Analysis of Bank Marketing Campaign, Predicting Outcomes of Customer Purchase Intention and Household Electricity Consumption using Machine Learning

Kenechukwu Otito Ajufo
School of Computing
National College of Ireland
Dublin, Ireland
x19190174

***Abstract*—Machine Learning is a prominent field of computer science, it can easily identify patterns and trends in data, requires little human interference, handles large volume of data (big data), plus it can be used in various areas of human life. In this project, attempts are made to use machine learning algorithms explore three diverse data to gain insights to them. The first dataset is related to a bank marketing campaign, there Logistic Regression and Decision Tree models to find the feature that make customer opt for a term deposit. The second is on purchase intention of an ecommerce website. Naïve Bayes Classifier and Decision Trees are employed to determine the important feature which lead to revenue generation on the ecommerce website. Finally, Random Forest and Multiple Regression are used to find weather factors and temperature interconnected with energy consumption in a house. The results from the analysis showed that the Decision Tree performed best for both the bank marketing and online purchase intention datasets while the Random Forest Regressor was the better algorithm for Household energy consumption dataset.**

***Index Terms*—machine learning, logistic regression, decision tree, naïve bayes, random forest, multiple linear regression**

I. Introduction

There is a growing need for data mining and machine learning. This demand is due to enormous amount of existing data, and data we produce daily. On their own, patterns in the data can be inconspicuous and hard to deduce insights. Applying data mining and machine learning gives us the ability to transform this raw data into meaningful information, thereby enabling us to discover correlations to make informed assumptions. This project targets some niche areas where data mining and machine learning can be applied.

A. Analysis of Bank Marketing Campaign.

The Banking sector is a critical part of any country's economy. Marketing is extremely necessary for banks; the overall goal of any marketing campaign is to attract new customers and retain existing customers. Banks will try to use marketing campaigns in more efficient and appropriate ways because it is exhaustive, costly and time consuming to try to reach everyone. Data driven processes are key to designing personalized campaigns that target considerably less amount of people who would likely go for the product or service thus enabling the brand to optimize its marketing exposure

This project will try to find what personal features influences potential subscriber opt for a term deposit.

B. Predicting Outcomes of Customer Purchase Intentions.

Purchase intention can be defined as a customer's desire to get a specific good or service. Nowadays, customers have an array of options and variety of ways to purchase a good or service this make purchase intention an influential tool. Online platforms continue to gain acceptance making them substitutes to traditional methods of shopping. In contrast to traditional methods of shopping, a customer's activity online can be measured exactly and instantly; this makes consumer behavior and ultimately purchase intentions on these platforms crucial to understand, develop strategies which can be applied to improve returns.

This project will attempt to use machine learning methods to uncover what session components affect revenue to be able to make inference on the factors that determine customers purchase intentions.

C. Predicting Outcome of Household Electricity Consumption.

Energy is a fundamental part of everyday human activity. It performs a significant function in the development of numerous countries. It aids economies flourish and overall promotes better quality of life as electricity is needed for almost every sector of the economy, it powers businesses and factories, transportation, communication and at a the very basic level power appliances in our homes. Accessible information on energy consumption can help improve demand forecasting which leads to cost-effectiveness of energy for consumers, reduction of carbon emissions and energy waste, thereby boosting energy efficiency.

For this reason, the intention is to discover how weather, temperature and time are interconnected with the electricity

usage of a house. The goal here is to observe if these factors have a significant correlation with electricity usage in the house in order to use the same factors to predict electricity usage in the house.

D. Structure of the Paper.

The following describes how the report is structured, Part II discusses related works: these include analysis and review of the literature available to discover the types of machine learning techniques applied. Part III will describe the methodology used to carried out all the study. Part IV addresses the evaluation of how the machine learning algorithms have been conducted to achieve this project's objective. And at last, Part V will be the conclusion.

II. Literature Survey

The section reviews the available literature in the fields of the selected datasets. Machine learning has been applied to almost every aspect with promising results.

A. Analysis of Bank Marketing Campaign.

Banks and many other financial services-oriented companies use machine learning to perform many reasons. They can range from fraud detection and prevention, to predicting which customer is likely going to subscribe to an upcoming service or generally to just gain insights on developing investment opportunities.

Moro, Laureano and Cortez implemented the Cross-industry standard process for data mining (CRISP-DM) methodology with the intention of using data mining to explore three models: Naïve Bayes, Decision Tree and Support Vector Machines classifiers which explain the success of contact [1]. At the end of the first and second iterations of CRISP-DM processes, the authors were only able to fit the Naïve Bayes model, this is due to the large amount of viable output. However, the third iteration of the process had them refine their data understanding phase. Only half of the initial 58 attributes were used to model the data, the analysis produced results showing that the duration was the crucial attribute in determining customer's opting for a term deposit with Support Vector Machines (SVM) classifiers being the best model with an area under the curve of 0.938.

In [2], Elsalamony did a comparative analysis using four classification techniques: Multilayer perception neural network (MLPNN), Bayesian networks, Logistic regression, and Ross Quinlan new decision tree model (C5.0) with the aim of assessing the performance of these algorithms to identify the characteristics which persuaded customers to subscribe to a term deposit. He used three measures: accuracy, sensitivity and specificity to judge the performance of these algorithms and concluded that the C5.0 model produced slightly better results than the remaining. He also reported that duration was the important attribute.

Neuro-fuzzy systems was used in [3] to predict the outcome of bank marketing campaigns, the method was employed because fuzzy rules can be easily understood by humans. Neuro-fuzzy systems is a learning algorithm that derived from Neural Network theory, it determines fuzzy rules by processing the data. The process yielded a prediction rate.

In [4], Naïve Bayes and the C4.5 decision tree models were fitted to predict whether a client will subscribe a term deposit with the comparative performance of the two algorithms. The C4.5 model performed better.

Chi-Square selection feature method and a supervised machine learning technique: Naive Bayes were used in [5]. The result showed to that the attributes in the dataset had to be reduced to improve performance.

Finally, in [6] the authors built Naïve Bayes, KNN and SVM models to classify bank data. They sort to find customers who were presumably respond positively their bank marketing. By comparing the overall success of the models, they reported the Naïve Bayes fit the data most appropriately.

After a thorough review of related literature and examining the research question, Logistic Regression and Decision trees models were picked because it can help classify the data and thereby answering the question.

B. Predicting Outcomes of Customer Purchase Intentions.

Online stores use data mining and machine learning techniques to acquire data, process it to understand behaviour their customers. The main aim is to try to make every customer's experience tailored to them, therefore we get recommendations of products and services based on previous purchases. Breakthroughs in machine learning have impacted heavily on market campaign strategies.

There are few studies using machine learning to predict outcome of customer purchase intentions online. However, machine learning has been applied to predict customer purchase in physical stores. In [7] Naïve Bayes classifiers, Support Vector Machine (SVM), linear discriminant analysis and logistic regression were implemented on RFID data gotten from the path individuals took in a Japanese supermarket to predict the shopper purchase intention. The result revealed that SVM showed better performance at forecasting and determined the time customer spent in the supermarket helped understand customer behaviour.

The purpose in [8] was to predict online consumer repurchase intentions, to achieve this purpose a machine learning approach was employed, using techniques like Random Forest, AdaBoost, Decision trees (C5.0), support vector machine and neural network. It starts by identifying the characteristics for repurchase intentions and then proceeds to fit models. The result show that AdaBoost surpasses the other methods in performance.

Reference [9] attempts to build a classification model to predict whether a user will be interested in buying a products that are placed in the shopping cart online using statistical and machine learning algorithms like C4.5 decision trees, random forest, random tree and Multilayer Perceptron. After the algorithms with the best predictive precision selected, seven in number, and then transferred to a voting algorithm.

This essentially improved the final model making it better than the individual models passed to the voting algorithm.

[10] considers the use of an item's popularity and details from users visits to infer purchase intentions of anonymous users. They applied these models to two real world datasets from two ecommerce retail companies. They used different classifier and boosting methods and propose that their model spots online purchasing behaviour signals which is then used for prediction of online purchases.

[11] addresses the significance of shopping intentions online, goes on to classify shopping intentions using machine learning method to classify using clickstreams. Furthermore, it discusses limitations like the model been site dependent and consideration of purchasing power of the customers.

Finally, to analyse the intention of the user to purchase in social commerce was the sole aim of [12]. Their idea of that exploring customer's desire to buy in social commerce is relevant because of the significant position the consumers play in marketing. A sentiment analysis is performed as it is seen as a way of extracting consumer opinions then Naïve Bayes and Support Vector Machine classifier models were fitted.

After the overall review of these related work, to answer the proposed research question, the decision to use machine learning classification techniques is validated.

C. Predicting Outcome of Household Electricity Consumption.

Forecasting energy consumption studies have been conducted extensive, therefore various perspectives to how to predict household consumption exist, the work of [13] produces confirmation that the use of machine learning methods by the construction industry is valuable. It made used of Artificial Neutral Networks to predict a building usage and compared this with predictions based on traditional physics calculations.

Weather's influence on energy consumption was analysed in [14], it proposed an optimum regression approach by the combining regression algorithms like Artificial Neural Network (ANN) and Support Vector Machine (SVM). The authors report that the combination of the two techniques yielded better results than the single ANN or SVM.

Random Forest algorithm was utilized to predict the hourly usage of electricity in two residential buildings in [15]. This was then compared with Regression trees and Support Vector Regression to discover which performed better. The authors report that the Random Forest perform better and is not particularly sensitive to the number of variables.

Random Forest was also used in [16], the authors try to study the effect of eight input variables on the heating and cooling load of some residential buildings. They showed that Random Forest can provide more reliable forecasts of heating and cooling loads compared to linear regression approaches.

In [17], the study comprised of the application of linear regression and nonlinear regression (Support Vector Machine) to investigate the relationship between resident's behaviours and their energy use. The authors stated that the linear model produces better results. They developed a web application to enable users to receive feedback. The authors in [18] also tried to find the model with the highest accuracy in predicting total energy consumption using linear regression, support vector regression (SVR) and Random Forest. They reported the SVR has the best accuracy in predicting consumption at the end of the month while Random Forest was better predicting the next day's consumption.

Reference [19] the researchers proposed models that can forecast power consumption using domestic buildings as case point. To carry out evaluation of the models, they used R square and Mean Square Error (MSE).

After examining the available literature and with the objective in mind, the decision to use regression analysis was made.

III. Data Mining Methodology

Knowledge Discovery in Database (KDD) is an iterative process to find useful knowledge from data, the process involves steps which follow a sequentially order, it also permits going back to steps in the process. This methodology is applied to all three datasets.

A. Analysis of Bank Marketing Campaign.

Logistic Regression and Decision trees models are fitted. These two supervised machine learning techniques were selected to try to classify the features that influence the target variable. The KDD methodology used is outlined in the following steps.

1) Data Selection: The data set used for the Bank Marketing campaign was retrieved from the UCI Machine Learning Repository.[1] It contains information from the telephone marketing efforts of an unnamed Portuguese bank whilst trying to make customers subscribe to a term deposit. It contains 17 columns and 45,211 instances.

2) Data Preprocessing: This stage involves the processing the selected data, cleaning it in order to get consistent data. The data set was scrutinized for missing values. There were no values missing in the entire dataset.

3) Data Transformation: This stage the data suitable to carry out the Logistic regression and Decision Trees are chosen. However, in this analysis all the attributes in the dataset is used to perform the machine learning algorithms on.

4) Data Mining: Before any algorithms can be applied, the data set is split into train and test subset. The train data subset is 80 percent of the total dataset. The models will be applied to the train data and then used to predict the test data. The summary of the Logistic Regression and Decision Tree models are shown in tables I and II. In addition, the decision tree diagram is shown in figure 1.

[1] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

TABLE I
LOGISTIC REGRESSION MODEL RESIDUALS

Deviance Residuals				
Min	1Q	Median	3Q	Max
- 6.0667	-0.2610	-0.1775	-0.1269	3.5846

TABLE II
DECISION TREE MODEL SUMMARY – BANK MARKETING

Decision Tree Model Summary			
Variables Used in Tree	*Number of Nodes*	*Residual Mean Deviance*	*Misclassification Error Rate*
"Duration', 'poutcome', 'month', 'contact'	9	0.4884	0.1098

TABLE III
DECISION TREE MODEL SUMMARY – ONLINE INTENTION

Decision Tree Model Summary			
Variables Used in Tree	*Number of Nodes*	*Residual Mean Deviance*	*Misclassification Error Rate*
"PageValues", "Month", "BounceRates"	6	0.4808	0.1007

TABLE IV
NAÏVE BAYES CLASSIFIER SUMMARY

A-Priori Probabilities	
FALSE	TRUE
0.8452555	0.1547445

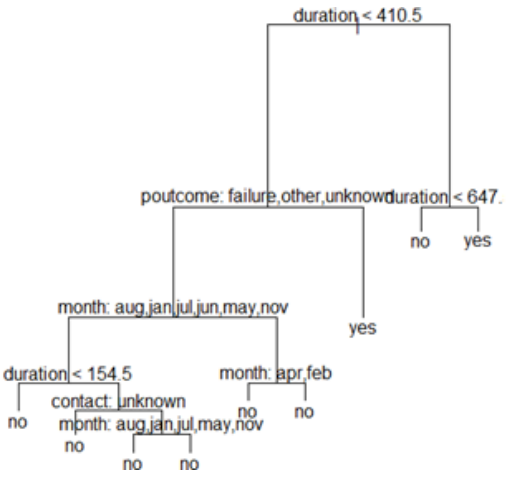

Fig. 1. Decision Tree for Bank Marketing Dataset

B. Predicting Outcomes of Customer Purchase Intentions.

The aim here is to predict the outcomes of customer purchase intentions and to achieve this goal two classification machine learning algorithms model were generated. Naïve Bayes and Decision Trees were used. The process followed the steps described below

1) Data Selection: The data set used was obtained from the UCI Machine Learning Repository.[2] It has information on sessions of various customers on an ecommerce website. It also contains whether the session lead to shopping and comprises 18 attributes and 12,330 instances.

2) Data Preprocessing: After loading the dataset, it was checked for inconsistency, there were no missing values in the dataset.

3) Data Transformation: All the attributes are essential to find the factors have the most influence on revenue generation. The only transformation done on the data in this stage was changing the target variable into a factor.

4) Data Mining: The dataset is split into train and test data. The train data is applied to the model and then the test data is used as a way of knowing if the algorithms are generalizable. The summary for the Decision Tree Model and Naive Bayes Classifiers can be viewed in Tables **??** and **??**. The A-Priori probabilities show the frequency of each attribute occurring in the training data. Figure 2 displays the Decision Tree model diagram.

C. Predicting Outcome of Household Electricity Consumption.

The objective with this dataset was to find how factors like weather, temperature and time are interconnected with the electricity usage of the home. With that in mind Multiple Regression and Random Forest models were fitted to attain the objective. The KDD process is as follows.

1) Data Selection: The data set used was gotten from Kaggle.[3] The dataset contains reading over the span of one minute of household appliances in kilowatts and the weather

[2]archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset [3]www.kaggle.com/taranvee/smart-home-dataset-with-weather-information

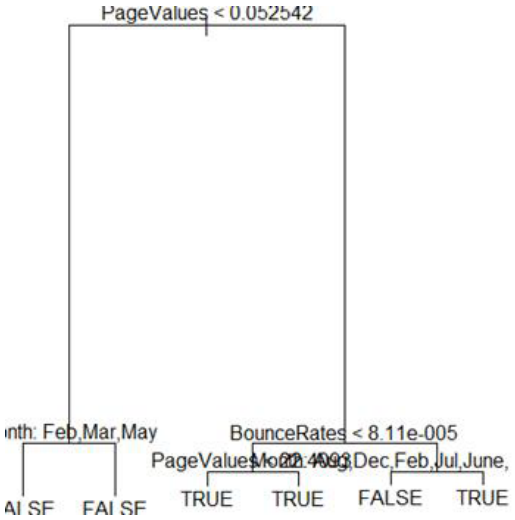

Fig. 2. Decision Tree for Online Intention Dataset

TABLE V
MULTIPLE REGRESSION COEFFICIENT TABLE

Coefficient	Value
(Intercept)	-2.6694
temperature	0.033914
humidity	0.091396
visibility	-0.000064
apparentTemperature	-0.030790
pressure	0.003123
windSpeed	-0.008618
windBearing	0.000140
precipIntensity	7.106847
dewPoint	0.002425
precipProbability	-0.408001

conditions for the house. These readings were recorded by a smart home meter. It has 32 columns and 503,911 instances.

2) Data Preprocessing: The data set was checked for missing values after loading it. There was one row in the data set that had missing values, I dealt with it by dropping the row as it did not count for 1% and hence cannot influence the data.

3) Data Transformation: To make the columns in the dataset easier to access, the column names were changed. Since the algorithms proposed were Regression and not classification, the categorical variables not needed for the regression analysis were dropped as well.

4) Data Mining: The dataset is split into train and test data. Due to the size of the dataset, I was not able to fit the models in R. Eventually, Python was used to fit the models. The equation for the Multiple Linear Regression is shown in table V

IV. Evaluation and Results

The evaluation methods are applied to the performance of the machine learning methods in the section above. The more accurate model is, the better inferences we can get. For the Classification tasks, Accuracy is used to while for the Regression R squared values are used.

A. Analysis of Bank Marketing Campaign.

Accuracy is used to check the performance of the models here; it is the ratio of correct predictions the model made to the total number of instances in the sample. Tables VI, **??** and VIII contain the evaluation results.

TABLE VI
LOGISITIC REGRESSION EVALUATION SUMMARY

Accuracy	Misclassification Error
0	1

TABLE VII
DECISION TREE MODEL CONFUSION MATRIX

	Actual No	Actual Yes
Predicted No	7548	549
Predicted Yes	445	501

TABLE VIII
DECISION TREES EVALUATION SUMMARY

Accuracy	Misclassification Error
0.89	0.11

B. Predicting Outcomes of Customer Purchase Intentions.

Accuracy is also used here to check the performance of both models. The results are shown in tables IX through XII

TABLE IX
DECISION TREE MODEL CONFUSION MATRIX

	Actual False	Actual True
Predicted False	1942	141
Predicted True	148	235

TABLE X
DECISION TREES EVALUATION SUMMARY

Accuracy	Misclassification Error
0.88	0.12

TABLE XI
NAÏVE BAYES CLASSIFIER CONFUSION MATRIX

	Actual No	Actual Yes
Predicted No	8753	630
Predicted Yes	1669	1278

TABLE XII
NAÏVE BAYES EVALUATION SUMMARY

Accuracy	Misclassification Error
0.81	0.19

C. Predicting Outcome of Household Electricity Consumption.

TABLE XIII
MULTIPLE REGRESSION EVALUATION SUMMARY

R2	MAE	MSE	RMSE
0.0052	0.576	1.0831	1.0407

The R Square, Mean Absolute Error, Mean Squared Error and Root Mean Squared Error are compared.

From the tables above it is easy to make comparisons of the Machine Learning algorithms, The Decision Tree performed best for both the bank marketing and online purchase intention datasets while the Random Forest has the higher R square value for the Household energy consumption.

V. Conclusion and Future Work

This project applied different machine learning techniques to real world data in order to deduce the prominent features or in some cases the relationship between features of the data. The features that influences a potential customer to opt for a term deposit in the bank marketing dataset was last contact duration, the outcome of the previous marketing campaign, month of last contact and the mode of contact. Furthermore, to make inference on the factors that determine customers purchase intentions, the most important session component that lead to revenue in the Online Purchase Intention dataset was the page value i.e. Google Analytics metric for page on the e-commerce website. Lastly the Random Forest had a higher R Square value for the Household energy consumption.

There are numerous algorithms left for the future consideration. The models here can be scaled, Clustering and Artificial Neutral Networks can be applied to all three datasets to better explore the data and discover insights which were not gotten from here.

TABLE XIV
RANDOM FOREST REGRESSION EVALUATION SUMMARY

R2	MAE	MSE	RMSE
0.5626	0.340	0.476	0.690

References

[1] S. Moro, R. Laureano, and P. Cortez, "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology," 2011.

[2] H. A. Elsalamony and A. M. Elsayad, "Bank Direct Marketing based on Neural Network and C5.0 Models," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 2, no. 6, 2013.

[3] M. Scherer, J. Smolag, and A. Gaweda, "Predicting Success of Bank Direct Marketing by Neuro-fuzzy Systems," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2016, pp. 570–576.

[4] M. Karim and R. Rahman, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing," *Journal of Software Engineering and Applications*, 2013.

[5] T. Parlar and S. K. Acaravci, "Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data," *International Journal of Economics and Financial Issues*, vol. 7, no. 2, p. 692, 2017.

[6] T. Das, "A Customer Classification Prediction Model based on Machine Learning Techniques," in *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2015, pp. 321–326.

[7] Y. Zuo, K. Yada, and A. S. Ali, "Prediction of Consumer Purchasing in a Grocery Store Using Machine Learning Techniques," in *2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*. IEEE, 2016, pp. 18–25.

[8] A. Kumar, G. Kabra, E. K. Mussada, M. K. Dash, and P. S. Rana, "Combined artificial bee colony algorithm and machine learning techniques for prediction of online consumer repurchase intention," *Neural Computing and Applications*, vol. 31, no. 2, pp. 877–890, 2019.

[9] E. G. Boroujerdi, S. Mehri, S. S. Garmaroudi, M. Pezeshki, F. R. Mehrabadi, S. Malakouti, and S. Khadivi, "A study on prediction of user's tendency toward purchases in websites based on behavior models," in *2014 6th Conference on Information and Knowledge Technology (IKT)*. IEEE, 2014, pp. 61–66.

[10] O. Mokryn, V. Bogina, and T. Kuflik, "Will this session end with a purchase? inferring current purchase intent of anonymous visitors," *Electronic Commerce Research and Applications*, vol. 34, p. 100836, 2019.

[11] F. Shi and C. Ghedira, "Intention-based online consumer classification for recommendation and personalization," in *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*. IEEE, 2016, pp. 36–41.

[12] R. Chang, X. Shen, B. Wang, and Q. Xu, "A novel method for software defect prediction in the context of big data," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. IEEE, 2017, pp. 100–104.

[13] W. Kim and S. Katipamula, "A review of fault detection and diagnostics methods for building systems," *Science and Technology for the Built Environment*, vol. 24, no. 1, pp. 3–21, 2018.

[14] Q. Zeng, N. Zhang, Y. Wang, Y. Liu, C. Kang, Z. Zeng, W. Yang, and M. Luo, "An optimum regression approach for analyzing weather influence on the energy consumption," in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 2016, pp. 1–6.

[15] Z. Wang, Y. Wang, R. Zeng, R. S. Srinivasan, and S. Ahrentzen, "Random forest based hourly building energy prediction," *Energy and Buildings*, vol. 171, pp. 11–25, 2018.

[16] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings*, vol. 49, pp. 560–567, 2012.

[17] C. Chen and D. J. Cook, "Behavior-based home energy prediction," in *2012 Eighth International Conference on Intelligent Environments*. IEEE, 2012, pp. 57–63.

[18] J. M. M. Arce and E. Q. B. Macabebe, "Real-time power consumption monitoring and forecasting using regression techniques and machine learning algorithms," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*. IEEE, 2019, pp. 135–140.

[19] A. Mastrucci, O. Baume, F. Stazi, and U. Leopold, "Estimating energy savings for the residential building stock of an entire city: A gis-based statistical downscaling approach applied to rotterdam," *Energy and Buildings*, vol. 75, pp. 358–367, 2014.