Practical exercises of Pycaret, DagsHub, MLFlow and DVC

Model development

Context

Customer retention is one of the most important key indicators for companies with a subscription-based business model.

Exercise

You will need to train a **classification model** to predict the **churn** variable with **Pycaret**. You must apply the necessary preprocessing to the dataset and store the experiment in **MLFlow locally**. Dataset: https://raw.githubusercontent.com/srees1988/predict-churn-py/main/customer_churn_data.csv

Solution: *Churn_model_development.ipynb*

Repository with DagsHub

Context

You will need to create an online repository to be able to work with your data science team

Exercise

You will need to create an **online repository** for your model on **DagsHub**. To do this you will need to create an account on DagsHub, create a repository and configure access through **Git**. You will need to initialize Git, add the model and notebook files to your local repository and finally **upload it to DagsHub**.

Data Bootcamp
BEST DATA TRAINING

Versioning dataset with DVC

Context

For your experiments to be reproducible, you must version your dataset.

Exercise

You will need to develop **code** that will get the **raw data** and store it in a local csv as raw in the **data folder**. Then you will need to preprocess the data and store it as well.

You will then need to configure **Git** to ignore the data folder, and program **DVC** to version the data. Finally, you must upload the datasets to **DagsHub**.

Solution : *data_download.ipynb*

Data Bootcamp

BEST DATA TRAINING

Shared MLFlow

Context

In order for your team to collaborate, you must register the experiments in a shared MLFlow environment.

Exercise

You will need to schedule access to the **MLflow server** hosted **on DagsHub**. For this you will have to configure the access **credentials** to MLFlow and then develop several experiments with your churn model. Register the experiments in MLFlow and finally use MLFlow to select the best experiments and models.

Solution: *Churn_model_development.ipynb*

Data Bootcamp
BEST DATA TRAINING