Approaching Competitive Data Science

Date : 19th November 2022 | Speaker : Ayon Roy |
Venue : Ab Data Bolega Online Session by ITM Khargar's Data Freak Community

Visit - AYONROY.ML

Hello **Buddy!**

I am **Ayon Roy**

Executive Data Scientist @ NielsenIQ

Z by HP Global Data Science Ambassador

Mentored/Judged **100+** Hackathons

Delivered **60+** Technical Talks

Brought **Kaggle Days Meetup** Community in India for the 1st time

If you haven't heard about me yet, you might have been living under the rocks. Wake up !!

Agenda

- What is Competitive Data Science (CDS) ?
- Why should you try CDS at least once?
- How should you start CDS? (Approach & Start)
- Where to & How to get involved with core level CDS? (Launch)
- Is CDS everything?
- What other than CDS you should focus on to become hirable?

What is
Competitive Data Science ?

A great opportunity to

- **Sharpen your programming & analtyical skills**
- **Enhance domain knowledge**
- **Learn more about practical applications of data science & machine learning algorithms**

by participating in some real world Data Science Competitions hosted by organizations on various platforms.

But why

Competitive Data Science

is gaining traction in 2022?

It’s possibly due to the

Organizations are having hard time to solve so many data science problems while their data science team is busy with other projects. So hosting a data science competition on certain platform can help & is helping.

Data science competitions help organizations solve complex business problems while enabling data scientists to learn from the experience and win awards.

Organizations need to define the problem, provide data and put a prize on the challenge. Competing data scientists build and present different algorithms to be the winner.

Why should you try Competitive Data Science at least once?

To avoid situations like

And to

- **Understand how to solve predictive modeling competitions efficiently** and learn which of the skills obtained can be applicable to real-world tasks.
- **Learn how to preprocess the data and generate new features** from various sources such as text and images.
- **Be taught advanced feature engineering techniques** like generating mean-encodings, using aggregated statistical measures, or finding nearest neighbors as a means to improve your predictions.
- **Be able to form reliable cross validation methodologies that help you benchmark your solutions** and avoid overfitting or underfitting when tested with unobserved (test) data.

- **Gain experience in analyzing and interpreting the data**. You will become aware of inconsistencies, high noise levels, errors, and other data-related issues such as leakages and you will learn how to overcome them.
- **Acquire knowledge of different algorithms** and learn how to efficiently tune their hyperparameters and achieve top performance.
- **Master the art of combining different machine learning models** and learn how to ensemble.
- **Get exposed to past (winning) solutions** and codes and learn how to read them.

How should you start your Competitive Data Science journey?

The only thing you need to know Before Starting your CDS journey

"For participating in data science competitions, you only need an urge to constantly learn and improve. Getting a good ranking will follow."

Initial steps to start your CDS Journey

- Make sure your basics about Python & Mathematical concepts are clear enough.
- Focus on understanding core Data Science & Machine Learning algorithms
- Try to make self projects with the concepts you learned

Get the entire guide to start ML, cracking internships, etc at https://ayonroy.ml/help

Visit - AYONROY.ML

The next steps

- Try participating in Kudos/Knowledge Competitions (Like Titanic etc.)
- Then try to learn about the approaches from other's notebooks
- Try to apply your learnings from those approaches in Featured/Prized Competitions
- Try exploring variety of techniques you can use to get better results

How to approach a Competitive Data Science Problem?

1. **Start with a very simple baseline model**. Just have a look at the data, then create a model without any data cleaning or feature engineering.
2. **Understand the problem and data to create a good validation set**. A good validation set is a must. Only then can you can trust your local results. Otherwise, be prepared for a private leaderboard shakeup.
3. **Try Feature engineering**. Good features always differentiate between a winner and a top 100 finish.
4. **Try building a variety of models** like Gradient Boosting Models, Neural Nets, etc.
5. **Try stacking or blending of these results using Ensembling.** It gives you the edge to win a competition. Therefore, it's a tool you will always want to keep handy.

Time is a very crucial factor in any data science competition.

You should not waste your time writing the same snippets from scratch again and again in multiple competitions. Instead, focus your valuable time on doing something new and better

Where to get involved with Competitive Data Science ?

My personal suggestions

- https://www.kaggle.com/
- https://www.crowdanalytix.com/community
- https://zindi.africa/about
- https://datahack.analyticsvidhya.com/
- https://www.crowdai.org/challenges
- https://tianchi.aliyun.com/competition/gameList/activeList
- https://www.datasciencechallenge.org/
- https://www.drivendata.org/

Know a few more platforms to kick start your CDS journey here

Be a part of Communities like

1. Kaggle Days
2. Women in Machine Learning & Data Science
3. ODSC
4. Data Freak Community & a lot more....

Be a part of as many hackathons as you can

who wants to miss networking, free food & swags alongside unlimited learning

How to get involved more with Competitive Data Science ?

1. Do such courses where the skills learnt in them can be used in Competitions.
2. Publish your competition research, approaches on the forum & do write about the things that you want to share with others via blog etc.
3. Participate in Discussion forums, share your knowledge through answering questions & asking genuine questions.
4. Make notebooks & share them along with great EDA, feature engineering etc so that others can learn from it.
5. Try to reproduce interesting kernels.
6. Be consistent in whatever you are trying to share with the CDS community.

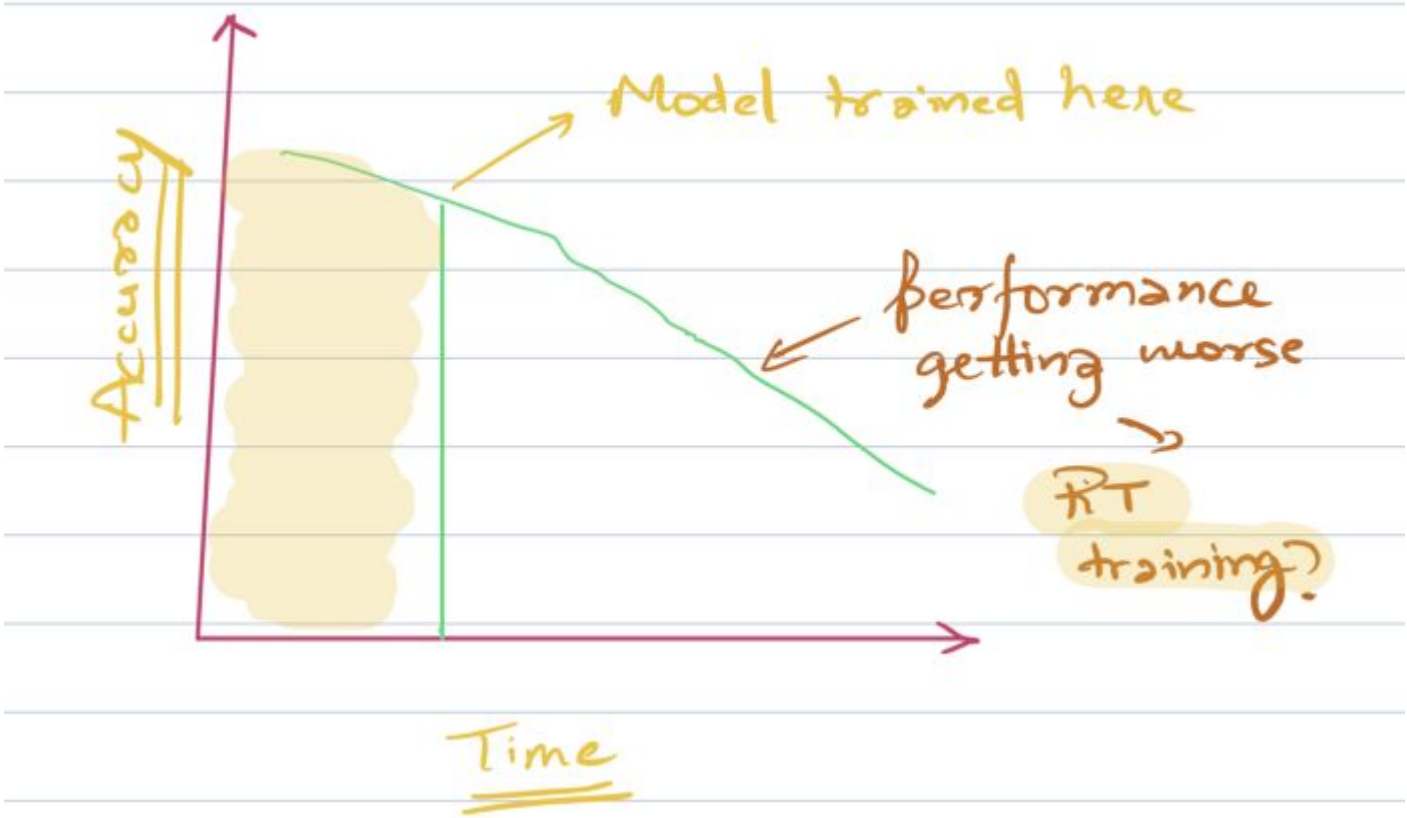
Visit - AYONROY.ML

Is

Competitive Data Science

everything what the industry requires?

What you will do in such a scenario ?



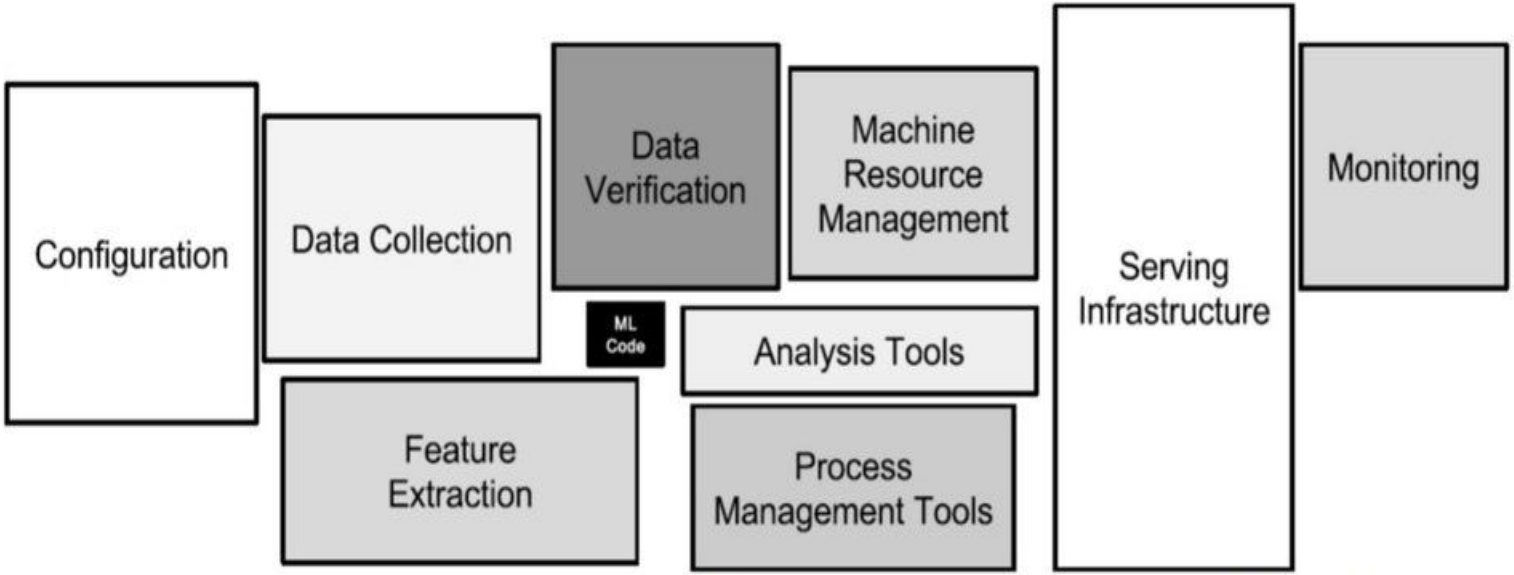


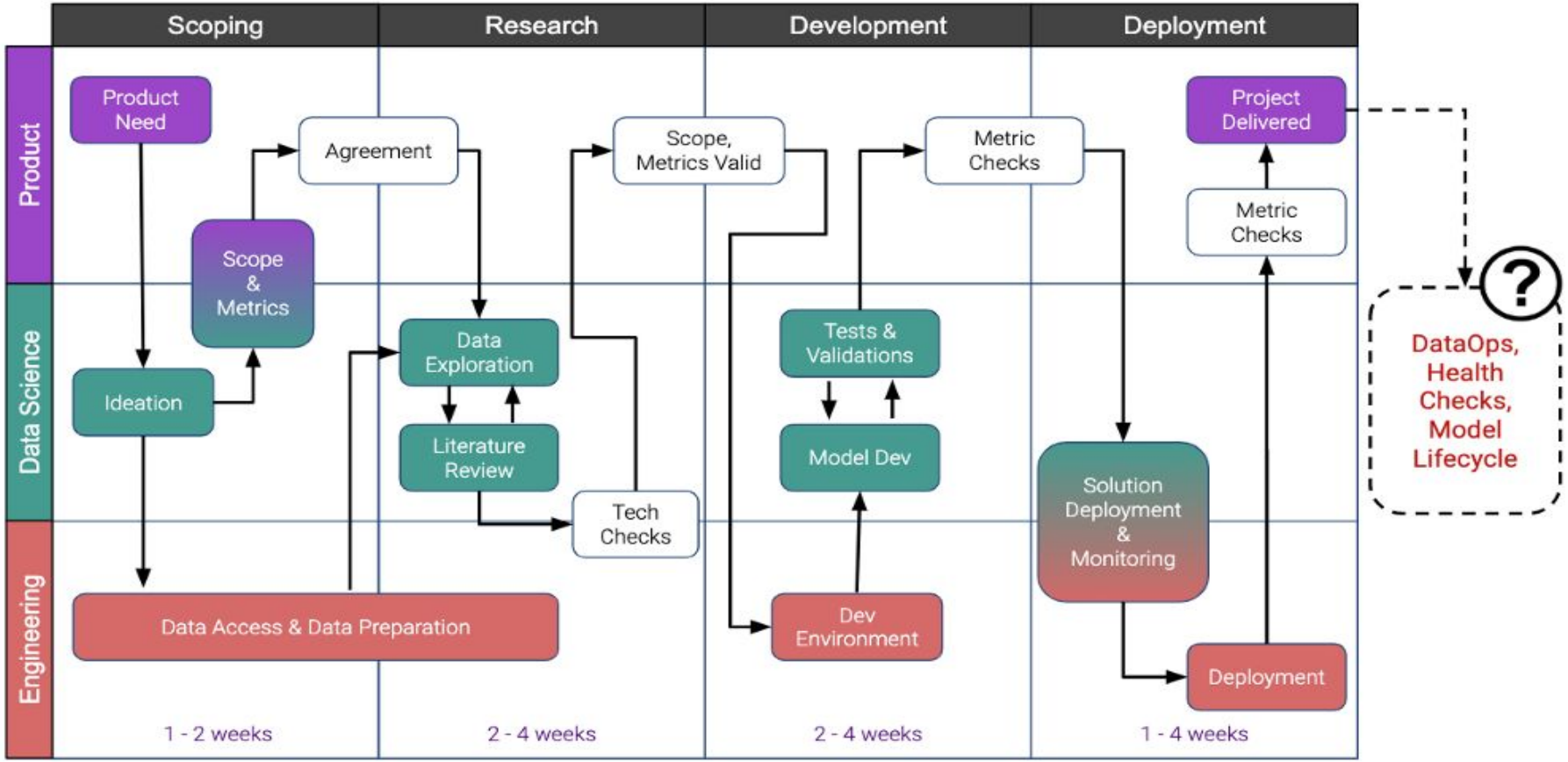
Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

View the Google's Research Paper here

What other than

Competitive Data Science

you should focus on, to become hirable?



Things to focus on while making a Data Science Project

Analytics Project Life Cycle

The 5 Phases

UNDERSTAND & PLAN

Translate the business question to analytics question

DATA

Work cross functionally to gather and process data

ANALYSIS

Solve problems

PRESENTATION

Present results and tell a story

DOCUMENTATION & REFLECTION

Make your project a happy ending

How to organize your 1st Data Science Project?

Local Project Directory	Github Repository

▪ Project plans/objectives
▪ Project datasets
▪ Project codes
  ○ Jupyter notebook
  ○ R scripts
  ○ Python scripts
▪ Output files
  ○ Visualizations
  ○ Tables
  ○ Other useful outputs
▪ Project report
	▪ README file
	▪ Project datasets
	▪ Project codes
	  ○ Jupyter notebook
	  ○ R scripts
	  ○ Python scripts
	▪ Output files
	  ○ Visualizations
	  ○ Tables
	  ○ Other useful outputs
	▪ Project report

https://gist.github.com/ericmjl/27e50331f24db3e8f957d1fe7bbbe510

But why organize your 1st Data Science Project?

- Organization increases productivity as avoid wasting time searching for project files such as datasets, codes, output files, and so on.
- A well-organized project helps you to keep and maintain a record of your ongoing and completed data science projects.
- Completed data science projects could be used for building future models.
- A well-organized project can easily be understood by other data science professionals when shared on platforms such as Github.

A few important pointers to keep in mind

1. Focus on understanding what business use case you are trying to solve before applying Data Science, Machine Learning.
2. Focus on Communication Skills to convey the result of your Data Science concepts to the business stakeholders.
3. Focus on DevOps to make your models production ready.
4. Focus on networking & showcasing your work to the community.

A few useful resources

1. https://towardsdatascience.com/use-kaggle-to-start-and-guide-your-ml-data-science-journey-f09154baba35
2. https://www.coursera.org/learn/competitive-data-science#syllabus
3. https://towardsdatascience.com/how-to-successfully-manage-a-data-science-delivery-pipeline-33bdec1a9a27
4. http://kaggle.com/learn

Let me answer your Questions now.

Finally, it’s your time to speak.

Visit - AYONROY.ML

Danke Schoen

Questions ? Any Feedbacks ? Did you like the talk?
Tell me about it.

If you think I can help you,
connect with me via

Email : ayon-roy@outlook.com

LinkedIn : https://www.linkedin.com/in/ayon-roy

Website : https://AYONROY.ML/