of interpretation. An illustrative example is the Decision Tree Regression algorithm whose process is exemplified in the figure below:

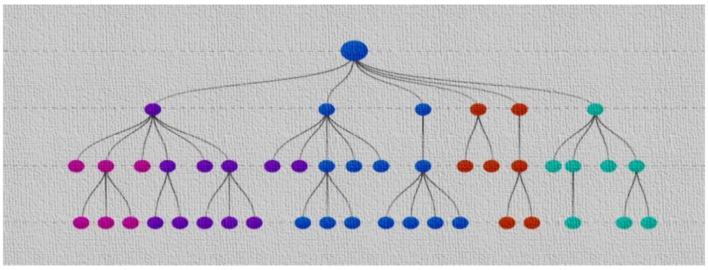


Fig. 1. Example decision tree

- *Ensemble methods* – involve a group of predictive models to produce an optimal model that achieves better accuracy and stability. Some examples of such models are Random Forest, Gradient Boosting, Extreme Gradient Boosting, LightGBM, CatBoost. Models trained using these methods try to achieve a balance between bias and variance errors, that is, it is a way to analyze a trade-off of underfitting and overfitting as can be seen in the figure below:

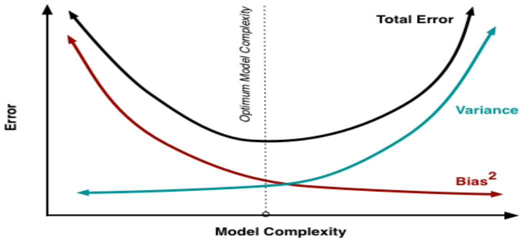


Fig. 2. Managing the bias-variance trade-off [4]

2.2 Results obtained

The problem was proposed on the Analytics Vidhya portal in the Black Friday contest[4] and the best RMSE score obtained at the current moment is **2372.** It is not specified what method was applied to achieve this metric but given a large enough data set it is very likely that a multi-layer neural network was constructed.

There are some public solutions on GitHub and of these we took the source of Shweta Chandel as a reference[5]. Several algorithms were applied, and the obtained results are in the table below:

Table 1

Comparison of applied models based on the same raw and with the same font

Algorithm	RMSE
LinearRegression	4609.92
DecisionTreeRegressor	3786.33
RandomForestRegressor	2786.27
AdaBoostRegressor	3855.36
GradientBoostingRegressor	2829.88
XGBRegressor	2591.85

As it can be seen from the table above the XGB method has been validated and it achieves the best performance. Also, to improve the accuracy the respective model was trained on the entire data set and achieved a score of 2574.95 (on the test set used to evaluate the models in the competition)

On the other hand, Extreme Gradient Boosting (XGB) is the core algorithm for winning competitions on the Kaggle platform[6]. In the literature review, there are many methods proposed to analyse and predict sales using various machine learning techniques, so below we will summarize a few of these approaches.

Firstly, Mohamed Leila applied Deep Learning models to the problem and saw how would they perform with help of the tensorflow library. He increased the

[4] https://datahack.analyticsvidhya.com/contest/black-friday/
[5] https://github.com/shwetachandel/Black-Friday-Dataset
[6] https://www.kaggle.com/getting-started/145362

embeddings dimensions for some columns, applied features crossed tehnique, added additional layers and increase the number of units per layer and so obtained the 2640 value for RMSE metric. [5]

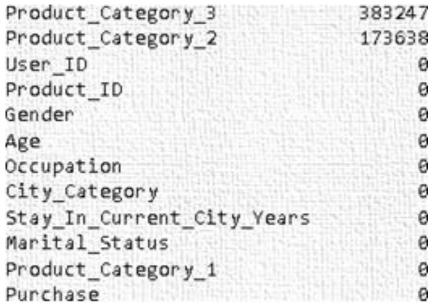
Secondly, Rising Odegua applied three machine learning algorithms namely K-Nearest Neighbor (KNN), Random Forest (RF) and Gradient Boosting (GB) to build a sales forecasting model using the data provided by Data Science Nigeria. In order to evaluate the performance of models, he calculated the MAE metric (Mean Absolute Error). The best model was with RF algorithm with MAE rate of 0.409178. [6]

Thirdly, the authors of the article [7] used the Black Friday Sales Dataset from Kaggle platform and build many predictive models based on algorithms such Linear and Ridge Regression, Extreme Gradient Boosting Regression (XGB), Decision Tree, Random Forest and Rule-Based Decision Tree (RBDT). As a result based on the RMSE metric, the last algorithm obtaines the best score with a RMSE of 2291.

3. Data set

The data set[7] contains a summary of purchases from a store and given that these data were collected during Black Friday, the number of records is very large, namely 550068. In addition to product details such as id, category, purchase price records contain customer demographic information like age, gender, marital status, city stability, etc.

An important first step is to check for missing data, and we notice that we have two features out of 12 with a very high percentage of the total number of records 0.69%, respectively 0,31%.



```
Product_Category_3              383247
Product_Category_2              173638
User_ID                              0
Product_ID                           0
Gender                               0
Age                                  0
Occupation                           0
City_Category                        0
Stay_In_Current_City_Years           0
Marital_Status                       0
Product_Category_1                   0
Purchase                             0
```

Fig. 3. Missing data

As we have a very large number of missing data, we cannot remove the two columns from the figure above, as we would have a compromise of the model

[7] https://www.kaggle.com/competitions/gb-black-friday-sales/data

(under fit). Also, as there is interconnection between the 3 types of products categories, we chose to fill undefined values with the value 0 over an average or frequency (mean, mode) strategy.

A distribution of buyers visiting Black Friday stores according to gender and geographical area was made and it was found that we have 3 times more men than women and most customers come from area B as can be seen from the figure below:

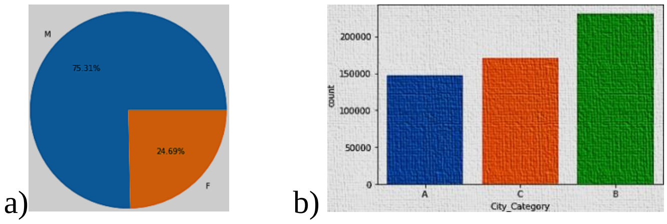


Fig. 4. Distribution of customers by (a) gender (b) geographical area

Although we previously noticed that we have about 50,000 more customers from zone B than the next ranked zone, those from zone C occupy the first position with a percentage of 34.98% in relation to the purchase price, which means that in area B we have more purchases but with lower prices and in area C the opposite. Also, the purchase percentages by area are roughly the same so there is no clear distinction between the source and target area.

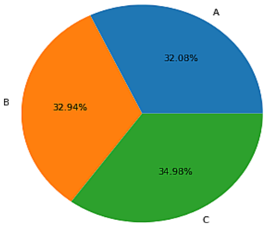


Fig. 5. Purchase amount distribution by geographic area

Following an analysis of the number of buyers according to 3 representative fields - age, gender, and marital status, it is observed that in raport with the number of purchases young people (26-35 years old) predominate with a weight of ~40%. As was expected people in the category 0-17 years had the fewest purchases due to the low economic power (they don't have a job). Moreover, in this category (0-17) there are sections only for single people (0), but this makes sense, and married people shop less than single people. Also since there is no category that explicitly describes the type of products by gender it is possible that the figure below incorrectly reflects the fact that men shop more than women.

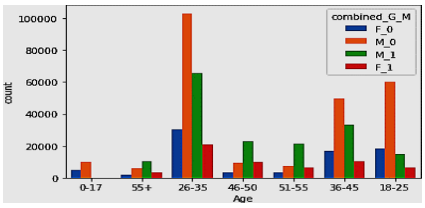


Fig. 6. Determining the number of buyers according to age, gender, and marital status

Although the figure above shows that we have a high percentage of young customers (26-35 years old) the purchase value is uniformly distributed by age as can be seen from the box plot diagram below:

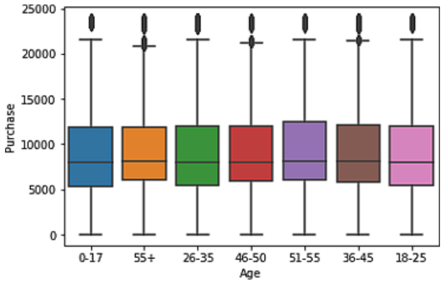
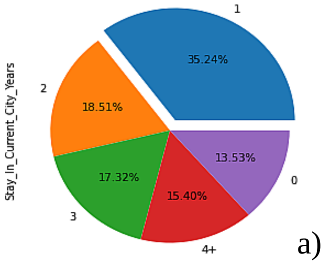
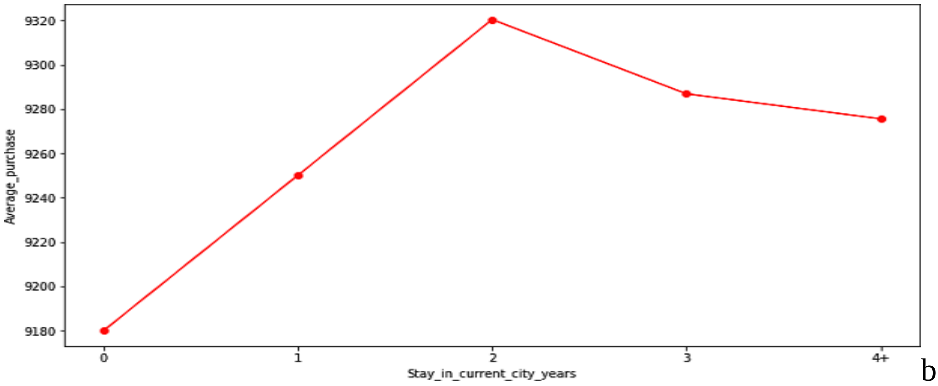


Fig. 7. Purchase distribution by customer age

Another significant information is the distribution of buyers and average purchase value by city stability. It is noted that most customers stayed in the current city for only one year and the weight of this majority is clearly superior to the other categories, but in relation to the target value, those who have stability for 2 years paid more on average although are half of those who stayed 1 year.





b)

Fig. 8. Stability distribution in the city according to (a) no. of buyers (b) average purchase value

Black Friday is a famous time for discounted products; therefore, many people buy in large numbers and the purchase amount is repeated in many cases. This aspect translates into a multinomial distribution sales distribution as shown in the graph below:

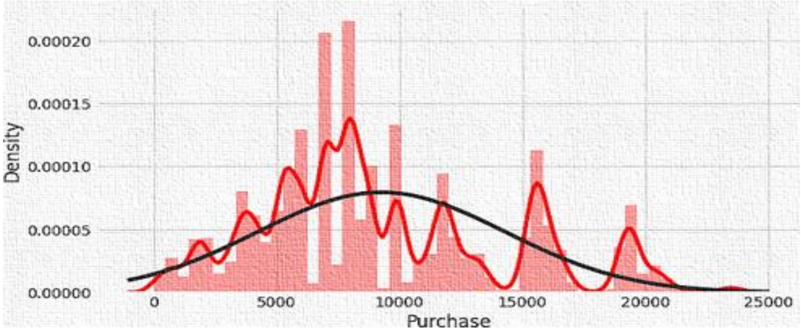


Fig. 9. Distribution sale

Another aspect to consider is building the correlation matrix to consider the best columns that have an impact on the target value. To improve the accuracy of the model two additional columns were added:

- Category_Count – number of unique categories purchased by each user (1, 2 or 3);
- Product_Score – the ratio of the product frequency to the maximum frequency in the data set.

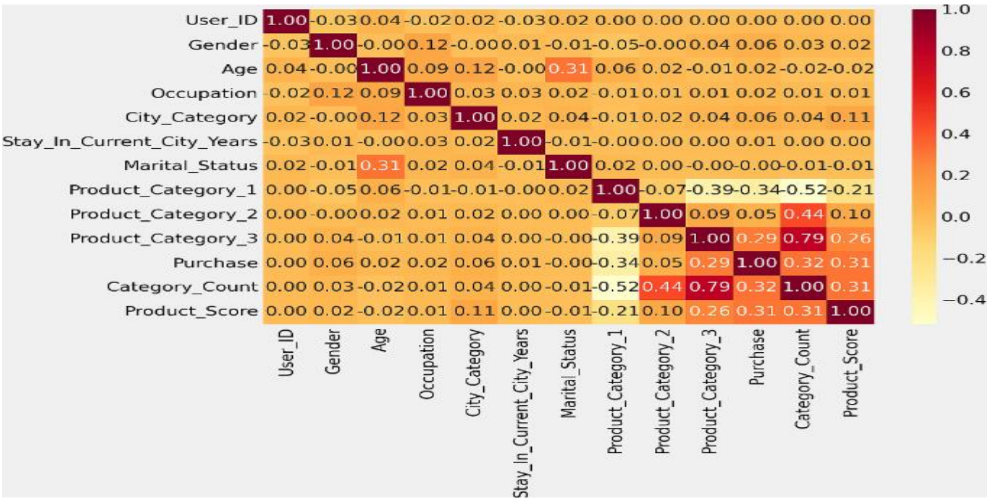


Fig. 10. Pearson correlation matrix

As it can be seen above the dependent characteristic "Purchase" is well negatively correlated with the independent variable ”Product_Category_1” and positively with ”Product_Category_3” and the other two added.

Additionally in the preprocessing step each age range was mapped to the mean value, object-type attributes were converted to numeric attributes using Label Encoder and redundant columns were removed like ”Product_ID” and ”Marital_Status”. The column ”User_ID” was not removed even though we have a (linear) 0 correlation with the target value as it was observed that better metrics are obtained for several algorithms used in terms of the RMSE metric and the results are stable for different seeds. One explanation for this process is that a non-linear correspondence is made between the independent and the dependent variable.

4. Proposed methods

Before building the models, the data set was split into 80% training data and 20% test data. To ensure that the two subsets of data are statistically comparable the ”train_test_split” function from the scikit-learn library was used which randomly splits the data.

Several methods have been applied:

- ***Linear Regression* [8]**
- Linear model that determines y from a linear combination of input variables;
- The "method of least squares" is mainly used to determine the linear function that best approximates the list of (y, x) pairs;
- It requires elimination of collinearity (no two independent variables should be highly correlated) to avoid overfitting the data.
- ***K-Nearest Neighbors Regression* [9]**

- It uses the idea of similarity (proximity) and the most used distance function is the Euclidean;
- The target variable is predicted by interpolating the results associated with a narrow neighborhood;
- The "GridSearchCV" technique is used to determine the optimal value of the number of neighbors (K).

▪ ***Decision Tree Regression*** **[10]**

- Uses a set of binary rules to calculate the target value;
- Requires little data preprocessing to find a sine curve based on local linear regressions;
- It is based on heuristic algorithms so the determination of the global optimal decision tree cannot be guaranteed.

▪ ***Random Forest Regression*** **[11]**

- Uses multiple decision trees and applies an averaging function to find a better overall model;
- Models using the bagging technique (there is no interaction between trees);
- It works efficiently on large data, allows the identification of important variables in the regression but unlike the Decision Tree it requires more computing power.

▪ ***Extreme Gradient Boosting Regression*** **[12]**

- Effective implementation of the Gradient Boosting algorithm being used in many competitions due to its superior performance and execution speed;
- Suitable for tabular data, there is also an optimized DMatrix structure that provides efficiency.

▪ ***Light Gradient Boosting Machine Regression (LGBM)*** **[13]**

- Converges much faster than XGB (about 7 times) due to avoiding depth traversal and is suitable for large data sets;
- Continuous values are replaced by discrete values and in consequence the memory usage is reduced;
- To avoid overfitting, the appropriate choice of the parameter max_depth is recommended.

▪ ***CatBoost Regression*** **[14]**

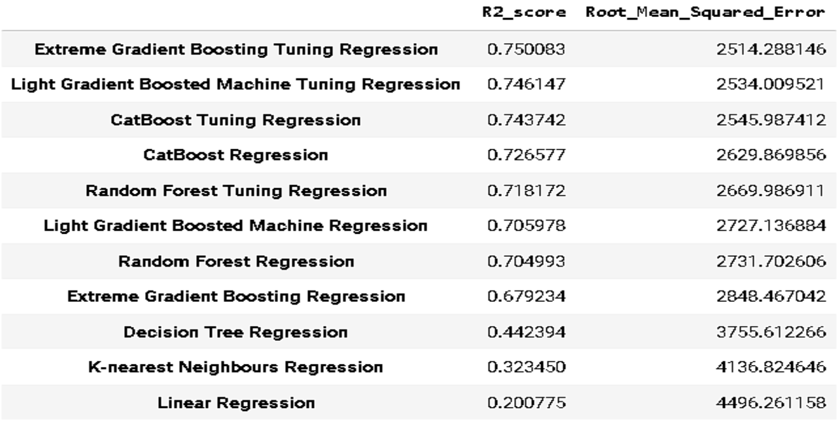
- Overfitting is controlled by building a balanced (symmetric) tree structure unlike XGB and LGBM where asymmetric trees are used;
- It uses the process of ordered boosting that is a permutation-based approach (the model is trained on a subset of the data and the residuals on another subset);
- The default settings provide very good performance (by adjusting the hyper parameters, results close to the initial ones are obtained).

5. Results

For the evaluation part of the models, we do not only want to visualize the results but also to determine some quantitative measures that appreciate their performance and accuracy. In statistics two representative metrics are often used for regression problems:

- R_2 (coefficient of determination) – measure to indicate the percentage of variation in the response variable;
- RMSE (root mean square error) – measure that indicates the square root of the average of the squared errors to show how close the estimated value is to the real one.

To choose the best model a high value (as close as 1) of the R_2 score is sought, respectively a small value (as close as possible to 0) of the RMSE metric. In order to obtain better metrics some of the models described in the previous chapter were optimized using the "RandomizedSearchCV" technique.



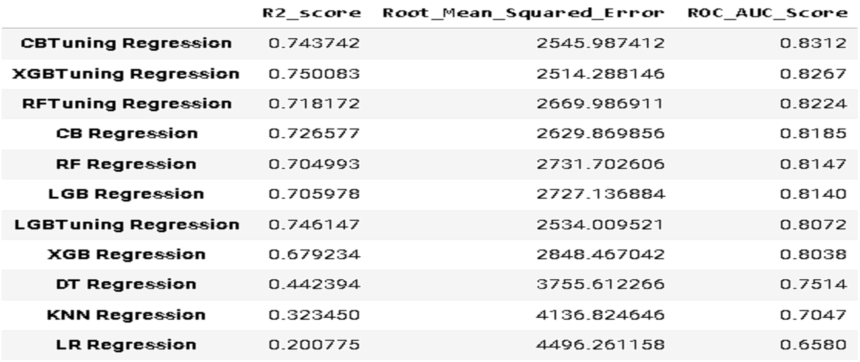
	R2_score	Root_Mean_Squared_Error
Extreme Gradient Boosting Tuning Regression	0.750083	2514.288146
Light Gradient Boosted Machine Tuning Regression	0.746147	2534.009521
CatBoost Tuning Regression	0.743742	2545.987412
CatBoost Regression	0.726577	2629.869856
Random Forest Tuning Regression	0.718172	2669.986911
Light Gradient Boosted Machine Regression	0.705978	2727.136884
Random Forest Regression	0.704993	2731.702606
Extreme Gradient Boosting Regression	0.679234	2848.467042
Decision Tree Regression	0.442394	3755.612266
K-nearest Neighbours Regression	0.323450	4136.824646
Linear Regression	0.200775	4496.261158

Fig. 11. Evaluation of metrics

As can be seen from the table above the weakest model is the linear one as it has the lowest R_2 score, and the highest root mean square error so no linear function can be found that approximates well the purchase price of a product.

An interesting thing to note is that parameter tuning in the case of the CatBoost method does not bring significant performance, as expected given the fact described in the previous chapter that the initial settings of the method provide very good performance. Also, the best metrics are given by the optimized XGB model, but unlike the less expensive LGBM variant, no notable differences are brought (0.01 better R_2 and 20 better RMSE) hence from the point of view of the time and memory used, we would opt for this light model. We continued the process of comparing the trained models considering the roc_auc_score metric which for regression problems is identified as the probability of obtaining an estimated target value X for an input sequence higher than another estimated target value Y under the conditions that the real target value X is higher than Y (a max_iter number of input sequences is randomly chosen and checked how many times the below condition is met. [15]

$$\left(y_{pred}[i] - y_{pred}[j]\right) / \left(y_{test}[i] - y_{test}[j]\right) \geq 0 \tag{4}$$



	R2_score	Root_Mean_Squared_Error	ROC_AUC_Score
CBTuning Regression	0.743742	2545.987412	0.8312
XGBTuning Regression	0.750083	2514.288146	0.8267
RFTuning Regression	0.718172	2669.986911	0.8224
CB Regression	0.726577	2629.869856	0.8185
RF Regression	0.704993	2731.702606	0.8147
LGB Regression	0.705978	2727.136884	0.8140
LGBTuning Regression	0.746147	2534.009521	0.8072
XGB Regression	0.679234	2848.467042	0.8038
DT Regression	0.442394	3755.612266	0.7514
KNN Regression	0.323450	4136.824646	0.7047
LR Regression	0.200775	4496.261158	0.6580

Fig. 12. Comparison of trained methods

Regarding the AUC score, values between 0.8 and 0.9 are considered good, hence considering the 3 metrics and the results in the figure above we can admit that the optimized XGB model is the right model to predict the price of purchase of a product for the data set considered. As a comparison with the results obtained in the reference source mentioned in Chapter 2 we note that we obtained a better RMSE metric for each individual model as shown in the table below:

Table 2

RMSE metric comparison with reference values based on the same raw and with the same font

Algorithm	RMSE ref	RMSE
LinearRegression	4609.92	4496.26
DecisionTreeRegressor	3786.33	3755.61
RandomForestRegressor	2786.27	2669.98
XGBRegressor	2591.85	2514.28

From a graphical point of view, we see that the values predicted by the best XGB model and the reference values are about the same, but there is certainly room for improvement.

For an explanation of why people frequent stores during Black Friday is the result below which illustrates that ”Product_Category_1” is the feature with the highest regression coefficient (>0.7). In other words, the products in this category are indispensable and now they are offered at a promotional price.

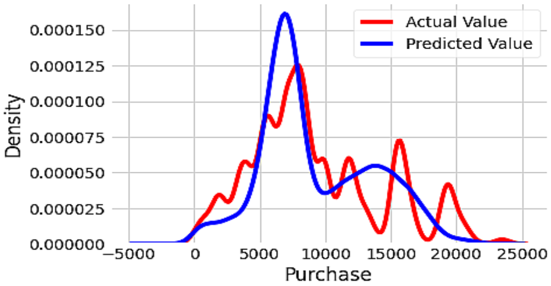


Fig. 13. Graphical visualization of differences between estimated and initial data

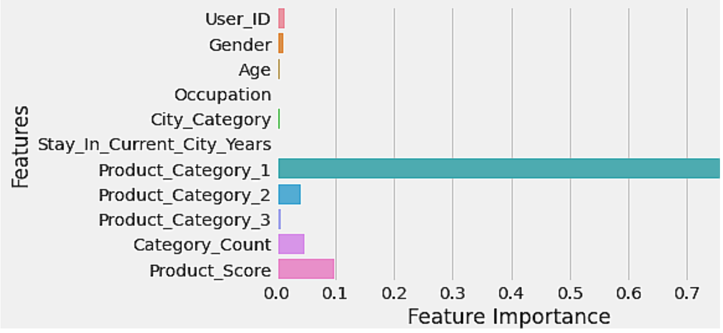


Fig. 14. Importance of independent variables for optimal model training

The model was also validated on the test used on the competition platform and it obtained the public score 2513.77. The first place in the ranking has a score of 2372 at this moment[8] so a difference of only 5.97%.

6. Conclusions

Artificial intelligence technologies consist of expert systems, fuzzy logic artificial, neural networks, machine learning and genetic algorithms [16]. The technology-based workplace scenarios and training concepts, the embedded case studies and experience-based learning in intelligent assistance systems create a suitable basis for working out exemplary needs in Teaching and Learning [17].

Machine learning is used for extracting correlations and insights for effective real-life problem solving. And machine learning models have more excellent properties and have shown better results compared with traditional statistical methods [18].

Other systems developed based on actual technology are:

1. Speech enhancement system using Fischer discriminative dictionary learning FDDL where the speech enhancement is defined as the task of extracting clean speech signal from a noisy mixture [19].

[8] https://datahack.analyticsvidhya.com/contest/black-friday/lb

2. Restful services for orientation and accessibility in a university campus - where the finding the right navigation route in a university campus represents a regular need for various stakeholders who participate to the academic life [20].

In this project, 7 regression algorithms were analyzed and applied to obtain the best model for predicting the purchase price of a product in a store during Black Friday.

In this article we considered the following:

- analysis of the data set, filling the missing values and discarding the columns that are in a weak correlation with the target value;
- extracting the most important features in relation with the products price;
- adding new attributes to improve the accuracy of mathematical models;
- identification of the representative class of buyers (who buys more, who frequents more depending on gender, occupation, marital status, age);
- application of standard regression models and optimized algorithms (XGB, CatBoost, LightGBM);
- choosing appropriate hyperparameters for model tuning using GridSearchCV and RandomizedSearchCV techniques;
- the application of the Cross-validation method for the stability of the models.

To determine the most suitable model the results were compared according to 3 metrics (R_2, RMSE and ROC). The optimized XGB method (we get the optimal values for hyperparameters using the cross-validation technique) recorded the best scores reaching a performance very close to the best solution in the competition (only 5.97% difference). An interesting result observed is that the large number of customers are those who have recently settled in the city, but those who spend more are people who have been residents for 2 years. Also, the 25-35 age group represents a percentage of 40% of buyers and 75.31% of their total number are men. Afterwards, considering the large size of the data set ($0.5 \cdot 10^6$ records) we can opt to build a fully connected neural network with multiple layers to achieve better performance.

REFERENCES

[1] Coursera, «3 Types of Machine Learning You Should Know,» 16 06 2023. [Online]. Available: https://www.coursera.org/articles/types-of-machine-learning.

[2] «Brands Overcome Inventory Issues on Black Friday,» 27 11 2021. [Online]. Available: https://www.businesswire.com/news/home/20211127005197/en/.

[3] *A. Wolfewicz*, «Deep Learning vs. Machine Learning – What's The Difference?,» 15 02 2023. [Online]. Available: https://levity.ai/blog/difference-machine-learning-deep-learning. [Consultato il giorno 13 01 2024].

[4] «Tree Based Algorithms: A Complete Tutorial from Scratch,» 02 03 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/.

[5] *M. Leila*, «Predict customer purchases on Black Friday using deep learning,» 06 08 2018. [Online]. Available: https://towardsdatascience.com/predict-customer-purchases-on-black-friday-using-deep-learning-8f9547c1d18d. [Consultato il giorno 13 01 2024].

[6] *R. Odegua*, «Applied Machine Learning for Supermarket Sales Prediction,» 2020. [Online]. Available: https://www.researchgate.net/profile/Rising-Odegua/publication/338681895_Applied_Machine_Learning_for_Supermarket_Sales_Prediction/links/5e6fd2e6299bf14570f2622d/Applied-Machine-Learning-for-Supermarket-Sales- Prediction.pdf. [Consultato il giorno 13 01 2024].

[7] *S. Ramasubbareddy, T. Srinivas, K. Govinda, E. Swetha*, «Sales analysis on back friday using machine learning techniques,» Intelligent System Design: Proceedings of Intelligent System Design: INDIA 2019, n. Springer Singapore, pp. 313-319, 2021.

[8] *J. Brownlee*, «Linear Regression for Machine Learning,» 15 08 2020. [Online]. Available: https://machinelearningmastery.com/linear-regression-for-machine-learning/.

[9] «What is the k-nearest neighbors algorithm?,» [Online]. Available: https://www.ibm.com/topics/knn.

[10] «Decision Trees,» [Online]. Available: https://scikit-learn.org/stable/modules/tree.html.

[11] S. E. R, «Understand Random Forest Algorithms With Examples,» 15 10 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.

[12] *J. Brownlee*, «XGBoost for Regression,» 07 03 2021. [Online]. Available: https://machinelearningmastery.com/xgboost-for-regression/

[13] *E. Khandelwal*, «Which algorithm takes the crown: Light GBM vs XGBOOST?,» 22 05 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/.

[14] *B. John*, «When to Choose CatBoost Over XGBoost or LightGBM,» 07 08 2023. [Online]. Available: https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm.

[15] *S. Mazzanti*, «You Can Compute ROC Curve Also for Regression Models,» 16 09 2021. [Online]. Available: https://towardsdatascience.com/how-to-calculate-roc-auc-score-for-regression-models-c0be4fdf76bb.

[16] *Ayberk Yilmaz, Hatice Yilmaz Alan, Özlem Faydasiçok, Lidya Amon Susam, Rüya Şamli, Baki Akkuş, Aydın Erol, Ertan Güdekli, Çisem Ilayda Inci, Mehmet Erhan Emirhan*, Deep neural networks as a tool to estimation of cosmic radiation dose received on flight, U.P.B. Sci. Bull., Series A, Vol. 84, Iss. 2, 2022 ISSN 1223-7027, https://www.scientificbulletin.upb.ro/rev_docs_arhiva/fulld17_218825.pdf 2022.

[17] *Elisabeth Lazarou, Cristian Mustata, Cristian Dragomirescu*, Working and learning in industry 4.0 environments, U.P.B. Sci. Bull., Series D, Vol. 81, Iss. 4, 2019 ISSN 1454-2358, https://www.scientificbulletin.upb.ro/rev_docs_arhiva/full9e3_652738.pdf, 2019.

[18] *A Susan (Sixue) Jia, Pengling Yu*, Machine learning approach for identifying relative poverty of urban households: A case study of China, U.P.B. Sci. Bull., Series C, Vol. 85, Iss. 3, 2023 ISSN 2286-3540, https://www.scientificbulletin.upb.ro/rev_docs_arhiva/fulld1a_926594.pdf, 2024.

[19] *Dima Shaheen, Oumayma AL-Dakkak, Mohiedin Wianakh*, Speech enhancement system using fischer discriminative dictionary learning FDDL, U.P.B. Scientific Bulletin., ISSN 2286-3540, Series C, Vol. 80, Iss. 1, https://www.scientificbulletin.upb.ro/rev_docs_arhiva/full2d1_160534.pdf, 2018

[20] *Ioan Damian, Anca Ionita*, Restful services for orientation and accessibility in a university campus, U.P.B. Scientific Bulletin., ISSN 2286-3540, Series C, Vol. 83, Iss. 1, https://www.scientificbulletin.upb.ro/rev_docs_arhiva/full3ad_145718.pdf, 2021.