Profiling Top Kagglers: Leustagos, Current #7 / Highest #1

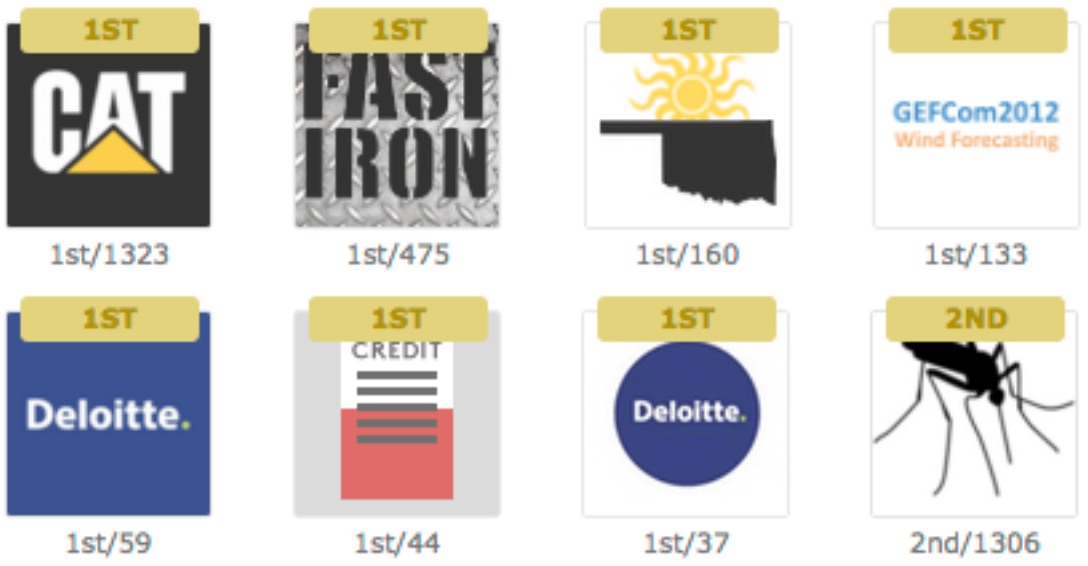
Next up in our series Profiling Top Kagglers is Lucas Eustaquio Gomes da Silva (better known as Leustagos on Kaggle). Leustagos is one of only 13 data scientists to ever hold the #1 spot on Kaggle, and he has been a consistent face at the top of our user rankings since joining the community four years ago. In this blog, Leustagos shares what he's learned in his years competing, his typical approach to a new competition, and also how Kaggle has helped him develop professionally as a data scientist. Leustagos is currently busy kicking some serious cancer ass so we want to send him extra thanks for taking the time to share this interview with us.

How did you start with Kaggle competitions?

I started back in the second half of 2011 just after doing the Andrew Ng Coursera class. I found people talking about ML competitions on the course forum and got curious about it. I tried taking part in one competition and placed pretty badly. After this I really wanted to learn about the magic used by the winners.

What is your first plan of action when working on a new competition?

Check the dataset to understand how to build a validation set.



Leustagos' top 8 finishes

What does your iteration cycle look like?

1. Understand the dataset. At least enough to build a consistent validation set.
2. Build a consistent validation set and test its relationship with the leaderboard score.
3. Build a very simple model.
4. Look for approaches used in similar competitions in the past.
5. Start feature engineering, step by step to create a strong model.

6. Think about ensembling, be it by creating alternate versions of the feature set or using different modeling techniques (xgb, rf, linear regression, neural nets, factorization machines, etc).

What are your favorite machine learning algorithms?

Gradient boosted trees by far! I like GBT because it gives pretty good results right off the bat. Look at how many competitions are won using them!

What are your favorite machine learning libraries?

xgboost, scikit-learn and pandas

What is your approach to hyper-tuning parameters?

I look for params used in previous competitions and tweak them one by one until the outcome stops improving.

What is your approach to solid CV/final submission selection and LB fit?

We are usually asked to choose 2 submissions. My first submission is always the one that performed more consistently both on cv and the leaderboard. Not necessarily the best leaderboard score if I think its overfit. The second submission is usually one that performs good enough but its somewhat diverse from the first one.

In a few words: What wins competitions?

Good feature engineering, strong ensembling skills and, many times, team work.

What is your favourite Kaggle competition and why?

Global Energy Forecasting Competition 2012 - Wind Forecasting. It was the first competition that I was a prize winner and 1st place at that. It felt so good!

Have you ever hit a plateau during a competition? What, if anything, helped you move beyond this?

Many times! I usually try to take another look at the dataset to check for other angles. Seek information in the forums, scripts, and more importantly, when this happens I try and brainstorm with my fellow teammates if I'm on a team. If I'm not, and none of the previous options worked, I look for a team.

What field in machine learning are you most excited about?

I like working with time series and classification tasks. Not exactly a field, but I'm not

tied to any industry in particular. I just like a good challenge.

Which machine learning researchers do you study?

I don't study anyone specifically, but I'm willing to learn anything that works in practice. For that I look into papers. Of course, when dealing with a new problem I look for the state of the art too. Nowadays its very hard to find good and useful papers because researchers are forced to write many of them so the quality goes down.

Can you tell us something about the last algorithm you hand-coded?

I don't remember hand coding anything more complex than a linear regression. I'm more empirical than theoretical and given my background in software developing I try to avoid coding from scratch as much as possible. Its very error prone. I prefer to understand algorithms at a high level and to know the task they are best at solving. After that is just putting together pieces of a puzzle.

How important is domain expertise for you when solving data science problems?

It helps, but its not mandatory.

What do you consider your most creative trick/find/approach?

I like calculating smoothed likelihoods to replace categorical variables with high cardinality. I know some ways of doing so. Applying mathematical transformations to output to more closely match the error function can be very tricky too.

How are you currently using data science at work and does competing on Kaggle help with this?

I'm working with threat detection in enterprise networks. We analyse network traffic and use machine learning to identify malicious behaviour, like a command and control attack. Competing on Kaggle, besides being addictively fun, helps to broaden my skills and approaches. There is always something to learn.

What is your opinion on the trade-off between high model complexity and training/test runtime?

I really depends on the application. In competitions, better performance outweighs everything. For real applications, the solution I seek for is usually the simplest solution that meets the client's performance criteria. It must be at least good enough in performance, and it must be easy to use and maintain. Many clients cannot afford much trickery for feature engineering or complex algorithms in their workflow.

What is your view on automating Machine Learning?

It can work. I don't know how it will shape the data scientist job, but I think there will always be a need for one. A automated system is useless if there is no one that understands how to operate it and interpret its outcome.

How did you get better at Kaggle competitions?

Trial and error, reading the forums and copying successful approaches, learning from team mates and always, always, trying to be very creative on top of all that. I like to to add my own personal touch.

What have you learned from doing Kaggle competitions?

I've learned how to do competitive machine learning, which is a bit different from practical machine learning. But I've learned how to do the practical too, as those simple concepts of dealing with data and algorithms are needed at every corner in competitive machine learning.

Bio

Lucas Eustaquio Gomes da Silva, also known as Leustagos, is a mostly self taught data scientist who graduated in control and automation engineering. He currently works in the data mining field using machine learning techniques to identify and alert clients about malicious traffic in corporate networks.