Sales Forecasting using Machine Learning

Tula Snigdha
Computer Science
University of central Missouri
sxt58770@ucmo.edu

Anitha Anumanthula
Computer Science
University of central Missouri
axa24670@ucmo.edu

Venkat Sai Jagini
Computer Science
University Of Central Missouri
vxj48600@ucmo.edu

Shiva Kandagatla
Computer Science
University Of Central Missouri
sxk52450@ucmo.edu

Abstract:

The Friday following Thanksgiving in the United States marks the start of the busiest shopping period of the season. With the help of machine learning, businesses can accurately predict a consumer's potential spending by analyzing their past purchase behavior. This allows manufacturers to develop more efficient marketing strategies to target specific customer segments based on their past actions and motivations. Over the past fifty years, this one-day celebration has evolved into a month-long sales season, providing international business owners with an ideal opportunity to connect with potential clients and increase visibility through Black Friday deals and promotions. Large online retailers like Amazon and Flipkart offer discounts on a variety of products, including clothes, electronics, cookware, and home decor. Researchers have conducted studies to forecast consumer spending using three analytical strategies. The Black Friday Sales Database on Kaggle is being analyzed and predicted using extreme gradient boosted trees algorithm, providing a practical solution to a compelling problem in the business market.

Keywords: Black Friday, Regression Algorithm, Random Forest, XGBoost, Mean Square Error, RMSE, R2_score, Feature Engineering, Machine Learning.

I. INTRODUCTION

The rise of the Internet has transformed the retail industry, with online shopping becoming a preferred method for many due to its accessibility, lower costs, greater variety, and convenience. The COVID-19 pandemic has further increased the trend of online sales, with Black Friday becoming a key event for retailers to boost their revenue. Marketers utilize data analysis to target specific customer groups and understand their purchasing behavior and interests, allowing for the creation of personalized marketing campaigns. Ethnicity can also be a factor in offering discounts. Data science plays a crucial role in maintaining the reputation of e-commerce businesses, including identifying fraudulent activity and predicting costs and trends. Black Friday and Cyber Monday are popular discount events that drive consumer enthusiasm for purchasing. Researchers have found that combining feature extraction with hyperparameter adjustment can improve the efficacy of boosting and bagging algorithms in predicting consumer spending behavior. The Random Forest Regressor method was found to have the lowest Mean Squared Error value and be the most effective in predicting consumer spending behavior. With the help of data science and machine learning techniques, e-commerce businesses can personalize the shopping experience for their customers. By analyzing customer behavior and preferences, businesses can provide personalized product recommendations, offer tailored discounts and promotions, and improve overall customer satisfaction. Social media platforms have become an integral part of e-commerce, allowing businesses to connect with customers, build brand awareness, and drive sales. Platforms like Instagram and Facebook have introduced shopping features that allow users to purchase products directly from their feeds, making it easier for businesses to sell their products and reach new customers. Retailers are experimenting with emerging technologies such as augmented reality, virtual reality, and chatbots to enhance the shopping experience. These technologies can help customers visualize products, get personalized recommendations, and receive support from virtual assistants. Consumers are becoming more environmentally conscious, and retailers are responding by adopting sustainable practices. This includes using eco-friendly materials, reducing waste, and implementing sustainable supply chain practices. Sustainable practices not only benefit the environment but can also attract environmentally conscious consumers. With the rise of e-commerce, retailers are adopting an omni-channel approach to meet customer needs. This involves integrating online and offline channels to provide a seamless shopping experience. For example, customers can order products online and pick them up in-store or return products bought online in-store. Mobile devices are increasingly becoming the primary way that customers interact with e-commerce businesses. Mobile commerce allows customers to shop on-the-go and makes it easier for businesses to reach customers anytime, anywhere. Overall, the retail industry is constantly evolving, and e-commerce is playing a major role in this transformation. By leveraging data science and emerging technologies, businesses can improve the shopping experience for customers, increase revenue, and stay competitive in a rapidly changing market.

II.MOTIVATION

Understanding consumer behavior and catering to their needs is critical for companies to achieve profits and enhance consumer experience across all industries. To evaluate a person's purchase history, factors such as age, gender, occupation, and location need to be considered. Younger people tend to spend more on shopping than older ones, while individuals with higher incomes tend to spend more than those with lower incomes. Moreover, people living in urban areas tend to spend more on shopping than those in rural areas. Segmentation models can help us understand and forecast customer behavior, thereby enabling companies to anticipate their needs and improve transportation and delivery reliability, leading to increased revenue. Additionally, by using machine learning algorithms, companies can achieve higher accuracy rates in forecasting consumer behavior and provide sustainable product choices to customers.

Main contribution and Objectives:

- Assisting the retail shops to fix a price for the products so that they can earn profits.
- Revealing and recognizing the crucial points from the dataset along with age, gender.
- Establishing a quantitative effect of elements that are selected.
- how they influence a customer's buying decision by getting to know them personally.
- The Retailers can stock up the inventory of the particular product category by the sales prediction done by the algorithm so there wouldn’t be any downtime in the business.

III.RELATED WORK

When analyzing Black Friday sales, advertising is a crucial factor to consider as it plays a significant role in attracting customers and driving sales. Various machine learning methods have been employed to forecast Black Friday sales revenue, including linear regression, decision tree, random forest, gradient boosting, and deep learning. Researchers have used different datasets such as the Black Friday Sales Dataset from analytics Vidhya, Data Science Nigeria, and Kaggle to train and test their models. Evaluation metrics like Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) have been used to assess the accuracy of these models. Some models have outperformed others, such as the Random Forest algorithm achieving a precision of 93.53% and an RMSE score of 2730. Additionally, researchers have compared the effectiveness of hold-out validation and K-Fold cross-validation, with the latter showing better accuracy results for some methods. These machine learning techniques can help companies forecast sales, identify consumer expectations, and develop effective advertising strategies to boost revenue. Aaditi Narkhede et al. [7], these results of variations here constitute an efficient method for information filtering with decision-making. has been using machine learning algorithm in forecasting sales in locations such building structural Big Mart can forecast consumer expectations but also oversee its administration for inventories appropriately. innovative technologies to determine customer expectations more accurately as well as create more effective strategies to boost revenue.

M.Sahaya Vennila et al. [8] has digested, implemented, & evaluated machine learning methods that forecast revenue. This Black Friday Sales Dataset from Kaggle is indeed the information being used testing and evaluation. Initial preprocessing of the information. This information is trained and tested datasets using the K-Fold algorithm. Linear Regression, Decision Tree, Random Forest, plus Gradient Boost are all used to create the forecasting models. The reliability measurement methods employed include Mean Absolute Error (MAE) but also Root Mean Squared Error (RMSE). The Random Forest considerably outperformed other models in the trial, achieving an efficiency of 77%, an RMSE score of 2730, and an MAE value of 2349.

S. Yadav et al [4] has examined & contrasted overall effectiveness of both the hold-out validation as well as K-Fold cross-validation. this same outcome from experiments showing that kfold cross-validation produces better high accuracy. For a certain collection of methods.

Consumer Expectations:

Since most customers wait all year for such discounts, needs of customers regarding Black Friday have increased. The stores' inventory is where the primary issue arises. Because of the large number of purchases made during such deals, items frequently sell in a short period of time, leaving customers disappointed but also ruining your entire workday. At this point many customers' aspirations were salvaged either by innovative internet shopping.

Customers don't need to stand in massive queues ahead of establishments or endure the arduous task from in purchasing on Black Friday because customers can purchase the exact goods for the exact huge discount online. This latest epidemic didn't adequately derail customers' needs because merchants have e - retailing websites that allowed individuals buy products but have them transported while still locations remained shut.

Promotions:

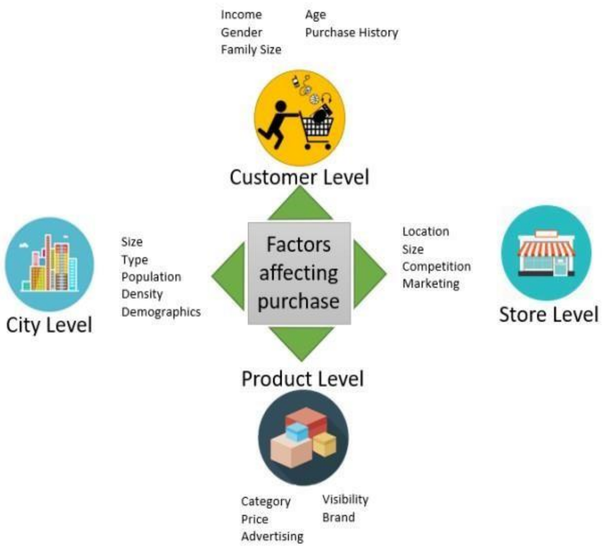
Another of the important factors to consider while analyzing your Black Friday deals is advertising. Its popularity of Black Friday sales in the USA is primarily due to such deals. Customers are targeted with appealing advertising during Black Friday sales; as just a result, they risk ending up acquiring the items when there's no real a need them, such as in case of TV as well as mobile phone updates. Regarding the client, expectations, and the importance in advertising on Black Friday, you may anticipate that now most customers anticipate purchasing specified items or, but at very least, will searching after items marketed during an alluring discount due to BF's concept.

IV.PROPOSED FRAMEWORK

This database we are dealing with employed quantitative information for data exploration.

Data exploration employs a graphic approach to evaluate large datasets and emphasize certain key features. Data analysis, including the iris database and scoreboard info, are involved in empirical information. Python, pandas, matplotlib, NumPy array, and seaborn seem to be the applications was using to possible methods.

Given the extensive range of modules it contains and the simplicity with which optimization techniques may be implemented, Python is a widely utilized language inside the data science industry. Becoming an interpreter, it really has gained notoriety when processing big files. Python excels at data munging. Pandas is indeed an incorporated indexed, quick but also efficient Data Frame object enabling data manipulation.
This supports text and CSV files.



1.Linear regression:

Linear regression is a machine learning algorithm based on supervised learning. Another example of a prediction model is one in which the predicted output is continuous. Utilizing a specific independent variable, linear regression makes predictions for the dependent variable (y)
(x). The analysis reveals that the variables have a linear relationship.

y = Mx + C is the equation of a line which is fitted to the data which is the linear regression function. Therefore, x seems to be the input variable, y is just the estimated value, and M, is the slope which signifies the estimated change in y for a 1-unit increase of x. C can be defined as the Y- intercept which means its value is equal to value of y when x is 0. The purpose of this method would be to compute as well as identify the lines that fit the predicted value as well as the independent variable very most.

2.Decision tree:

Decision Tree approach is used to build the regression or classification models, which creates a tree-like architecture. Merely divided up into smaller parts, the dataset. Depending on the characteristics of a specific characteristic, the dividing node inside a DT includes a data element along both branches. The controlling instructions as well as values serve as the foundation for diverging. In selecting the root node, a feature extraction method is crucial.

In decision trees, Information Gain is the splitter characteristic which is determined by the quantity of data needed to explain the trees. We also calculate the entropy to decide the root node at every branch. Gain Ratio is the instance selection metric that chooses characteristics with a lot of information.

3.XGBoost Learning Model:

Several of the strongest together some effective ways to apply its Gradient Boosted Trees algorithm to everyone supervised learning problems called XGBoost. The ML technique XGBoost is also quite successful, resulting in it being frequently utilized during contests & contests. In comparison to certain other gradient boosting approaches, XGBoost is nearly ten times quicker and offers strong predictive ability. Additionally, this contains a range of regularizations that lessen fitting problem & enhance efficiency. It really is focused on function optimization by utilizing suitable regularization techniques and improving algorithms. Let's specify certain footnotes as well as parameters prior to analyzing algorithms as well as normalization procedures.

Random Forest algorithm:

Given various samples,
it constructs decision trees but also uses its mean both classification but also overwhelming opinion during prediction. The Random Forest Algorithm's ability to deal with large datasets with both dependent variables, just like in regression, plus predictor variable, as seen in classifications, is among its most crucial qualities. In terms of categorization issues, this delivers superior outcomes. As just an approximate solution, Random Forest was possible to perform between the classification and regression problems. The Random Forest was built just on philosophy that integrating numerous DTs even though opposed to relying solely across one DT.

Figure lists some characteristics which have an impact here on Random Forest Regressor.

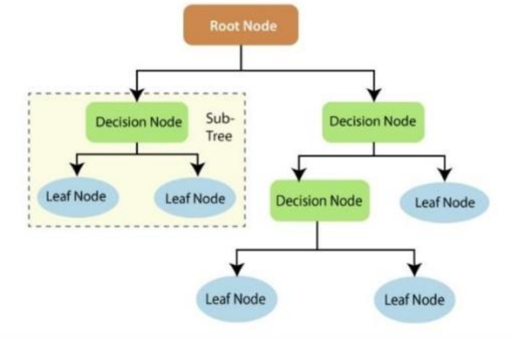
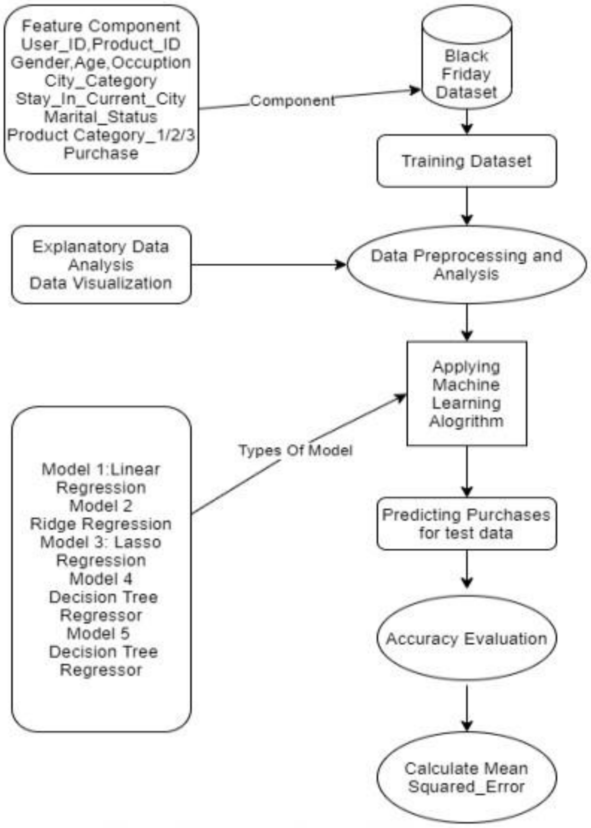


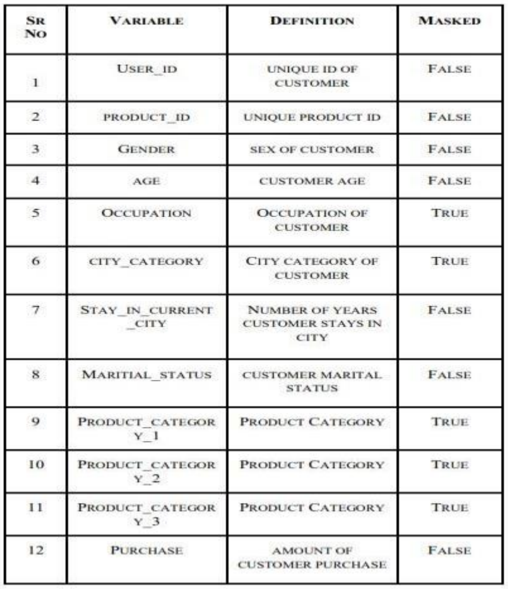
Fig 3

Another fundamental distinction seen between random forest method and the decision tree algorithm is that neither arbitrarily selects its root nodes then groups those nodes. For produce the necessary forecast, the random forest uses the bagging approach.

Data Description:

Delivering specific estimates about occurrences outside of those included in the training set seems to be the core aim of machine learning models. We can split a piece of something like the information over that we currently have the solution as just a stand-in again for unknown information to measure the accuracy of algorithms' forecasts having respect to that information.


Feature Component User_ID,Product_ID Gender,Age,Occuption City_Category Stay_In_Current_City Marital_Status Product Category_1/2/3 Purchase
Component
Black Friday Dataset
Training Dataset
Explanatory Data Analysis Data Visualization
Data Preprocessing and Analysis
Applying Machine Learning Alogrithm
Types Of Model
Model 1:Linear Regression Model 2 Ridge Regression Model 3: Lasso Regression Model 4 Decision Tree Regressor Model 5 Decision Tree Regressor
Predicting Purchases for test data
Accuracy Evaluation
Calculate Mean Squared_Error



SR NO	VARIABLE	DEFINITION	MASKED
1	USER_ID	UNIQUE ID OF CUSTOMER	FALSE
2	PRODUCT_ID	UNIQUE PRODUCT ID	FALSE
3	GENDER	SEX OF CUSTOMER	FALSE
4	AGE	CUSTOMER AGE	FALSE
5	OCCUPATION	OCCUPATION OF CUSTOMER	TRUE
6	CITY_CATEGORY	CITY CATEGORY OF CUSTOMER	TRUE
7	STAY_IN_CURRENT_CITY	NUMBER OF YEARS CUSTOMER STAYS IN CITY	FALSE
8	MARITIAL_STATUS	CUSTOMER MARITAL STATUS	FALSE
9	PRODUCT_CATEGORY_1	PRODUCT CATEGORY	TRUE
10	PRODUCT_CATEGORY_2	PRODUCT CATEGORY	TRUE
11	PRODUCT_CATEGORY_3	PRODUCT CATEGORY	TRUE
12	PURCHASE	AMOUNT OF CUSTOMER PURCHASE	FALSE

Table 1

Next, we assess how accurately the analysis shows actual information. Observational are usually included in trained data to customize teaching methods but instead adjust hyperparameters. Selections of information from the testing dataset are utilized to objectively assess how well the learning approach performed just on training as well as to forecast how much each client will spend during the Black Friday sales. Companies will be able to study and tailor offerings for even examples. Variables in the database include user

occupation, etc. Table 1 includes information about the database construction. The Black Friday Sales dataset is often used to training a variety of machine learning techniques more consumers' favorite items using the purchasing forecast provided. This predictor variable is going to become the purchasing indicator. The Purchasing Indicator would forecast how much a throughout other environments via pip. In the initial

V.RESULT

consumer would spend during in the Black Friday

discounts. With order to develop machine

The following table provides a comparative of the MSE values including all methods. As shown by Table 2, it is obvious that using XGBoost performed best than that of the Random Forest and decision tree regressor machine learning algorithms. The Random Forest Regressor's MSE rate is 2879.5, making it so much more appropriate when it comes to the classification algorithm. However, algorithms that utilize ensemble learning work effectively with these data type. We can all concur with running algorithms upon smaller segments of both the information has produced good outcomes, using the effectiveness using random forest as just an instance, like the bagging-based strategy. Off an estimated RMSE of 4611 SVM to that of 2911 by XGBoost in the testing dataset, for example, the 70 | 30 division strategy, has shown a noticeable difference. All experimental findings result showed comparable benefits.

Nevertheless, if we aggregate several poor trainees in bagging, that simulation would have been poor. This is how the boosting concept enters the equation because

learning algorithms like Random Tree Regressor, and XGBoost.

This outcome is displayed with said amounts representing the root mean square error whenever analyzed quantitatively sampling database. This roots for mean squared error, which could be derived from sklearn.metrics, was employed to find RMSE (Root Mean Square Error). Sklearn metrics has already been deployed in Anaconda

three situations, existing Analytics Vidhya training dataset is utilized as its specimens collected, while fresh learning and testing databases then were produced based on the conditions. n such situations, the user id, product id, and buy elements are added to the test database, and even the purchasing feature being

every learner tries to lessen the mistakes of both the ones before them

subsequently removed from the testing set.

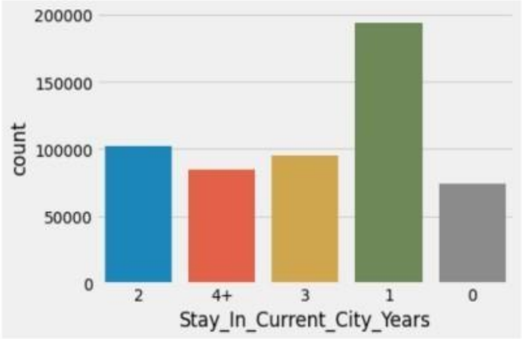
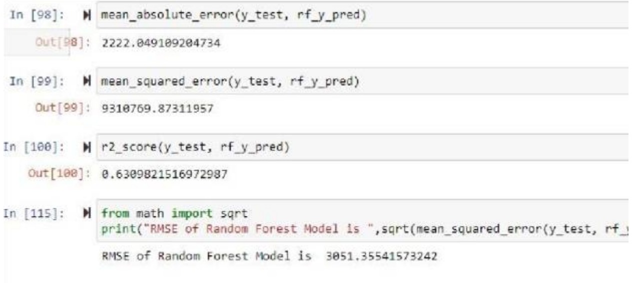


Fig 4

Above figure show the number of people lived for the respective year period in the city.

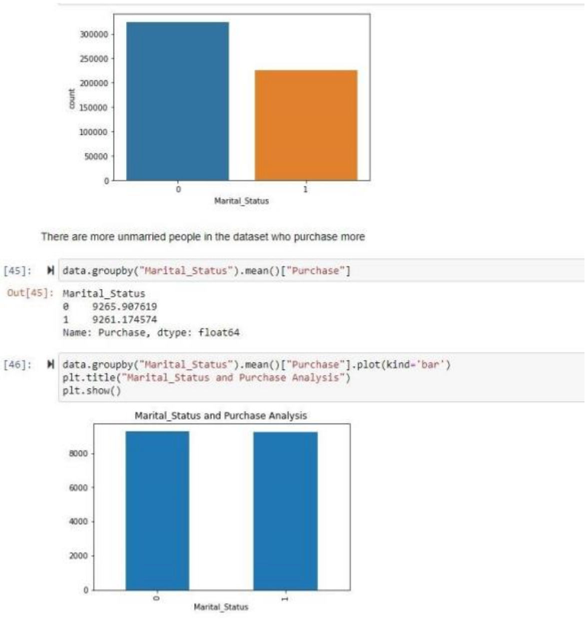


```
In [98]:  mean_absolute_error(y_test, rf_y_pred)
Out[98]:  2222.049109204734

In [99]:  mean_squared_error(y_test, rf_y_pred)
Out[99]:  9310769.87311957

In [100]: r2_score(y_test, rf_y_pred)
Out[100]: 0.6309821516972987

In [115]: from math import sqrt
          print("RMSE of Random Forest Model is ",sqrt(mean_squared_error(y_test, rf_

          RMSE of Random Forest Model is  3051.35541573242
```

Fig 5

Above RMSE value for the Random Forest

Model is 3051.35. In fig 6, that there are a greater number of unmarried people in the dataset who purchased more rather than married.



There are more unmarried people in the dataset who purchase more

```
[45]:    data.groupby("Marital_Status").mean()["Purchase"]
Out[45]: Marital_Status
         0    9265.907619
         1    9261.174574
         Name: Purchase, dtype: float64

[46]:    data.groupby("Marital_Status").mean()["Purchase"].plot(kind='bar')
         plt.title("Marital_Status and Purchase Analysis")
         plt.show()
```

Fig 6

References:

[1] C. M. Wu, P. Patil and S. Gunaseelan, "Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 16-20, doi: 10.1109/ICSESS.2018.8663760.

[2] Odegua, Rising. (2020). Applied Machine Learning for Supermarket Sales Prediction.

[3] Purvika Bajaj1, Renesa Ray2, Shivani Shedge3, Shravani Vidhate4, Prof. Dr. Nikhilkumar Shardoor5,"SALES PREDICTION USING MACHINE LEARNING ALGORITHMS",International Research Journal of Engineering and Technology (IRJET) ,Vol 7 Issue 6,2020,eISSN: 2395-0056 p-ISSN: 23950072.

[4] Ramasubbareddy S., Srinivas T.A.S., Govinda K., Swetha E. (2021) Sales Analysis on Back Friday Using Machine Learning Techniques. In: Satapathy S., Bhateja V., Janakiramaiah B., Chen YW. (eds) Intelligent System Design. Advances in Intelligent Systems and Computing, vol 1171. Springer, Singapore. [5]Potturi, Keerthan, "Black Friday A study of consumer behavior and sales predictions" (2021).
Creative Components.

[6] Swilley, Esther, and Ronald E. Goldsmith. "Black Friday and Cyber Monday: Understanding consumer intentions on two major shopping days." Journal of retailing and consumer services, vol. 20,1,2013, pp.43-50.

[7] Fischer, Eileen, and Stephen J. Arnold. "Sex, gender identity, gender role attitudes, and consumer behavior." Psychology & Marketing, vol.11, 2, 1994, pp.163-182.

[8] Song, Ji Hee, and Jason Q. Zhang. "Why do people shop online?: Exploring the quality of online shopping experience." American Marketing Association. Conference Proceedings. 2004.

[9] Vijayasarathy, Leo R. "Predicting consumer intentions to use on-line shopping: the case for an augmented technology acceptance model." Information & management, vol. 41,6, 2004, pp.747-762.

[10] Simpson, Linda, et al. "An analysis of consumer behavior on Black Friday." American International Journal of Contemporary Research, 2011.

[11] Bellizzi, Joseph A., and Robert E. Hite. "Environmental color, consumer feelings, and purchase likelihood." Psychology & marketing, vol.9, 5, 1992, pp.347-363

[12] Tashakkori, Abbas, Charles Teddlie, and Charles B. Teddlie. Mixed methodology: Combining qualitative and quantitative approaches. Vol. 46. Sage, 1998.

[13] Ma, Xiaogang. "Linked Geoscience Data in practice: Where W3C standards meet domain knowledge, data visualization and OGC standards." Earth Science Informatics,vol. 10, 4,2017, pp. 429-441.

[14] Badie, Bertrand, Dirk Berg-Schlosser, and Leonardo Morlino, eds. international encyclopedia of political science. Vol. 1. Sage, 2011.

[15] Hammersley, Martyn. "The relationship between qualitative and quantitative research: paradigm loyalty versus methodological eclecticism." ,2002, pp. 159-174.

[16] Mladenoff, David J., et al. "A regional landscape analysis and prediction of favorable gray wolf habitat in the northern Great Lakes region." Conservation Biology, vol. 9,2 ,1995, pp. 279-294

[17] Donovan, Robert J., et al. "Store atmosphere and purchasing behavior." Journal of retailing, vol.70, 3, 1994, pp. 283-294.