WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Python: Benefits of Using Advance Python libraries for Data Science

T. Rajasekhar, N. Bhavana

Master of computer Applications, Annamacharya Institute of Technology and science, Tirupati, Andhra Pradesh

ABSTRACT:

In last few years, there has been advancement in programming languages due to different libraries that are introduced. All the developers in this modern era prefer programming language that provides a built-in module/library which can make their work easy. This paper describes the advancement of one such language "Python" and it's increasing popularity through different statistical data and graphs. In this paper, we explore all the built-in libraries for all different computer science domains such as Data Science.

Keywords: Advanced python libraries for data science development.

Introduction:

Python has become a popular programming language for data science, and for good reason. The benefits of using Python for data science are manifold. Firstly, Python provides a wide range of powerful libraries and frameworks, such as NumPy, Pandas, and SciPy, which offer extensive functionality for data manipulation, analysis, and modeling. Its simplicity and readability make it an accessible language for beginners, while its versatility allows experienced data scientists to build complex algorithms and workflows.

Additionally, Python has a vast and active community that contributes to a rich ecosystem of resources, tutorials, and support. Its integration capabilities with other languages and tools, along with its scalability and compatibility with various platforms, make Python a flexible choice for data science projects. Overall, Python empowers data scientists with the tools and resources they need to efficiently explore, analyze, and derive insights from large and diverse datasets. Now that we know the benefits, let us look at the top 20 Python libraries for data science.

Top 20 Python Libraries for Data Science

- TensorFlow
- NumPy
- SciPy
- Pandas
- Matplotlib
- Keras
- SciKit-Learn
- PyTorch
- Scrapy
- BeautifulSoup
- LightGBM
- ELI5
- Theano
- NuPIC
- Ramp

- Pipenv
- Bob
- PyBrain
- Caffe2
- Chainer

1. TensorFlow

The first in the list of python libraries for data science is TensorFlow. TensorFlow is a library for high-performance numerical computations with around 35,000 comments and a vibrant community of around 1,500 contributors. It's used across various scientific fields. TensorFlow is basically a framework for defining and running computations that involve tensors, which are partially defined computational objects that eventually produce a value.

Features:

- Better computational graph visualizations
- Reduces error by 50 to 60 percent in neural machine learning
- Parallel computing to execute complex models
- Seamless library management backed by Google
- Quicker updates and frequent new releases to provide you with the latest features

Tensor Flow is particularly useful for the following applications:

- Speech and image recognition
- Text-based applications
- Time-series analysis
- Video detection

2. SciPy

SciPy (Scientific Python) is another free and open-source Python library for data science that is extensively used for high-level computations. SciPy has around 19,000 comments on GitHub and an active community of about 600 contributors. It's extensively used for scientific and technical computations, because it extends NumPy and provides many user-friendly and efficient routines for scientific calculations.

Features:

- Collection of algorithms and functions built on the NumPy extension of Python
- High-level commands for data manipulation and visualization
- Multidimensional image processing with the SciPy ndimage submodule
- Includes built-in functions for solving differential equations

Applications:

- Multidimensional image operations
- Solving differential equations and the Fourier transform
- Optimization algorithms
- Linear algebra

3. NumPy

NumPy (Numerical Python) is the fundamental package for numerical computation in Python; it contains a powerful N-dimensional array object. It has around 18,000 comments on GitHub and an active community of 700 contributors. It's a general-purpose array-processing package that provides high-performance multidimensional objects called arrays and tools for working with them. NumPy also addresses the slowness problem partly by providing these multidimensional arrays as well as providing functions and operators that operate efficiently on these arrays.

Features:

- Provides fast, precompiled functions for numerical routines
- Array-oriented computing for better efficiency
- Supports an object-oriented approach
- Compact and faster computations with vectorization

Applications:

- Extensively used in data analysis
- Creates powerful N-dimensional array
- Forms the base of other libraries, such as SciPy and scikit-learn
- Replacement of MATLAB when used with SciPy and matplotlib

4. Pandas

Next in the list of python librabries is Pandads. Pandas (Python data analysis) is a must in the data science life cycle. It is the most popular and widely used Python library for data science, along with NumPy in matplotlib. With around 17,00 comments on GitHub and an active community of 1,200 contributors, it is heavily used for data analysis and cleaning. Pandas provides fast, flexible data structures, such as data frame CDs, which are designed to work with structured data very easily and intuitively.

Also Read: What is Data Analysis: Methods, Process and Types Explained

Features:

- Eloquent syntax and rich functionalities that gives you the freedom to deal with missing data
- Enables you to create your own function and run it across a series of data
- High-level abstraction
- Contains high-level data structures and manipulation tools

Applications:

- General data wrangling and data cleaning
- ETL (extract, transform, load) jobs for data transformation and data storage, as it has excellent support for loading CSV files into its data frame format
- Used in a variety of academic and commercial areas, including statistics, finance and neuroscience
- Time-series-specific functionality, such as date range generation, moving window, linear regression and date shifting

5. Matplotlib

Matplotlib has powerful yet beautiful visualizations. It's a plotting library for Python with around 26,000 comments on GitHub and a very vibrant community of about 700 contributors. Because of the graphs and plots that it produces, it's extensively used for data visualization. It also provides an object-oriented API, which can be used to embed those plots into applications.

Features:

- Usable as a MATLAB replacement, with the advantage of being free and open source
- Supports dozens of backends and output types, which means you can use it regardless of which operating system you're using or which output format you wish to use
- Pandas itself can be used as wrappers around MATLAB API to drive MATLAB like a cleaner
- Low memory consumption and better runtime behavior

Applications:

- Correlation analysis of variables
- Visualize 95 percent confidence intervals of the models

- Outlier detection using a scatter plot etc.
- Visualize the distribution of data to gain instant insights

Also Read: Exploring The Data Science Learning Path

6. Keras

Similar to TensorFlow, Keras is another popular library that is used extensively for deep learning and neural network modules. Keras supports both the TensorFlow and Theano backends, so it is a good option if you don't want to dive into the details of TensorFlow.

Also Read: Keras vs Tensorflow vs Pytorch

Features:

- Keras provides a vast prelabeled datasets which can be used to directly import and load.
- It contains various implemented layers and parameters that can be used for construction, configuration, training, and evaluation of neural networks

Applications:

- One of the most significant applications of Keras are the deep learning models that are available with their pretrained weights. You can use these models directly to make predictions or extract its features without creating or training your own new model.

7. Scikit-learn

Next in the list of the top python libraries for data science comes Scikit-learn, a machine learning library that provides almost all the machine learning algorithms you might need. Scikit-learn is designed to be interpolated into NumPy and SciPy.

Applications:

- clustering
- classification
- regression
- model selection
- dimensionality reduction

8. PyTorch

Next in the list of top python libraries for data science is PyTorch, which is a Python-based scientific computing package that uses the power of graphics processing units. PyTorch is one of the most commonly preferred deep learning research platforms built to provide maximum flexibility and speed.

Applications:

- PyTorch is famous for providing two of the most high-level features
- tensor computations with strong GPU acceleration support
- building deep neural networks on a tape-based autograd system

9. Scrapy

The next known python libraries for data science is Scrapy. Scrapy isone of the most popular, fast, open-source web crawling frameworks written in Python. It is commonly used to extract the data from the web page with the help of selectors based on XPath.

Applications:

- Scrapy helps in building crawling programs (spider bots) that can retrieve structured data from the web
- Scrappy is also used to gather data from APIs and follows a 'Don't Repeat Yourself' principle in the design of its interface, influencing users to write universal codes that can be reused for building and scaling large crawlers.

10. BeautifulSoup

BeautifulSoup - the next python library for data science. This is another popular python library most commonly known for web crawling and data scraping. Users can collect data that's available on some website without a proper CSV or API, and BeautifulSoup can help them scrape it and arrange it into the required format.

If you wish to learn all about python libraries, python and other programming languages, and get a hang of the data science field, explore our exclusive data science career resource page today!

11. LightGBM

The LightGBM Python library is a popular tool for implementing gradient-boosting algorithms in data science projects. It provides a high-performance implementation of gradient boosting that can handle large datasets and high-dimensional feature spaces.

Features:

- The LightGBM Python library is easy to integrate with other Python libraries, such as Pandas, Scikit-Learn, and XGBoost.
- LightGBM is designed to be fast and memory-efficient, making it suitable for large-scale datasets and high-dimensional feature spaces.
- The LightGBM Python library provides a wide range of hyperparameters that can be customised to optimise model performance for specific datasets and use cases.

Applications:

- Anomaly detection
- Time series analysis
- Natural language processing
- Classification

12. ELI5

ELI5 is a Python library for debugging and visualising machine learning models. It provides tools to help data scientists and machine learning practitioners understand how their models work and diagnose potential problems.

Features:

- ELI5 provides a range of techniques for interpreting machine learning models, such as feature importance, permutation importance, and SHAP values.
- ELI5 provides tools for debugging machine learning models, such as visualising misclassified examples and inspecting model weights and biases.
- ELI5 can generate human-readable explanations for how a model makes predictions, which can help communicate with non-technical stakeholders.

Applications:

- Model interpretation
- Model debugging
- Model comparison
- Feature engineering

13. Theano

Next in the list of python librabries is Theano. Theano is a Python library for numerical computation designed for deep learning and machine learning applications. It allows users to define, optimise, and gauge mathematical expressions, which includes multi-dimensional arrays - the fundamental building blocks of many machine learning algorithms.

Features:

- Theano is designed to efficiently perform numerical computations on both CPUs and GPUs, which can significantly speed up the training and testing of machine learning models.

- Theano provides automatic differentiation functionality, making it easy to compute gradients and optimise parameters while training machine learning models.
- Theano allows users to optimise expressions for speed, memory usage, or numerical stability, depending on the requirements of their machine learning task.

Applications:

- Scientific computing
- Simulation
- Optimisation
- Deep learning

14. NuPIC

NuPIC (Numenta Platform for Intelligent Computing) is an open-source Python library for building intelligent systems based on the principles of neocortical theory. It is designed to simulate the behaviour of the neocortex, the part of the brain responsible for sensory perception, spatial reasoning, and language.

Features:

- NuPIC implements a biologically inspired HTM algorithm to learn temporal patterns in data and make predictions based on those patterns.
- NuPIC is designed to process streaming data in real-time, making it well-suited for anomaly detection, prediction, and classification applications.
- NuPIC provides a flexible and extensible network API, which can be used to build custom HTM networks for specific applications.

Applications:

- Anomaly detection
- Prediction
- Dimensionality reduction
- Pattern recognition

15. Ramp

Ramp is an open-source Python library for building and evaluating predictive models. It provides a flexible and easy-to-use framework for data scientists and machine learning practitioners to train and test machine learning models and compare the performance of different models on various datasets and tasks.

Features:

- Ramp is designed to be modular and extensible, allowing users to build and test different predictive model components easily.
- Ramp supports multiple input formats for data, including CSV, Excel, and SQL databases, which makes it easy to work with different types of data.
- Ramp provides a collaborative environment for data scientists and machine learning practitioners to work together on building and evaluating predictive models.

Applications:

- Building predictive models
- Evaluating model performance
- Collaborating on machine learning projects
- Deploying model in diverse environments

16. Pipenv

Pipenv is a popular tool used for managing Python dependencies and virtual environments. It provides developers with a simple and efficient way to handle dependencies for their Python projects. It is especially useful for data science projects, often involving working with many different libraries.

Features:

- Pipenv manages dependencies for your Python projects, including packages from PyPI and those installed from other sources such as GitHub.
- Pipenv creates a virtual environment for your project and installs the necessary packages inside it. This ensures that your project's dependencies are isolated from other Python installations on your system.
- Pipenv generates a Pipfile.lock file that records the exact versions of each package installed in your project's virtual environment. This ensures that your project always uses the same dependencies, even if newer versions of those packages are released.

Applications:

- Managing dependencies
- Streamlining development
- Ensuring reproducible results
- Simplifying deployment

17. Bob

Next in the list of python librabries is Bob. Bob is a collection of python data science libraries that provide a range of tools and algorithms for machine learning, computer vision, and signal processing. Bob is designed to be a modular and extensible platform that allows researchers and developers to build and evaluate new algorithms for various tasks easily.

Features:

- Bob supports reading and writing data in various formats, including audio, image, and video.
- Bob includes pre-implemented facial recognition, speaker verification, and emotion recognition algorithms and models.
- Bob is designed to be modular and extensible, allowing developers to add new algorithms and models easily.

Applications:

- Face recognition
- Speaker verification
- Emotion recognition
- Biometric authentication

18. PyBrain

PyBrain is a python data science libraries for building and training neural networks. It provides a wide range of tools and algorithms for machine learning and artificial intelligence tasks, including supervised, unsupervised, reinforcement, and deep learning.

Features:

- PyBrain provides a flexible and extensible architecture allowing users to build and customise neural network models easily.
- PyBrain includes a wide range of algorithms for machine learning tasks, including feedforward neural networks, recurrent neural networks, support vector machines, and reinforcement learning.
- PyBrain includes tools for visualising the performance and structure of neural networks, making it easier to understand and debug your models.

Applications:

- Pattern recognition
- Time-series prediction
- Reinforcement learning
- Natural language processing

19. Caffe2

Caffe2 is a Python library for deep learning designed to be fast, scalable, and portable. It is developed by Facebook and used by many companies and research organisations for machine learning tasks.

Features:

- Caffe2 is designed to be fast and scalable, making it ideal for training large-scale deep neural networks.
- Caffe2 provides a flexible architecture allowing users to customise and extend deep neural networks easily.
- Affe2 supports multiple platforms, including CPU, GPU, and mobile devices, making it a versatile tool for machine learning tasks.

Applications:

- Object and image recognition
- Recommender systems
- Natural language processing
- Video analysis

20. Chainer

Chainer is a Python library for building and training deep neural networks. It was developed by the Japanese company Preferred Networks and is designed to be both powerful and flexible.

Features:

- Chainer uses a dynamic computation graph, which allows for more flexible and efficient training of deep neural networks.
- Chainer supports many neural network architectures, including feedforward, convolutional, and recurrent neural networks.
- Chainer includes built-in optimisation algorithms, such as stochastic gradient descent and Adam, which can be used to train neural networks.

Applications:

- Video analysis
- Robotics
- Research and development
- Natural language processing