The Winton Stock Market Challenge

Join a multi-disciplinary team of research scientists

$50,000 · 1,300 teams · 2 years ago

rmldj

7th place

Options

Local CV and LB

posted in The Winton Stock Market Challenge 2 years ago

1

I wonder what is your experience in comparing local CV scores with the LB one. I do not mean in absolute value but rather in relation to the all zero benchmark.

I made a 5-fold CV of my model. Not surprisingly the local scores of the different CV folds had some standard deviation (around 13.0) but in each fold the score was consistently better/smaller by 5-7 from the corresponding score in each fold of the all zero benchmark.

However when I submitted the predicitions for the test set it scored around 1775 - worse by 5 than the all zero benchmark!

I wonder if you encountered something similar?

I am a bit amazed as seeing a consistently better perfomance on each of the 5-folds I would expect also better behavior on the test set...

I may have some bug in my code but I did not manage to spot it so far - that's why I am asking (NB. my scoring code gives 1773.92 for the all zero benchmark on the whole training set).

Comments (31) Sort by Hotness

Please sign in to leave a comment.

rmldj • (7th in this Competition) • 2 years ago • Options

3

@woshialex - that is interesting..

I made now a more refined test. I repeated now the local 5-fold CV with random shuffling:

```
kf=KFold(len(yall), n_folds=5, shuffle=True)
```

a hundred times. This generated in total 500 scores from which I subtracted the score of the zero benchmark in the corresponding fold.

The results are as follows:

1. In *all* cases the result was better than the all zero benchmark
2. The difference was -6.9 with a standard deviation of order 1.5

Irrespective of the details of the model this seems to strongly indicate that the test set is taken from a different distribution than the training set (like different market conditions???).

Could the organizers clarify whether the split into the training and test set was just a random split of some big common dataset or not?

If it really is a random split then I must have either some data corruption in the test set or a bug. If the train and test distributions are indeed different then the competition really seems hopeless as one would have to optimize the public LB...

alexeymosc • (612th in this Competition) • 2 years ago • Options

1

> ***Lê wrote***
>
> Very interesting results! Would you please give me the EUR/USD forex quotes? I would like to give it a try on forex. Thank you very much!

Hi Le,

I used this data feed: https://www.dukascopy.com/swiss/english/marketwatch/historical/

It allows one to go as deep as tick level, but I used minutes.

wawltor • 2 years ago • Options

1

> ***rmldj wrote***
>
> I guess that coming to grips with the train/test data dichotomy here is exactly the key to this competition..
>
> I never encountered a discussion of such a scenario in the literature and so far all my various attempts at quantifying this failed. However judging by the scores at the top of the leaderboard a solution must exist..

I guess the Top 3 have find the way to solve the problem,But It is a key to competition.Thank for your request! Good luck!

JWJAnderson • 2 years ago • Options

1

> ***Sef wrote***
>
> Would it be possible to have an additional validation data set, one with labels but uncorrelated to the training set?
>
> It would help us choosing the models instead of relying on the LB, even if it is just as small as 1000 rows of data.

Hi Sef. This is what the training set is intended for. Part of the challenge is to figure out how to do model selection with the data provided.

JWJAnderson • 2 years ago • Options

1

> ***rmldj wrote***
>
> Do you mean that:
>
> 1. You expect that cross-validation on the training set need not give insight into model behavior on the test set
> 2. Consequently the test set may be quite different from training set but you refrain from making any statement about that? (so figuring how to deal with that is part of the difficulty of the competition?)

Hi rmldj. It's a bit tricky to go into more details about this without giving possible hints away here, as we would have to start revealing sampling details, so apologies. But we have understood your concern, and do not think there are any issues with the integrity of the challenge.

stud3nt • 2 years ago • Options

1

I had some experience in predicting stock market behaviour and I can tell you one thing, if data is divided sequentially (for example days 1-n for training and days n+1-m for testing) then CV can give you totally wrong results.

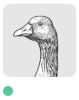
For example, if you have dataset od 10 training examples where examples 2,4,6,8 and 10 are used for training and 1,3,5,7,9 are used for validation then, to predict the value of example 3, you already know value of example 4,6,8 and 10 which is not case in reality.

On the other hand, stock returns are pretty much similar to Gaussian noise and it is very hard to spot patterns.

woshialex • (16th in this Competition) • 2 years ago • Options

2

I guess lots of rows in the train set or test set are highly correlated but not among each other. So the local CV does not work. :) too bad that the organizers do not give us a date/time related column for us to do some smart/consistent CV split....



JWJAnderson • 2 years ago • Options

0

HI all. Some interesting discussion here for sure.

> ***rmldj wrote***
>
> No - I just do random splits (80% training 20% test). However I did a lot of those random splits (500 of them) and always got better results than the benchmark.

With respect to the cross-validation, we can confirm we do not have any further concerns with the data.

> ***rmldj wrote***
>
> > ***rcarson wrote***
> >
> > Thank you all for sharing. But I'm wondering if this is allowed to be shared. I hope Admin can clarify.
>
> Indeed clarification would be great. I hope that this topic nevertheless abides by the rules as no details of any model are given and the results of the CV experiment only serve to justify the question to the organizers about the split into the train and test datasets as the above CV experiment suggests that these sets have statistically significant different properties which is quite different from most competitions on Kaggle I guess (at least from the ones in which I participated...).

With respect to whether or not this discussion is allowed to be shared, there is (as this thread shows) quite a grey area between asking questions about the data and discussing problem strategy. We'll try and be proactive about saying when we think things have crossed into giving too much away, but would urge people to err on the side of caution. This thread is probably approaching the territory of too much strategic discussion, but hopefully we have now answered your question.

alexeymosc • (612th in this Competition) • 2 years ago • Options

-1

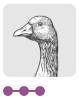
Yesterday I spent some hours to replicate the work of my proprietary algorithm on EUR/USD forex quotes in predicting a 64-minute ahead direction. Of course I knew the validation sample's targets, which was the core goal of this effort.

Not much surprisingly I managed to get a 60% accuracy of prediction on a teach+test sample (24 000 instances) and it spiraled down to 55% accuracy on a validation sample (2 250 000 instances, covering 6 years), with ME of a trade without transaction cost equal to 0.0002 and 0.0001 respectively (those familiar with FX can note the model is hardly profitable out of sample).

That assured me that my algo works on a single symbol at least, generating some positive ME, better that zero.

I guess here we deal with a much more delicate model that is a mix of different stocks behaviour.

Just trying to understand where I am.



redstr • 2 years ago • Options

0

I also had this. You've missed something important. Look harder, recheck your assumptions.

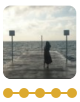
ebdd0304e745... • (174th in this Competition) • 2 years ago • Options

0

I got the same problems, local cv score is better than benchmark, but on LB score, it perform slightly worse than benchmark.

I'm curious about how they sampling the data, but they're professional in this field, so I think they know what to do to prepare data.

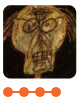
If all the traning set is derieved from the same time span an the test set from another time span that could be the problem, because you lose ability to use local cv to validate the model.



Jiwei Liu • (647th in this Competition) • 2 years ago • Options

0

Thank you all for sharing. But I'm wondering if this is allowed to be shared. I hope Admin can clarify.



rmldj • (7th in this Competition) • 2 years ago • Options

0

> ***rcarson wrote***
>
> Thank you all for sharing. But I'm wondering if this is allowed to be shared. I hope Admin can clarify.

Indeed clarification would be great. I hope that this topic nevertheless abides by the rules as no details of any model are given and the results of the CV experiment only serve to justify the question to the organizers about the split into the train and test datasets as the above CV experiment suggests that these sets have statistically significant different properties which is quite different from most competitions on Kaggle I guess (at least from the ones in which I participated...).

Nick D. Haynes • (84th in this Competition) • 2 years ago • Options

0

> ***rmldj wrote***
>
> I wonder if you encountered something similar?
>
> I am a bit amazed as seeing a consistently better perfomance on each of the 5-folds I would expect also better behavior on the test set...

I'll second this. I'm seeing a puzzling difference between CV scores for my different models and how they get scored on the leaderboard.

As far as I can tell, it doesn't seem like the train and test data are different, statistically speaking. If you run some simple descriptive statistics on the features that are available in the train and test sets, they match up. Similarly if you plot histograms. So I'm stumped for now.

SM • 2 years ago • Options

0

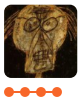
rmldj do you control fraction of 61-62 in your cv?

rmldj • (7th in this Competition) • 2 years ago • Options

0

> ***Sergey Makarevich wrote***
>
> rmldj do you control fraction of 61-62 in your cv?

No - I just do random splits (80% training 20% test). However I did a lot of those random splits (500 of them) and always got better results than the benchmark.



rmldj • (7th in this Competition) • 2 years ago • Options

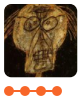
0

> ***JWJAnderson wrote***
>
> HI all. Some interesting discussion here for sure.
>
> > ***rmldj wrote***
> >
> > No - I just do random splits (80% training 20% test). However I did a lot of those random splits (500 of them) and always got better results than the benchmark.
>
> With respect to the cross-validation, we can confirm we do not have any further concerns with the data.

Thanks for the reply - however let me clarify whether I understood your answer correctly:

Do you mean that:

1. You expect that cross-validation on the training set need not give insight into model behavior on the test set
2. Consequently the test set may be quite different from training set but you refrain from making any statement about that? (so figuring how to deal with that is part of the difficulty of the competition?)



rmldj • (7th in this Competition) • 2 years ago • Options

0

Thanks a lot for the clarification!

Sef • 2 years ago • Options

0

Would it be possible to have an additional validation data set, one with labels but uncorrelated to the training set?

It would help us choosing the models instead of relying on the LB, even if it is just as small as 1000 rows of data.

LLMSI • 2 years ago • Options

0

I can also testify large discrepancies between the local cv loss and the lb one.

Perhaps this data collides with the standard machine learning where the joint distribution of the predictors and the target variables is the same for the both the training and the testing set. In such deformed data learning using out-of-box methods does not help the inference accuracy on test.

Probably the admin will not be able to provide an input, nevertheless someone experienced with stock data could briefly answer if i) this phenomenon is usual with other stock data, or ii) is it just this specific dataset that has a synthetically-implanted discrepancy between the train and test sets. If it is i), then it would motivate people to research how to deal with discrepant data. If ii) is the case, then I fear the challenge looses interest.

alexeymosc • (612th in this Competition) • 2 years ago • Options

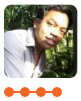
0

> if i) this phenomenon is usual with other stock data,

I trade currencies (forex) for years and am familiar with stock data. It is very widespread that models generated on some time period totally fail to work on the other time period due to the time series non-stationary nature (even volatility is volatile, as one top person said). As a hint, it is possible to estimate distribution params on a part of training data and then compare these to the left out part of data to see how much they change.

> or ii) is it just this specific dataset that has a synthetically-implanted discrepancy between the train and test sets.

It is not disclosed.

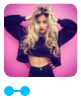


Mendrika Rama... • (3rd in this Competition) • 2 years ago • Options

0

> ***rmldj wrote***
>
> I wonder what is your experience in comparing local CV scores with the LB one. I do not mean in absolute value but rather in relation to the all zero benchmark.
>
> I made a 5-fold CV of my model. Not surprisingly the local scores of the different CV folds had some standard deviation (around 13.0) but in each fold the score was consistently better/smaller by 5-7 from the corresponding score in each fold of the all zero benchmark.
>
> However when I submitted the predicitions for the test set it scored around 1775 - worse by 5 than the all zero benchmark!
>
> I wonder if you encountered something similar?
>
> I am a bit amazed as seeing a consistently better perfomance on each of the 5-folds I would expect also better behavior on the test set...
>
> I may have some bug in my code but I did not manage to spot it so far - that's why I am asking (NB. my scoring code gives 1773.92 for the all zero benchmark on the whole training set).

My CV score for the all-zeros benchmark is 1773.92, same as yours. My current best scores ~1772 on my local CV. I have a model that scores way better than 1772 on my local CV but scores in the 1800s on the LB. I'm finding cross-validation quite challenging also.

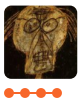


wawltor • 2 years ago • Options

0

> ***Mendrika Ramarlina wrote***
>
> > ***rmldj wrote***
> >
> > I wonder what is your experience in comparing local CV scores with the LB one. I do not mean in absolute value but rather in relation to the all zero benchmark.
> >
> > I made a 5-fold CV of my model. Not surprisingly the local scores of the different CV folds had some standard deviation (around 13.0) but in each fold the score was consistently better/smaller by 5-7 from the corresponding score in each fold of the all zero benchmark.
> >
> > However when I submitted the predicitions for the test set it scored around 1775 - worse by 5 than the all zero benchmark!
> >
> > I wonder if you encountered something similar?
> >
> > I am a bit amazed as seeing a consistently better perfomance on each of the 5-folds I would expect also better behavior on the test set...
> >
> > I may have some bug in my code but I did not manage to spot it so far - that's why I am asking (NB. my scoring code gives 1773.92 for the all zero benchmark on the whole training set).
>
> My CV score for the all-zeros benchmark is 1773.92, same as yours. My current best scores ~1772 on my local CV. I have a model that scores way better than 1772 on my local CV but scores in the 1800s on the LB. I'm finding cross-validation quite challenging also.

I think that the train and test data came from two model ,or two different periods.But I do not know how to deal with problem.I am the new kaggler and a student ,I really want to somebody to give me a help, some hints or some paper is ok.



rmldj • (7th in this Competition) • 2 years ago • Options

0

I guess that coming to grips with the train/test data dichotomy here is exactly the key to this competition..

I never encountered a discussion of such a scenario in the literature and so far all my various attempts at quantifying this failed. However judging by the scores at the top of the leaderboard a solution must exist..

PandaBambu • (671st in this Competition) • 2 years ago • Options

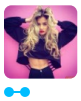
0

Same thing here. CV results better than benchmark, but worse on the LB. Quickly loosing interest in the problem. If test comes from a different distribution, organizers should at least give a hint.

SM • 2 years ago • Options

0

Sorry, wrong competition)

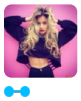


wawltor • 2 years ago • Options

0

> ***PandaBambu wrote***
>
> Same thing here. CV results better than benchmark, but worse on the LB. Quickly loosing interest in the problem. If test comes from a different distribution, organizers should at least give a hint.

Maybe the organizers just want us to solve the this problem,Maybe predicting the change of stock‘ price is not important！ Just my guess.



wawltor • 2 years ago • Options

0

> ***Sergey Makarevich wrote***
>
> Sorry, wrong competition)

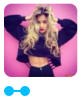
Thank you for your advice ,indeed.This competition makes me to think,not just the feature or model .I am Chinese，My english is poor，forgiving me ! :-) :-)

SM • 2 years ago • Options

0

> ***wawltor wrote***
>
> > ***Sergey Makarevich wrote***
> >
> > Sorry, wrong competition)
>
> Thank you for your advice ,indeed.This competition makes me to think,not just the feature or model .I am Chinese，My english is poor，forgiving me ! :-) :-)

I wrote recommendation for other competition Rossmann, so had deleted it and wrote Sorry)



wawltor • 2 years ago • Options

0

> ***Sergey Makarevich wrote***
>
> > ***wawltor wrote***
> >
> > > ***Sergey Makarevich wrote***
> > >
> > > Sorry, wrong competition)
> >
> > Thank you for your advice ,indeed.This competition makes me to think,not just the feature or model .I am Chinese，My english is poor，forgiving me ! :-) :-)
>
> I wrote recommendation for other competition Rossmann, so had deleted it and wrote Sorry)

haha it still works to me

Lê Quang Nam • 2 years ago • Options

0

> ***alexeymosc wrote***
>
> Yesterday I spent some hours to replicate the work of my proprietary algorithm on EUR/USD forex quotes in predicting a 64-minute ahead direction. Of course I knew the validation sample's targets, which was the core goal of this effort.
>
> Not much surprisingly I managed to get a 60% accuracy of prediction on a teach+test sample (24 000 instances) and it spiraled down to 55% accuracy on a validation sample (2 250 000 instances, covering 6 years), with ME of a trade without transaction cost equal to 0.0002 and 0.0001 respectively (those familiar with FX can note the model is hardly profitable out of sample).
>
> That assured me that my algo works on a single symbol at least, generating some positive ME, better that zero.
>
> I guess here we deal with a much more delicate model that is a mix of different stocks behaviour.
>
> Just trying to understand where I am.

Very interesting results! Would you please give me the EUR/USD forex quotes? I would like to give it a try on forex. Thank you very much!

© 2018 Kaggle Inc