Analyzing previous marketing data of a bank to maximize its term deposit subscribers in future marketing campaigns

Dimuthu Wickramanayake[1] and Prabhashi Meddegoda[1]

[1]Department of Computer science and Engineering, Faculty of Engineering, University of Moratuwa
[2]dimuthu.20@cse.mrt.ac.lk
[2]prabhashi.20@cse.mrt.ac.lk

***Abstract*— Here we have taken a data set which is related with direct marketing campaigns (phone calls) for bank term deposits of a Portuguese banking institution[SMR14]. For this we have designed a simple website with graphical visualization of the analysis results that enables the bank to identify the clients they need to focus on in a marketing campaign, thus increase the profit with less effort. This website also enables the employees to predict whether a client will subscribe a term deposit or not.**

I. Introduction

The data is related with direct marketing campaigns for bank term deposits of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. So the data set consist of 20 variables and one class variable. This data set can be used to determine the success of banking related campaigns.

The goal of this research is to develop an interactive web site to visualize the relationships of each variable with class variable. Using the visualization determine anomalies in the data and find the correlations of the variables. Then final goal for this research is to have a simple form where a banking person would enter details of a person and find out whether that person would say "yes" or "no" using our prediction algorithm. We have tested five machine learning algorithms to get the one with highest precision.

First part of the paper focus on descriptive analysis of data set using visual tools and then in the next part we use diagnostic analysis to determine more relationships between data and find out the reason for the behavior of the data. Then in the next part we do a predictive analysis of the data.

Here when building the web application we have used Django framework to get the most out of python for the data set handling. In the client side we have used ChartJS library to visualize the data. In the server side along with Django modules like Pandas, Sklearn and etc have been used.

II. Attribute Information

Bank client data:

- Age (numeric)
- Job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- Marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed)
- Education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- Default: has credit in default? (categorical: 'no', 'yes', 'unknown')
- Housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
- Loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

Related with the last contact of the current campaign:

- Contact: contact communication type (categorical: 'cellular','telephone')
- Month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- Day of week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- Duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

- Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- Previous: number of contacts performed before this campaign and for this client (numeric)
- Poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes

- Emp.var.rate: employment variation rate - quarterly indicator (numeric)

- Cons.price.idx: consumer price index - monthly indicator (numeric)
- Cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- Euribor3m: euribor 3 month rate - daily indicator (numeric)
- Nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

- y - has the client subscribed a term deposit? (binary: 'yes', 'no')

III. Descriptive analysis

This data set contains 41188 instances ordered by date (from May 2008 to November 2010); 36548 records with label 'no' and 4640 records with label 'yes'. There are missing values in 4 categorical attributes, all coded with 'unknown' label. These missing values were treated as separate feature values. This is because the missing values are not random and may themselves be information.

First the data types were identified by evaluating the data set. In the data set there were categorical nominal values, categorical ordinal values, and metric discrete and metric continuous. Categorical variables are Jobs, Marital, Education, Default, Housing and Loan. Out of these Education could be put in to the categorical ordinal values as the education level can be identified from lowest point to higher point. So in order to have a better look at the data set, each variable is analysed.

A. Age

Age is of metric discrete data type. For the first part of the descriptive analysis wee processed the data of the age column and found followings,

* Minimum Age : 17

* Maximum Age : 98

* Mean Age : 40

* Standard deviation of the age : 10.4

In this section age data plot with number of "yes" and "no" responses.

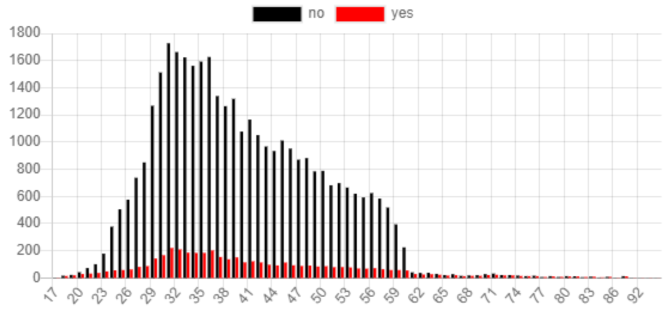


Fig. 1. Age Chart

.

Here it can be clearly seen that target of the telemarketers of the bank were people from early thirties late thirties. Although the number of people who said yes are high in these region, the number of people who said no is also high. But when it comes to people who are above 60 has same number of "yes" and "no"s.

B. Job

Job is of categorical nominal data type. Here we have taken the frequency of each class occurrences and then plot it with different jobs.

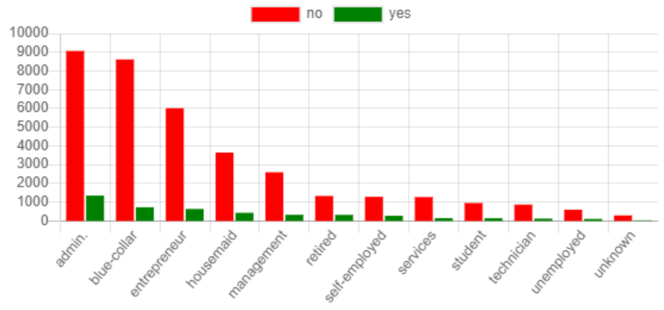


Fig. 2. Job Chart

.

Here by looking at the data we can see people with admin jobs were targeted most. Although they have higher number of "yes" amount still the ration between "yes" and "no" is very low. Best ratio can be seen in people who are retired and self employed. If we consider the age data distribution it was clear that people over 60 tend to say yes more and here it has been restated.

C. Marital

Marital status is of categorical nominal data type. Here we have taken the frequency of each class occurrences and then plot it with different relationship types.

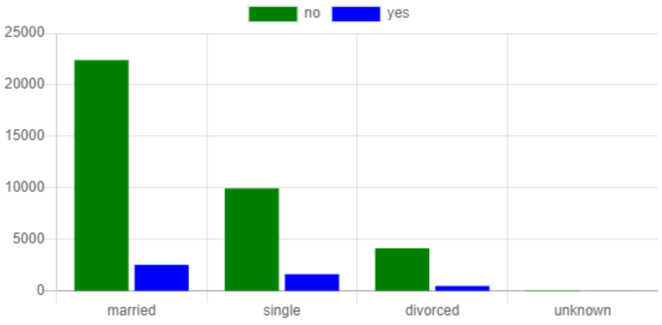


Fig. 3. Marital Chart

.

Out of the people who were contacted, most of them are married people and it's obvious because bank has contacted more people in their mid thirties. Although they have the higher "yes" percentage still the ratio between "yes" and "no" is very low. But single people and divorced people have higher ratio compared to married people.

D. *Education*

Education is of categorical ordinal data type. Here we have taken the frequency of each class occurrences and then plot it with different education levels.

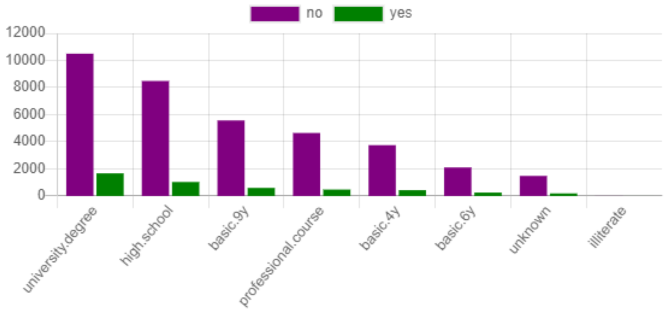


Fig. 4. Education Chart

.

Out of the people who were contacted, most of them are people who has a university level education.

E. *Default, Housing and Loan*

Default, Housing and Loan are of categorical nominal data type. Here we have taken the frequency of each class occurrences and then plot it with the corresponding feature values.

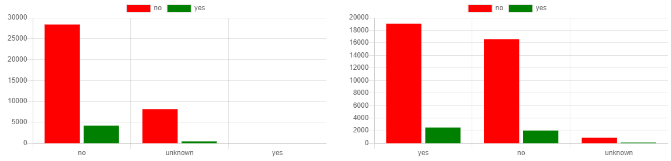
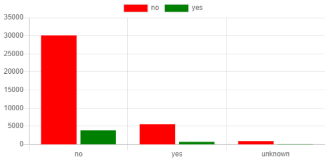


Fig. 5. Loan Chart

.



.

When considering the default data we can see there are lot more unknown data. People with no default credit has said yes and people with credit haven't respond at all. According to the data people with default credit haven't been taken in this campaign.

When considering the Housing loan data the people with an already housing loan have said yes more. Here the selection of people to be called must had a favour for people with loans.

When considering the Personal Loan Data the people with an already personal have said yes more.

F. *Duration*

Duration is of metric continuous data type. So we have find following information about this data column

* Maximum duration : 82 minutes
* Minimum duration : 0 minutes
* Mean duration : 4.3 minutes
* Standard deviation of the duration : 4.3 minutes

Here we have taken the frequency of each class occurrences and then plot it with different relationship types. Duration means the last contact duration. According to the data set information this is a very important field.

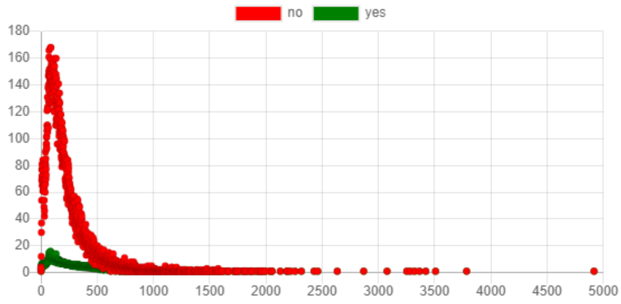


Fig. 6. Duration Chart

.

Here we can clearly see call with an optimal length about one and half minutes has taken more "yes" results. Although the number of "yes" are higher in other hand number of "no" is also high at his point.

G. *Day of week*

This is the last contact day of the week. Here this is an categorical nominal value type. Here the categories are weekdays. So we have plot a graph with weekdays and number of "yes" and "nos"s.

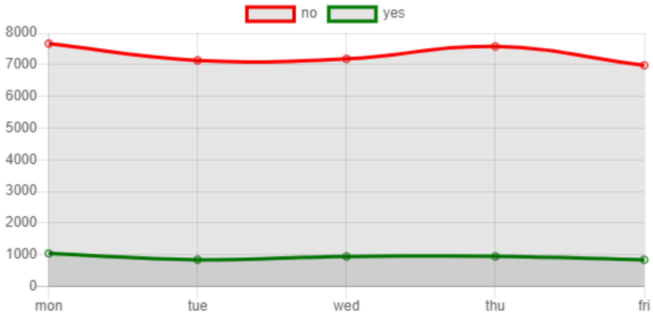


Fig. 7. Day of week chart

.

By looking at the data distribution we can see a cyclic pattern in the data where we have more results and more calls on Mondays and Thursdays.

H. *Month*

This is the last contact month of year. This data type is categorical nominal. Here the categories are months in the year. Here we have plot a graph with months and number of "yes" and "nos"s.

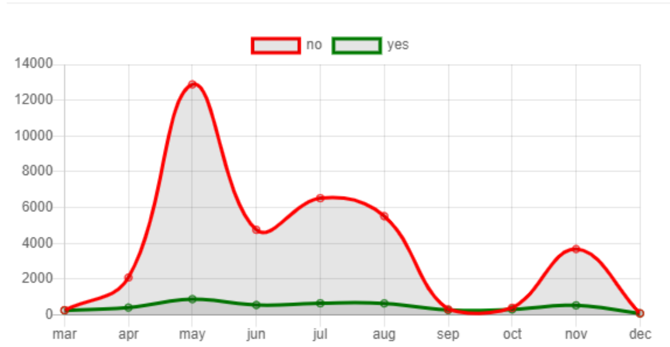


Fig. 8. Month Chart

.

Here there are no any call in the first quarter of the year. But then from march to may there is a sudden spike in the number of calls. Although the number of "yes" in may is higher the ratio between the "no" is lesser. But in September and October no of "yes" and "no" are going together.

I. Contact

This is the type of the contact whether the call was taken to the mobile or land line. The data type is categorical nominal and categories are cellular and telephone. The plot show cellular and telephone with frequency of "yes" and "nos"s.

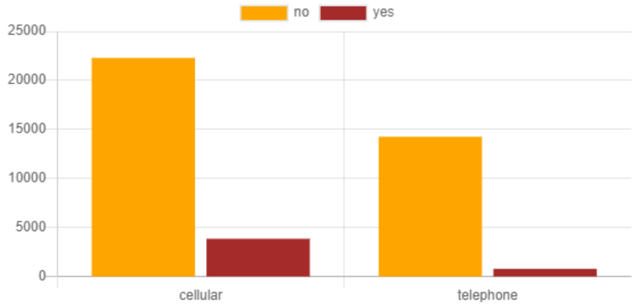


Fig. 9. Contact Chart

.

Here we can see higher number of calls were taken to mobile and higher number of "yes" response has been taken from mobile.

J. Campaign

This is number of contacts performed during this campaign and for this client. This is of metric discrete data type. Following information were found by analyzing these data column.

* Maximum number of contacts : 82

* Minimum number of contacts : 0

* Mean number of contacts : 4.3

* Standard deviation of the number of contacts : 4.3

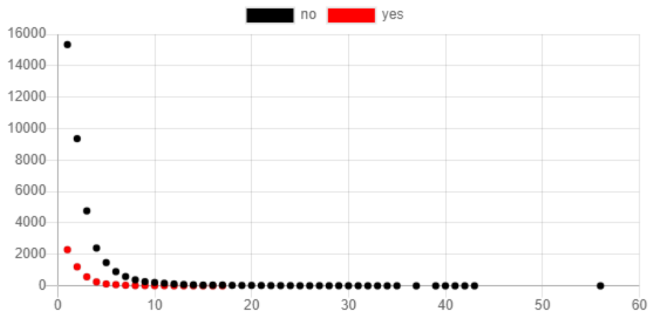


Fig. 10. Campaign Chart

.

Then the number of contacts in the campaign is plotted with number of "yes" and "no" frequency.

Here there are more "yes" result when the number of contacts is one. But the number of "no" is also high at this point.

K. Number of days that passed by after the client was last contacted from a previous campaign

Here Number of days that passed by after the client was last contacted from a previous campaign is of metric discrete type. Following data were revealed when evaluating the data.

* Maximum number of days that passed by after the client was last contacted : 82

* Minimum number of days that passed by after the client was last contacted : 0

* Mean number of days that passed by after the client was last contacted : 4.3

* Standard deviation of the number of days that passed by after the client was last contacted : 4.3

Then the number of days that passed by after the client was last contacted is plotted with number of "yes" and "no".

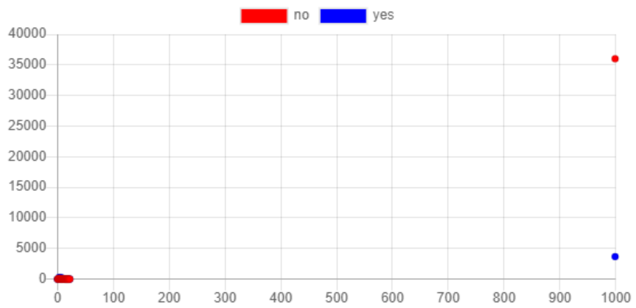


Fig. 11. PDays Chart

.

Here the plot shows a cluster of points around 0 -30 and then two dots in 999, since 999 is set as the value in this feature if the client was not previously contacted.

L. Previously contacted

Here the number of contacts performed before this campaign and for this client is of metric discrete type. Following data were revealed when evaluating the data.

* Maximum number of contacts performed before this campaign and for this client : 82

* Minimum number of contacts performed before this campaign and for this client : 0

* Mean number of contacts performed before this campaign and for this client : 4.3

* Standard deviation of the number of contacts performed before this campaign and for this client : 4.3

Then the number of contacts performed before this campaign and for this client is plotted with number of "yes" and "no".

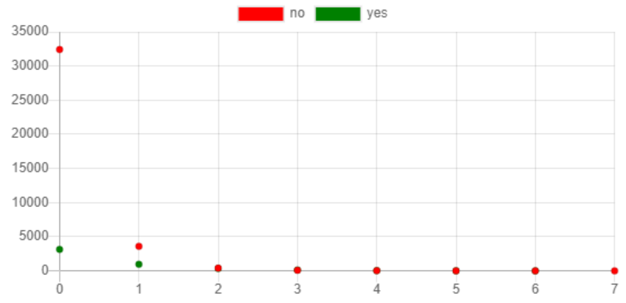


Fig. 12. Previously contacted Chart

.

Here the the people who were not contacted before this campaign have shown more results. But the "no" response is also high but people who were contacted one time has less difference in between people who said yes and who said no. So we can see effectivity of the contacting beforehand.

M. *Outcome of the previous marketing campaign*

Outcome of the previous marketing campaign is an categorical nominal data type value. Here we have plot the outcome data with number of "yes" and"no" for the response.

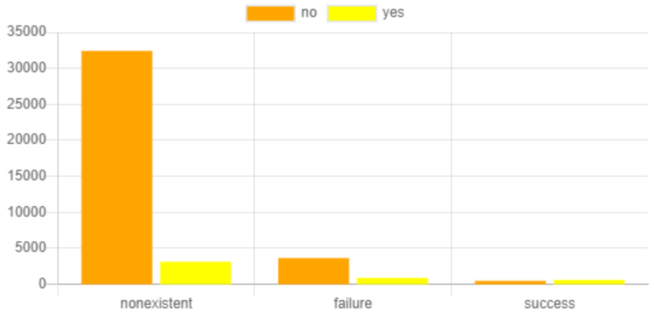


Fig. 13. Outcome of the previous marketing campaign

.

When examining the plot we can see there are many non existing data. This is because most of these people involved in this campaign haven't been involved in a campaign before.

IV. Diagnostic Analysis

Diagnostic analysis is used to identify the reasons for the results be to the way it is. Through this analysis, anomalies in data and the causal relationships between the features are identified.

Here, time series analysis is carried out to find why the subscription to term deposits are higher in some months / days of the week than the other months / days of the week along three consecutive years / three consecutive months. By using this, we can identify if there are any anomalies in the data or if there are patterns.

A. *Time series analysis of data*

Following is a time series analysis of the data set from 2008 - May to 2010 November. Here in the plot we can't see a trend in the behavior of the data but we can see a cyclic pattern in the number of "yes" responses of the data. Actually if we study this further we can think of this as a seasonal pattern.

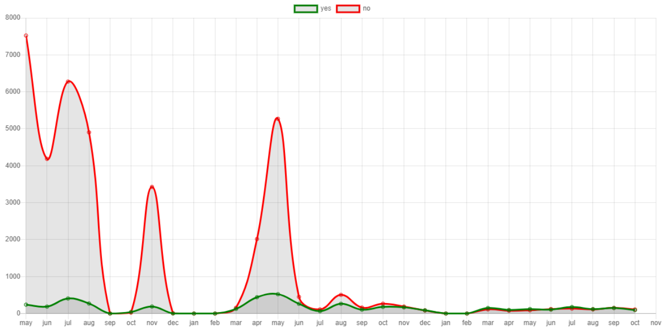


Fig. 14. Time series analysis for months from 2008 - 2010

.

To understand closely how the data have been behaving in a time series a time of 4 weeks have been selected and a graph has been plot.

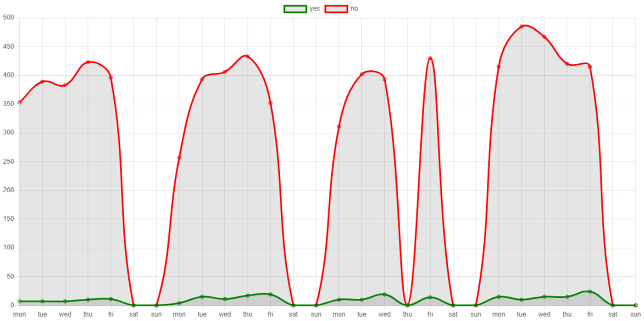


Fig. 15. Time series analysis for 4 weeks

.

Here we can't see a trend in the "yes" data. But here we can see a cyclic pattern. But as the days are not the same in this cyclic pattern we can't conclude this as a seasonal pattern.

But there are many causes that could let these behavior. Mainly when we see the monthly time series there in 2008 more call were taken but when it goes to 2010 number of calls are less. But as of then as they have improved their calling behavior we can see number of "yes" and "no" are going together in 2010. Calling agents must have improved their knowledge by then would be a better reason.

For this seasonal pattern which we can see in monthly time series we can conclude that early months should be vacation time for caller agents and around may every year they get active.

V. Predictive Analysis

A. Methodology

Read the data set into a data frame. The data set consists of 41188 records, with 20 features and the output. Separate the data set to features (data frame) and output (series)

B. Preprocessing

There are missing values in four categorical features; job, default (has credit in default?), housing (has housing loan?), loan (has personal loan?), represented by string 'unknown'. These were treated as separate values instead of removing / imputing. According to the descriptive analysis done, it was obvious to see there were many 'unknown' values. It gives us the sense that these values might themselves be an information. Removing those records with missing values will also reduce the size of the data set significantly and it is not worth it. Imputing may also result in faulty records.

The data set contains a feature 'duration' which is the last contact duration, in seconds. Since the value is not known before a call is performed, and since the output value (yes / no) is known once the call is ended, this feature is removed for the predictive modelling.

The labels (non-numerical) are encoded to numerical labels

The data set consists of features with different data types; numeric and categorical.

The numerical features were standardize by removing the mean and by scaling to unit variance (StandardScaler)

There was one feature that can be considered as having Categorical Ordinal data type; education. This feature was converted to numerical representation using OrdinalEncoder (this results in single column of integers per feature. Integers are 0 to (no of categories - 1).

All the other categorical features were nominal. So they were encoded to numeric using OneHotEncoder. The features cannot be converted to numeric values since these values do not have an order. Therefore, here a binary column is created for each nominal value.

These nominal and categorical preprocessing is added to a pipeline, where the classifier is added as the next step.

C. Classification

The following classifiers were modeled to select the best classifier.

1) Logistic Regression Classifier: This is a statistical model. This is a linear model, but the predictions are transformed to values between 0 and 1 using logistic function

2) Decision Tree Classifier: This is a tree based model where the data are split according to the given parameters. For this problem, Gini impurity is used as the criteria to measure quality of a split, best split as the strategy to choose the split, maximum depth as 4, minimum number of samples required to split as 2 and Complexity parameter used for Minimal Cost-Complexity Pruning as 0.002 (The subtree with the largest cost complexity that is smaller than 0.002 will be chosen)

3) K Nearest Neighbors Classifierr: This is a neighbor based method. This works directly on learned samples instead of creating rules. This method assigns the class of the majority of its k neighbors as the class of the test instance. For this problem, the best k value was 5.

4) Naïve Bayes Classifier: This is a probabilistic classifier based on Bayes theorem of probability. Here, Gaussian is used as the probability distribution.

This classifier assumes that all features are independent of each other (which may reduce its accuracy for this problem).

5) Support Vector Machine: This method finds a hyperplane in n-D space (n: no. of features) that distinctly classifies the data points. Here, stochastic gradient descent (SGD) is used as the learning method. In SGD, the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing learning rate.

The data set is an imbalanced data set with 36548 records with label 'no' and 4640 records with label 'yes'. Some of the classifiers (Naïve Bayes, k Nearest Neighbors as examples) are much sensitive to class ratio. Handling class imbalance is also important to the consistency of the accuracy scores. In order to ensure the same ratio of classes for both training and testing data set, stratification is used.

The classifiers are made aware that the data are imbalanced (if needed), by setting their class weight property to balanced; meaning that the classifier should give higher weight to the minority class and lower weight to minority class (adjust weights inversely proportional to class frequencies).

The training data is fit to the classifiers and the performance metric values are obtained.

D. Performance Evaluation

Since this is an imbalanced data set, confusion matrix, precision, recall and F1 score are used along with the accuracy score, since accuracy score alone can be misleading

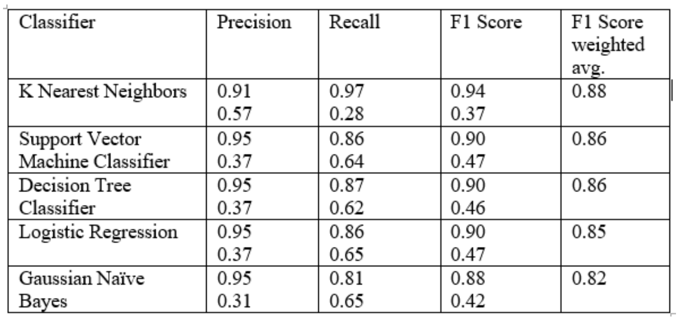
After the detection was done, then we took the relative velocity of the ground target. Furthermore, we got the real velocity of ground target from a smart phone fixed to the ground target. Some of the readings are shown in the following table.

In this table, each cell for Precision, Recall and F1 score gives the scores for label 'no' and 'yes' respectively.

F1 Score is used to select the best classifier since these scores are for an imbalanced data set and weighted average of F1 Score is used since it accounts this imbalance in its calculation. According to these results, accuracy score is proportional to the F1 Score weighted average. This is because the class imbalance was taken into consideration.

TABLE I
PERFORMANCE EVALUATION

Classifier	Accuracy Score	Confusion Matrix
K Nearest Neighbors	0.895	$\begin{bmatrix} 7114 & 196 \\ 672 & 256 \end{bmatrix}$
Support Vector Machine Classifier	0.836	$\begin{bmatrix} 6297 & 1013 \\ 338 & 590 \end{bmatrix}$
Decision Tree Classifier	0.838	$\begin{bmatrix} 6332 & 978 \\ 353 & 575 \end{bmatrix}$
Logistic Regression	0.834	$\begin{bmatrix} 6269 & 1041 \\ 326 & 602 \end{bmatrix}$
Gaussian Naïve Bayes	0.796	$\begin{bmatrix} 5950 & 1360 \\ 322 & 606 \end{bmatrix}$



Classifier	Precision	Recall	F1 Score	F1 Score weighted avg.
K Nearest Neighbors	0.91 0.57	0.97 0.28	0.94 0.37	0.88
Support Vector Machine Classifier	0.95 0.37	0.86 0.64	0.90 0.47	0.86
Decision Tree Classifier	0.95 0.37	0.87 0.62	0.90 0.46	0.86
Logistic Regression	0.95 0.37	0.86 0.65	0.90 0.47	0.85
Gaussian Naïve Bayes	0.95 0.31	0.81 0.65	0.88 0.42	0.82

Fig. 16.

.

According to the above results, the highest F1 Score weighted average is scored by k Nearest Neighbors (n = 5). The 2nd and 3rd best classifiers are Support Vector Machine (with SGD) and Decision Tree Classifier (C4.5 algorithm). Based on these results, k Nearest Neighbors with k=5 was selected as the classifier for the predictions in the designed application.

VI. CONCLUSION

In this research we have understood the data set using descriptive analysis on each variable. This was done by applying statistic operations to numeric (metric) values and then by using visualization methods. Here all the back end work was done using python (Django framework) and client side was handled using ChartJS and other JS libraries. Next a diagnostic analysis was done on the data and here we used time series analysis for the data monthly and daily basis. Here some hypothesis has been presented about this cyclic and seasonal pattern of the data. Next using five pattern recognition algorithms the data was trained and searched for the method with best accuracy. So the KNN method was identified as the best method. So this method has been used in the predic section of the web application where when a user insert some data about a person it would generate a response by predicting "yes" or "no".

REFERENCES

[SMR14] P. Cortez S. Moro and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems. 2014.