INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Performance analysis of the Machine Learning Classifiers to predict the behaviour of the customers, when a new product is launched in the market

Poulomi Saha
poulomisaha075@gmail.com

ABSTRACT

Here in this study, I will analyze the correlation of the buyers with their age and salary using various machine learning classifier algorithms. This study will predict, who will buy a new item faster as soon as it is launched in the market and how it will be related to the age and salary of the people, who are buying it. The aim of this study is to investigate six different types of Machine Learning, Classifier algorithms (namely Logistic Regression, SVM, Naive Bayes, KNN,Decision Tree, Random Forest and to show their comparative analysis and to predict whether a person will buy a certain product as soon as it is launched in the market. Experiments are performed on the Social_Network_Ads data set which is sourced from Kaggle the online community of data scientists and machine learning engineers. The performance of all the above algorithms is evaluated on the various metrics like recall, precision, F1_score and confusion matrix. Results are then compared.

Keywords— *Classification, Decision Tree, K-Nearest Neighbor (K-NN), Logistic regression, Machine learning, Naive Bayes, Random forest, Support Vector Machine (SVM)*

1. INTRODUCTION

The increasing amount of data by various organizations over the last decade has led to great progress in the field of Machine Learning [1]. There are various techniques that are used by different companies to produce customer friendly products from time to time, but for that, at first, they carry out various researches. Day by day, to manage the increasing amount of data properly, the researchers are pushing the capability of the boundary of Machine Learning and also Artificial Intelligence [2]. Companies are using the data that are available to take better decision on day to day basis, which will help the companies to increase the sale of various products and also to produce the user-friendly product. This is referred to as Business Intelligence [3].

Machine Learning techniques have been an excellent support for making a prediction of a particular system by training. It is about learning structures from the data that is provided. [4] Machine Learning in recent years has been reliable, evolving and supporting in all the domains and especially in the Business domain. Machine Learning techniques are used to classify who are the customers who used to buy the products consistently as soon as they are launched in the market and the correlation between the age, salary and the product. This classification helps the company to predict and launch the products accordingly, based on the rate of sales and also the loss or gain of the company.

Classification is a supervised learning approach in which the computer learns from the data input given to it and then uses this learning to classify new observations. In this work, we propose several methods based on Machine Learning techniques to find the correlation between the age, salary and the customer of the newly launched product. SVM, Naive Bayes, Logistic Regression, KNN, Decision Tree and Random Forest Machine Learning Classifiers are used and evaluated on Social_Network_Ads data set and the experimental performance of all the six classification algorithms are compared on various metrics.

2. LITERATURE SURVEY

Patrik Norlin and Viktor Paulsrud conducted a study to identify new customers buying a product using Machine learning Classifiers. It is also found that when we use more variables, it is used to improve the model, both in higher and lower dimensional space.

"Misbehaving: The making of Behavioural economics", by, Richard H. Thaler, in this book, he has given various examples to show how humans do not behave according to economic models. This book shows us, that most of the buying decisions of human

are not based on well-defined logic. Culture, intuition, age, communication, trust plays a major role in one's product buying decision.

3. OBJECTIVE

With the increase in population, the number of products brought increased at a rapid rate, which resulted in an increase of data in various companies that are manufacturing and selling the products. This also led to increased data in the medical, banking and various other fields. This increase in data at an alarming rate needed great attention and thus several Machine Learning Classification techniques are used to take proper product manufacturing decision and thus increase their customers.

Machine Learning has provided promising support for predicting the co-relation between the age, salary, taste of the person, using various techniques for the betterment of the companies, so that they can manufacture customer friendly products.

In this study we mainly analyze the performance of the Machine Learning Classifiers (SVM, Naive Bayes, Logistic Regression, KNN, Decision Tree and Random Forest) through their implementation on the Social_Network_Ads data set, that will help us to understand the correlation between age, salary and taste of the various people and to produce products accordingly.

4. METHODOLOGY

4.1 Classifier Description: The following classifiers has been implemented for the prediction of the category of buyers and their relation with age and salary.

4.1.1Support Vector Machine (SVM): Support Vector Machine is a machine learning algorithm that constructs a hyper plane or set of hyper planes in a higher dimensional space (here in figure 1 is a hyper plane in 2D represented by a line and hyper plane in 3D is represented by a plane) which is used in classification.

The parameter in SVM:

(a) Kernel: The classification of the hyper plane in linear SVM is done by transforming the problem using some linear algebra and in this, the kernel plays its role.

(b) Regularization: The regularization parameter (often termed as C parameter in python's sklearn library) is used to tell the SVM optimization, how much we want to avoid misclassifying each training example.

(c) Gamma: The gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'.

4.1.2 Naive Bayes: A Naive Bayes classifier is a probabilistic machine learning model that is used for classification. This classifier works based on the Naive Bayes theorem:

Bayes Theorem

$$P(A|B) = P(B|A)*P(A)/P(B)$$

4.1.3. Logistic Regression: It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured in which there are only two possible outcomes. The goal of the Logistic regression is to find the best fitting model to describe the relationship between the independent variables.

4.1.4. K-Nearest Neighbour: The k-Nearest Neighbour is a supervised classification algorithm. It takes a bunch of labelled points and uses them to label other points. A new point is assigned to the class which is most common among its K Nearest Neighbour.

4.1.5. Decision Tree: Decision tree builds a classification or regression model in the form of the tree structure. It splits down a data set into smaller and smaller subsets and at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes, it represents a classification or decision. The topmost decision node corresponds to the best predictor called root node. Decision trees can handle both categorical and numeric data.

4.1.6. Random Forest: Random Forest or random decision forest is an ensemble learning method that is used for classification, regression and other tasks, that is used to operate by constructing a multiple of decision trees at the training time and then gives the output of the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

4.2 Dataset Description

The proposed methodology is evaluated on Social_Network_Ads data set which is obtained from the Kaggle data set repository and is a standard data set.

The considered data set consists of 400 different instances and the customers in the data set consists of females with the minimum age of 18 to the maximum age of 60. The data set consists of males with the same range of age as females. There salary ranges from 15000 to 150000. The binary target variable takes (0 or 1) values, while '0' implies a negative result for the customers who didn't buy the product, and '1' indicates a positive result.

The five features were chosen in order to predict the number of customers buying a new product, those significant variables were used for the prediction using the classifier.

The five features present in the data set are:

- The User Id of the customer. (F1 feature: USER ID)
- The Gender of the customer is used to find the gender. (F2 feature: Gender)
- The Age column is used to find the age of the customer who is buying the product mostly. (F3 feature: Age)
- The Estimated salary column is used to find the range of salary of the customers who are buying the product. (F4 feature: Estimated Salary)
- The Purchased column is used to show whether a customer has brought the product or not. (F5 feature: Purchased)

4.3. System Requirement

4.3.1 Software Requirement

(a) 1.Anaconda-4.4

(b) 2.Jupyter Notebook

4.3.2 Hardware Requirement

Processor: Intel®Core™ i5 6200U CPU@ 2.30GHz 2.40GHz

Ram: 8.00GB

System type: 64-bit Operating System, x64based processor.

4.4. Evaluation Parameters

4.4.1 Confusion Matrix: A confusion matrix is a table (as shown in the figure) that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

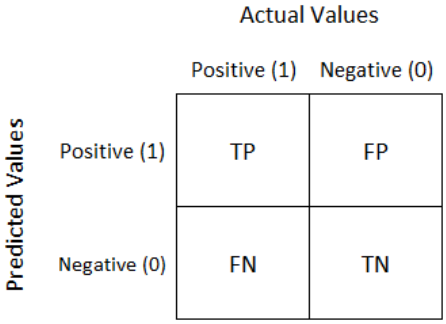

Fig. 1: Confusion Matrix

True Positive (TP): These are the correctly predicted positive values which mean the value of the actual class is positive and the value of the predicted class is positive.

True Negative (TN): These are the correctly predicted positive values which mean the value of the actual class is negative and the value of the predicted class is negative.

False Positive (FP): When the actual class is negative and the predicted class is positive. This is also known as type-1 error.

False Negative (FN): When the actual class is positive and the predicted class is negative. This is also known as type-2 error.

4.4.2 Accuracy: Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observation.

$$\text{Accuracy} = TP+TN/(TP+TN+FP+FN)$$

4.4.3 Recall (Sensitivity): Recall is the ratio of correctly predicted positive observations to all observations in the actual class.

$$\text{Recall} = TP/(TP+FN)$$

4.4.4 Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\text{Precision} = TP/(TP+FP)$$

4.4.5 F1-Score: F1 Score is the weighted average of precision and recall. Therefore, this score takes both false positives and false negatives into account.

$$\text{F1-Score} = 2*(\text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision})$$

5. RESULTS AND DISCUSSION

5.1 The performance of the six classifiers are evaluated on the various metrics

Accuracy, Precision, Recall, F1-Score using 90:10 split. The table below shows the comparative performance of the Classification algorithms.

Table 1: Comparative Performance of Classification Algorithms on Various (90:10) split of data set

Classifier Metrics	SVM	Naive Bayes	Logistic Regression	KNN	Decision Tree	Random Forest
Accuracy	0.93	0.90	0.89	0.93	0.91	0.92
Recall	0.90	0.78	0.75	0.90	0.90	0.90
Precision	0.87	0.89	0.88	0.87	0.82	0.85
F1-Score	0.89	0.83	0.81	0.89	0.86	0.87

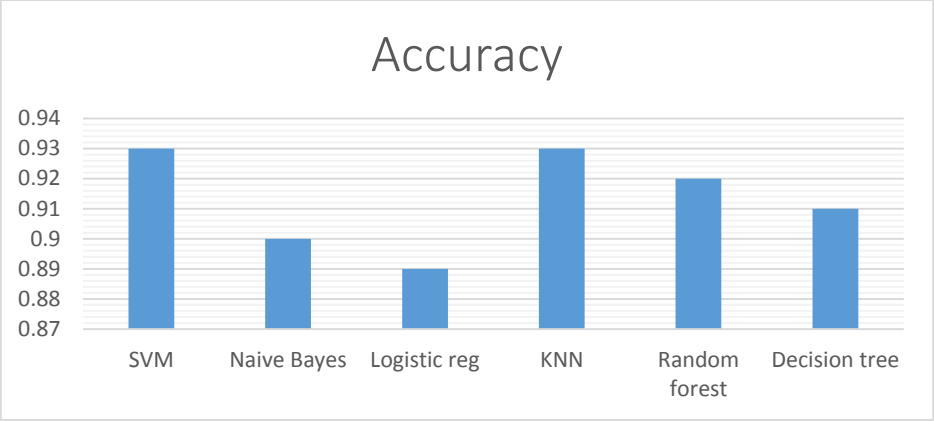

Fig. 2: Accuracy of various Classifiers

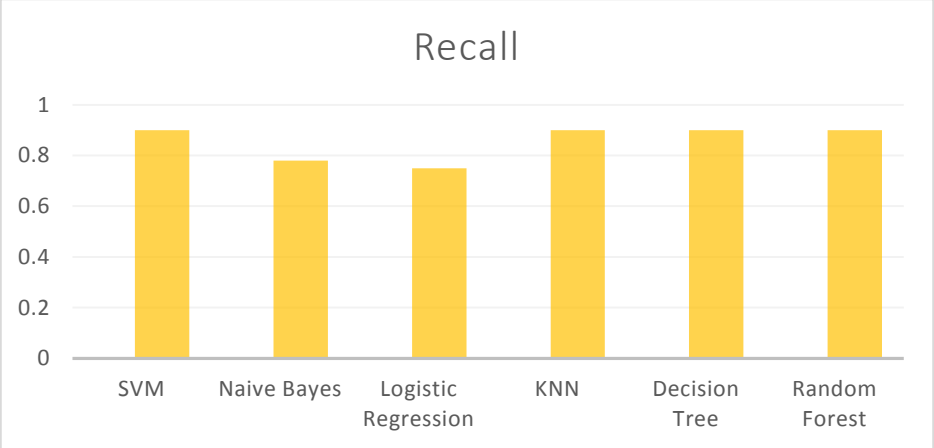

Fig. 3: Recall (Sensitivity) of various Classifiers

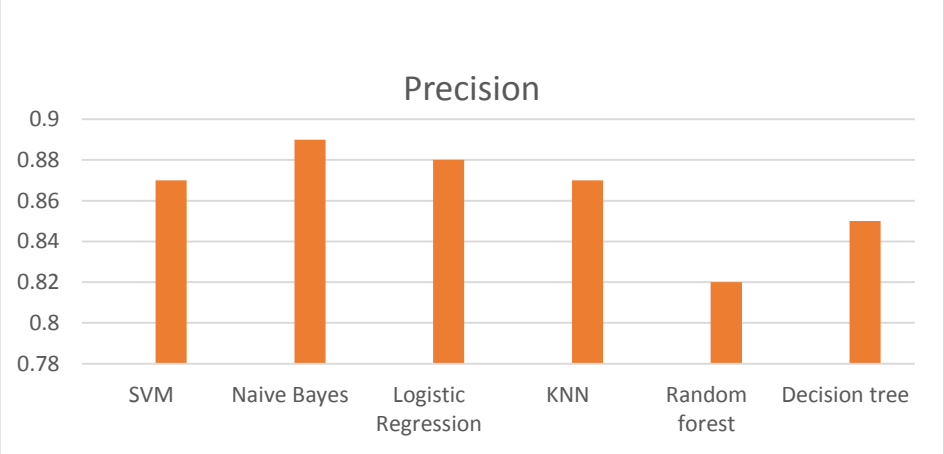

Fig. 4: Precision of various Classifiers

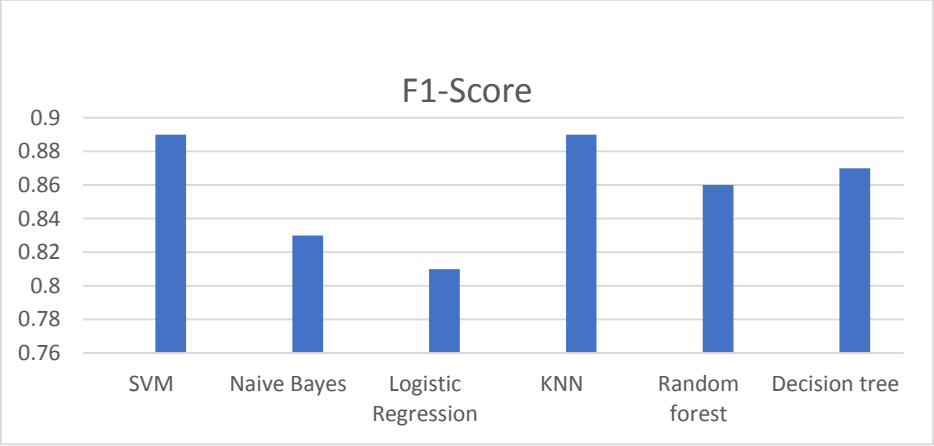

Fig. 5: F1-Score of various Classifiers

The 90:10 split of data set yield the best result for different metrics when applied on various classifiers. We obtained the best accuracy of 0.93 and the best F1-Score of 0.89 from SVM and KNN. The remaining algorithms obtained an accuracy ranging from 0.89 to 0.92, F1-Score ranging from 0.81 to 0.87.

Therefore, it can be concluded that the SVM classifier and KNN classifier outperformed all the other classifiers

6. CONCLUSION AND FUTURE WORK

One of the most important real-world implication is to find the correlation between the customers, their age and salary. This in many ways helps the various companies to launch different types of new products and to make their companies customer-friendly. During this work, six Machine Learning Classification algorithms were studied and evaluated on real-world Social_Network_Ads data set. The results determine the adequacy of the designed system in the 90:10 split as discussed above.

In the future work, we look forward to implementing the Machine learning classification algorithms on the designed system for more accurate prediction and by using more

7. ACKNOWLEDGEMENT

I am profoundly grateful to Dr Jitendra Kumar Rout for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

8. REFERENCES

[1] Mitra .R, 2018, Predicting buying behaviour using Machine Learning: A case study on Sales Prospecting (Part I).URL https://becominghuman.ai/predicting-buying-behavior-using-machine-learning-a-case-study-on-sales-prospecting-part-i-3bf455486e5d

[2] Columbus. L, 10 Ways Machine Learning Is Revolutionizing Sales,©2019 Forbes Media LLC. URL https://www.forbes.com/sites/louiscolumbus/2018/12/26/10-ways-machine-learning-is-revolutionizing-sales/#5150ee043fd1

[3] Norlin .P and Paulsrud .N, 2017, project, Identifying New Customers Using Machine Learning, pg(6-10).

[4] Scikit-learn - machine learning in python. URL https://scikit-learn.org/stable/

