*Machine learning

*Mini Project

Name – Ranvijay Patel Amarnath

Roll No. – 1901156

Email - ranvijay.amarnath@iiitg.ac.in

Topic :

Direct marketing campaigns (via phone calls). Goal is to predict whether the client will subscribe to a term deposit.

Introduction

In this project I will demonstrate how to build a model Predicting client subscribing or not to a term deposit. For that I will use the following steps-

- Project definition
- Data exploration
- Feature engineering
- Building of train/validation/test sample
- Model selection
- Model Evaluation

Relevant Information :

The data is related with direct marketing campaigns of a portugues banking institution. The marketing companies were based on phone calls. More than one contact were required, in order to access if the product (term of deposit) would be (or not) subscribed.

Project definition :

Predict if a client will subscribe (yes / not) to a term deposit. This is defined as a classification problem because either the client will subscribe the term or will not subscribe so we can say that it is classification problem that has two class yes or no(1/0).

Data Exploration :

The data that is used in this project originally comes from the UCI machine learning repository. That is totally based on banking institution. The marketing company were based on phone calls.

So, now we are going to explore our given dataset.

1. Total number of samples/patterns are 41188.
2. Total number of features are 21.
3. One output named “y”.

Attribute information :

Input variables :

1. age : (Numeric)
2. job : type of job : (categorical : blue-collar, technician, management, services, retired, admin, housemaid, unemployed, entrepreneur, self-employed, unknown, student)
3. marital : marital status : (categorical : married, single, divorced, unknown)
4. education : (categorical : basic.4y, unknown, university.degree, high.school, basic.9y, professional.course, basic.6y, illiterate)
5. default : has credit in dafault ? (binary : unknown, no, yes)
6. 6. housing : has housing lone ? (binary : yes, no, unknown)

Related with the last contact of the current campaign :

7. loan : has personal loan ? (binary : no, yes, unknown)

8. contact : contact communication type (categorical : cellular, telephone)

9. month : last contact month of the year (categorical : jan, feb, mar,….. nov, dec)

10. day_of_week : last contact day of week (categorical : mon, tue, wed, thu, fri, sat, sun)

11. Duration : last contact duration (numerical)

12. campaign : Numerical

13. Pdays : Number of days that passed by after client was last contacted(Numeric : -1 means the client was not previously contacted)

14. previous : Number of contact perform before this compaign and for this client (Numeric)

15. poutcome : Outcome of the previous marketing compaign (Categorical : nonexistent, success, failure)

16. emp_var_rate : Numeric data

17. cons_price_idx : Numeric data

18. cons_conf_idx : Numeric data

19. euribor3m : Numeric data

20. nr_employed : Numeric data

21. y : output (binary : 0, 1)

• About Features and Output

- After analyzing the columns we can clearly say that there are two types of columns. Numerical and categorical columns.

- The most important column is y. Which is the output/target variable . This will tell us if the client subscribed to a term deposit or not.(Binary : ‘yes’ or ‘no’).

- So, we can say that the output variable can be used for binary classification. We will try to predict if the client will subscribe to term deposit.

After deeply analyze the columns we see there are a mix of categorical(non-numerical) and numerical data .So, we can say that---

- We have some values for every columns
- age, duration, campaign, pdays, previous, poutcome, emp_var_rate, cons_price_idx, cons_conf_idx, euribor3m, nr_employed are the numerical columns/variables.
- job, marital, education, default, housing, loan, contact, month, day_of_week are the non-numerical/categorical columns/variables.
- Default, housing and loan have three values each(yes, no and unknown)
- Output (y) has only two value "yes" and "No".

Feature Engineering :

Feature Engineering is classifying features such as numerical and categorical features . As we know that our model can not work with the categorical data. So, we have to first convert the categorical data to numerical data by a special technique called **One-HotEncoding** , then we will proceed to train our model.

In this section, we will create features for our predictive model. For each section, we will add new variables to the dataframe and then keep track of which columns of the dataframe we want to use as part of the predictive model features.

We will first extract the numerical and categorical data

Numerical Data

These are the numerical columns that we will use :

- Numerical columns are “campaign, pdays, previous, emp_var_rate, cons_price_idx, cons_conf_idx, nr_employed, age, euribor3m”

Categorical Data

- Categorical variables are non-numeric data such as job and education.
- To turn these non-numerical data into variables, the simplest thing is to use a technique called one-hot encoding, which will be explained below.

- The categorical columns are “job, marital, education, default, housing, loan, contact, month, day_of_week, poutcome”

One-Hot Encoding

- To convert our categorical features to numbers, we will use a technique called one-hot encoding. In one-hot encoding, you create a new column for each unique value in that column. Then the value of the column is 1 if the sample has that unique value or 0 otherwise.

- We will perform this technique for non-numeric data colums to convert it into the numeric data.

- After applying One-Hot Encoding technique there are total of 53 number of new columns generated. So, total number of comuns are 64 now.

- In the next slide we will see all the generated columns….

Total Columns

```
['campaign',
'pdays',
'previous',
'emp.var.rate',
'cons.price.idx',
'cons.conf.idx',
'nr.employed',
'age',
'euribor3m',
'job_admin.',
'job_blue-collar',
'job_entrepreneur',
'job_housemaid',
'job_management',
'job_retired',
'job_self-employed',
'job_services',
'job_student',
'job_technician',
'job_unemployed',
'job_unknown',
'marital_divorced',
'marital_married',
'marital_single',
'marital_unknown',
'education_basic.4y',
'education_basic.6y',
'education_basic.9y',
'education_high.school',
'education_illiterate',
'education_professional.course',
'education_university.degree',
'education_unknown',
'default_no',
'default_unknown',
'default_yes',
'housing_no',
'housing_unknown',
'housing_yes',
'loan_no',
'loan_unknown',
'loan_yes',
'contact_cellular',
'contact_telephone',
'month_apr',
'month_aug',
'month_dec',
'month_jul',
'month_jun',
'month_mar',
'month_may',
'month_nov',
'month_oct',
'month_sep',
'day_of_week_fri',
'day_of_week_mon',
'day_of_week_thu',
'day_of_week_tue',
'day_of_week_wed',
'poutcome_failure',
'poutcome_nonexistent',
'poutcome_success']
```

Building Training, Validation and Test Samples

- In this section we are going to split our dataset into training, validation and test set. The major advantage behind the splitting is that how well our model can perform on the unseen data.

- So, we will split into three part

- Training samples: these are samples from the data set used to train the model. It can be 70% of the data.

- Validation samples: these are samples used to validate or make decisions from the model. It can be 15% of the data.

- Test samples: these are samples used to measure the accuracy or performance of the model. It can be 15% of the data.

- In this project, we will split into 70% train, 15% validation, and 15% test.
- We have successfully split our dataset into 70% of training set, 15% of validation and 15% of testing data.
- Now, we are ready to train our model and calculate the score and accuracy of the model.
- Based on the Accuracy and score we can say that which model is performing well for this datasets.

Model Selection

This section allows us to test various machine learning algorithm to see how our independent variables accurately predict our dependent “y” output variable. We will then select the best model based on performance on the validation set.

we will first compare the performance of the following machine learning models using default hyper parameters:

- Logistic regression
- Stochastic gradient descent
- Single layer perceptron
- Multilayer perceptron

1. Logistic Regression

- Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), can take only discrete values for a given set of features(or inputs).

- Logistic regression uses a soft function called sigmoid/logistic function to predict the output.

- Logistic regression uses a line (Sigmoid function) in the form of an "S" to predict if the dependent variable is true or false based on the independent variables. The "S-shaped" curve (on the line graph) will show the probability of the dependent variable occurring based on where the points of the independent variables lands on the curve.

- In this case, the output (y) is predicted by the numerical and categorical variables such as age, education and so on. Logistic regression is best used for classifying samples.

- Score of Logistic regression model : 0.909841372612496

- logistic regression model
- Training
  - AUC : 0.933
  - Accuracy : 0.910
  - recall : 0.393
  - precision : 0.668
  - specificity : 0.975 f1
  - score : 0.495
  - Confusion Matrix :

```
[[24942   635]
 [ 1974   1280]]
```

- Validation
  - AUC : 0.940
  - Accuracy : 0.915
  - recall : 0.411
  - precision : 0.676
  - specificity : 0.976
  - f1 score : 0.511
  - Confusion Matrix : [[5375 132]
[396 276]]

- Testing
  - AUC : 0.929
  - Accuracy : 0.910
  - recall : 0.380
  - precision : 0.704
  - specificity : 0.979
  - f1 score : 0.493
  - Confusion Matrix : [[5350 114]
[443 271]]

2. Stochastic Gradient Descent

- Stochastic Gradient Descent analyzes various sections of the data instead of the data as a whole and predicts the output using the independent variables. Stochastic Gradient Descent is faster than logistic regression in the sense that it doesn't run the whole dataset but instead looks at different parts of the dataset.

- It uses the log loss function while fitting the model.

- raining and evaluating Stochastic Gradient Descent model performance.

- After the prediction of the model using SGD Algorithm, Following are the performance.

- Score of the Stochastic gradient descent model : 0.8844286176756232

- Stochastic gradient descent model

  - Training
    - AUC : 0.786
    - Accuracy : 0.887
    - recall : 0.000
    - precision : 0.000
    - specificity : 1.000
    - f1 score : nan
    - Confusion Matrix :

      -
        ```
        [[25577     0]
         [ 3254     0]]
        ```

- Validation
  - AUC : 0.788
  - Accuracy : 0.891
  - recall : 0.000
  - precision : 0.000
  - specificity : 1.000
  - f1 score : nan
  - Confusion Matrix : [[5507 0] [672 0]]
- Testing
  - AUC : 0.762
  - Accuracy : 0.884
  - recall : 0.000
  - precision : 0.000
  - specificity : 1.000
  - f1 score : nan
  - Confusion Matrix : [[5464   0]
    [714   0]]

3. Multi layer perceptron model (MLP)

- Multilayer perceptron model comes under the artificial neural network (ANN) .

- A **multilayer perceptron** (**MLP**) is a class of feedforword artificial neural network (ANN).

- MLP utilizes a supervised learning technique called backpropagation for training.

- It uses sigmoid function as activation function.

- Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

- There are mainly three layer in MLP. Input layer, Hidden layer and Output layer.

- The multilayer feed-forward neural network model consists of multilayer of computational units. Each units in one layer known as a perceptron can compute a continuous output using the logistic function by using the outputs of last layer as its inputs.

- So, in the working of the model, First it takes input and there is some weights associated with the neurons, then it perform the model based on some activation function and produce the output.

- After that we calculate the error based on the actual output and predicted output. This technique is called the forward propagation.

- After calculating the error back-propagate to update the weights. Then calculate the accuracy of our model.

Performance of MLP

- Score of the MLP Model : 0.9014244091939139

- Training
  - AUC : 0.989
  - Accuracy : 0.967
  - recall : 0.768
  - precision : 0.925
  - specificity : 0.992
  - f1 score : 0.839
  - Confusion Matrix :
    - [[25374 203]
    - [755 2499]]

- Validation
  - AUC : 0.917
  - Accuracy : 0.901
  - recall : 0.476
  - precision : 0.554
  - Specificity : 0.953
  - f1 score : 0.512
  - Confusion Matrix : [[5249 258]
[352 320]]
- Testing
  - AUC : 0.912
  - Accuracy : 0.901
  - recall : 0.475
  - precision : 0.592
  - specificity : 0.957
  - f1 score : 0.527
  - Confusion Matrix : [[5230 234]
[375 339]]

Comparison of the model :

Comparison of Accuracy Score of Logistic Regression, Stochastic gradient descent and Multilayer Perceptron Model.

Model	Training Data	Validation	Test Data
Logistic Regression	0.91	0.915	0.91
SGD Model	0.887	0.891	0.884
MLP Model	0.967	0.901	0.901

Accuracy in Percentage :

Model	Training Data	Validation	Test Data
Logistic Regression	91%	91.5%	91%
SGD Model	88.7%	89.1%	88.4%
MLP Model	96.7%	90.1%	90.1%

Conclusion

- Through this project, we created a machine learning model that is able to predict how likely clients will subscribe to a bank term deposit.

- The best model was Logistic regression model. The model's performance is 91%. That is highest among all compare to other machine Learning model.

- So, We should focus on targeting customers with high cons.price.idx (consumer price index) as they are high importance features for the model.

*Machine learning

*Mini Project

Name – Ranvijay Patel Amarnath

Roll No. – 1901156

Email - ranvijay.amarnath@iiitg.ac.in

Litrature Survey

Research Paper Survey :

1. http://www.columbia.edu/~jc4133/ADA-Project.pdf

- In this article it uses the following Machine learning model to predict the model.
  - logistic regression
  - neural network
  - random forest
  - KNN model

After training the datasets using these model, these are the following results has been recorded.

1. Logistic regression :

- We practiced 10-fold cross validation to tune parameters for logistic classifier. After we got a tuned model, we tested its performance on the testing set and the accuracy achieved 0.9132.

- Resampling results:

AUC	Sens	Spec	Accuracy	Kappa
0.936	0.973	0.424	0.912	0.471

2. Neural Network (MLP)

- A 5-fold cross-validation uses to select best parameters for neural network

- AUC values are 0.9516 and 0.9423 for training set and test set.

3. Random forest :

AUC for training data is 0.991, for test data is it is 0.9427.

4. KNN :

- AUC score for K- NN is 0.8531.

Comparison :

	Random Forest	Neural Network	k-NN	Logistic Regression
AUC	0.9427*	0.9423	0.8531	0.9364
Test accuracy	0.9136	0.9145*	0.9018	0.9137
FPR at TPR=0.99	0.2642	0.2566*	0.3054	0.3809

(* means this algorithm is the best in the given measurement)

Conclusion :

- MLP is performing best among all these.

- Neural network dominates in two measurements and ranked 2nd in AUC, so it’s the most powerful model.

- In the light of overall test accuracy and AUC, and FPR at TPR=0.99, the best model is neural network. It has the most powerful prediction ability

2. https://research.ijcaonline.org/volume85/number7/pxc3893218.pdf

- In this article, it uses the following machine learning model to train our model and predict the accuracy that how well the model is performing.
  - multilayer perception neural network (MLPNN)
  - tree augmented Naïve Bayes (TAN)
  - logistic regression (LR)
  - decision tree model (C5.0)

Table measurement of MLPNN, TAN, LR AND C5.0

Model	Partition	Accuracy	Sensitivity	Specificity
MLPNN	Training	90.92%	65.66%	93.28%
MLPNN	Testing	90.49%	62.20%	93.12%
TAN	Training	89.16%	55.87%	91.97%
TAN	Testing	88.75%	52.19%	91.73%
LR	Training	90.09%	64.83%	91.76%
LR	Testing	90.43%	65.53%	92.16%
C5.0	Training	93.23%	76.75%	94.92%
C5.0	Testing	93.23%	59.06%	93.23%

Conclusion :

- the specificity measure has C5.0 with the highest values in training samples 94.92% and 93.23% for testing samples. From the previews, C5.0 is the best in accuracy, sensitivity, and specificity analysis of training sample.

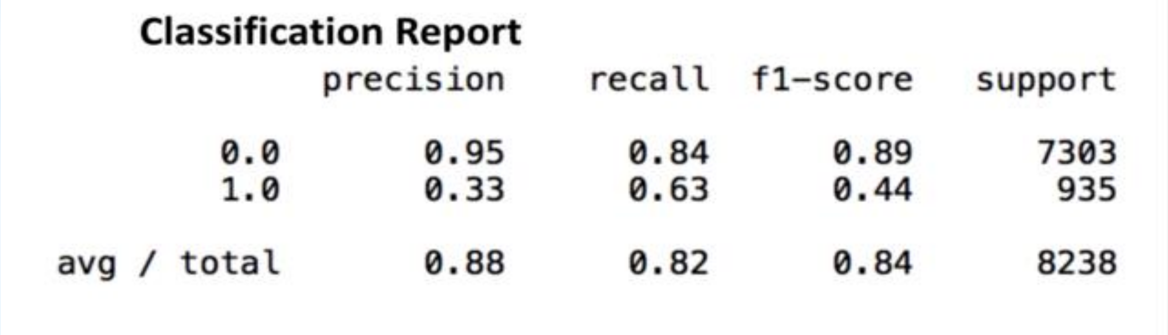
- However, the MLPNN is the best for accuracy; LR takes the best percentage for sensitivity, and C5.0 return to be the best in specificity analysis of testing samples.

3. https://medium.com/analytics-vidhya/term-deposit-opening-decision-f57776b88453

- In this article it uses thee model to train our model.
  - Logistic regression
  - K-NN
  - Random forest

1. Logistic Regression

Accuracy score : 0.815



Classification Report

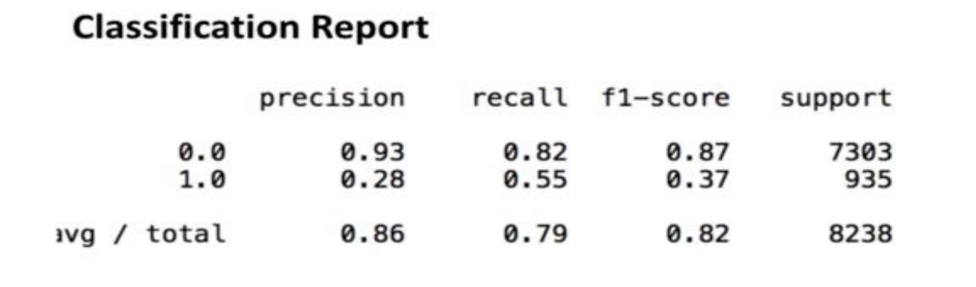
```
             precision    recall  f1-score   support

        0.0       0.95      0.84      0.89      7303
        1.0       0.33      0.63      0.44       935

avg / total       0.88      0.82      0.84      8238
```

2. K-NN :

- Accuracy Score = 0.789



Classification Report

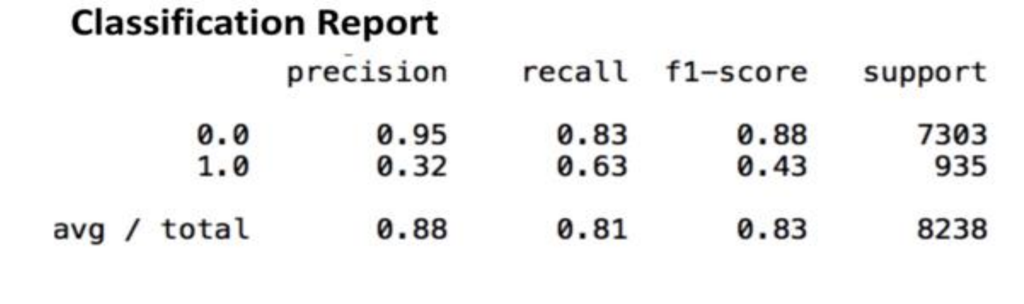
```
             precision    recall  f1-score   support

        0.0       0.93      0.82      0.87      7303
        1.0       0.28      0.55      0.37       935

avg / total       0.86      0.79      0.82      8238
```

3. Random forest :

Accuracy Score : 0.807



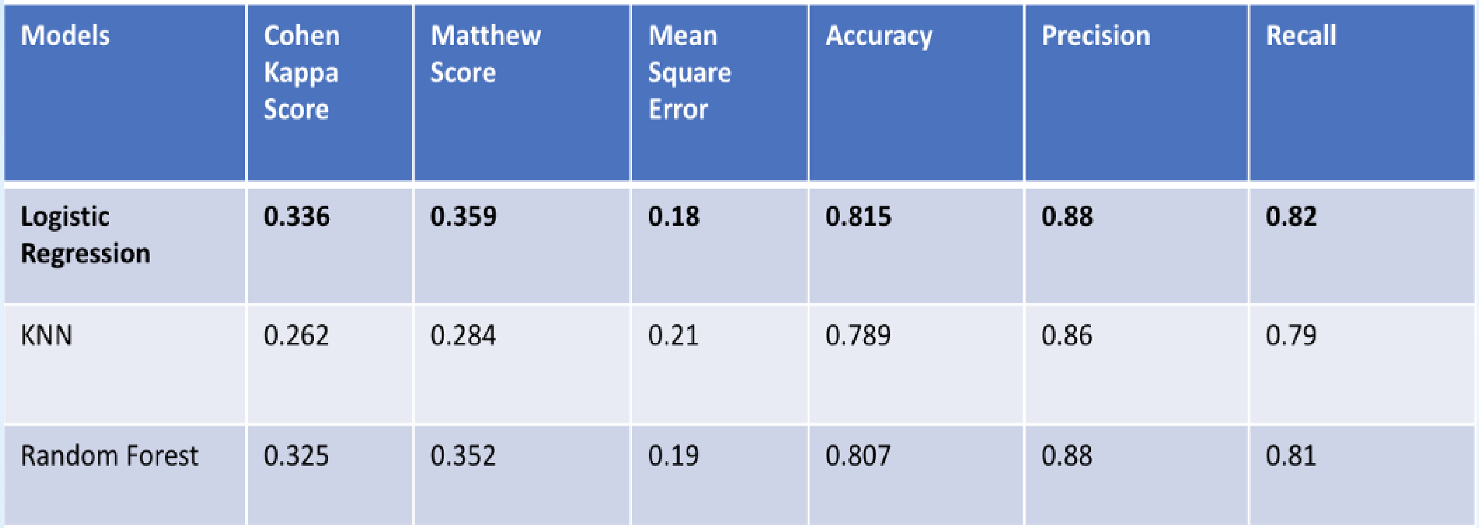
Classification Report

```
             precision    recall  f1-score   support

        0.0       0.95      0.83      0.88      7303
        1.0       0.32      0.63      0.43       935

avg / total       0.88      0.81      0.83      8238
```

Summery :



Models	Cohen Kappa Score	Matthew Score	Mean Square Error	Accuracy	Precision	Recall
Logistic Regression	**0.336**	**0.359**	**0.18**	**0.815**	**0.88**	**0.82**
KNN	0.262	0.284	0.21	0.789	0.86	0.79
Random Forest	0.325	0.352	0.19	0.807	0.88	0.81

- After according to the table we can say that the Logistic regression model is the best fit model among all three model.

Final Conclusion :

- After Analysing all three articals
- In first article MLP is performing the best out of all.
- According to second article the random forest is best.
- According to 3rd the MLP is the best predictor.