A picture containing text, clipart

Description automatically generated

**Student Project Proposal**

|  |  |
| --- | --- |
| Project Title | Predicting Turbo Fan Failure |
| Industry Sponsorship (if Any) | Booz-Allen-Hamilton |

**Project Description**

**Problem definition**

*[50-100 word description of the problem which you will solve]*

|  |
| --- |
| The U.S. DoD’s Joint Artificial Intelligence Center has designated Predictive Maintenance as one of its two founding National Mission Initiatives (NMIs). The challenge we are trying to solve is accurately predicting the remaining useful life (RUL) of turbofan engines measured in operations cycles. RUL is equivalent of number of flights remaining for the engine. |

**Key Research Questions/ Technological constraints that the Project will Answer**

|  |
| --- |
| The primary objective of the project will be to provide the DoD a model that can be used to reduce the likelihood of turbofan engine in their military fleet of flights. In order to do this we will build a machine learning model to predict the probability of turbo fan failure at any given point in time, as well as provide a model to then determine how many flights are left in the turbofan before failure (RUL). |

**Final deliverables at the end of the project**

*[Please list the desired technical deliverables from the project team in as much detail as possible]*

|  |
| --- |
| 1. A complete exploratory analysis using Jupiter notebook, including and executive summary, key findings, constraints, challenges, data discovery, feature engineering, feature selection, machine learning model training and testing, pipeline development (pre-processing, modelling, etc.), next steps. 2. A public Github repository containing the entire project workflow: Jupiter notebook for exploratory analysis, read me section for introducing the project workflow and results, raw data, training data, Airflow DAG schedulers. 3. A production ready machine learning model with in a cloud environment with a PowerBI front end for user interaction/decision making. |

**Key activities/ technologies the project team may be expected to undertake/ work with**

*[E.g. What kind of technology stack will you work with, the datasets you may need to work on, what kind of analysis you may be expected to undertake, etc.]*

|  |
| --- |
| 1. Create project folder scaffolding locally for the project. I will use the Python library *cookiecutter* to create or leverage a project template design. 2. Create public GitHub repository for the project 3. Create the ‘Read me’ file to introduce the project, data, modelling process, etc. 4. Extensive exploratory data analysis using Jupyter notebook which will contain a executive summary, key findings, constraints, challenges, data discovery, feature engineering, feature selection, machine learning model training and testing, pipeline development (pre-processing, modelling, etc.), next steps. 5. Create the Airflow operators to get data from NASA/Kaggle, write python module to handle data pre-processing jobs using Scikit-learn pipelines (from EDA step). This includes missing value handling, feature engineering, feature selection. Next I will create the ML model module (chosen from EDA step). Both of these steps will then have an Airflow task assigned. 6. Deploy model to cloud environment for model training 7. Create front end application in Power BI for users to query and generate predictions |
|  |

**Expected learning outcomes**

*[What do you expect to learn from the project? Please mention the technical skills you will imbibe over the project.]*

|  |
| --- |
| 1. Generally speaking, end-to-end machine learning working on more than just the modelling stage: data pipelines, orchestration design (tasks, dependencies, etc.), 2. Orchestration using Airflow 3. Deploying models in the cloud for scalable machine learning applications that are enterprise ready |

|  |  |
| --- | --- |
| Team Size: | 1 |
| Member names: | Dinush De Alwis |

**Tentative Time plan**

Submit a tentative time plan (table/chart or text) regarding breakdown of the work that will be conducted between in the second half of your cohort, from week 6 onward.

**Week 6:**

* Build project folder scaffolding
* Download dataset to raw data folder (many text files)
* Create python script to append text files
* Start EDA in Jupyter Notebook

**Week 7:**

* Write read me file introduction to the project
* Finish EDA
* Create Scikit learn pre-processing pipelines (NA handling, feature engineering, feature selection, training/test set creation)

**Week 8:**

* Conclude on which model to use
* Incorporate model into existing Scikit learn pipeline
* Conclude accuracy metrics to report

**Week 9:**

* Write Airflow operator to get training data files, append them for final training data set
* Write python modules for pre-processing and model pipeline
* Write Airflow operator for above step and set execution order

**Week 10 & 11:**

* Cloud deployment
* ML performance monitoring setup

**Week 12:**

* Build front-end application in PowerBI

**Week 13:**

* Build presentation deck

**System Design**

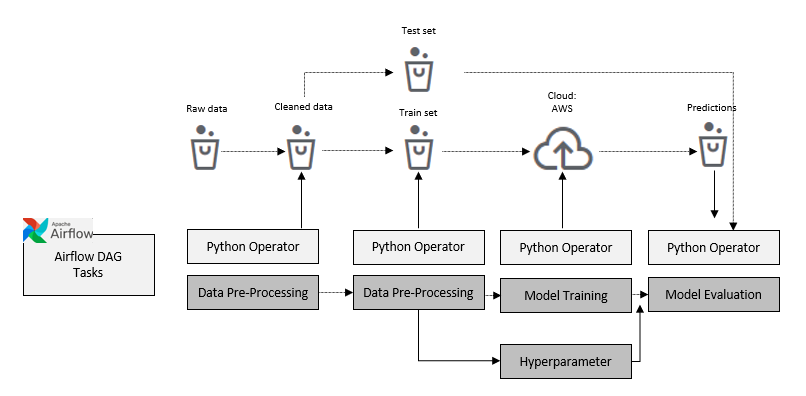
From the System design perspective, outline the following:

* Data
* Process (Models, iterations)
* Outcome (output and recommendations)

What are the system design considerations for your deployable ML model? Describe the iterations, delivery formats and limitations you may face and some solutions to overcome the limitations

* Should the model be deployed to run in batch, or to be hit from an api or some sort of streaming process as events are generated?
* What sort of infrastructure will be required for training? If it is a model that requires a lot of resources, where is the best place to train?

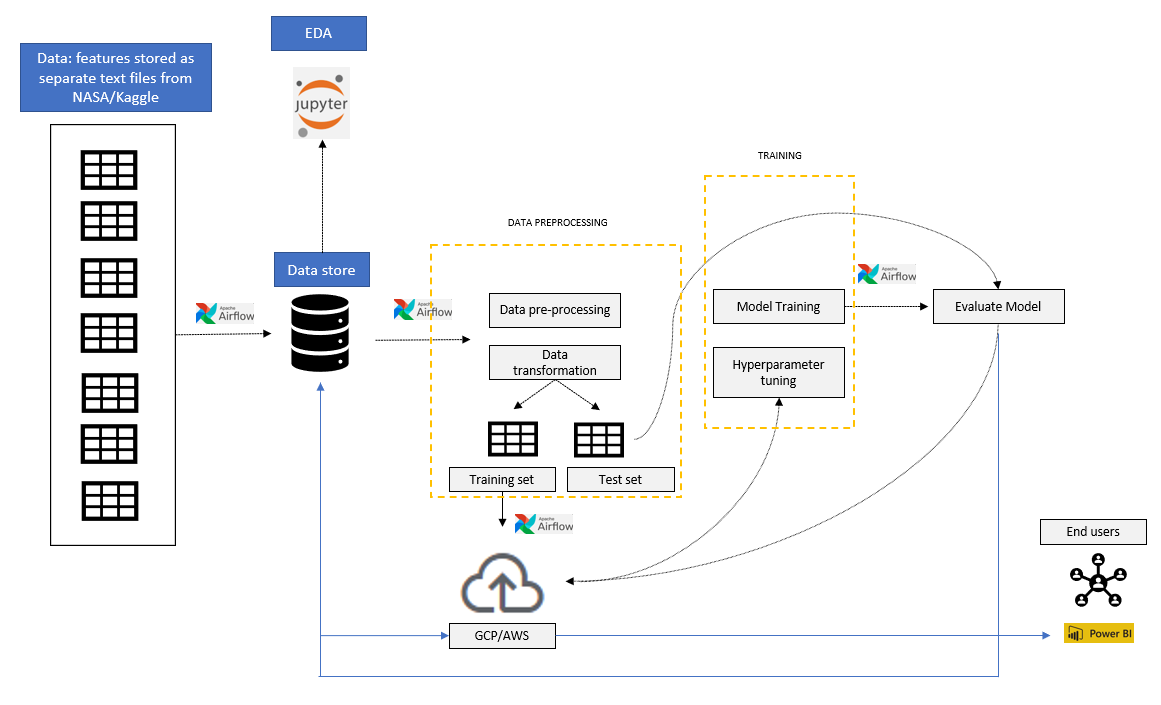
High level Airflow orchestration tasks:



High-level system design:

1. Aggregate features into a single file/SQL table. This will be driven by an Airflow operator.
2. Bring data into Jupyter notebook for exploratory data analysis (EDA) and work to drive creation of preprocessing modules (missing value handling, feature engineering and selection*; preprocessing.py*)
3. Send training data to cloud for model training, hyperparameter tuning
4. Evaluate model results
5. Repeat steps 3 and 4 until model results are acceptable
6. Send final model back to cloud or to SQL database to serve as back-end for application (PowerBI)

See system design diagram below.



**Ethical Considerations**

Are there any ethical considerations of your project? Consider the data source, the intended outcome, and/or the eventual use cases.

* Did you modify anything about your plan based on these considerations?
* Can you anticipate any issues that might arise during the process?

No meaningful ethical concerns in this project with respect to the data. There is no personal information and it has been provided by NASA on their own.