

DataCraft Application Installation and Operation Guide

Jia Chen 201768810, Jiageng Zhu 201770214, Junqin Cheng 20178863,
Luyuan Huang 201769063, Xiankai Shi 201769457, Yifei Chen 201768854

Application Introduction

DataCraft is a comprehensive, interactive tool designed for professionals in machine learning and data modeling to streamline the feature engineering process. Data cleaning is just one of its many functionalities, with the tool also encompassing data exploration, feature selection, and model building.

Code Storage Location

The complete source code for DataCraft can be accessed via the following GitHub link: <https://github.com/JiagengZhu21/Team26>

Installation Steps

Download and Extract the Code

Visit the GitHub link, download the project file, and extract it to a specified directory.

Install Python and Dependencies

Before starting the project, it is advisable to create a virtual environment. This helps to isolate dependencies of different projects to prevent potential version conflicts and keeps the global environment clean. Within the virtual environment, use pip to install all necessary Python packages. Sometimes, certain packages may not support newer or older versions of Python.

Ensure that Python 3.6 or higher is installed and install all dependencies via the command line.

Startup Guide

Configure and Run DataCraft in PyCharm

Open PyCharm, import the project, and set up the Python interpreter. Start the application by right-clicking main.py and selecting Run 'main'.

Feature Usage Guide

Data Loading and Cleaning, Feature Engineering Operations, Model Building and Evaluation

Advanced Features and Customization

Customize Feature Transformation Rules

Allows users to define and save specific transformation rules for reuse.

Performance Optimization Options

Supports multithreading and distributed processing to optimize handling of large datasets.