Predict the traffic volume

Max. Marks: 100

Indian metro cities are famous for their notoriously varied traffic volume that is experienced by commuters every day. People plagued with traffic jams often raise concerns over poor traffic management systems. But with so much technology at our disposal, why should we continue to deal with traffic management in obsolete ways?

In this modern and advanced era, why the problems regarding traffic management should be dealt with in obsolete ways?

Your city's Traffic Police department has decided to use Machine Learning and Artificial Intelligence techniques to solve their traffic problems. They have collected traffic volume patterns and climate conditions that have been observed for 4 years. They want to be able to forecast the traffic volume.

Task: Your task is to predict the traffic volume for given time duration and climate conditions.

Data description

Columns	Description
date_time	Date, time, and hour of the data that is collected in the local IST time
is_holiday	Categorical Indian national holidays combined with regional holidays
air_pollution_index	Air Quality Index (10-300)
humidity	Numeric humidity in Celcius
wind_speed	Numeric wind speed in miles per hour
wind_direction	Cardinal wind direction (0-360 degree)
visibility_in_miles	Visibility of distance in miles
dew_point	Numeric dew point in Celcius
temperature	Numeric average temperature in Kelvin
rain_p_h	Numeric amount in mm of rain that occurred in the hour
snow_p_h	Numeric amount in mm of snow that occurred in the hour
clouds_all	Numeric percentage of cloud cover
weather_type	Categorical short textual description of the current weather
weather_description	Categorical longer textual description of the current weather
traffic_volume	Numeric hourly traffic volume bound in a specific direction

Data Set

- **Train.csv**: 33750 x 15
- **Test.csv**: 14454 x 14
- **Submission.csv**: 14454 x 2

Sample submission format

date_time	traffic_volume
1969-05-17 21:00:00	500
1969-05-17 21:00:00	530

1969-05-17 21:00:00	545
1969-05-17 22:00:00	750

Please Note:

- Only the team leaders are supposed to give the submission for the challenge and only these submissions will be considered valid for the final shortlist.
- Team members can collaborate offline with the leader for the submissions but need not register and participate in the challenge.

Evaluation criteria

$$leaderboard\ score = max(0, (100 - rmse))$$

Download dataset

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 (99) SORT BY: Relevance

Join Discussion...

Cancel Post

singhsimran39 Edited 5 days ago

Coordinators/Moderators a few questions:
1. 9th August 2014 to 10th June 2015 data is missing in the train set. Is that a mistake of genuinely there is no data?
2. At a lot of places the test data has multiple occurrences for same date+time combination (although values in other columns are different). Any specific reason?
3. Also what exactly is traffic volume?? Number of vehicles in some unit space or something??
Interesting things to notice:
1. Does the whether type haze cause traffic or does traffic cause haze? Not sure how useful this feature might be. Any thoughts??
Any other out of the ordinary things anyone has noticed?

11 votes · Reply · Message · Permalink

Anoubhav Agarwaal 5 days ago

Point #1 is the biggest issue. And #2 looks like a mistake from there end

1 vote · Reply · Message · Permalink

Anoubhav Agarwaal 5 days ago

The test file has so many repeated timeslots with different feature values. I am not sure what people are predicting for this. If the time slot is wrong in the test file itself, no matter what you predict it will be wrong

0 votes · Reply · Message · Permalink

Anoubhav Agarwaal 5 days ago

The test data set has 501 unique dates. On an hourly account, there will be 501*24 = 12024 rows in the test set. However, there are 14454 rows in the test set. Thus, 2430 date-time index is being repeated randomly in that test set however with different attribute values. So you are making predictions are filling it in random time slots. Can somebody please confirm if the test set has issues? Thanks

1 vote · Reply · Message · Permalink

G. Kranthi Kiran 5 days ago

How about you remove the duplicate time-indexes and then predict the non-duplicate indexes and then fill-in same for the duplicated time-indexes in the submission file.

3 votes · Reply · Message · Permalink

HARISH. L 5 days ago

Yup actually that is a good thing maybe they gave it on purpose for us to discover and act on it accordingly. Because pre processing is the most important phase as we all know

2 votes · Reply · Message · Permalink

singhsimran39 5 days ago

Try this at home: build a model without considering time as a feature. Then build

another model with time as one of the features among others. Compare the results for the same time in test data. That might just tell you the importance of time in predicting traffic (a lot of off the shelf libraries do that for you though)

0 votes • Reply • Message • Permalink

Hemanth Kumar Edited 6 hours ago

But build a model using other feature it give accuracy but no sense bcz traffic will dependent on time that means AM & PM and weekdays & weekand's .so we need build model using both time series and some other features .

0 votes • Reply • Message • Permalink

Anoubhav Agarwaal 5 days ago

As far as I know removing fields/rows in a given test set is unheard of. Also I mentioned that only the time indexes are duplicate, the remaining attributes and features are all unique. Is your model not taking into account these other given attributes? Coz if it does, each row is unique.

0 votes • Reply • Message • Permalink

G. Kranthi Kiran 5 days ago

I didn't check the other features but the target for every duplicate time-index was same. So I thought just keeping the first one would be best.
Its totally upto you to decide how to treat them in the end.
Mine was just a suggestion that I thought how I would treat them.

2 votes • Reply • Message • Permalink

HARISH. L Edited 5 days ago

I checked the unique values by making a list of the date in test data set and i got that 2468 values are duplicate

0 votes • Reply • Message • Permalink

Hemanth Kumar 6 hours ago

u can use drop_dupicates('date_time') it will automatically drop the repeating time values

0 votes • Reply • Message • Permalink

Sangeet Kumar Mishra 6 days ago

Why am I getting exactly zero scores, even if my predictions look alright? I mean even if it's bad, how can it be exactly zero, it's an MSE after all.

4 votes • Reply • Message • Permalink

Mr_KRAKEN 6 days ago

same problem getting exactly zero scores

0 votes • Reply • Message • Permalink

Salil Mishra 6 days ago

The evaluation metric is maybe \max(0, 100-MSE), so if your MSE is greater than 100, your score will be zero.

3 votes • Reply • Message • Permalink

Mohit Uniyal 6 days ago

I am also having the same problem,
and I have ensured that my MSE is not more than 100, I tried it using train_test split. MSE is near 10, so 100-10 => 90. My score should be 90, but it is showing 0. is it the fault at their end while calculating the score?

0 votes • Reply • Message • Permalink

shelldragoon1104 6 days ago

Its possible that your validation set might not be mimicking the actual test set.

2 votes • Reply • Message • Permalink

G. Kranthi Kiran 5 days ago

@mohituniyal2010
Nope. You're model is just overfitting. Try a good validation metric which is better for Time-Series or just do a manual split as "valid-train" : "last 1 year-remaining all years".
Try this and check your errors.

2 votes • Reply • Message • Permalink

Jayaraj Mudaliar 6 days ago

The holiday list is of US Holidays. Are the data is of Indian Metro traffic? Are we suppose to include indian festival/holidays

7 votes • Reply • Message • Permalink

Kundan Kumar 6 days ago

No external data should be used.I saw the file the holiday section is most probably useless it can be dropped.

0 votes • Reply • Message • Permalink

Souren Hazra 6 days ago

Test data also consists of us holidays, so I think we can assume it is not a indian metro trafic dataset

1 vote • Reply • Message • Permalink

shobhit upadhyaya Edited 3 days ago

Hi All,
Earlier people were finding hard to score more than zero, now I can see 99 . What changed ?
If you think logically scoring logic now is = max(0, 100 - RMSE)
To achieve 99 you should get RMSE = 1
If you square RMSE you will get MSE that will be 1
So the people who have scored now 99 they should have got 99 before also by previous scoring logic = 100 - MSE
Or there is a data leak what people have found ;)
OR scoring is RMSLE instead of RMSE
max(0, 100 - RMSLE)

3 votes • Reply • Message • Permalink

Nitin kshatriya Edited 3 days ago

There is no data leakage. I started today only so no idea what the error was, but as most people have mentioned max(0, 100- RMSLE) looks like evaluation metric. For my local validation, I am using RMSLE, which is corresponding to the leaderboard.

4 votes • Reply • Message • Permalink

shobhit upadhyaya 3 days ago

Thanks @nitin for letting me know :)

1 vote • Reply • Message • Permalink

Garnepudi Venkatesh 5 days ago

Even if the predicted values differ by 0.1, based on MSE formula for the test size of 14454, it would give 14454* (0.1^2) -> 144.54 which is still greater than 100. This (max(0,100 - mean__squared__error(actual values, predicted values)) might be the evaluation metrics. Given such evaluation metrics, I think this problem is a difficult one to

solve.
5 votes • Reply • Message • Permalink

Abhijit Das 5 days ago
Based on the given metric , my model is generating the error in the order of "-3000s". They must definitely change the metric.
1 vote • Reply • Message • Permalink

Shayantan Banerjee 6 days ago
When are we able to see the score in the leaderboard?
4 votes • Reply • Message • Permalink

Aditya Vikram SIngh 5 days ago
Just out of curiosity , how can someone forecast weather for 1 year in advance ??
2 votes • Reply • Message • Permalink

G. Kranthi Kiran 5 days ago
By data.
1 vote • Reply • Message • Permalink

Manan Gupta 5 days ago
Guys , can anyone please help me out with the submission of the test predictions . As I try to upload , it says 'Runtime Error - FILE_NOT_OK
3 votes • Reply • Message • Permalink

Akash Gupta 3 days ago
same problem
0 votes • Reply • Message • Permalink

Parijat Mishra 2 days ago
Is this problem resolved ? I am facing the same error .
0 votes • Reply • Message • Permalink

Ishan Nangia 2 days ago
Have you guys made sure that when you save your file an extra index column isn't being made ? Also are there any other details being given for the error ??
0 votes • Reply • Message • Permalink

Chandrashekhar Yadav 12 hours ago
same problem.
0 votes • Reply • Message • Permalink

G. Kranthi Kiran 5 days ago
How about changing the evaluation score to :
Score = 100 - mean__squared__error(np.log(actual values), np.log(predicted values))
Which are giving me scores in the range of ~99.8-99.9.
3 votes • Reply • Message • Permalink

Xyz Abc 4 days ago
simple analysis:
Creating fake true values and prediciton by difference of 11 from the original values
#Fake true
random_y = np.array(range(1,test.shape[0]+1))
#Fake prediction by difference of 11
random_y_pred = random_y - 11
#Calculate mean squared error
mean_squared_error(random_y, random_y_pred)
output: 121.0
Scoring logic :
100 - 121 = -21 < 0
There is no way without changing the scoring logic we can achieve > 0
2 votes • Reply • Message • Permalink

Salil Mishra 4 days ago
Exactly, the easiest thing they can do now is to convert MSE to MSLE(mean squared log error). It will solve the problem and people can at least see the results for their submission.
0 votes • Reply • Message • Permalink

singhsimran39 Edited 4 days ago
I guess they have changed the evaluation metric to 100 - mean__squared__error(actual values, predicted values). But still the leaderboard has all scores as 0
0 votes • Reply • Message • Permalink

Arnab Biswas 4 days ago
I don't really appreciate the fact that "Evaluation criteria" has been silently changed in the middle of the competition. Not sure if the data has also been changed or not. A small note or an email broad cast could have been a better way to maintain transparency.
1 vote • Reply • Message • Permalink

G. Kranthi Kiran 3 days ago
Same thoughts and I think there is an type in the evaluation metric :
I think it should be
max(0, (100-rmsle))
1 vote • Reply • Message • Permalink

Anuran Chakraborty 2 days ago
I cannot submit any file even though I have submitted only 5 times today. Is anyone else having the problem?
1 vote • Reply • Message • Permalink

G. Kranthi Kiran 2 days ago
Same here. The instructions page states that maximum submission limit for each day is 10.
1 vote • Reply • Message • Permalink

Akash Gupta 6 days ago
i got an error while uploading Submission File -- "File does not contain prediction for 2017-05-18 00:00:00"
1 vote • Reply • Message • Permalink

Pranay Raj K 5 days ago
just the change the date time format to the mentioned format either in python or in excel, it will work
0 votes • Reply • Message • Permalink

Parijat Mishra 2 days ago
Is this problem resolved ? I am facing the same error .What is the solution ?

0 votes · Reply · Message · Permalink

Akash Gupta a day ago

plz tell me how to solve it

0 votes · Reply · Message · Permalink

G. Kranthi Kiran 21 hours ago

The error statement basically sums it up. Just check your submission file if it contains the prediction for the particular date it throws the error for.

And also check the shape of the submission file and check if the number of rows are equal to 14454?

0 votes · Reply · Message · Permalink

Tadasha Nayak 6 days ago

Is this a time-series problem?

1 vote · Reply · Message · Permalink

Arnab Biswas 2 days ago

Yes. It is.

0 votes · Reply · Message · Permalink

Abhijit Das Edited 6 days ago

On cross validating my model offline , I am getting large negative values... maybe that's why the score is 0 for everyone on the leaderboard. I think the metric defined is not well suited for this problem statement.

1 vote · Reply · Message · Permalink

Shayantan Banerjee 5 days ago

How is it that no one is able to score on the leader board? They should have come up with some other metric to evaluate. However bad my model is, I should have a way of knowing this.

0 votes · Reply · Message · Permalink

Aditya Vikram SIngh Edited 5 days ago

as errors are being squared I think they is going well above 100

1 vote · Reply · Message · Permalink

Kabir Nagpal 5 days ago

mean squared error can be anything even greater than 100
is it mean error we need to consider or absolute mean error?

1 vote · Reply · Message · Permalink

R SHRIPRASAD 4 days ago

Moderators,please change the metric for evaluation. The current metric only allows us to be around 10 units off from the actual values on an average. No one can be that close to the solution.

1 vote · Reply · Message · Permalink

Sachin Mukherjee 3 days ago

Few People started getting non zero score. What was you approach?

0 votes · Reply · Message · Permalink

G. Kranthi Kiran 3 days ago

The evaluation metric has been changed. Check that once.

1 vote · Reply · Message · Permalink

Sachin Mukherjee 3 days ago

Can I submit the same file or do I need to predict once again

0 votes · Reply · Message · Permalink

G. Kranthi Kiran 3 days ago

Submit again.

0 votes · Reply · Message · Permalink

Sachin Mukherjee 3 days ago

I have submitted it again and it shows error message as input contains NaN, infinity or a value too large for float64.

0 votes · Reply · Message · Permalink

Nitin kshatriya 3 days ago

check min and max values predicted by your model. I guess it's giving a large prediction for few cases.

0 votes · Reply · Message · Permalink

Sachin Mukherjee 3 days ago

The problem is that one model is giving me a place in leaderboard while the other is giving Runtime error while submitting the file. As I'm new to this field I don't know why this is happening. The model due to which I'm getting runtime error has a good cross-validation scrore than the first model. Any Idea why this is happening?

0 votes · Reply · Message · Permalink

Bhuvana Kundumani a day ago

Hi Sachin
How did you solve run time error you mention above?

0 votes · Reply · Message · Permalink

Gowtham Raj 2 days ago

hello, Kiran can I know what kind of problem is that? regression problem ah?

0 votes · Reply · Message · Permalink

G. Kranthi Kiran a day ago

Time-series problem.

0 votes · Reply · Message · Permalink

Revant Tiwari 2 days ago

How do I get started with time series data?

0 votes · Reply · Message · Permalink

G. Kranthi Kiran a day ago

Quite a few good articles are there at Analytics vidhya.
Simple google search would've done.

1 vote · Reply · Message · Permalink

Nitin kshatriya Edited a day ago

As per 24872'th row in training data, even after 9.8 meters of hourly rain, 5535 was the traffic volume :D

1 vote · Reply · Message · Permalink

G. Kranthi Kiran a day ago

Lol
0 votes • Reply • Message • Permalink

Lakshmi Narayan Sahu a day ago
Getting following error while submitting the file:
/hackerearth/PYTHON3_4692_5a35_8ccc_8d8c/__pycache__/s_5551_0774_c4f3_5ce3.cpython-35.py:32: RuntimeWarning: divide by zero encountered in log1p
/hackerearth/PYTHON3_4692_5a35_8ccc_8d8c/__pycache__/s_5551_0774_c4f3_5ce3.cpython-35.py:32: RuntimeWarning: invalid value encountered in log1p Input contains NaN, infinity or a value too large for dtype('float64').
1 vote • Reply • Message • Permalink

coding chef a day ago
getting same error
0 votes • Reply • Message • Permalink

Lakshmi Narayan Sahu a day ago
I think, I have identified the issues. The reason is, few values are come up with less than or equal to zero value.
0 votes • Reply • Message • Permalink

SHABIR JAMEEL FAROOK 6 days ago
Hi Team,
For some reason, I could not download the dataset as it always says as "Failed - Download error"
While I look into figure out the concern, please share your thoughts.
Thanks,
Shabir
0 votes • Reply • Message • Permalink

Kundan Kumar 6 days ago
I downloaded the data set and uploaded it on my drive, download it from here:
https://drive.google.com/file/d/1JY2VAhl5Rcb6EsiR6b1O4g5rBxdM955I/view?usp=sharing
0 votes • Reply • Message • Permalink

SHABIR JAMEEL FAROOK 6 days ago
Thanks Kundan.
A restart of my machine helped me to download with no issues.
Thanks again for your time and help on this.
0 votes • Reply • Message • Permalink

Shaikh Safiya Naaz Abdul Hakeem. 6 days ago
Prediction file is the output in .csv ??? I am confused..or my code?
0 votes • Reply • Message • Permalink

Kundan Kumar 6 days ago
Prediction file is the output in .csv and the source files are your code.
0 votes • Reply • Message • Permalink

Shaikh Safiya Naaz Abdul Hakeem. 6 days ago
Okay Thank you
0 votes • Reply • Message • Permalink

Shashank Agrawal 6 days ago
I have entered my teammate's name in the form, but can he enter separately for the purpose of viewing the problem, leaderboard etc ??
0 votes • Reply • Message • Permalink

Shubhankar Mrinal 6 days ago
does the laptop hang on testing and training??
0 votes • Reply • Message • Permalink

Shivendra Sharma 4 days ago
Dear competitors,
How have you all addressed the phase from 2014 to 2015? Deleting those rows seems senseless, some imputing surely needs to be done.
0 votes • Reply • Message • Permalink

Shubhankar Mrinal 4 days ago
sm1 pls help. how to proceed
0 votes • Reply • Message • Permalink

Balram 4 days ago
How to create the team? I don't see any such options. Plz help!
0 votes • Reply • Message • Permalink

Balram 4 days ago
NM
0 votes • Reply • Message • Permalink

Anish Shukla 2 days ago
I am getting an error while uploading Submission File -- "File does not contain prediction for 2017-08-28 03:00:00"
0 votes • Reply • Message • Permalink

Bhaskar Goonisetty 2 days ago
Change the column format for the datetime column to yyyy-mm-dd hh:mm:ss in excel and reupload.
0 votes • Reply • Message • Permalink

Kriti Sahu 2 days ago
The dataset is very counter intuitive and requires no learning at all as it is mostly deterministic
0 votes • Reply • Message • Permalink

G. Kranthi Kiran a day ago
How do you justify that?
0 votes • Reply • Message • Permalink

Kriti Sahu a day ago
Can not justify before the competition is over.
0 votes • Reply • Message • Permalink

G. Kranthi Kiran 21 hours ago
Bohut thanda!
0 votes • Reply • Message • Permalink

Nitin Kshatriya 2 days ago

Can anyone explain how are they using wind speed and wind directions?

0 votes · Reply · Message · Permalink

coding chef a day ago

hey admin my new submissions aren't being accepted. It's just taking long time and then puff..... nothing recorded. It was working fine yesterday please look into it asap.

0 votes · Reply · Message · Permalink

Umang Bhalani a day ago

Does wind direction matter?

0 votes · Reply · Message · Permalink

Syed Areeb Wadood a day ago

getting runtime error don't know why any help would be appreciated

0 votes · Reply · Message · Permalink

Bhuvana Kundumani a day ago

did you solve runtime error?

0 votes · Reply · Message · Permalink

Mritunjoy Das a day ago

this is the error I am getting even Though I checked my data everything was okay
/hackerearth/PYTHON3_c9da_f865_24e2_f958/__pycache__/s_9aec_ee68_4003_6d7b.cpython-35.py:32: RuntimeWarning: invalid value encountered in log1p Input contains NaN, infinity or a value too large for dtype('float64').

0 votes · Reply · Message · Permalink

Surabhi Pandey 4 hours ago

it is mainly because you might be having some 0 or negative values in submission file check and submit again

0 votes · Reply · Message · Permalink

Rohit Shukla 9 hours ago

Hi All, i have made my submission file and when i am uploading it, i am getting error that " File does not contain prediction for 2017-12-01 08:00:00" every time i am getting this error with date changed, although i have cross checked my data, in that no prediction is missing for any dates.

0 votes · Reply · Message · Permalink

Surabhi Pandey 5 hours ago

getting this error RuntimeWarning: invalid value encountered in log1p Input contains NaN, infinity or a value too large for dtype('float64'). everytime I am trying to upload

0 votes · Reply · Message · Permalink

Rajnish 38 minutes ago

I am unable to submit my csv file.
error showing:File does not contain prediction for 2017-10-26 04:00:00
Runtime Error - FILE_NOT_OK
THIS time stamp is present in my date time column in submitted file(checked manually)

0 votes · Reply · Message · Permalink