Top 5 Open-Source Projects where Data Scientist can Contribute

One of the most crucial aspects of landing your desired role in data science is building a strong, potent, eye-catching portfolio that proves your skills and shows that you can handle large-scale projects and play nicely in a team. Your portfolio needs to prove that you spent the time, effort, and resources to hone your skills as a data scientist.

Proving your skills to someone who doesn't know you, especially in a short time frame — the average time a recruiter spends on a resume or a portfolio is 7~10 seconds — is not easy. However, it's not impossible either.

A good portfolio should include various types of projects, projects about data collecting, analytics, and visualization. It should also contain projects of different sizes. Dealing with small projects is very different than dealing with large-scale ones. If your portfolio has both sizes, then it means you can read, handle and debug all size software, which is a skill required for any data scientist.

That may lead you to wonder how you would find good open-source data science projects that are easy to get into and look great on your portfolio. And that's a great question, but with the exploding number of data science projects out there, finding good ones that could be the thing that lands you the job is not the easiest of tasks.

When you try looking up data science projects to contribute to, you will often come across the big ones, like Pandas, Numpy and Matplotlib. These giant projects are great, but there are less known ones that are still used by many data scientists and will look good on your resume.

№1: Google's Caliban for Machine Learning

Let's kick this list off with a project from the tech giant Google. Often when building and developing data science projects, you may find it difficult to build a test environment that will show you your project in a real-life situation. You can't predict all scenarios, and make sure to cover all edge cases.

Google offers Caliban as a potential solution for that problem. Caliban is a testing tool that tracks your environmental properties during execution and allows you to reproduce specific running environments. Researchers and data engineers developed this tool at Google that performs this task on a daily basis.

№2: PalmerPenguins

Next on our list is PalmerPenguins, a dataset that was only recently open-sourced. This dataset was built and developed to replace the very well-known and used Iris dataset. The reason behind Iris's fame is its simplicity of use for beginners and also the wide variety of its possible applications.

PalmerPenguins offers an amazing dataset that you can use for data visualization and classification applications with the same ease as you would use Iris, but with much more options. One more great aspect of this dataset is that it offers art to teach data science concepts.

№3: Caffe

Next up, we have one of the promising frameworks for deep learning out there, Caffe. Caffe is a deep learning framework that was designed and built with speed, modularity, and expression as priorities. Caffe was originally developed by a team of researchers from the UC Berkeley AI lab and the vision and learning community.

After only one year of releasing Caffe as an open-source project, it was forked by more than 1000 researchers and developers around the world. It helped transform research topics and build new startups and industrial forces. The Caffe community is one of the welcoming, supportive open-source communities to join.

№4: NeoML

Machine learning is probably the heart of data science applications, so I had to have at least one open-source project solely for machine learning. NeoML is a machine learning framework that allows the user to design, built, test, and deploy machine learning models hassle-free with a collection of more than 20 traditional machine learning algorithms.

It includes materials that support natural language processing, computer vision, neural networks, and image classification and processing. This framework is written in C++, Java, and Objective-C and can run on any platform from Unix-based ones, macOS, and Windows.

№5: Kornia

We'll conclude our list with Kornia. Kornia is a supporting computer vision library for PyTorch. It includes various routines and differentiable that can be used to solve some generic computer vision problems. Kornia is built upon PyTorch and heavily depends on its efficiency and CPU power to compute complex functions.
Korina is more than just a package; it is a set of libraries that can be used together to train models and neural networks and perform image transformation, image filtering, and edge detection.

Final Thoughts

So you made it through the maze that is data science job hunting, you managed to decipher the job role's names and figure out which role fits your skills better and what you would like to do, it's time to think of how to make your portfolio land you that job with no delay.

You have probably gone through many projects during your data science learning journey, from smaller ones with a few lines of code to relatively large ones with hundreds of lines. But, to really prove your skills and knowledge level, you need to have some contributions that will make you stand out in the applicants' pool.

One way you can catch recruiters' eyes is by contributing to large-scale projects used by many data scientists all over the world.

Original. Reposted with permission.

Related:

- Getting Started in AI Research
- Facebook Open Sources Blender, the Largest-Ever Open Domain Chatbot
- Uber's Ludwig is an Open Source Framework for Low-Code Machine Learning

---

<= **Previous post**

Top Stories Past 30 Days

Most Popular	Most Shared
1. **A Guide On How To Become A Data Scientist (Step By Step Approach)**	1. **A Guide On How To Become A Data Scientist (Step By Step Approach)**
2. **Data Scientist, Data Engineer & Other Data Careers, Explained**	2. **Data Scientist, Data Engineer & Other Data Careers, Explained**

3. **Vaex: Pandas but 1000x faster**
4. **Data Preparation in SQL, with Cheat Sheet!**
5. **Top Programming Languages and Their Uses**

3. **How to Determine if Your Machine Learning Model is Overtrained**
4. **DeepMind Wants to Reimagine One of the Most Important Algorithms in Machine Learning**
5. **Essential Linear Algebra for Data Science and Machine Learning**

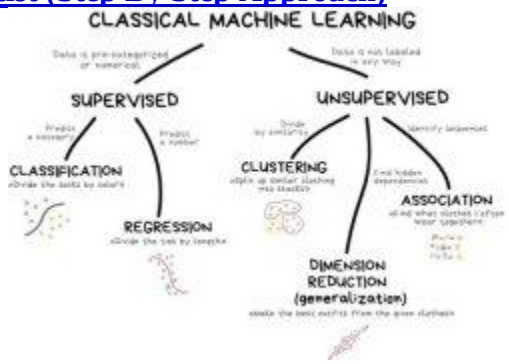
Latest News

- 5 Data Science Open-source Projects You Should Consider...
- How to Fine-Tune BERT Transformer with spaCy 3
- Top Stories, May 31 – Jun 6: A Guide On How To Be...
- PyCaret 101: An introduction for beginners
- 5 Tasks To Automate With Python
- Beyond Brainless AI with a Feature Store

Top Stories Last Week

Most Popular

1. **A Guide On How To Become A Data Scientist (Step By Step Approach)**
2. **5 Tasks To Automate With Python**
3. **How I Doubled My Income with Data Science and Machine Learning**
4. **Will There Be a Shortage of Data Science Jobs in the Next 5 Years?**
5. **How to Make Python Code Run Incredibly Fast**



Most Shared

1. **How I Doubled My Income with Data Science and Machine Learning**
2. **5 Tasks To Automate With Python**
3. **How to Create and Deploy a Simple Sentiment Analysis App via API**
4. **Will There Be a Shortage of Data Science Jobs in the Next 5 Years?**
5. **How to Make Python Code Run Incredibly Fast**

More Recent Stories

- Beyond Brainless AI with a Feature Store
- 10 Deadly Sins of Machine Learning Model Training
- BigQuery vs Snowflake: A Comparison of Data Warehouse Giants
- How a Data Scientist Should Communicate with Stakeholders
- Will There Be a Shortage of Data Science Jobs in the Next 5 Ye...
- Similarity Search: Euclid of Alexandria goes shoe shopping
- Machine Learning Model Interpretation
- Stop (and Start) Hiring Data Scientists
- How to Make Python Code Run Incredibly Fast
- How to Create and Deploy a Simple Sentiment Analysis App via API
- How I Doubled My Income with Data Science and Machine Learning **[Gold Blog]**
- Overcoming the Simplicity Illusion with Data Migration
- Make Pandas 3 Times Faster with PyPolars
- Top 4 Data Extraction Tools
- Top Stories, May 24-30: A Guide On How To Become A Data Scient...
- Supercharge Your Machine Learning Experiments with PyCaret and...
- State of Mathematical Optimization Report, 2021
- Essential Math for Data Science: Basis and Change of Basis
- 4 Tips for Dataset Curation for NLP Projects
- Choosing the Right BI Tool for Your Business