Winning solutions - Data Hackathon 3.x

kunal Sep '15

This is a thread to share the winning entries for the weekend hackathon.

@phani_srinath **@binga** will share their approaches here.

This is to make sure we beat them in next 3 days.

Regards,
Kunal

phani_srinath Sep '15

@here: Here we go. As kunal said, here is the approach that i have followed.

Data Pre-Processing:
1. The moment i received the data, i realised that there were a lot of columns with missing values and a lot of levels
2. I also realised that there were a few date columns.So extracted the year, month, age etc. As far as the Lead_Creation date is concerned, i extracted the month and as it has only ~80 levels in the data, I didnt drop that column but as DOB has a lot of levels i dropped it aftr extracting the year , month and age
3. As far as missing values are concerned, because there were a lot of missing values, I replaced all those with a placeholder value lik -9999 so that the algorithm treats these missing values as a seperate level
4. For attributes like City, Employer_name, as they had a lot of levels again, i merged all the rare levels.

Modeling:
1. I started with a RandomForest. Performed a 4 fold CV on the data without the CIty and Employer_name and tuned it to get to ~0.84 on the LB
2. Later tried XGBoost (XGBClassifier a Scikit learn wrapper of XGBoost) and a 4 fold CV on the same data and tuned it to hit ~0.859 on the LB. But the biggest problem for me was the CV(scikit learn implementation) was still ~0.848 and was struggling to get a better CV.
Though I was on top of the leaderboard I was worried as my CV wasn't great
3. Then I tuned another XGB model after including city and employer_name which gave me a CV of ~0.858 and I could hit on the LB ~0.8622 with just this model. I was a lot more confident with this one.
4. Now averaged both the best submissions (2 and 3) to get to 0.8639 on the LB. But I was worried if i was overfitting as the first one wasnt really a good one because the Cv was on the lower side. Hence i did a weighted average of the two submissions by giving a lower weight to the first submission. And thats it. 0.8644 on the LB.

5. I strongly believe that the weighted average saved my day.

Varnikaa Sep '15

@phani_srinath - Thank you very much for sharing the approach and congrats on winning. I have few basic questions. Very very basic actually.

Data Pre-Processing:

Point no.2 - Could you please elaborate what is mean by dropping the variables after extracting the year, month and age? I understand that after doing a exploratory analysis on the "dob" you dropped it. Is this correct?

Point 3 - How effective replacing a missing values with mean/median or any other method of handling missing values rather than imputing "9999". If there are lot of missing values and if you replace this with 9999, will this not become an outlier or dominate the dataset?

Point 4- Is there any thumb rule you maintain for merging the rare levels?

Modelling:

Point 4,5 - What is mean by averaging of the submissions? I hope it is not ensembling? Does that mean the simple average of the output?
If you can elaborate it will be helpful.

Thank you.

binga Sep '15

Hello everybody,

I have finished 2nd on the private LB for the weekend competition

My public score: 0.86126 and private score: 0.84138. Missed out on the 1st by a very tiny 0.0004.

Anyways, coming to the approach,

- I used Python completely.
- I have initially started off with only a few set of variables. Mostly numeric variables. I have decided to use 4-fold CV till the end of the competition (4 because my laptop has 4 cores and anything more than would be a computationally expensive and slow iteration time). Treated missing values as -3.14.
- Tried out LR, RF and XGB. Found out that RF CV was around 0.83 CV and 0.844 LB and XGB was able to reach 0.84 CV and 0.849 public LB.
- On Saturday evening, the others overtook me quite easily and then I realized that they must be using the other variables. So, I decided to use dates (day+month+year etc.), cities and finally employer names. I removed rare levels.
- That's when I could push XGB to 0.8557 and public LB 0.8557. It was an Aha! moment because my CV was exactly my LB score. But then, I was stuck there for a long time. Then, I realized I made a mistake tuning the XGB and instantly saw an improvement to 0.859 CV and 0.8594.
- On Sunday, I realized I couldn't push the XGB any further and I thought of ensembling. Sadly, my

RF which was performing upto 0.83 the previous day wasn't even scoring 0.79 CV which was a big disaster on Sunday afternoon. I just couldn't push it till the end of Sunday.

- Instead of spending time on that, I tried other algorithms and luckily a simple Logistic Regression scores 0.836 on LB. I was pleasantly surprised (shocked too!) to see that it performed well. So, I decided to use an ensemble of XGB + LR and the public score went to 0.8612. I couldn't cross-validate due to lack of time and decided on some weights for each of them.

I was pretty sure that my score wouldn't drop by a fair margin because I used an ensemble. Unfortunately, I just fell behind my a tiny score. Well, I should have cross-validated for the right weights.

Big mistake: I should have saved my RF code/features/params. That was a bad mistake.
Big learning: Logistic Regression was working well. I was lucky to try it in the last few hours and add it to the ensemble
What didn't work: KNN - 0.79 CV. I think it can be done better than this.

Tuning and CV strategy for XGB:
Typically, people use 5 folds. You can make a choice. To see the reliability of CV estimate, a few guys use 10-fold as well.
Steps: 1. Decide 'n' in n-fold. Stick to it for complete analysis.
2. Create a baseline score using a simple model.
3. Now, use XGBoost default settings and establish another XGB baseline score.
4. Put num_trees at 10000 and a tiny learning rate of 0.01.
5. Try step(4) for various max_depth.
6. While doing step(4), monitor the progress. Note at what tree# is the model overfitting
7. After you're done with 1-6, you would have reached a saturation score
8. Now comes some magic! Start using subsample and tada, your score improves.
9. Use colsample_bytree, then scale_pos_weight, improve your score
10. Try using max_delta_step and gamma too (a little tricky to tune)

Modekurthy_Venkata_S Sep '15

@Varnikaa
I can clarify few points , others can disagree if I am wrong

Point 2 : I am not clear on this . I derived age of person on the day Lead creation date , Datediff(Lead Creation Date , DOB) . Then I dropped Lead creation date and dob. Lead creation date has no use as it is sample of past 3 months data .

Point 3 : As we are going by decision trees model, model will take care missing values .
To run code , we pass arguments to function . One of the argument is missing value .As data is dense matrix , NA is masked by 9999. So function interpret 9999 as missing value. Make sure , data does not consists this number ,before we replace NA .

Point 4 ,5 : Averaging (Probability of individual disbursed of model 1 and model 2)
Ensemble is process of adding weak models. so he took average .

---

phani_srinath Sep '15

Excellent questions **@Varnikaa** , not many people think so much. OKie to start with your 1st question...
From the DOB as for different customers you can have different DOB's, i might not have a lot of people

withthe same DOB.Which means I might not really have a pattern there infact it would result in a lot of variance and when i convert that to dummy variables, it would result in a lot of dimensionality. Hence what we do is we extract the age(numeric), year of birth, month of birth, as they possess finite number of levels. Then I went ahead and dropped the DOB column. But for the Lead_creation_data, i extracted the month, of lead creation, but as there were only 4 months where the lead creation has happened for ~87K customers, there are good number of leads created on each day. So there is a chance of each of these dates resulting in some kin of a pattern, hence, despite extracting the month and other stuff i didnt delete the Lead_creation_date column.

1. Well, for the missing values, there are various techniques that you can use. The reason i chose to fill it with a place holdervalue instead of a mean/median is,I wanted that to be considered a level in itself. Yes that might be an outlier for numeric attributes where you need to handle it slightly differently (your mean or median imputation might work better in case of numeric) but for categorical attributes i am more interested in the pattern that each of these levels results in. So I deliberately considered that a pattern. Not having a value for me is a pattern. Instead of looking it as amissing value, i look at it as a pattern in itself. Think about it. It was a trial that was made and it worked for me. In fact one of the most imp attribute that RF threw was CIty-9999

2. For merging the levels, I generally look at the counts of the levels and try to merge levels with less counts. I havent tried any other types of mergining.In my previous firm people followed an approach where they used to merge those levels whose class distributions are similar. Just another approach. Not sure how effective that could be

3. Averaging submissions is also a kind of ensembling. Average that i mentioned is a simple average. As both my models were XGB models (structurally same) i could simply average the submissions for an ensemble. but if my models were different then I might have to look into other ensembling techniques than averaging or weighted averaging. Probabilities from different models have a different behaviour for example RF very rarely gives you a probability 1. it very often throws values like 0.95 for extremely confident predicitons too. thats an inherent bias that the algo possesses. To nullify this you need better ensembling techniques than simple averaging.

I hope i have been able to answer all your questions.

manaman Sep '15

Most of the missing values in this data set(Loan Amount Submitted , Loan Tenure Submitted, Interest rate) are numerical. What are the categorical missing values are you talking about?

Varnikaa Sep '15

@phani_srinath **@Modekurthy_Venkata_S** - Awesome answers. Thank you very much for clarifying me. Extraction of the age is a technique that i had never heard of. I should try that.

One more question. RF didnt show me the DOB as an important variable. How was that for you? Was that showing as an important variable after the conversions that you made?

One of my initial hypothesis is that lead creation date should not make any impact for any of the models. When i compared the lead create date to the target variable, all the dates

> were almost equally distributed or no significant variance seen among the lead create date and target variable. Hence i did not include the variable in any of my analysis. Is it right to do this way?
>
> Imputing with 9999 was one another interesting stuff that i learned here. Have you noticed the variable "Var1" with value HBXX. It didnt had any interest rate, processing fee and loan EMI etc. How you managed to handle this? I did a subset of this part of the dataset and I was fighting to build a separate model by itself and combining the results with the remaining part of the data.
>
> Thanks again for your valuable inputs. Happy to compete with you guys 😄

manaman Sep '15

Most of the missing values in this data set(Loan Amount Submitted , Loan Tenure Submitted, Interest rate) are numerical. What are the categorical missing values are you talking about?

Amit_Mohanty Sep '15

good points here.

Varnikaa Sep '15

@phani_srinath **@Modekurthy_Venkata_S** - Awesome answers. Thank you very much for clarifying me. Extraction of the age is a technique that i had never heard of. I should try that.

One more question. RF didnt show me the DOB as an important variable. How was that for you? Was that showing as an important variable after the conversions that you made?

One of my initial hypothesis is that lead creation date should not make any impact for any of the models. When i compared the lead create date to the target variable, all the dates were almost equally distributed or no significant variance seen among the lead create date and target variable. Hence i did not include the variable in any of my analysis. Is it right to do this way?

Imputing with 9999 was one another interesting stuff that i learned here. Have you noticed the variable "Var1" with value HBXX. It didnt had any interest rate, processing fee and loan EMI etc. How you managed to handle this? I did a subset of this part of the dataset and I was fighting to build a separate model by itself and combining the results with the remaining part of the data.

Thanks again for your valuable inputs. Happy to compete with you guys 😄

Arunkumar_t Sep '15
Awesome Hackers!

For lead date -> month can be extracted and its proving useful in my test runs.
For DOB -> DOB should be converted to age or extract only year from DOB so that variable becomes more meaningful.

Both the above variables are important after treating them appropriately.

ML Algorithm should take care of Var1 with HBXX on its own while forming the tree.

Others can validate this info.

---

abhijit7000 Sep '15

First of all Congrats to **@phani_srinath** and kudos to **@Varnikaa** for some excellent questions.

@phani_srinath Can you further clarify on your approach on **"Imputing the Missing value"**. As also highlighted by **@manaman** those variables are numerical and not categorical.

I also wanted missing values to be considered a level in itself. But I could not decide between -9999/0/9999/999999. Because, if these missing values were significant then i would be inadvertently attaching more weight to lower percentile, median or upper percentile,respectively of the numerical distribution.

I also entertained the idea of imputing values with 0, and creating a separate flag variable to indicate missing. But that increased the dimensionality and complexity. So, please do clarify.

Thanks in advance for the precious learning experience.

---

abhijit7000 Sep '15

I have another doubt regarding **"Outlier detection and treatment"**
How did you guys took care of outlier values for variables like Monthly_Income , Loan_Amount_Applied?

I am able to see the outlier present using plots and boxplots, but how to code it dynamically? Following the boxplot logic, outlier is any value greater than [Q3+ (IQR)*c], where c is usually 1.50 .

Value of ‘c’ found after trial n run:-
- Monthly_Income - 12
- Loan_Amount_Applied – 5
- EMI_Loan_Submitted – 6

Is it a must to treat outliers before applying Random Forest? Is there a better way to treat the outliers in R?

Also the outlier should be capped up with what value? is it [Q3+ (IQR)*1.5]?

---

aayushmnit Sep '15

Hi **@abhijit7000**,

Generally tree based ensembling methods are robust towards outlier , but in my personal experience if you treat outliers and then run these models, your model will perform better. In terms of how to treat them, say if monthly income is going above 1000000 , i will modify it to 1000000.

Hope this helps.

Regards,
Aayush

Kalyan Sep '15

Thanks All for the great question and answers.

if it is possible could you please share your code on Git. this really helps.

Thanks

---

abhijit7000 Sep '15

Hi **@aayushmnit** ,

Thanks for the help.
I am able to understand your approach.

But I have a followup question. How did you decide the cut off for Monthly_Income is 1,000,000. Is it by manually observing the values, or are you using a statistical formula which could be replicated to any variable? If you could share that formula or approach, it would be great.

Regards,
Abhijit

---

aayushmnit Sep '15

Hi **@abhijit7000**,

I looked into the data and found that anybody with income more than that has not been disbursed. This could be because of many reasons, either monthly income is mispunched or may be because of something else. So I restricted it to a point i.e 1,000,000. I created a complementary factor of EMI / Income ratio so that these kind of instances can be tackled.

Regards,
Aayush

---

abhijit7000 Sep '15

@aayushmnit thanks a lot for sharing your approach.

---

aayushmnit Sep '15

Thanks **@abhijit7000**,

If you want to read my full approach with all the codes. Please click here , its my Github link.

Regards,
Aayush

---

abhijit7000 Sep '15

@aayushmnit , Thanks a lot for sharing your codes 😀

I really learnt a lot from your codes. In return I would like to highlight a slight scope of improvement in your code.😉

It is easier to append the Train & Test data into one data frame, then apply all preprocessing on this single data frame and finally break it back into 2 using row indexes. This removes the redundancy of writing the same code twice for Train & Test separately.

Hope this helps
Regards,
Abhijit

Kalyan Sep '15

@aayushmnit Thanks a lot for sharing your codes. Really Helpful.

---

aayushmnit Sep '15

@Kalyan and **@abhijit7000** ,

I have also forked codes from binga and manaman, they both finished above me. I think you guys are not on slack portals where all the codes sharing is happening. So you can look for there codes also in my profile.

Regards,
Aayush

Kalyan Sep '15

@aayushmnit how to join slack team? I am already in slack, but dont know the team name. Thanks..

---

aayushmnit Sep '15

I think you need invitation from **@kunal**

---

abhijit7000 Sep '15

Thanks @aayushmnit for the info.
@kunal can you invite me and @Kalyan to the slack portal.
I am also on slack but dont know the team name.

Thanks,
abhijit

---

abhijit7000 Sep '15

Hi @Kalyan ,

If you happen to get any info on joining the slack portal, would you give me a heads up. Likewise, I would update you if I get in.

Thanks,
Abhijit

---

praveen766 Sep '15

@phani_srinath can you share your github link or codes here

---

kunal 🛡 Sep '15

@Kalyan @abhijit7000

We have sent the slack invites to you multiple times on the email registered with us.

I have resent them today - pelase see if you have got them.

Regards,
Kunal

Amit_Mohanty Sep '15

@kunal .. can you please send me slack invite??

---

abhijit7000 Sep '15

Thanks @Kunal , for the re-invite 😄

---

numb3r303 Sep '15

I would also add one thing which I found out that we could do some unsupervised learning on the dataset to get into more detail about the data, just running a KMeans Clustering on the data gave some insights into the structure and we could use these labels as features in our learning models, gave me a considerable boost to my model.

Shyam_Naren Sep '15

@kunal - can you send the slack invite to me as well ? Thanks !

Raghuvaran_Raghu Sep '15

It would be nice if anyone uploads solution procedure for Hansa Cequity Hiring Hack.

zxy Oct '15

@kunal can you send me a slack invite. new to AnalyticsVidhya. Thanks !

jeevanananda Oct '15

@kunal Could you please send me an invite to join on Slack . Thanks