Exclusive Interview & AMA with Data Scientist – Rohan Rao (Analytics Vidhya Rank 4)

Introduction

There are several aspects to learning a new technical skill. You obviously need to learn the technical stuff, the applications, the hacks and obviously the science. But, in addition to these, you need mentors. You need people who have travelled the path before you, to make sure you learn from them.

Rohan Rao

Interacting with industry experts is one way to do so. With this in mind, we asked one of our top data scientist, **Mr. Rohan Rao**, to do an interview with us and an AMA with participants of "The Ultimate Student Hunt". Rohan not only agreed to give his time, but also made sure he answered all the questions by spending extra time with the participants of the hackathon.

For those of you who don't know Rohan, he completed his Post-Graduation in Applied Statistics from IIT-Bombay and is deeply passionate about machine learning & numbers. He is currently ranked 4th on Analytics Vidhya and 105th on Kaggle (at time of publishing this interview).

Rohan is also a high achiever in the world of Sudoku and Puzzles. He is a 4-time winner of National Sudoku Championship, 4-time winner of National Puzzle Championship and 3-time winner of the prestigious Times National Sudoku Championship. He spends his free time reading and discussing about latest developments in the data science industry on his blog.

Here are some excerpts from our interview and AMA with Rohan.

KJ : First of all, I would like to sincerely thank you for devoting time for this interview. To start with, please tell us how did your journey in analytics begin ?

Rohan Rao : I've always liked math and numbers. I pursued it through my studies, and analytics seemed a natural choice for me when I began working.

With interest, passion and perseverance, I picked up and specialized in building end-to-end Machine Learning solutions, which I currently enjoy doing.

KJ : No journey can reach this level with some challenges and sacrifices. What were the obstacles you faced and how did you overcome them ?

Rohan Rao: The biggest one is opportunity cost. When you really want to pursue something, you have to give up on other things. I always found it hard to sacrifice things to pursue Sudoku or Machine Learning, but I tried doing it as best as I could. Now, when I look back, all those sacrifices were worth it. Also, there were times when I wasn't performing well and wasn't able to improve. I think everyone goes through good and bad phases, it's best to just take some time off and come back recharged, harder and stronger. It worked for me.

Always pursue your passion, whatever it may be.

KJ : How did you start participating in the competitions and hackathons? When did the first win come your way ?

Rohan Rao: I began competing in ML competitions in late-2013, as a means to improve my general know – how in the field. It quickly caught on to me and I started enjoying it with every new competition and challenge.

I won my first contest in my first ever Kaggle competition (it was a hackathon!), but credit goes to my team-mates and colleagues, who did most of the work. But since then, I've never looked back.

KJ : How do you decide which competitions to participate ?

Rohan Rao : Data consistency + Domain + Time + Personal value / growth from the competition.

I think time is a crucial component. Competitions require a lot of investment of time and effort, and one needs to make a decision whether it would be worth it. When I compete, I go all in, never do it half-heartedly.

KJ : According to you, how should people approach problems in data science competitions ?

Rohan Rao: Let me put it down in 9 basic steps:

1. Understand the problem / objective you are trying to solve.
2. Understand and summarize what data you have / need.
3. Carefully read about the evaluation metric.

4. Explore and visualize the data, build simple, base models for benchmark.
5. Setup a robust / thorough validation framework consistent with the evaluation conditions.
6. Work on feature engineering and optimizing algorithms.
7. Try out as many different models / ideas as you can.
8. Ensemble / Blend / Stack multiple models.
9. Never hesitate in asking questions, taking help or even teaming up with others.

Just like a Sudoku has 9 digits to work on, I would recommend these 9 steps to work on as much as possible in every DS competition.

KJ : How do you gauge the complexity of the problem at the start of the competition ?

Rohan Rao : Explore the data as much as possible. Plot features, summarize columns, build benchmark models, and during the process, get a sense of the problem, data, time, complexity, etc. And then slowly build a good solid concrete solution by working on one idea after another.

KJ : Which programming language do you prefer to work in ?

Rohan Rao : Tools, languages, models, algorithms are developing fast. The more you know, the better. Python and R are the best to start off and master, and I use them both.

For quick summarization, plots and prototypes, I use R.

For text mining, production models and scalable solutions at work, I've generally used Python.

KJ : Which is your favorite machine learning algorithms and why ?

Rohan Rao : XGBoost will take the top spot. The main reason being, it is powerful, robust, fast and a clever algorithm.

The beauty of ML solutions is in understanding the problem and data, and exploring it in detail for feature engineering. Traditionally, when I just started out, I was spending far too much time trying out different algorithms and tweaking parameters. There's not much fun and skill in that. With XGBoost, getting a solid base algorithm became a matter of few minutes, and my focus naturally started shifting to feature engineering, which I believe is the most wonderful and challenging part of building ML solutions today.

And of course, it's been a part of my winning solution for most of the contests I've

done well in, so a big thanks to the community who are actively developing and improving it each day.

I also like Collaborative Filtering techniques, which I've implemented very often in my work. With a rich dataset, CF-based algorithms can give exceptionally good results for Recommendation Engines.

KJ : Why is it important to know the exact functioning each algorithm we are working with ?

Rohan Rao : A lot of people are building models as blackboxes. It's easy and it can take you to some level. If you really want to grow and get better, it's useful to know how the algorithms work so that you can tie in feature engineering into the models and make better progress.

KJ : What are the techniques you follow for feature engineering ?

Rohan Rao : I use Excel a lot, very useful for really quick summaries and visuals. If data is huge, I work on a subset, and most of my feature ideas are via plotting or visualization. The exact feature to engineer becomes a bit dependent on the model, but once you spot a trend/pattern, converting it to a feature is a skill that gets developed over time.

KJ : What are some good ways for feature selection ?

Rohan Rao : My thumb rule of feature selection is based on CV or Val scores. If selecting a feature improves CV score, I use it, else discard. For large number of features, I usually build small quick models and check variable importance or information gain, and select the top-x from them.

KJ : How do you deal with high cardinality categorical variables ?

Rohan Rao : First try out Label Encoding and One Hot Encoding. In most datasets in my experience, one of these two is good enough. Also, in the meantime, things are developing. Earlier, R's Random-Forest couldn't handle over 32 factors. Today, H2O's RF handles a categorical variable with 1000+ levels (if I remember right). Maybe someday, that number will be 1 lakh. Who knows!

KJ : How to avoid over-fitting ?

Rohan Rao : There are lots of methods like outliers-removal, regularization, bagging, etc. These days many of the latest models have parameters to control overfitting besides the fact that algorithms themselves are getting smarter by the day.

KJ : What are some points to remember to prepare a robust cross-validation set ?

Rohan Rao : Preparing a robust validation set is very important. It's best to replicate the system that is used on the Leaderboards or for the evaluation criteria. There are times when CV and test scores do not work in sync, and often these are uncontrollable due to the nature of the data.

KJ : Who is your favorite Data Scientist or role model ?

Rohan Rao : There are a few. Hard to choose one. I haven't met many of them in person. But there is one who is really special to me. Shashishekhar Godbole (former Kaggle Top-20). He taught me a lot of ML during my initial years. One of the best Data Scientists I know.

KJ : You just joined a high prize competition, Who are the 2 other data scientists you would want to see in the competition with you ?

Rohan Rao : Wow! I've never thought of that before. Let's see, I'd pick Marios (a.k.a. KazAnova) and Owen.

KJ : What is that one advice would you like to give to your younger self ?

Rohan Rao : Interesting, I'd tell myself to broaden my DS network sooner.

KJ : What advice would you give to fresher's to get their first break in data science / analytics industry ?

Rohan Rao : Data Science is vast. Follow the divide-and-conquer philosophy. Pick small elements and problems, read/research about them in detail, work on ideas, build end-to-end solutions from scratch (including coding, analysing, presenting, etc.) and understand the entire scope and flow of the projects. Slowly and steadily, pick harder problems, pick more challenging competitions and start exploring/improving in areas you enjoy the most.

So, LEARN and PRACTISE. Then PRACTISE and LEARN.

We decided to add some fun in the AMA with a quick Rapid Fire round. Rohan was expected to answer the first thing that popped in his head after listening to the question. And here we go:

KJ : Sudoku or Data Science ?

Rohan Rao : Both

KJ : 4 GB RAM Mac Book Air vs. 128 GB RAM instance on AWS ?

Rohan Rao : 128 GB RAM on AWS/GCP

KJ : In a Hackathon – team or individual ?

Rohan Rao : Individual

KJ : Mumbai or Bangalore ?

Rohan Rao : Mumbai

KJ : Most memorable competition ?

Rohan Rao:

On AV : Seer's Accuracy.

In general : Telstra on Kaggle.

KJ : One secret winning recipe you haven't shared with anyone till now ?

Rohan Rao : (If I share it, it won't be a secret anymore) Ok, but here's one. I love dropping features. Sometime, less is better.

KJ: Thanks Rohan for that awesome interview and AMA. I am sure the community will benefit tremendously by this. We wish you all the success in your upcoming World Championships this month and hope to see you on top in Knocktober 2016 later this month.

For those of you, who want to continuously learn from top data scientists and learn by doing data science – check out our latest hackathons here.

You can test your skills and knowledge. Check out Live Competitions and compete with best Data Scientists from all over the world.