02 : 02 : 55 : 45
DAY HRS MIN SEC

Predict the Criminals

LIVE

Dec 20, 2017, 06:00 PM IST - Apr 01, 2018, 10:30 PM IST

Predict the Criminals

Max. Marks: 1

Problem Statement

There has been a surge in crimes committed in recent years, making crime a top cause of concern for law enforcement. If we are able to estimate whether someone is going to commit a crime in the future, we can take precautions and be prepared. You are given a dataset containing answers to various questions concerning the professional and private lives of several people. A few of them have been arrested for various small and large crimes in the past. Use the given data to predict if the people in the test data will commit a crime. The train data consists of 45718 rows, while the test data consists of 11430 rows.

(Update - 6 March, 2018)

Please note the evaluation metric has been changed from Precision to Mathews Correlation Coefficient due to ineffectiveness of Precision metric in this dataset.

The evaluation metric is Matthews correlation coefficient

Download Dataset

Join our slack channel to discuss ML and DL here.

Data Description

You are given three files to download: train, test and sample submission.

Variable Name	Description
PERID	Person ID
IFATHER	FATHER IN HOUSEHOLD
NRCH17_2	RECODED # R's CHILDREN < 18 IN HOUSEHOLD
IRHHSIZ2	RECODE - IMPUTATION-REVISED # PERSONS IN HH

NO of children under 18

Variable Name	Description
IIHHSIZ2	IMPUTATION INDICATOR
IRKI17_2	IMPUTATION-REVISED # KIDS AGED<18 IN HH
IIKI17_2	IRKI17_2-IMPUTATION INDICATOR
IRHH65_2	REC - IMPUTATION-REVISED # OF PER IN HH AGED>=65
IIHH65_2	IRHH65_2-IMPUTATION INDICATOR
PRXRETRY	SELECTED PROXY UNAVAILABLE, OTHER PROXY AVAILABLE?
PRXYDATA	IS PROXY ANSWERING INSURANCE/INCOME QS
MEDICARE	COVERED BY MEDICARE
CAIDCHIP	COVERED BY MEDICAID/CHIP
CHAMPUS	COV BY TRICARE, CHAMPUS, CHAMPVA, VA, MILITARY
PRVHLTIN	COVERED BY PRIVATE INSURANCE
GRPHLTIN	PRIVATE PLAN OFFERED THROUGH EMPLOYER OR UNION
HLTINNOS	COVERED BY HEALTH INSUR
HLCNOTYR	ANYTIME DID NOT HAVE HEALTH INS/COVER PAST 12 MOS
HLCNOTMO	PAST 12 MOS, HOW MANY MOS W/O COVERAGE
HLCLAST	TIME SINCE LAST HAD HEALTH CARE COVERAGE
HLLOSRSN	MAIN REASON STOPPED COVERED BY HEALTH INSURANCE
HLNVCOST	COST TOO HIGH
HLNVOFFR	EMPLOYER DOESN'T OFFER
HLNVREF	INSURANCE COMPANY REFUSED COVERAGE
HLNVNEED	DON'T NEED IT
HLNVSOR	NEVER HAD HLTH INS SOME OTHER REASON
IRMCDCHP	IMPUTATION REVISED CAIDCHIP
IIMCDCHP	MEDICAID/CHIP - IMPUTATION INDICATOR
IRMEDICR	MEDICARE - IMPUTATION REVISED
IIMEDICR	MEDICARE - IMPUTATION INDICATOR
IRCHMPUS	CHAMPUS - IMPUTATION REVISED
IICHMPUS	CHAMPUS - IMPUTATION INDICATOR
IRPRVHLT	PRIVATE HEALTH INSURANCE - IMPUTATION REVISED

?

Variable Name	Description
IIPRVHLT	PRIVATE HEALTH INSURANCE - IMPUTATION INDICATOR
IROTHHLT	OTHER HEALTH INSURANCE - IMPUTATION REVISED
IIOTHHLT	OTHER HEALTH INSURANCE - IMPUTATION INDICATOR
HLCALLFG	FLAG IF EVERY FORM OF HEALTH INS REPORTED
HLCALL99	YES TO MEDICARE/MEDICAID/CHAMPUS/PRVHLTIN
ANYHLTI2	COVERED BY ANY HEALTH INSURANCE - RECODE
IRINSUR4	RC-OVERALL HEALTH INSURANCE - IMPUTATION REVISED
IIINSUR4	RC-OVERALL HEALTH INSURANCE - IMPUTATION INDICATOR
OTHINS	RC-OTHER HEALTH INSURANCE
CELLNOTCL	NOT A CELL PHONE
CELLWRKNG	WORKING CELL PHONE
IRFAMSOC	FAM RECEIVE SS OR RR PAYMENTS - IMPUTATION REVISED
IIFAMSOC	FAM RECEIVE SS OR RR PAYMENTS - IMPUTATION INDICATOR
IRFAMSSI	FAM RECEIVE SSI - IMPUTATION REVISED
IIFAMSSI	FAM RECEIVE SSI - IMPUTATION INDICATOR
IRFSTAMP	RESP/OTH FAM MEM REC FOOD STAMPS - IMPUTATION REVISED
IIFSTAMP	RESP/OTH FAM MEM REC FOOD STAMPS - IMPUTATION INDICATOR
IRFAMPMT	FAM RECEIVE PUBLIC ASSIST - IMPUTATION REVISED
IIFAMPMT	FAM RECEIVE PUBLIC ASSIST - IMPUTATION INDICATOR
IRFAMSVC	FAM REC WELFARE/JOB PL/CHILDCARE - IMPUTATION REVISED
IIFAMSVC	FAM REC WELFARE/JOB PL/CHILDCARE - IMPUTATION INDICATOR
IRWELMOS	IMP. REVISED - NO.OF MONTHS ON WELFARE
IIWELMOS	NO OF MONTHS ON WELFARE - IMPUTATION INDICATOR
IRPINC3	RESP TOT INCOME (FINER CAT) - IMP REV
IRFAMIN3	RECODE - IMP.REVISED - TOT FAM INCOME
IIPINC3	RESP TOT INCOME (FINER CAT) - IMP INDIC
IIFAMIN3	IRFAMIN3 - IMPUTATION INDICATOR
GOVTPROG	RC-PARTICIPATED IN ONE OR MORE GOVT ASSIST PROGRAMS
POVERTY3	RC-POVERTY LEVEL

?

Variable Name	Description
TOOLONG	RESP SAID INTERVIEW WAS TOO LONG
TROUBUND	DID RESP HAVE TROUBLE UNDERSTANDING INTERVIEW
PDEN10	POPULATION DENSITY 2010
COUTYP2	COUNTY METRO/NONMETRO STATUS
MAIIN102	MAJORITY AMER INDIAN AREA INDICATOR FOR SEGMENT
AIIND102	AMER INDIAN AREA INDICATOR
ANALWT_C	FIN PRSN-LEVEL SIMPLE WGHT
VESTR	ANALYSIS STRATUM
VEREP	ANALYSIS REPLICATE
Criminal	Target Variable

Upload Prediction File

Please upload the prediction file in the format as stated in the problem.

Choose file No file chosen

Submit & Evaluate

Upload Source Files

You need to submit a zip or tar archive consisting of a text file explaining your approach, details about feature engineering, tools you used and the relevant source files.

Choose file No file chosen

Upload

COMMENTS (175) SORT BY: Relevance

Join Discussion...

Cancel Post

Nirmala B 3 months ago

@admin, Can you please explain me some of the variables. For instance, what does the IFATHER represents?

▲ 3 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

It shows weather the father of the interviewee is present in the household.

▲ 0 votes ● Reply ● Message ● Permalink

Nirmala B 3 months ago

But it has values of range -1 to 4. So what it represents?. Kindly explain.

▲ 7 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Any -1 values represent missing data. The values represent different things such as father is in the household, father is not, do not know whether father is in household, skipped the question, etc.

▲ 0 votes ● Reply ● Message ● Permalink

Anandakrishnan Rajaram 19 days ago

Can we get a detailed data dictionary? so 1-father in the household,2-father is not,3-do not know,4-skipped the question,-1-missing data?

▲ 2 votes ● Reply ● Message ● Permalink

Sunny Prajapati 3 months ago

then the number should be binary right?

▲ 1 vote ● Reply ● Message ● Permalink

Sreeram TP 3 months ago

@admin, please explain this.!!

▲ 2 votes ● Reply ● Message ● Permalink

Nirmala B 3 months ago

Its been 3 days I am waiting for a response from admin team. Would appreciate if the clarification is given atleast now.

▲ 12 votes ● Reply ● Message ● Permalink

Hari Prasath N 3 months ago

Yeah. Al-most all of the features are unclear.

How so many success submission with this feature description :?

▲ 6 votes ● Reply ● Message ● Permalink

Sunny Prajapati 3 months ago

Hi Admin,

There are many coders here who want to have more description about variables and their values. Please provide us some explanation about values of features as well. That would be help us to understand the data.

▲ 14 votes ● Reply ● Message ● Permalink

Jasneet Singh Sawhney Edited 3 months ago

Please give more details about the features of this data, it is becoming very arduous to understand it.

▲ 3 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Any specific attribute?

▲ 0 votes ● Reply ● Message ● Permalink

Jayaram G 3 months ago

Lots of them need explanation.

To start with: ANALWT_C, VESTR and VEREP??

▲ 3 votes ● Reply ● Message ● Permalink

?

Abhishikth Sagar 3 months ago

ANALWT_C is a sample weight while the other two are estimation variables.

▲ 0 votes ● Reply ● Message ● Permalink

Apurva Khemka 2 months ago

TROUBUND, TOOLONG
has 1,2 98,99 as values. Many features having values like , 1,2,3, 98,99. so should values as 98, 99 be considered as outliers.

▲ 6 votes ● Reply ● Message ● Permalink

Siddharth 3 months ago

why it is saying always : File does not contain prediction for 10060480
can any one help me?

▲ 3 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Make sure that number is present in your PERID field.

▲ 0 votes ● Reply ● Message ● Permalink

Siddharth Edited 3 months ago

i am getting PERDI for 10060480 pred but still getting the same message..should i mail you my code?

▲ 3 votes ● Reply ● Message ● Permalink

Dan Ofer 3 months ago

Same issue, "10060480", binary predictions

▲ 2 votes ● Reply ● Message ● Permalink

Gaurav Chavan 3 months ago

make sure it does not contain quotes " " or ' '

▲ 0 votes ● Reply ● Message ● Permalink

Raman Goyal 3 months ago

it also happened to me bz i was submitting proba_ rather than binary value to criminal

▲ 4 votes ● Reply ● Message ● Permalink

Swapnil P Bhosale 3 months ago

I am getting the same error inspite of submitting binary values.

▲ 0 votes ● Reply ● Message ● Permalink

Amogh Badugu 3 months ago

Make sure the Criminal column dtype is int

▲ 0 votes ● Reply ● Message ● Permalink

Farhan Jailani 3 months ago

What if the Person Satisfies all the probabilities of being a criminal but has a pure heart and is not a criminal >D

▲ 8 votes ● Reply ● Message ● Permalink

Suraj Bonagiri a month ago

Humourous. But for newbies and ppl who took this seriously, remember that this is a classification problem and any classification model gives only the probability of a person being a criminal.

▲ 0 votes ● Reply ● Message ● Permalink

Supriyo Banerjee 3 months ago

Train, validation, and 1st level testing data seem highly imbalanced. In these cases sensitivity rather than accuracy should be the measure. I got 93.305% accuracy just by submitting all zeroes. But did I predict even one criminal ? No !

So maybe in the future you need to re think on how to evaluate a model if the classes are imbalanc ?

● Reply ● Message ● Permalink

4 votes

Abhishikth Sagar 3 months ago

It is an imbalanced classification problem. We are not using accuracy score. The evaluation metric is precision score.

2 votes ● Reply ● Message ● Permalink

Supriyo Banerjee Edited 3 months ago

Then I should get zero precision when I am submitting an all 'zero' file. Since precision is - (no of 1's predicted correctly) divided by (no of ones predicted). But I got 93% score. Correct me if i am wrong here. :)

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

It is because the average is set to 'micro'. You would get a 0 if the average was 'binary'.

0 votes ● Reply ● Message ● Permalink

Shiva Katepally 2 months ago

Can you exactly differentiate precision with average set to'micro' and accuracy score? Precision with average set to 'micro' is fancy alternative to accuracy score in binary classification. In what way evaluating model with a precision score with average set to 'micro' differs from evaluating model with accuracy score?

0 votes ● Reply ● Message ● Permalink

Nishant sethi 3 months ago

could you please give details of how to submit the result for evaulation

1 vote ● Reply ● Message ● Permalink

Shubham Naik Edited 3 months ago

Bro after performing predictions on file called as criminal test.csv,see there is a sample submission file which tells you to submit in
This format
Perid,criminal
So just open new Excel sheet copy column one from ur predicted file paste it in ur new file then copy predicted criminal values(0s or 1s) and paste it in column 2 of ur new file.press save and submit that file by choosing prediction file option in submission box.
U also have to submit ur approach and source file by compressing it in .zip or .tar format in other option of submission window.

4 votes ● Reply ● Message ● Permalink

Nishant sethi 3 months ago

Thanks

0 votes ● Reply ● Message ● Permalink

Manish Kumar Jha 3 months ago

I do not see how just having the options marked in a survey is helpful. Of course the problem is still doable, but the expectation that we should not apply any common sense to figure out the meaning of variables, and rely only on the data is absurd. Please provide the proper meaning of all the variables and what does the data contained in each of them signifies.

4 votes ● Reply ● Message ● Permalink

Prashant Kikani a month ago

Any one can achieve place in top 100 in just 20 minutes. Just like me. Just few things do like regularization, labeling, removing columns etc. It will boost your score 0.9535+. At last, don't forget to do ensembling !!

Best luck to all newbies like me.. :)

3 votes ● Reply ● Message ● Permalink

Suraj Bonagiri a month ago

Hey,
Do you mind expanding little more on the terms "labeling", "ensembling"and "removing colour ?

Newbie here. Like, newbie newbie. :P
I know one could google it but I'd like to know how you used it in this problem. So that I can get a feel.
Thanx man!!
▲ 0 votes ● Reply ● Message ● Permalink

Rahul Kumar a month ago
Bro you have used all the techniques that a pro use.
▲ 1 vote ● Reply ● Message ● Permalink

Nikhil Kumar Singh 3 months ago
Output should contain probabilities or {0,1} ?
▲ 1 vote ● Reply ● Message ● Permalink

BalajiVarsh 3 months ago
It should contain {0,1} not the probabilities.
▲ 2 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago
The output should consist of {0,1}.
▲ 0 votes ● Reply ● Message ● Permalink

Partha pratim Neog 3 months ago
Which average is being used while calculating precision_score? Macro? Micro? Weighted? None?
▲ 1 vote ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago
Binary
▲ 1 vote ● Reply ● Message ● Permalink

Partha pratim Neog 3 months ago
What is Binary? in sklearn the function precision_score takes values "macro" "micro" "weighted" and "None" for the parameter "average"
▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago
Binary is the default parameter for average.
▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago
The new average is micro.
▲ 1 vote ● Reply ● Message ● Permalink

Hemanth Bannu 2 months ago
precision_score with micro is exactly equal to accuracy scorce in binary classfication.
▲ 0 votes ● Reply ● Message ● Permalink

Abhijay Vuyyuru a month ago
Gareebon ke Kaggle, 100 USD ka kya karein?
▲ 3 votes ● Reply ● Message ● Permalink

DEEPAK AHIRE 3 months ago
@Admin, it's taking more time than usual to evaluate
▲ 2 votes ● Reply ● Message ● Permalink

Sangarshanan Veera 3 months ago
could you please post the sample code asap
▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago
We shall post it soon.

?

0 votes ● Reply ● Message ● Permalink

Soumya Shankar Banerjee 3 months ago

...still waiting for your sample code

2 votes ● Reply ● Message ● Permalink

Sreeram TP 2 months ago

Competition have been up for about 60days now.. Should we wait for the starter code.??

1 vote ● Reply ● Message ● Permalink

Kevin Dudeja a month ago

ye abhishikth launde se na ho paye

1 vote ● Reply ● Message ● Permalink

Lalit Mohan Sharma Edited a month ago

@admin PRXRETRY ,SELECTED PROXY UNAVAILABLE, OTHER PROXY AVAILABLE? What does it mean? Features are unclear,please give more description.

0 votes ● Reply ● Message ● Permalink

AnandKumar a month ago

https://www.icpsr.umich.edu/icpsrweb/NAHDAP/data/variables.jsp

Here you can refer the meaning of the features provided in the problem. Search the feature in this site

2 votes ● Reply ● Message ● Permalink

Kunapuli seshadri sastry 23 days ago

admin can you explain why score are changed overnight

1 vote ● Reply ● Message ● Permalink

Manimaran P 23 days ago

Change of eval metric

1 vote ● Reply ● Message ● Permalink

Kunapuli seshadri sastry 23 days ago

what are you using now as new evaluation metric

1 vote ● Reply ● Message ● Permalink

Manimaran P 22 days ago

Matthews correlation coefficient see Problem statement

1 vote ● Reply ● Message ● Permalink

Surendra Pratap Singh 21 days ago

@admin, submissions are getting queued up, with no progress. Some dude has solution stuck for the last 6 hours. Please check

2 votes ● Reply ● Message ● Permalink

Nagavelu V S 3 months ago

Any starter code

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

We will post a starter code soon.

1 vote ● Reply ● Message ● Permalink

Nethaji Chowdary 3 months ago

What is this??

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

This is a supervised machine learning problem.

1 vote ● Reply ● Message ● Permalink

Partha pratim Neog 3 months ago

It looks like a lot of people have scored 1 and will do that by the time the contest ends. How will they be ranks work among the people who score 1? Will the first person to score 1 always be on top? Please explain.

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar Edited 3 months ago

There is something wrong with the evaluation. We are looking into this. No one has scored 1 yet. Will update the leader board to reflect correct scores soon.

▲ 1 vote ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

This has been fixed.

▲ 0 votes ● Reply ● Message ● Permalink

dust 3 months ago

what does the value -1 indicate in features ??

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

You can assume them as missing values.

▲ 1 vote ● Reply ● Message ● Permalink

Harshal Patil 3 months ago

which technique is suitable- classification or regression

▲ 0 votes ● Reply ● Message ● Permalink

Puneeth 3 months ago

@Harshal Patil this is a classification problem.Here the target variable is a categorical variable(0 - is not a criminal and 1 - is a criminal).

▲ 0 votes ● Reply ● Message ● Permalink

Harshal Patil 3 months ago

Thanks Puneeth ..i used logistic regression but its functionality is not working properly so which algorithm suits.

▲ 0 votes ● Reply ● Message ● Permalink

Puneeth 3 months ago

@Harshal Patil there are many algorithms you can use. If you use python, sklearn provides many algorithms like ensembles, svm(Support Vector Machines), etc

▲ 1 vote ● Reply ● Message ● Permalink

pratyush sarangi 3 months ago

The features are quite unclear..can someone explain those..like the imputation indicator..

▲ 1 vote ● Reply ● Message ● Permalink

Shubham Naik 3 months ago

Imputation indicator is the indicator of the source of the imputation revised data. Eg: Questionare, General Statistics, etc
Imputation revised is when the input data has been merged with data from the imputation source which is represented by the Imputation indicator.
Recoded variables are recoded variables that were created from one or more of the edited or imputed variables from the preceding sections.

▲ 0 votes ● Reply ● Message ● Permalink

Sandeep Khan 3 months ago

Please someone provide more detailed description of the variables. For a starter, like myself, it's very hard to make out what is what

▲ 1 vote ● Reply ● Message ● Permalink

Swapnil P Bhosale 3 months ago

Submission ID: 14029367
Error : File does not contain prediction for 10060480
The submission file contains PERID 10060480 and its binary value. I am not submitting proba.
Please help me resolve this issue as soon as possible.

▲ 1 vote ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Send me your file.

▲ 0 votes ● Reply ● Message ● Permalink

pengoo yahoo 3 months ago

CELLNOTCL NOT A CELL PHONE,
what this feature means in context to the dataset ?? please anyone.

▲ 1 vote ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Do they possess a phone which is not a cell phone

▲ 0 votes ● Reply ● Message ● Permalink

Jasneet Singh Sawhney 3 months ago

Please provide a beginner notebook explaining all the features and thier classifiaction values, otherwise it would be like letting loose an arrow in dark.

▲ 1 vote ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Will do it soon.

▲ 0 votes ● Reply ● Message ● Permalink

Debadri Dutta 3 months ago

In a few columns there are values like 1,2,3, and then 99,98. Are these outliers ? And please we need a bit more explanation of the features.

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Will post a detailed notebook soon.

▲ 0 votes ● Reply ● Message ● Permalink

subaashini krishnamoorthy a month ago

Hi
Is the notebook available now describing the values given for each feature?

▲ 1 vote ● Reply ● Message ● Permalink

Saptarshi Baisya 3 months ago

New to all this types of challanges can anyone help me simplifiying the Problem statements

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

You basically have to predict weather a person is a criminal or not by their answers to various questions.

▲ 1 vote ● Reply ● Message ● Permalink

Vishal Kumar chaudhary 3 months ago

my file does contain prediction for 10060480 but it is saying that prediction for this is not there in file

▲ 1 vote ● Reply ● Message ● Permalink

Ritu raj 2 months ago

Hllo sir everyone do you have comment but why don't you give the rply

▲ 1 vote ● Reply ● Message ● Permalink

?

Kevin Dudeja 2 months ago

why doesn't my rank update with my new submissions?

▲ 1 vote ● Reply ● Message ● Permalink

Imran Manzoor a month ago

I got 95% and my rank is 400. Seeing that you will have to manually evaluate the other 50% for all other submissions, is there a minimum rank to ensure it will be evaluated? The current Rank 1 probably wont be rank 1 for the entire test dataset. So what will happen?

▲ 1 vote ● Reply ● Message ● Permalink

Abhijit Pandey 23 days ago

Is there a problem with system ??? Submissions are taking forever to evaluate.

▲ 1 vote ● Reply ● Message ● Permalink

Prashant Gupta 21 days ago

@admin, solutions are not accepting now.
it just keeps on evaluating

▲ 1 vote ● Reply ● Message ● Permalink

Rakesh Lal 3 months ago

what format of file are we required to submit?

▲ 0 votes ● Reply ● Message ● Permalink

Vinay Sharma 3 months ago

You need to compress all the files you made in R or Python to .zip or .tar format

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

For evaluation submit a .csv file in the format specified in the sample submission.

▲ 0 votes ● Reply ● Message ● Permalink

Saurabh Dey 3 months ago

fine fine but how to download the question

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

There is a link to download the dataset.

▲ 0 votes ● Reply ● Message ● Permalink

Rohan Shingade 3 months ago

How are people getting a perfect score 1?

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

It is a problem with the verification code. It is fixed now. The scores will be re-evaluated soon.

▲ 0 votes ● Reply ● Message ● Permalink

Rohan Shingade 3 months ago

if possible can you post a little more info about the features? what's given isn't helping much.

▲ 0 votes ● Reply ● Message ● Permalink

Toppireddy 3 months ago

any formula??

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

We shall post a sample code soon.

▲ 0 votes ● Reply ● Message ● Permalink

?

Abhishek Kumar 3 months ago

Unable to submit any file since yesterday night. All I see is the green icon "Submitting"

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

It shows that if you have reached your limit for daily submissions.

▲ 0 votes ● Reply ● Message ● Permalink

Abhishek Kumar 3 months ago

Yeah but that was happening yesterday too and today too. So the day count is checked using the reference of my last
10th submission or the usual 24hrs format?

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

24hrs format

▲ 0 votes ● Reply ● Message ● Permalink

Vyom Bani 3 months ago

What is meant by imputation indicator and imputation revised?

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Imputation indicator is the indicator for that particular attribute. Imputation revised is the revised value after calculating with the imputation indicator.

▲ 0 votes ● Reply ● Message ● Permalink

Niyan 3 months ago

how so many people are getting 1 in the prediction?

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

It was a bug. It has been fixed.

▲ 0 votes ● Reply ● Message ● Permalink

Nikhil Kumar Singh 3 months ago

"Evaluation Log
File does not contain prediction for 10060480"
I have checked the submission CSV, it has predictions for each id. I am still getting this error. Kindly look into this

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

It might be because you are submitting probabilities instead of classes. The answers must be 0 or 1.

▲ 0 votes ● Reply ● Message ● Permalink

ujjwal bansal 3 months ago

I have just join this compition to learn data science . can anyone help me wat are the formost tools are required except pyhton because i know that .

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

A beginner knowledge about machine learning algorithms and some coding experience is required. We shall post a sample code soon.

▲ 0 votes ● Reply ● Message ● Permalink

DAYANANDA CHALLA Edited 3 months ago

What does Imputation indicator refer to ? does that mean it requires imputation? if so what si Imputation Revised?

▲ 0 votes ● Reply ● Message ● Permalink

?

Abhishikth Sagar 3 months ago

Imputation indicator is the imputation indicator of the given attribute. Revised is when the attribute has undergone imputation.

0 votes ● Reply ● Message ● Permalink

Puneeth 3 months ago

@admin what does the attribute VESTR mean?

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

VESTR is a general estimation variable.

0 votes ● Reply ● Message ● Permalink

adarsh.meher93 3 months ago

Hello.
Can you please explain what does the following mean in the respective columns these have been used : IMPUTATION REVISED? IMPUTATION INDICATOR? RECODE - IMPUTATION REVISED ?

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Imputation indicator is the indicator of the source of the imputation revised data. Eg: Questionare, General Statistics, etc
Imputation revised is when the input data has been merged with data from the imputation source which is represented by the Imputation indicator.
Recoded variables are recoded variables that were created from one or more of the edited or imputed variables from the preceding sections.

0 votes ● Reply ● Message ● Permalink

Palash Ghosh 3 months ago

99, 98 to be considered as missing data / no data available?

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Missing data is represented as -1.

0 votes ● Reply ● Message ● Permalink

Shubham Naik Edited 3 months ago

Can anyone tell me how score on leaderboard is calculated,is it based on accuracy of model.lets say if current highest score is 0.95,does it means that person has uploaded model with accuracy of 95%?

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

The score on the leader board is only evaluated for 50% of the submission file. The entire submission is evaluated after the test finishes. Yes, the leader board score is the precision score of their submission.

0 votes ● Reply ● Message ● Permalink

Bharath Sriraam 3 months ago

I'm not able to make my first submission today. Please look into this matter Abhishikth Sagar. Thank You.

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Please try again and contact me if the problem persists. The system is working fine.

0 votes ● Reply ● Message ● Permalink

pengoo yahoo 3 months ago

the descriptions need to be updated .

0 votes ● Reply ● Message ● Permalink

amar naik 3 months ago

@admin. i am getting an error " Runtime Error - FILE_NOT_OK" when i am submitting my prediction file. Evaluation log says prediction for one PRED id missing. but i looked at the file and it does have this PRED id. Submission ID: 14078528 Can you help

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

Make sure your predictions are in {0,1} and not probability values.

▲ 0 votes ● Reply ● Message ● Permalink

Amogh Badugu 3 months ago

What is the metric for evaluation??

▲ 0 votes ● Reply ● Message ● Permalink

Amogh Badugu 3 months ago

I dont think the metric of evaluation is precision(as said).It seems more like an accuracy score

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

It is precision but with average as 'micro'.

▲ 0 votes ● Reply ● Message ● Permalink

Seemant Singh 3 months ago

Why is it showing me a score of 0.93053 even if I am submitting the file with all zeros?
There is something wrong with this dude.

▲ 0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

There is not. As this is an imbalanced classification, there are more 0s than 1s. The score you are getting is the precision score for 0s. As the average is 'micro', it calculates precision for both values.

▲ 0 votes ● Reply ● Message ● Permalink

Seemant Singh 3 months ago

Oh! So I tried with all 1's and I am getting a score of 0.06947. Now it's making sense. Thanks! for the clarification.

▲ 0 votes ● Reply ● Message ● Permalink

Amogh Badugu 3 months ago

Can you mention the type of each feature ?

▲ 0 votes ● Reply ● Message ● Permalink

Aryan Bakshi 3 months ago

how to make the model into a file?
What is the meaning of all the variables?

▲ 0 votes ● Reply ● Message ● Permalink

Meet Goti 3 months ago

Submission ID: 14103803
Error : File does not contain prediction for 10060480
The submission file contains PERID 10060480 and its binary value. I am not submitting probablity.
Please help me resolve this issue as soon as possible.

▲ 0 votes ● Reply ● Message ● Permalink

Rohan Shingade 3 months ago

make sure data type of Criminal column is int

▲ 0 votes ● Reply ● Message ● Permalink

Meet Goti 3 months ago

It is int64

▲ 0 votes ● Reply ● Message ● Permalink

?

Harshad 3 months ago

I can't understand this @admin can u explain me it once what I should do.....

0 votes ● Reply ● Message ● Permalink

Gaurav Chavan 3 months ago

Which variables are continuous and which one are categorical ?

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

All variables are categorical.

0 votes ● Reply ● Message ● Permalink

Gaurav Chavan Edited 3 months ago

ANALWT_C is continuous @Abhishikt Sagar

0 votes ● Reply ● Message ● Permalink

Burhan Usman 3 months ago

Does accuracy and micro average precisoin mean the same thing, i.e, (TP0 + TP1)/(TP0 +TP1+FP0+FP1) ?

0 votes ● Reply ● Message ● Permalink

Imam Khursheed 3 months ago

Really! You really want to use precision as an evaluation metric. I mean I could just predict everyone as 0 and still get an accuracy of 93 percent. !

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

That is because the average is 'micro'. Getting 93 percent does not mean anything. ML problems are relative. Getting 90% in one problem may have the same significance as getting 60% in another. It only matters how close you are to 100%.

0 votes ● Reply ● Message ● Permalink

Dan Ofer 3 months ago

Are the "99", "98" values also missing values?

0 votes ● Reply ● Message ● Permalink

Abhishikth Sagar 3 months ago

No

0 votes ● Reply ● Message ● Permalink

Aayush Shrivastava 3 months ago

Can you Please enlighten me more about those "relevant source files"

0 votes ● Reply ● Message ● Permalink

Kushagradar 3 months ago

@admin Can you please upload a file giving the description of the features ??

0 votes ● Reply ● Message ● Permalink

pengoo yahoo 3 months ago

please explain this feature ""IRPINC3 RESP TOT INCOME (FINER CAT) - IMP REV""

0 votes ● Reply ● Message ● Permalink

Sreeram TP 3 months ago

any started code yet.??

0 votes ● Reply ● Message ● Permalink

Nandagopal M 2 months ago

I am getting this following error
"File does not contain prediction for 10060480"
even if I have that ID in my submission file. This is something I am getting now!

0 votes ● Reply ● Message ● Permalink

Nirmal Patel 2 months ago

same here !! same error i am getting right now ;(..
is there anyone who can help with this error

0 votes ● Reply ● Message ● Permalink

Nandagopal M a month ago

my problem was that the Criminal variable was of type "float". Converting it to int solved my issue!

0 votes ● Reply ● Message ● Permalink

Rahul Sarkar a month ago

I had same problem. Just write on the file again. so you have to write twice on the file. Worked for me. So you can try it.

0 votes ● Reply ● Message ● Permalink

Hasan Mesbaul Ali Taher Edited 2 months ago

Q 1. Can you please tell which features above are ordinal and which are nominal ?
It is difficult to figure out as the description of the features is not clear

0 votes ● Reply ● Message ● Permalink

Vivek_Chatt 2 months ago

cannot submit file. Error showing: File_Not_Ok, could not determine delimiter, although, I submitted it in csv and I even reimported the csv file in python notebook to test if it is fine or not.

0 votes ● Reply ● Message ● Permalink

Manish Sharma a month ago

What are the prerequisite for this type of competitions?Please help

0 votes ● Reply ● Message ● Permalink

MOHAMMAD DAUD a month ago

@admin , Server is down please look

0 votes ● Reply ● Message ● Permalink

Kaviyarasu a month ago

naive bayes algorithm

0 votes ● Reply ● Message ● Permalink

Arush Tahiliani Edited 24 days ago

Hi admin,
which Evaluation Metric you are using to measure the accuracy

0 votes ● Reply ● Message ● Permalink

Imran Manzoor 24 days ago

precision score with average = 'micro'

0 votes ● Reply ● Message ● Permalink

Karan Chadha 21 days ago

@admin, I am new to this platform. Can you please tell me where to write the code and how to submit

0 votes ● Reply ● Message ● Permalink

Lalit Mohan Sharma 20 days ago

@admin why it is taking too much time for evaluation after submission?

0 votes ● Reply ● Message ● Permalink

Arnab Thakuria 16 days ago

what classifier have you people used which are giving such high scores??
i used a naive bayes classifier and it gives only 55% accuracy
??? any suggestions

0 votes ● Reply ● Message ● Permalink

Sayar Banerjee 11 days ago

Try using more complex algorithms such as SVM, Regression with regularization, Knn.

▲ 0 votes ● Reply ● Message ● Permalink

Arnab Thakuria 16 days ago

will a support vector classifier work in this case ???

▲ 0 votes ● Reply ● Message ● Permalink

Sayoni Dutta Roy 13 days ago

how can I approach the problem without a data dictionary?

▲ 0 votes ● Reply ● Message ● Permalink

Sayar Banerjee 11 days ago

Try basic preprocessing techniques and use a complex algorithm. Should get you a decent score.

▲ 0 votes ● Reply ● Message ● Permalink

Sagar Kar 10 days ago

what kind of preprocessing @Sayar

▲ 0 votes ● Reply ● Message ● Permalink

Sayar Banerjee 9 days ago

That is for you to find out.

▲ 0 votes ● Reply ● Message ● Permalink

Devi Meenakshi 4 days ago

@admin,can you please explain me all the column attributes and how they are helpful to predict the criminals...Its very urgent as I need to present it tomorrow. so,kindly explain me please....

▲ 0 votes ● Reply ● Message ● Permalink