iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4483

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Predictive Analysis on Bank Marketing Campaign Using Machine Learning Algorithms

M. Chenna Keshava[1], K. Sai Chandana[2], K. Sai Sagar[3], M. Naga Harish[4]
[1]*Adhoc Lecturer,* [2,3,4]*Student, Dept of CSE, JNTUACEP, Pulivendula, AP, India*

Abstract: ***The main aim is to predict the best marketing campaign based on whether the customer of the bank subscribe for a term deposit or not. We will use different classification algorithms and find the algorithm that will make best prediction in this aspect. The key motivation of this prediction is to create and keep clear relationship and valued customers accompanied by innovative ideas which can be used as measures to meet their requirements. Accuracy is used as the evaluation metric and the best model is chosen based on the accuracy.***
Keywords: Prediction, Classification models, Accuracy

I. INTRODUCTION

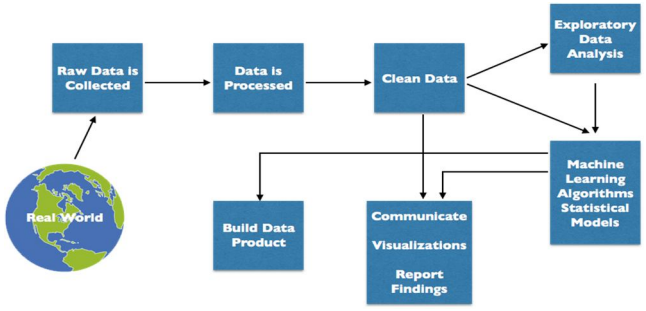
Banks store huge information about their customers. This information is collected so as to maintain good relationship between the customers and the bank in order to target them individually for definite products or banking offers. Usually, the selected customers are contacted directly through: personal contact ,telephone cellular, mail, and email or any other contacts to advertise the new product/service or give an offer, this kind of marketing is called direct marketing. The objective of direct marketing in retail banking is to attract new customers, to create a direct customer-bank communication to promote an offer or obtain customer information, and to strengthen a long-term relationship with the customer. Bank marketing has become very important for the banking industry. It is very important for a bank to develop good relationship with connection with the customers of the bank in order to target them individually for definite products or banking offers.

II. MOTIVATION

To understand the motivations and behavior, banks have over the period invested in different tools which gave them some information, but not exactly the one the customer expected. Banks needed to address issues like, how can they ensure long-term loyalty from its high-value customers? How can they attract and retain different types of customers and what additional product to sell? What rewards to target at its profitable customer? The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting it to predict future outcomes. Having the relevant data helped banks to target the right offers at the right time and make changes as and when required throughout the customer lifecycle. It all finally boiled down to data and its effective usage.

III. PROPOSEDWORK

By using data mining techniques predicting about the term deposit is a time consuming task. So in the proposed system we will use different supervised classification models to find the accuracy given by the each model and finally selects the best model which gives the highest accuracy. Some of the classification models which used are Logistic Regression, Random Forest and AdaBoost. The architecture of this application is



IV. EVALUATION METRIC

We will use accuracy score as evaluation metric to predict the target variable. It is defined as the number of correct predictions made as a ratio of all predictions made. Accuracy is most common evaluation metric for classification problems.

Number of correct predictions Accuracy= ----------

Total number of predictions

V. DATASET

The data set used for this project is well known as bank marketing from the University of California at Irvine (UCI). The dataset is comprise of multivariate data and it has both categorical and integer values. The dataset has 21 columns and 41188 rows, with 20 features, and one response variable. The number of subscribers is 4640 and the number of customers who did not subscribe is 36548.

The features which are used to predict the target variable are age, job, marital status, education, default ,housing ,loan ,pdays, contact, month, day_of_week, duration, campaign, previous, poutcome, emp. var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed and the target variable is y:has the client subscribed a term deposit?

VI. DATA PREPROCESSING

Some of the datasets contain irrelevant information and noisy data. These datasets should be handled properly to get a better result. Data preprocessing includes data cleaning, transformation and dimentionality reduction which convert the raw data into a form that is suitable for further processing.

The first step in data preprocessing is read the dataset, it is done by using read_csv(). And the shape of the dataset will helps to understand briefly about the size of the data.

```
In [35]: print("The training dataset has",full_data.shape[1],"columns and", full_data.shape[0],"rows")

The training dataset has 21 columns and 41188 rows
```

First five instances of the dataset can be viewed using head() function. This function is the core part of go-to python pandas functions for investigating our datasets.

```
# Print the first few entries of the Email Marketing Challenge data
display(full_data.head())
```

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	prev
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	...	1	999	0
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	...	1	999	0
2	37	services	married	high.school	no	yes	no	telephone	may	mon	...	1	999	0
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	...	1	999	0
4	56	services	married	high.school	no	no	yes	telephone	may	mon	...	1	999	0

5 rows × 21 columns

The info() function in pandas is used to have a view of the consise summary of the entire data. We use this function while performing exploratory analysis of data. This function gives summary about all the attributes and also used to find any missing values in the entire dataset. The output of this info() function gives the summary that includes list of all columns with their data types and the number of non-null values in each column. We also have the value of range index provided for the index axis.

```
full_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
age               41188 non-null int64
job               41188 non-null object
marital           41188 non-null object
education         41188 non-null object
default           41188 non-null object
housing           41188 non-null object
loan              41188 non-null object
contact           41188 non-null object
month             41188 non-null object
day_of_week       41188 non-null object
duration          41188 non-null int64
campaign          41188 non-null int64
pdays             41188 non-null int64
previous          41188 non-null int64
poutcome          41188 non-null object
emp.var.rate      41188 non-null float64
cons.price.idx    41188 non-null float64
cons.conf.idx     41188 non-null float64
euribor3m         41188 non-null float64
nr.employed       7763 non-null float64
y                 41188 non-null object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

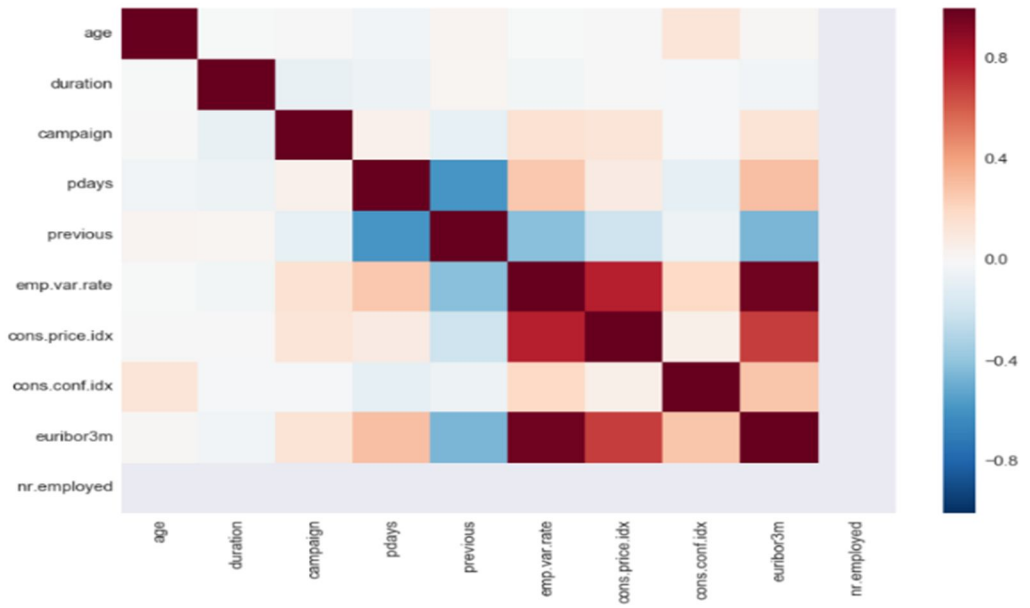
VII. DATA VISUALIZATION

Data visualization is a technique that uses an array of static and interactive visuals within a specific context to help us understand and make sense of large amounts of data. This provides an important suite of tools for gaining a qualitative understanding and also helpful when exploring and getting to know a dataset. We use this technique in identifying patterns, corrupt data and outliers.

By using data visualization we will easily find the correlation between different attributes in the dataset where correlation is an important tool for feature engineering in building machine learning models and also it refers to a mutual relationship or association between quantities. Correlation helps to predict one quantity from another.

```
In [45]: # visulize correlations of features with a heatmap
         plt.figure(figsize=(15, 10))
         sns.heatmap(full_data.corr())
         plt.xticks(rotation=90)
         plt.yticks(rotation=0)
```

```
Out[45]: (array([ 0.5,  1.5,  2.5,  3.5,  4.5,  5.5,  6.5,  7.5,  8.5,  9.5]),
          <a list of 10 Text yticklabel objects>)
```



The above heat map visualizes the correlation between different attributes in the dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

VIII. MODELLING

In order to model the data we will first divide the entire dataset into training and testing datasets. 70 percent of the data comes under training data and the rest of the data goes into testing dataset.

IX. ALGORITHMS AND TECHNIQUES

The supervised learning algorithms are used to train the machine using data which is well labeled that means some data is already tagged with correct answer. After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analysis the training data (set of training examples) and produces correct outcome from labeled data. The supervised algorithms used for prediction in this application are: Logistic Regression (Benchmark Model), Random Forest, Ada Boost.

A. Logistic Regression

Logistic regression is one of the supervised classification algorithm used for binary classification. It gives a discrete binary outcome between 0 and 1. This algorithm works by measuring the relationship between the dependent variable and one or more independent variables.

1) *Advantage:* Outputs have a nice probabilistic interpretation and the algorithm can be regularized to avoiding overfitting. It may handle nonlinear effects. You can add explicit interaction and power terms. There is no homogeneity of variance assumption. Normally distributed error terms are not assumed.
2) *Disadvantage:* Logistic regression tends to underperform when there are multiple or non linear decision boundaries. We cannot solve non linear problems with logistic regression since its decision surface is linear.

B. Random forest

Random Forest is a predictive modeling algorithm which is used for both classification and regression tasks. It works well with default hyper parameter. It can be used to rank the importance of variables in a regression or classification problem.

1) *Advantage:* Reduction in overfitting by averaging several trees, there is significantly lower risk of overfitting. Random forests also have less variance than a single decision tree. It means that it works correctly for a large range of data items than single decision trees. They also do not require preparation of the input data. You do not have to scale the data.
2) *Disadvantage:* It takes more time to train the samples.The main disadvantage of Random forests is their complexity. They are much harder and time-consuming to construct than decision trees.

C. Adaptive Boosting

Ada-boost classifier combines weak classifier algorithm to form strong classifier. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with selection of training set at every iteration and assigning right amount of weight in final voting, we can have good accuracy score for overall classifier. The metric we use to evaluate these models is accuracy score. The accuracy score is described as a ratio of the number of correctly predicted instances in divided by the total number of instances in the dataset multiplied by 100 to give a percentage (e.g. 95% accurate).Here we use 10-folds cross validation which means we split the entire data set into 10 parts i.e., perform training on 9 and testing on 1 and repeat this for all the combinations of train-test splits.

1) *Advantage:* The AdaBoost algorithm will select the weak classifier that works best at that round of boosting. This algorithm is very simple to implement and it adjusts adaptively the errors of the weak hypotheses by WeakLearn.
2) *Disadvantage:* Time and computation expensive. Hard to implement in real time platform and complexity of the classification increases. It can be sensitive to noisy data and outliers.

X. RESULTS

The accuracies given by the three algorithms mentioned above are:

Model Name	Accuracy Score
Logistic Regression	90.92
Random Forest	91.22
Adaptive Boosting	91.32

From the above results we conclude that Random Forest algorithm gives the best accuracy score in prediction and it is the best algorithm when compared with the others.

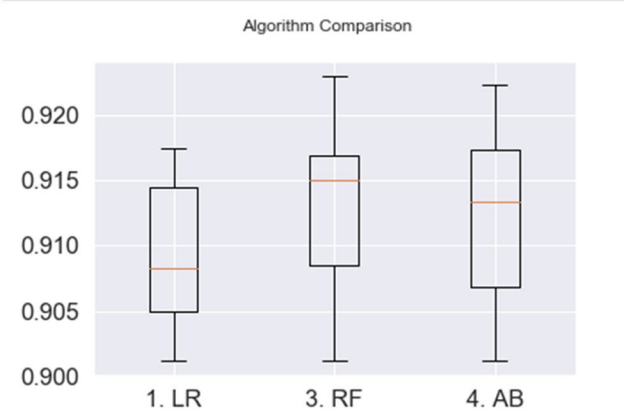
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

XI. REFINEMENT

We will refine the models by using GridSearchCV. And also tune the parameters 'random_state'=30 &'min_samples_split'=10 thereby increasing the accuracy of the tuned model. After tuning the parameters of random forest the final accuracy is 91.66.

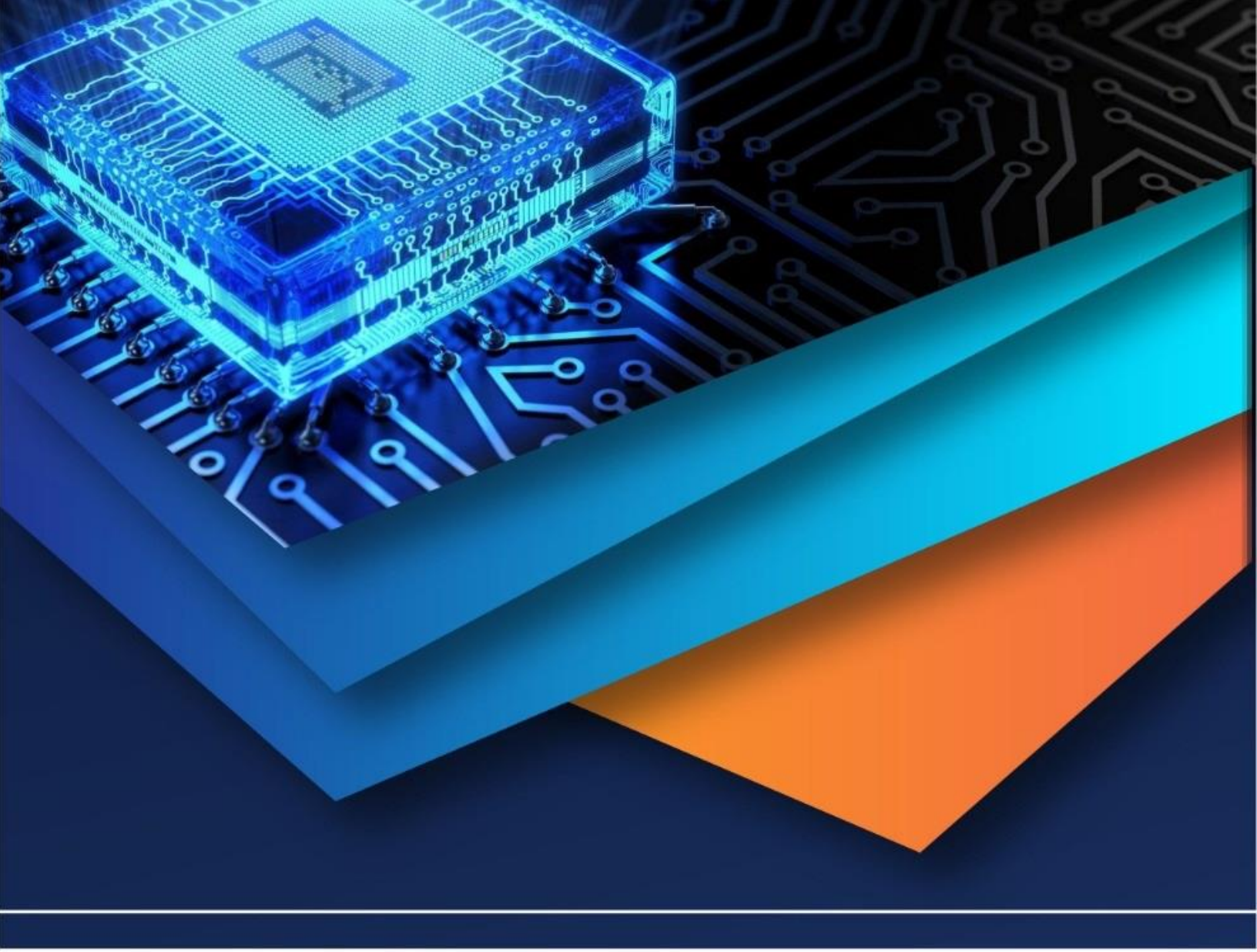
XII. CONCLUSION

By using machine learning algorithms we can easily perform predictive analysis on bank marketing campaigns. These models are trained with the dataset and gives the accuracy score as result. From all the algorithms Random Forest algorithm gives the best accuracy score in predicting whether the customer subscribes for a term deposit or not. So it is selected as the best algorithm when compared with others.



REFERENCES

[1] **https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd**
[2] https://www.tutorialspoint.com/bank_management/bank_management_marketing.htm
[3] http://www.mintel.com/comperemedia
[4] Eniafe Festus Ayetiran, "A Data Mining -Based Response Model for Target Selection in Direct Marketing",I.J.Information Technology and Computer Science, 2012,1, 9-18.
[5] https://archive.ics.uci.edu/ml/datasets/bank+marketing
[6] https://www.tutorialspoint.com/bank_management/bank_management_ marketing.htm
[7] https://www.retailbanking-academy.org/media/uploads/NEWBRANDING-DOCS/modules/RBII/RB_II_New_201_Small.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY