We've made a few improvements to the forums. You can read read more on the blog.

Forums / Data Analysis Assignment 1

How I achieved 81.5/85 - walkthrough of analysis with final report and code

Subscribe for email updates.

Sort replies by: Oldest first Newest first Most popular

assignment × Assignment1 × answer × +

Add Tag

Thia, Kai Xin · a month ago

Hi all,

update I have created a website to consolidate all the best learning materials for data analysis that I can find, it is still in beta but do check it out if you are interested to learn more about data analysis: http://www.whizage.org

I believe that sharing is the best way of learning and seeing that I did reasonably well, I hope the stuff I share here will help some people. This is by no means a perfect solution so feel free to comment if I made any mistakes so I can learn as well ^^. Thanks for reading!

Code (*code is fixed now. I have to move "best.iter" to come before iScore, somehow they were in the wrong places before): http://www.thiakx.com/misc/coursera/dataAnalysis/assignment1/assignment1Code_ThiaKaiXin.R

Report (*updated, miss out the square root part for my root mean square error): http://www.thiakx.com/misc/coursera/dataAnalysis/assignment1/assignment1_ThiaKaiXin.pdf

Other well written reports across the forum:

1) 82/85 using just linear regression and stuff taught in class: https://class.coursera.org/dataanalysis-001/forum/thread?thread_id=2681

2) 79/85 using quadratic fico score term: https://class.coursera.org/dataanalysis-001/forum/thread?thread_id=2698

3) 72/85 using linear models and SVD: https://class.coursera.org/dataanalysis-001/forum/thread?thread_id=2752

Step 1:Gathering background information and formulate base hypothesis

First, we should take a look at lending club's website and find out what is their official definition for interest rate: https://www.lendingclub.com/public/how-we-set-interest-rates.action

Summarizing the page, our base hypothesis can be: Interest rate = FICO score + Requested Loan Amount + Loan maturity + some credit attributes + base rate (5.05%)

Another interesting to note is that Amount Funded by Investors does not affect interest rate, as stated on lending club website; interest rate does not change once submitted for funding: http://www.lendingclub.com/kb/index.php?View=entry&EntryID=207

Step 2: Data Cleaning

This part is pretty straight forward. Drop me a message below if you need any detailed explanation for:

• Replace NA values with average

• Remove interest rate and debt to income ratio "%" sign and convert column to numeric

• Split FICO range into 2 columns: FICOMin and FICOMax. Create third column FICOMean by taking the average of FICOMin and FICOMax.

Step 3: Finding out the relative importance of each variable

Our hypothesis: Interest rate = FICO score + Requested Loan Amount + Loan maturity + some credit attributes + base rate (5.05%)

Is there a way to confirm our hypothesis above? Can we have a table that ranks the relative importance of each variable? Multiple linear regressions is probably not the best nor neatest way to do so. Instead, we should consider a non linear method, like gradient boosting from the R library gbm. R library: http://cran.r-project.org/web/packages/gbm/gbm.pdf

Ok, it is really hard for me to explain gradient boosting in a few words.

Basically, gradient boosting creates an ensemble of classifiers by resampling the data, which are then combined by majority voting. Resampling is strategically geared to provide the most informative training data for each consecutive classifier. Source: http://www.scholarpedia.org/article/Ensemble_learning#Boosting

While I will advice you to google around and find out more about GBM, it is ok if you have no idea what I just said. For a start, just know how to use it (sample working code only):

```
gbmMod<-gbm(Interest.Rate ~., loansData, n.trees=1000,
distribution="gaussian", interaction.depth=7, bag.fraction=0.9)

iScore <- summary(gbmMod, 1000)

View(iScore)
```

The code snippet above is meant for you to do a quick test and obviously not my final code (I adjusted a few extra variables for the gbm model, did 5 fold cross validation and obtain best iteration to prevent overfitting and calculated root mean square error on random sampling for verification).

The code basically fits a gradient boosted model on the training data and returns you a table that lists the reduction of squared error attributable to each variable. This helped me determine the relative contribution of each variable in determining the interest rate.

It was clear that the top three attributes – FICO Mean, Loan Length and Amount Requested contributed almost 90% in the prediction of interest rate and should be the key focus of my analysis.

For confonders, I used ANOVA and identified “Amount.Requested”, “Loan.Length”, “Loan.Purpose”, “Debt.To.Income.Ratio”, “Home.Ownership” and “Inquiries.in.the.Last.6.Months” as factors that are strongly associated with interest rate and fico mean. Going through a sanity check, however, it is likely that only “Debt.To.Income.Ratio” and “Inquiries.in.the.Last.6.Months” are logically and directly correlated with fico mean and might be confounders.

Step 4: Charting

This part is pretty straight forward, after identifying the top 3 variables, I simply plotted 3 scatter plots of interest rate against fico mean and color grouped it by 1) interest rate 2) loan length 3) amount requested. I picked 3 random points can called them Sally, Peter and Jack and described how the different FICO mean, loan length and amount requested seems to affect their interest rate base on the plots. To support my observations for Sally, Peter and Jack, I did linear regression based on the initial hypothesis to find out the effects on interest rate for every point gained in FICO mean, every $1,000 increase in requested loan amount or different loan maturity.

Step 5: Conclusion

I reconfirmed my final findings and talked about 3 other things:

• Possible explanation for why the other factors beyond the top 3 factors mattered so little

• Why I expect loan purpose, states and home ownership to play a bigger role as the company expands in future, although they are not statistically significant now

• How we can increase accuracy with more data

Whew~right that's about it. Feel free to ask questions, I will try my best to answer them.

Cheers,

Kai Xin

P.S If you are too shy to post your questions here, you can add me on linkedin and drop me a mail: http://www.linkedin.com/in/thiakx

^ 155 v

Sebastian Meznaric · a month ago

Well done, really nice assignment!

I just have one comment. When you remove NA values and replace them with averages, that will make sample variance a biased estimator. In a case of skewed data it can also alter your median and a bunch of other statistics. This also does not help the quality of the fit and can potentially skew your curve, especially if the NA value has extreme values for other data.

7

Thia, Kai Xin · a month ago

You are correct. I did a sum(is.na(loansData)) and found that there are only 7 NAs in the whole data set. I figured 7 out of 2500 observations with 15 variables is really negligible so I just use mean.

A better solution was in one of the peer assignments I marked. The author mentioned that Monthly.Income = NA might mean a lack of job, so he/she set those Monthly.Income NA columns = 0. If this was a serious assignment for work, we should go through every column and figure out the most sensible way to impute/ignore each NA value

6

Anonymous · a month ago

I think it is better to ignore data points with NA values rather than to make unjustified assumptions about them.

12

Juan C. Morales Brignac · a month ago

Thia,

I just wanted to mention that there were 77 n/a entries (not "NA" but "n/a") in the Employment Length column. These n/a entries are not identified by the "is.na" command. I noticed that the GBM model you used did not include (or it discarded) Employment Length as a factor. I used the function "step" a few days ago, after learning about it in last week's videos, and I had to first remove the 77 n/a's from the Employment Length column before it worked. I guess that GBM did not give you any problems with respect to the n/a's in the Employment Length column.

Nice writeup!

4

Michael Reach · a month ago

Juan, I do not think that the n/a in Employment Length are NAs in the sense you mean. I thought it meant, Not Applicable, meaning that the person wasn't currently working, either retired or unemployed. That's important info for a lender, and I wouldn't have wanted to lose it!

4

Juan C. Morales Brignac · a month ago

Michael, I agree with you. You can just leave n/a as type "factor" and I imagine that the GBM model would treat it as such and would not create an error. What happened to me was that as I worked on cleaning up the loansdata a few weeks ago, I also explored making the employment length a "numeric" variable (which ended up being a waste of time because it is a flat liner according to a boxplot). When I ran step() a couple of days ago, after watching Dr. Leeks video on Model Selection, the n/a's in the data are not numeric and the function step() quit on me. I eliminated these rows just to see how step() worked out. I did not use step() in the writeup I handed in because I did not know about it at that point. I only eliminated the 77 rows a couple of days ago just to see how step() worked out. Thank you.

1

Anonymous · 11 days ago

Hi, is it different to use * or + for adding covaraints ? since you have used * here.... lmLength <- lm(loansData$Interest.Rate ~ loansData$ficoMean*loansData$Loan.Length) thanks.

0

 Thia, Kai Xin · 11 days ago

Hi, when we do *, R does both + and *. In otherwords my code:

```
lmLength <- lm(loansData$Interest.Rate ~ loansData$ficoMean*loansData$Loan.Length)
```

is the same as

```
lmLength <- lm(loansData$Interest.Rate ~ loansData$ficoMean*loansData$Loan.Length + loansData$ficoMean +  loansData$Loan.Length)
```

0

+ Add New Comment

 Joel Pulliam · a month ago

Very helpful! Thanks for doing this. I'll definately look into GBM. Seems like 'step' but more powerful.

0

+ Add New Comment

 Jean-Charles Bagneris COMMUNITY TA · a month ago

Kai Xin, I don't see any problem in sharing your work now, after hard deadline. You might consider putting it somewhere and sharing the link instead of posting it on the forum, though, for two reasons:

- forums are no longer available once the course closes,
- it would probably not be convenient to post your full work and graphics here.

Only my 2 cents.

(And, by the way, congrats for your work!)

6

 Thia, Kai Xin · a month ago

Sure. Added my report and code. =)

0

Jörg Narr · a month ago

Hi Jean-Charles, from previous courses' experience I know that archives are at least available 6 months after completion. But your suggestion is definitely wise also.

0

Ziyad Saeed · 22 days ago

Put the code on Rpubs

0

+ Add New Comment

Charles Roth · a month ago

"For confonders, I used ANOVA and identified “Amount.Requested”, “Loan.Length”, “Loan.Purpose”, “Debt.To.Income.Ratio”, “Home.Ownership” and “Inquiries.in.the.Last.6.Months” as factors that are strongly associated with interest rate and fico mean".

My 2 cents on confounders. Confounders are variables that add noise and make a correlation

more difficult to understand. When doing a regression you expect strong associations. These are not confounders, they are in the variables that explain the differences in interest rates given the same FICO scores. I did not find any confounding variables in this analysis, only variables that helped explain variation or did not help.

▲ 6 ▼

Anne Paulson COMMUNITY TA · a month ago

In my analysis I speculated that Date of Loan Issuance is probably a confounder: if we knew it, we could get better predictions.

▲ 4 ▼

Charles Roth · a month ago

I agree completely. With 75% of the variance explained, it is possible that the dataes could have been tied to the prime rate to explain the other 25%.

▲ 1 ▼

Anne Paulson COMMUNITY TA · a month ago

Not only that, but we can be sure that the dates WERE tied to the prime rate or some other indication of prevailing rate. They had to be, unless we assume that the people at the Lending Club are idiots. The interest rate offered to a borrower has to be a function of the loan's credit characteristics and the prime rate.

And what the heck was going on with the 15 people who got loans despite being well under the Lending Club's cutoff, and got them at much better interest rates than would have been predicted from the rest of the data? I can only speculate that those loans were originated early in the Lending Club's history, when the interest rate was lower and when their lending standards were looser. Those 15 Subprime borrowers stuck out like a sore thumb every time I made a graph.

▲ 4 ▼

Gundas Vilkelis · 21 days ago

The funny thing that the unnamed first column (some kind of ID?) in loansData.csv positively correlated with the Interest Rate (after taking into account FICO and Loan Length). If that column is some kind of a sequential ID, that would imply that the Interest Rate was rising over time. But maybe too many guesses here :)

▲ 0 ▼

+ Add New Comment

ANIMESH KUMAR · a month ago

Thanks Kai Xin. This is very helpful and congrats!

0

+ Add New Comment

Anonymous · a month ago

for cleaning the data, I would suggest removing the data points where the interest rate was inferior or equal to zero, as we are interested in interest as an outcome, and those are either mistakes or people who did not get the loan.

4

+ Add New Comment

Alejandro Foullon · a month ago

Awesome collaborative effort Thia! Thanks for sharing! I will read it and come back with some constructive comments. Cheers.

0

+ Add New Comment

Anonymous · a month ago

Some comments:

You should not replace NA points with averages. This kind of interpolation only works when you have reason to believe a function is continuous and well-behaved in between the two (or more) points you use for interpolation. By performing this kind of linear interpolation you are biasing the results by making an unjustified assumption of linearity. It is better to remove the data points entirely.

Your analysis points out another glaring problem with the peer grading system. You used an analysis technique which was not covered in this course. Your peer graders were then asked to assess the correctness of your methods though they probably had no exposure to them before. I think for future assignments it would be best to stick to the methods taught in this class so that

your peer graders can give a fair and accurate assessment of your work.

I don't think state would be an important factor even if more data were available. A person's home state has no bearing on their creditworthiness.

17

Anonymous · a month ago

Then, "state" might be a confounder?

0

Christopher J House · 25 days ago

I apologize in advance if I offend but I disagree with your assertion and want to give my "two cents" worth on the topic of using methods which were not covered in the course. I think it is important to highlight that there are going to be people taking this course with all kinds of levels of experience and training. I think the critical thing here for peer review is good appropriate references for stats or functions used (also part of the assessment). Therefore, the reviewer can do a little due-diligence, review, at least briefly, this new topic, and then assess the appropriateness. Now, if there was no information or detail given in the report that would permit an assessment of appropriateness for the tests used it would be an easy mark-down for me as a reviewer.

1

+ Add New Comment

Arun Sar · a month ago

Thanks .. it really was very helpful!

0

+ Add New Comment

Michael Reach · a month ago

Good job. I did note that every single one of the FICO ranges was a range of 4, like 350-354. So there was no point in keeping both; you can do the analysis on the minimum or maximum, makes no difference.

PDF-XChange Click to buy NOW! www.docu-track.com

2

+ Add New Comment

 Anne Paulson COMMUNITY TA · a month ago

Nice job. I particularly liked the clarity of your explanation. Not only did you do the statistical analysis, but you persuaded the reader, through your thorough explanation, that your analysis made sense.

2

+ Add New Comment

 Shih-gian Lee · a month ago

Thank you for posting your work. It took courage to post it here for discussion. I know there are quite a few techniques that being used here were not covered in the lectures until week 5/6 or not covered at all. But, through discussions in this forum, I have already learned quite a bit. MSE is a pretty common method in Kaggle competition, even though not mentioned in the lecture. If the instructor can give his take on the analysis here, it will be a valuable learning experience for everyone.

Congratulations on your high score!

0

 Thia, Kai Xin · a month ago

Haha. Yes I am a top 250 Kaggle player =)

1

+ Add New Comment

Limarie Cabrera · a month ago

Great work! On a side note, I'm glad I'm not the only one who found 7 NA records in the dataset. I read not one but TWO analyses that remarked that they found only 2 NA records in the dataset. I began to wonder if I had done something incorrectly, although all evidence

indicated otherwise.

0

Juan C. Morales Brignac · a month ago

Limarie, maybe it referred to the fact that only 2 rows were affected? The 7 NA's were constrained to just two rows.

5

Limarie Cabrera · a month ago

Oh good, you can help me with this then, because it was really bothering me. I had done a subset of is.na(loans) and it pulled up 7 records, with NAs across the entire row. Argh, I should have done complete.cases, right?

0

Juan C. Morales Brignac · a month ago

If you use

> length(which(is.na(loansData))) it will tell you that seven records have an NA. These are records, not the rows themselves.

If you create a new dataframe using

> noNAloans <- na.omit(loansData) dim(noNAloans) the answer to the dimensions will be [1] 2498 15

2498 rows (instead of 2500) and 15 columns. It only eliminates two rows. The seven cases with NA are constrained to only the two rows that it eliminated.

I hope that this helps clarify.

I also checked NA's in each column while I was in week 1 of the project. Rows 367 and 1595 were eliminated. I am new to R and I have found that "playing" with it helps to understand it (I still have a lot to learn!). Dealing with a data set with NA's and n/a 's has required me to adjust my mentality regarding data sets.

1

+ Add New Comment

Dave Niesman · a month ago

Hi,

Thanks for posting your information. It's a fine analysis.

In my data set, I removed both unfunded and loan entries where the data contained a N/A. I was left using 2492 data points. Did you include an unfunded loan with all valid fields?

With regard to the gradient boosting, was this something I should have picked up in the lectures or did you bring this to the assignment from your background?

Thanks.

0

Thia, Kai Xin · a month ago

Hmm, I think since interest rate is set before the loans are released for funding and interest rate will not change even if the unfunded loan is put up for 2nd chance, I think we should keep the unfunded loan with all valid fields.

I learnt gradient boosting from an earlier coursera course: Machine learning by Andrew Ng. https://www.coursera.org/course/ml

I believe we should be taught gradient boosting in later parts of this course as well since it is a fairly common and powerful method =)

0

Shih-gian Lee · a month ago

Kai Xin, I searched the Machine Learning pdfs but could not find it. I also don't remember learning that in Andrew's course. I took that a while ago. It would be nice if I could review the concept again.

0

Anne Paulson COMMUNITY TA · a month ago

I too took the Coursera Machine Learning course from Andrew Ng. I do not remember learning about boosting from that (excellent) course. I went back and looked at my notes; I believe boosting was not covered in that class.

There is a set of lectures by Andrew Ng about machine learning floating around the net. They are the lectures from a more advanced machine learning class, not the one offered on Coursera. Perhaps Ng covered boosting in those more advanced lectures?

0

Thia, Kai Xin · 25 days ago

Sorry, my bad for not explaining clearly.

Theory wise: I meant gradient descent. The R GBM library function I use "implements extensions to Freund and Schapire's AdaBoost algorithm and J. Friedman's gradient boosting machine" and both AdaBoost and Friedman's model are based on gradient descent: http://www-stat.stanford.edu/~jhf/ftp/trebst.pdf.

Code wise: I am participate in Kaggle competition and GBM happens to be one of the most common library. I learnt to apply the GBM model in R by myself through experimentation and reading forums (like this) on Kaggle. So....yep this will take time, took me about a month to learn how to code GBM, randomforest, bagging etc in R and until today even after Andrew Ng lesson I am not fluent in explaining the theory/statistics behind in.

▲ 0 ▼

 Anne Paulson COMMUNITY TA · 25 days ago

I understand gradient descent-- it's a way to do optimization, and logistic regression for example uses gradient descent. What I don't understand is how gradient descent, or any optimization technique, is applicable to what you are doing. You say "The code basically fits a gradient boosted model on the training data." What kind of model? Is this a neural network with real-valued output using AdaGrad to train faster?

▲ 0 ▼

 Thia, Kai Xin · 25 days ago

Well, the R gbm library got many distributions to choose from. I used "gaussian" which bsaically employs least squares regression using J. Friedman's gradient boosting machine modeling (from the friedman paper above).

From what I understand, gbm basically fits a series of small decision trees (I chose to have 1000 tress) and we control the learning rate via "shrinkage" parameter. It will then follow a gradient descent way of making taking small progressive steps towards the global minimum (hopefully, so we need to give it enough trees to converge).

The cool part is at every progressive step it will learn from the previous step what are the poorly predicted obervations before recommending the next iterative step (hence we need to set the bag fraction)

So the final code I used for gbm is this:

```
gbmMod<-gbm(Interest.Rate ~., train,n.trees=1000,
        shrinkage=0.01, distribution="gaussian",
        interaction.depth=7,
        bag.fraction=0.9,
        cv.fold=5,
        n.minobsinnode = 50
)
```

▲ 0 ▼

+ Add New Comment

 Raja Doake · a month ago

Reading through your analysis and code was very instructive, and inspired me to post my assignment as well! I'll create a separate thread for it so as not to clutter yours.

The machine learning-style approach you used -- splitting data into training/testing, etc -- was effective and robust. In contrast, using standard multiple linear regression, I struggled to look for ways to validate my models. I'm looking forward to diving into this now that we're on to prediction. It's been a long time (wow, 9 years!) since I took a machine learning class and needed to think about data this way.

0

+ Add New Comment

Andrew Oswald · a month ago

Speaks volumes. Nice job!!

0

+ Add New Comment

 Jaiyon Han · a month ago

Thanks for sharing. It is one of great ways to learn something looking at the best job. Thanks~~

1

+ Add New Comment

Anonymous · a month ago

As the reasonable criticisms here demonstrate, part of the reason you got 81.5, instead of 83, or 78, was luck. Your efforts were clearly a step above the average student for the course, but there's a pretty sizable ε term in all the peer grades.

0

Thia, Kai Xin · a month ago

Agree with the luck part (such is life?). Looking beyond the absolute numbers, we are all here to learn something, so to me grades does not matter =P.

5

+ Add New Comment

Anonymous · a month ago

Nice work, thank you much for sharing.

On the second plot, you should have chosen different hues of red and green. I suffer from dichromacy like 5 - 10 % of all human males, and I have a hard time telling apart the colors in your plot.

Another curious observation: I identified and mentioned the same confounders and you did, but I did not use the word "confounder" in my paper. I ended up getting 0 for the "confounders" rubric.

1

Thia, Kai Xin · a month ago

Thanks for pointing out the issues with the hues, I will take note of that in my future work.

Cant help you with the confounder part though, I guess you have to make a separate section for confounder in your next report.

1

+ Add New Comment

⬇ scroll down for more ⬇

PDF-XChange
Click to buy NOW!
www.docu-track.com