Topic: AdaBoost- Extreme Gradient Boosting

Instructions

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

Name: Kavali Rakesh

Batch Id: DSWDMCON 18012022

Topic: Ensemble Techniques.

1. Business Problem

1.1. Objective

1.2. Constraints (if any)

2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.

Using R and Python codes perform:

3. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

3.2 Outlier Imputation

4. Exploratory Data Analysis (EDA):

4.1. Summary

4.2. Univariate analysis

4.3. Bivariate analysis

5. Model Building

5.1 Build the model on the scaled data (try multiple options)

5.2 Perform Bagging, Boosting, Voting, Stacking on given datasets

5.3 Train and Test the data, use grid search cross validation, compare accuracies using confusion matrix

5.4 Briefly explain the model output in the documentation

6. Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided

Note:

The assignment should be submitted in the following format:

- **R code**
- **Python code**
- **Code Modularization should be maintained**
- **Documentation of the model building (elaborating on steps mentioned above)**

Problem Statement: -

Instructions

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

Name: U Mukalingam

Batch Id: DSWDMCOH 190321

Topic: Ensemble Techniques.

7. Business Problem

7.1. Objective

7.2. Constraints (if any)

8. Work on each feature of the dataset to create a data dictionary as displayed in the below image:

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.

Using R and Python codes perform:

9. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

3.2 Outlier Imputation

10. Exploratory Data Analysis (EDA):

10.1. Summary

10.2. Univariate analysis

10.3. Bivariate analysis

11. Model Building

5.5 Build the model on the scaled data (try multiple options)

5.6 Perform Bagging Boosting (adaboost, fastadaboost, Xgboost), Stacking, Voting on the given datasets in Hands on Material

5.7 Train and Test the data and compare accuracies by Confusion Matrix and use different Hyper Parameters and also use GridSearchCV to improve your model performance

5.8 Briefly explain the model output in the documentation

12. Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided

Note:

The assignment should be submitted in the following format:

- **R code**
- **Python code**
- **Code Modularization should be maintained**
- **Documentation of the model building (elaborating on steps mentioned above)**

Problem Statement: -

Diabetes is disease caused by increase in Blood glucose levels in your body, Blood glucose is the main source of energy and it gets from the food you eat. Insulin a hormone, made by the pancreas helps to get glucose from your food reach every cell of your body for energy. Sometimes, your body does not make enough or use Insulin at all, due to this the glucose levels in your body increases which leads to severe health problems and moreover diabetes has no cure, you can only avoid taking sugar filled foods and take precautions. In Pregnant women's the rising of glucose levels is a

danger for to be mother and the baby, if we can predict accurately whether a pregnant women will become diabetic or not can be help doctors to treat patients in a much efficient way and also for

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38.0	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
16	7	100	0	0	0	30.0	0.484	32	1
17	0	118	84	47	230	45.8	0.551	31	1

pregnant women can avoid becoming diabetic by taking necessary steps during pregnancy. Build an ensemble model to correctly classify the outcome variable and improve your model prediction by using GridSearchCV. You must apply Bagging, Boosting, Stacking and Voting on the dataset.

Sol:

Business Objective: To predict the diabetes disease for the new patient using the ensemble techniques.

Constraints: Lack of analysis of the patient's data.

Data Types: given data and its data types are shown below

Name of feature	Description	Data type	Relevance
Number of times pregnant	Number of time the woman got pregnancy	Internal	Relevant
Plasma glucose concentration	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Ratio	Relevant
Diastolic blood pressure	Diastolic blood pressure (mm Hg)	Ratio	Relevant
Triceps skin fold thickness	Triceps skin fold thickness (mm)	Ratio	Relevant
2-Hour serum insulin	2-Hour serum insulin (mu U/ml)	Ratio	Relevant
Body mass index	Body mass index (weight in kg/(height in m)^2)	Ratio	Relevant
Diabetes pedigree function	Diabetes pedigree function	Ratio	Relevant
Age (years)	Age of patient	Ratio	Relevant
Classvariable	Variable for patiest having diabetes or not	Nominal	Relevant

Data Pre-Processing: the complete data can be used for preparing the models of Ensemble Techniques.

Ensemble Models: I have segregated the data into training and testing data set as 70% and 30% of the total data respectively. I have applied all the ensemble techniques of voting, bagging, boosting and stacking in all the techniques I have got the accuracy of approximately 75% for all the techniques.

Problem Statement: -

Most cancers form a lump called as tumour or a growth. But not all lumps are cancer. Doctors take out a piece of the lump and look at it to find out if it's cancer. Lumps that are not cancer are called benign (be-NINE). Lumps that are cancer are called malignant (muh-LIG-nunt). There are some cancers, like leukemia (cancer of the blood), that don't form tumour. They grow in the blood cells or other cells of the body. For instance, If a doctor tends to wrongly diagnose a benine tumour as a malignant tumour can a cause a overwhelming anxiety in patient which can lead to depression or much worse, a wrong diagnosis is a major problem in our health care sector, to improve their analysis build an ensemble model on the dataset which can accurately classify benine and Malignant tumours on the dataset given. Perform Bagging, Boosting, Stacking, Voting algorithm and provide your insights in the documentation.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity
1	87139402	B	12.320	12.39	78.85	464.1	0.10280	0.06981	0.039870
2	8910251	B	10.600	18.95	69.28	346.4	0.09688	0.11470	0.063870
3	905520	B	11.040	16.83	70.92	373.2	0.10770	0.07804	0.030460
4	868871	B	11.280	13.39	73.00	384.8	0.11640	0.11360	0.046350
5	9012568	B	15.190	13.21	97.65	711.8	0.07963	0.06934	0.033930
6	906539	B	11.570	19.04	74.20	409.7	0.08546	0.07722	0.054850
7	925291	B	11.510	23.93	74.52	403.5	0.09261	0.10210	0.111200
8	87880	M	13.810	23.75	91.56	597.8	0.13230	0.17680	0.155800
9	862989	B	10.490	19.29	67.41	336.1	0.09989	0.08578	0.029950
10	89827	B	11.060	14.96	71.49	373.9	0.10330	0.09097	0.053970
11	91485	M	20.590	21.24	137.80	1320.0	0.10850	0.16440	0.218800
12	8711003	B	12.250	17.94	78.27	460.3	0.08654	0.06679	0.038850
13	9113455	B	13.140	20.74	85.98	536.9	0.08675	0.10890	0.108500
14	857810	B	13.050	19.31	82.61	527.2	0.08060	0.03789	0.000692
15	9111805	M	19.590	25.00	127.70	1191.0	0.10320	0.09871	0.165500
16	925277	B	14.590	22.68	96.39	657.1	0.08473	0.13300	0.102900

Sol:

Business Objective: To predict the cancer disease for the new patient using the ensemble techniques.

Constraints: Lack of analysis of the patient's data.

Data Types: the complete data is about the cancer patients and it effects and measure in the body.

Data Pre-Processing: I have identified that id could not be used for the analysis on the data so I have dropped the column and remaining all the columns are used for the analysis.

Ensemble Models: I have segregated the data into training and testing data set as 70% and 30% of the total data respectively. I have applied all the ensemble techniques of voting, bagging, boosting and stacking in all the techniques I have got the accuracy of approximately 73% for all the technique's.

Problem Statement: -

A sample of global companies and their ratings are given for the cocoa bean production along with the location of bean being used by the companies. Identify the important features in the analysis and accurately classify the companies based on their ratings and draw insights from your model. Perform Ensemble methodology such as Bagging, Boosting, Stacking, voting algorithms on the dataset given.

Company	Name	REF	Review	Cocoa_Percent	Company_Locat	Rating	Bean_Type	Origin
A. Morin	Agua Grande	1876	2016	63%	France	3.75		Sao Tome
A. Morin	Kpime	1676	2015	70%	France	2.75		Togo
A. Morin	Atsane	1676	2015	70%	France	3		Togo
A. Morin	Akata	1680	2015	70%	France	3.5		Togo
A. Morin	Quilla	1704	2015	70%	France	3.5		Peru
A. Morin	Carenero	1315	2014	70%	France	2.75	Criollo	Venezuela
A. Morin	Cuba	1315	2014	70%	France	3.5		Cuba
A. Morin	Sur del Lago	1315	2014	70%	France	3.5	Criollo	Venezuela
A. Morin	Puerto Cabello	1319	2014	70%	France	3.75	Criollo	Venezuela
A. Morin	Pablino	1319	2014	70%	France	4		Peru
A. Morin	Panama	1011	2013	70%	France	2.75		Panama
A. Morin	Madagascar	1011	2013	70%	France	3	Criollo	Madagascar
A. Morin	Brazil	1011	2013	70%	France	3.25		Brazil
A. Morin	Equateur	1011	2013	70%	France	3.75		Ecuador
A. Morin	Colombie	1015	2013	70%	France	2.75		Colombia
A. Morin	Birmanie	1015	2013	70%	France	3		Burma
A. Morin	Papua New Guir	1015	2013	70%	France	3.25		Papua New Guin
A. Morin	Chuao	1015	2013	70%	France	4	Trinitario	Venezuela

Sol:

Business Objective: Segregating the companies data based on the ratings.

Constraints: Lack of analysis of the companie's data.

Data Types: given data and its data types are shown below

Name of feature	Description	Data type	Relevance
Company	Name of the comany	Nominal	Relevant
Name	Name of the person	Nominal	Relevant
REF	Reference id	Nominal	Relevant
Review	Review year	Ordinal	Relevant
Cocoa_Percent	Cocoa percentage in the product	Ratio	Relevant
Company_Location	Location of the company	Nominal	Relevant
Ratings	Ratings given by the company	Ratio	Relevant
Bean_Type	Type of bean used	Nominal	Relevant
Origin	Origin of the company	Nominal	Relevant

Data Pre-Processing: I have identified the categorical data in the data given for which I have converted numeric data by using factor function in R and label encoding in Python. For this data to do the analysis I have segregated the data based on ratings for which I have considered the ratings as high if it is greater than 3 and others as less respectively.

Ensemble Models: I have segregated the data into training and testing data set as 70% and 30% of the total data respectively for doing the analysis. I have applied all the ensemble techniques of voting, bagging, boosting and stacking in all the techniques I have got the accuracy of approximately 70% for all the techniques.

Problem Statement: -

Data privacy has been and is always an important factor that websites are very critical about, to safe guard their customers details from ethical hackers and other unsolicited misuse of data, users are required to use alpha numeric characters while creating their account for the first time for password strength. Perform Ensemble technique to classify the user's password strength of users, use Bagging, Boosting, Stacking, voting algorithms on the dataset given.

characters	characters_strength
kzde5577	1
kino3434	1
visi7k1yr	1
megzy123	1
lamborghin1	1
AVYq1IDE4MgA	1
u6c8vhow	1
v1118714	1
universe2908	1
as326159	1
asv5o9yu	1
612035180tok	1
jytifok873	1
WUt9IZzE0OQ7	1
jerusalem393	1
g067057895	1
52558000aaa	1
idofo673	1

Sol:

Business Objective: Segregating the passwords of the user as strong and week.

Constraints: Lack of analysis of the password data.

Data Types: given data and its data types are shown below

Name of feature	Description	Data type	Relevance
Characters	Password of the user	Nominal	Relevant
Password strength	Password strength of the user	Nominal	Relevant

Data Pre-Processing: I have segregated the password strength as strong and weak based on the characters used by the user and the final data is used for doing the analysis.

Ensemble Models: I have segregated the data into training and testing data set as 70% and 30% of the total data respectively for doing the analysis. I have applied all the ensemble techniques of voting, bagging, boosting and stacking in all the techniques I have got the accuracy of approximately 68% for all the techniques.