The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

Advanced Data Science Capstone Project
by Marco Alberto Camarena Ospino

1 ARCHITECTURAL COMPONENTS OVERVIEW

1.1 DATA SOURCE

1.1.1 Technology Choice

External data source.
CSV (coma separated values format) 92.3 MB of data.

1.1.2 Justification

The CSV file provided is a common format for table data.

1.2 ENTERPRISE DATA

1.2.1 Technology Choice

Not applicable.

1.2.2 Justification

Not applicable.

1.3 STREAMING ANALYTICS

1.3.1 Technology Choice

Not applicable.

1.3.2 Justification

Not applicable.

1.4 DATA INTEGRATION

1.4.1 Technology Choice

Watsons Studio, IBM Cloud Object Storage.

1.4.2 Justification

IBM Cloud Object Storage provides a free plan and is easy to integrate into Watson Studio projects (as of Nov. 2021).

1.5 DATA REPOSITORY

1.5.1 Technology Choice

IBM Cloud Object Storage.

1.5.2 Justification

IBM Cloud Object Storage provides a free plan and is easy to integrate into Watson Studio projects. (as of Nov. 2021).

1.6 DISCOVERY AND EXPLORATION

1.6.1 Technology Choice

Watson Studio, Jupyter Notebooks.

The data quality is assessed with EDA performed with pandas, pandas-profiler and matplotlib libraries.

1.6.2 Justification

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

1.7 ACTIONABLE INSIGHTS

1.7.1 Technology Choice

Watson Studio, Jupyter Notebooks, Python (pandas, sklearn).

Feature Engineering is based on date-based features. The variables are pre-processed to work with linear models.

1.7.2 Justification

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

1.8 APPLICATIONS / DATA PRODUCTS

1.8.1 Technology Choice

The chosen model is Linear Regression with L2 regularization (Ridge).

The model is assessed with Root Mean Squared Error (RMSE) score.

Jupyter Notebook with full pipeline: takes in raw data in CSV and outputs the fitted pipeline with trained sklearn model saved to pickle format and ready for production.

Model deployed as service with Dash application running on Flask REST API. Access through web-interface.

1.8.2 Justification

The model is chosen to fit the Watson Studio Lite (free) plan limitations. The tree-based boosting models would perform better but are too expensive to train.

RMSE is the standard for regression tasks. It allows to evaluate the model performance and penalizes big errors more than Mean Absolute Error.

The model is easy to access through web-interface and the service can be easily adjusted for the batch predictions or web-application response.

1.9 SECURITY, INFORMATION GOVERNANCE AND SYSTEMS MANAGEMENT

1.9.1 Technology Choice

Not applicable.

1.9.2 Justification

Not applicable.