**Milestone Report – 4**

**Capstone Project: News Article Classification**

* **Dayakar & Krupasindhu (31st**
* **Oct 2021)**

1. **Project Architecture Design Completed** (Attached .PPT document for reference)
2. **ML Approach Documentation completed** (Attached .PPT document for reference)
3. **Developed Python script for Data preparation**
4. **Install python mongodb connectors for connecting with Pyspark**

**\* !pip3 install pymongo[srv]**

**\* !pip3 install Pyspark**

**\* !pip3 install mlflow**

**\* Load the data from mongodb**

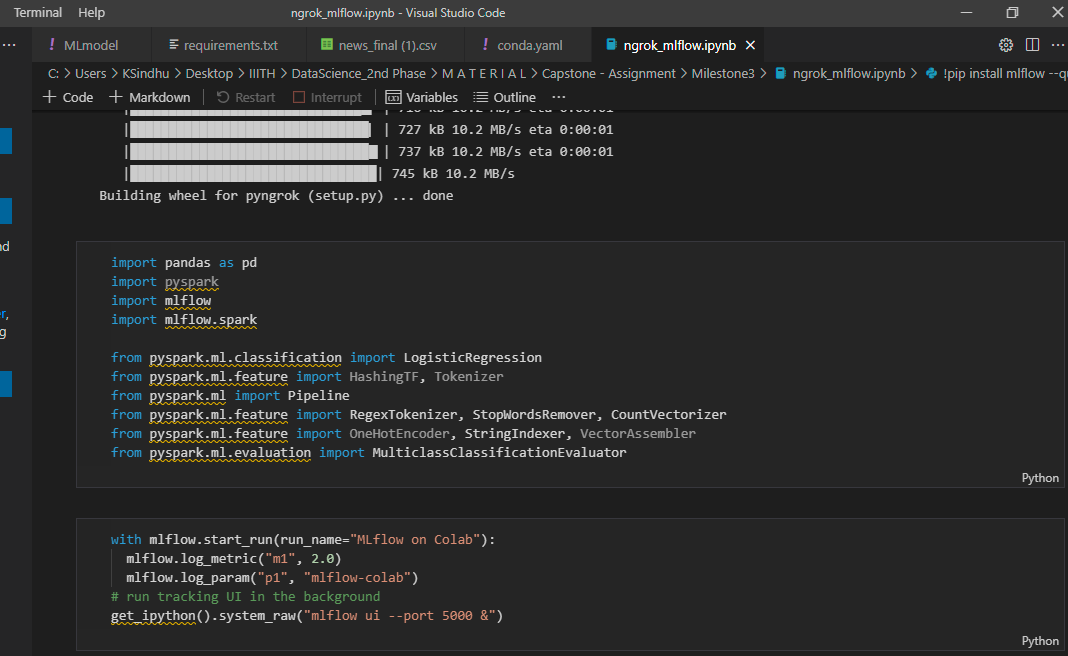
**\* Perform EDA to study the data**

**\* Select the relevant columns for modeling**

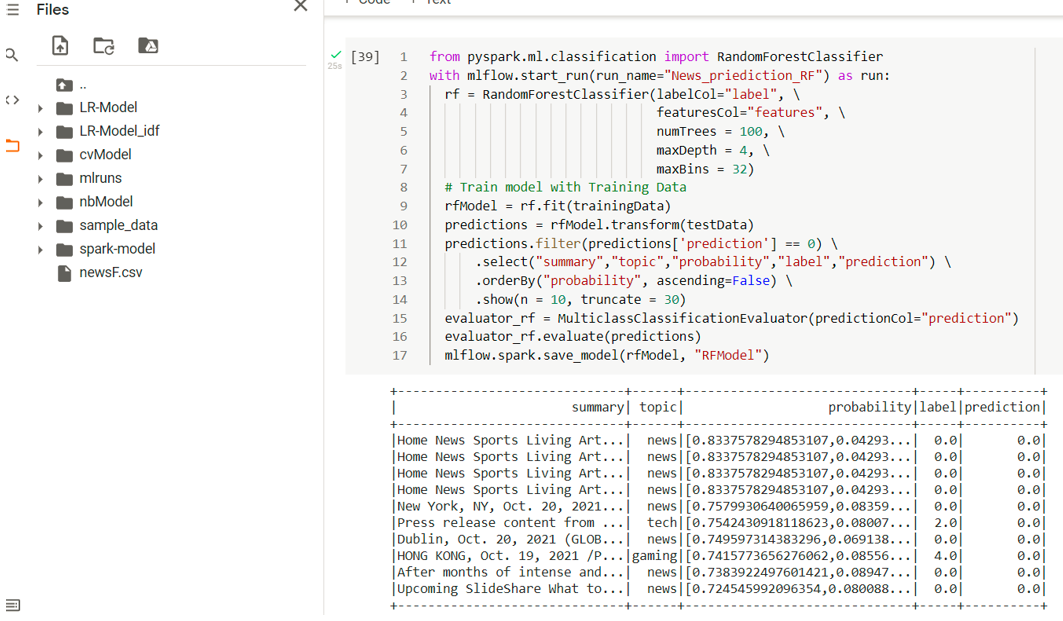
1. **Integrate Model with Model tracking APIs**

* **Split the data for train and test**
* **Model the data with different algorithms**
* **Evaluate the models and select the best performing and optimized model**
* **Track and log models with MLflow**
* **Register models with the Model Registry**
* **Describe models and make model version stage transitions**
* **Integrate registered models with production applications**
* **Search and discover models in the Model Registry**

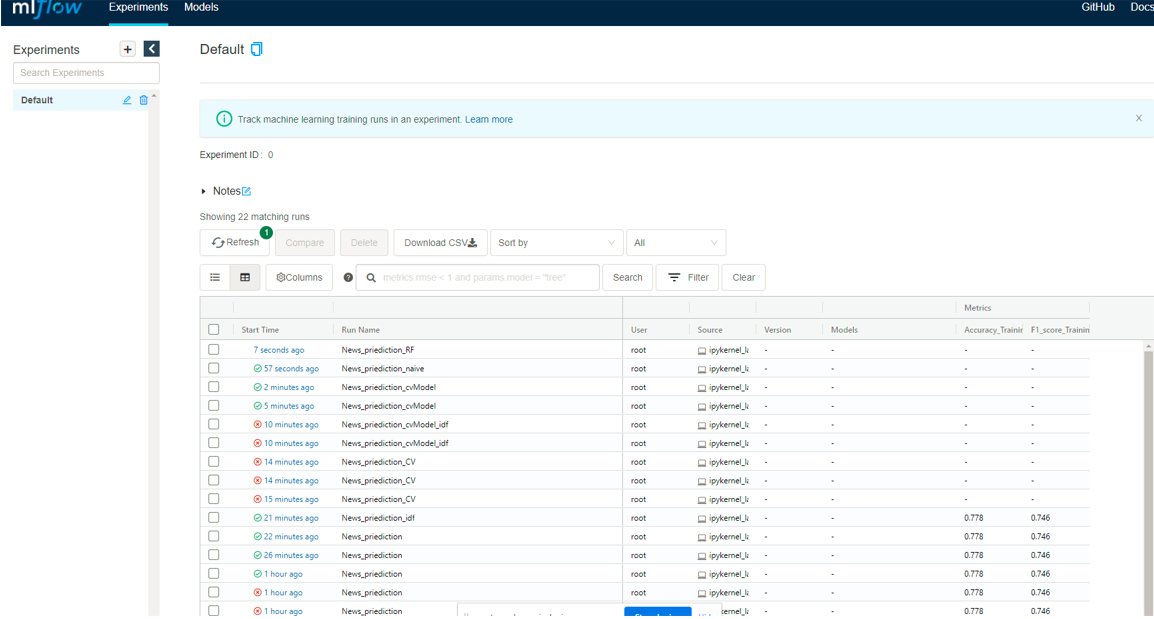
1. **MLFlow Libraries View:**



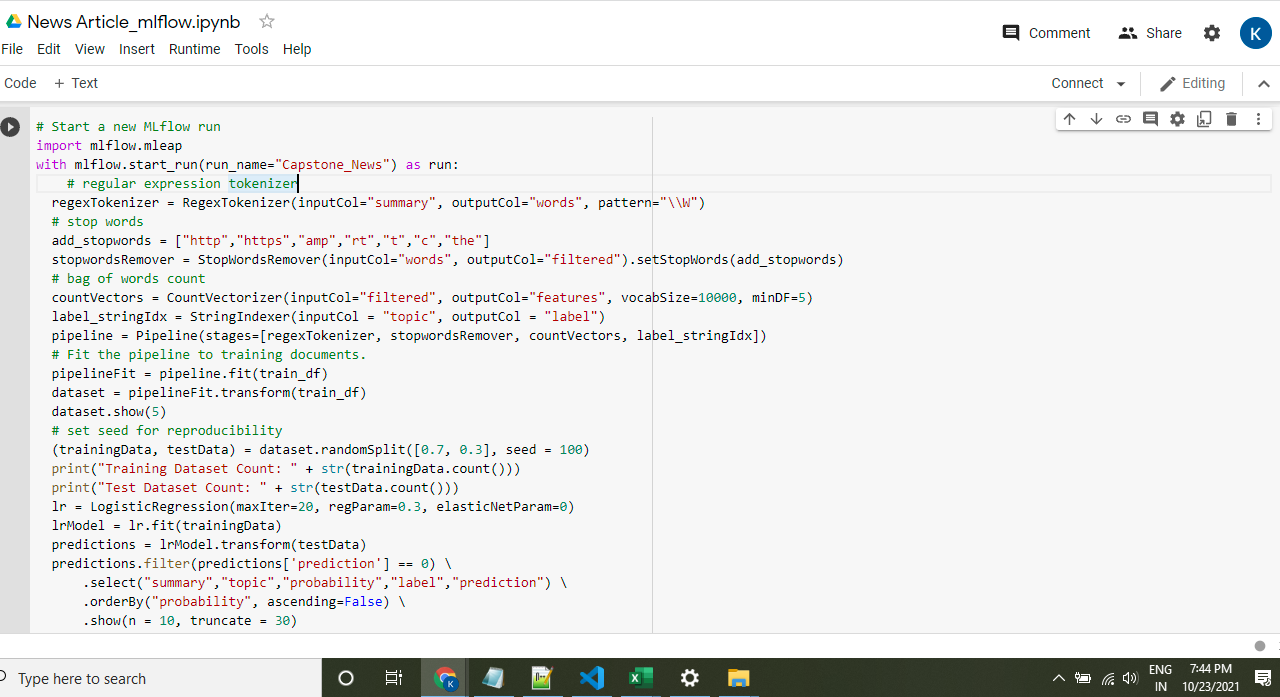
1. **Model View:**



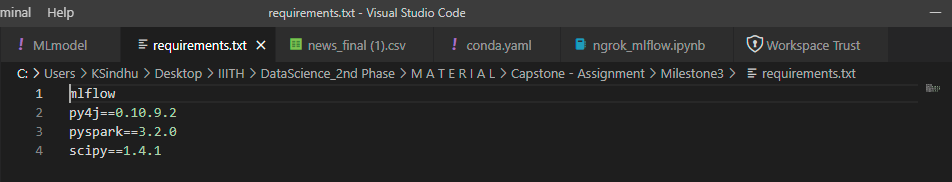
1. **Model Accuracy View:**



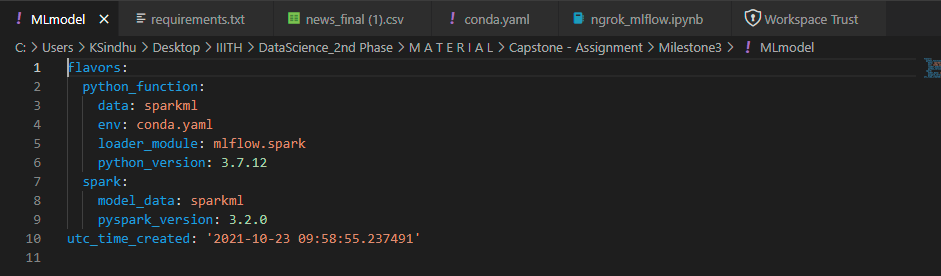
1. **ML Registry & Pipeline:**



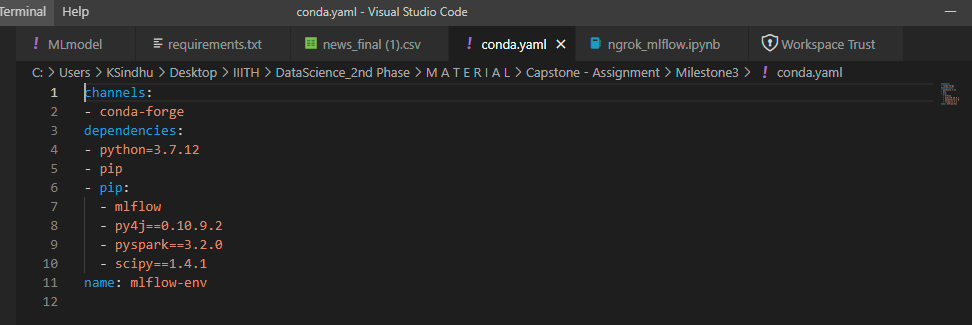
1. **Requirements.txt view:**



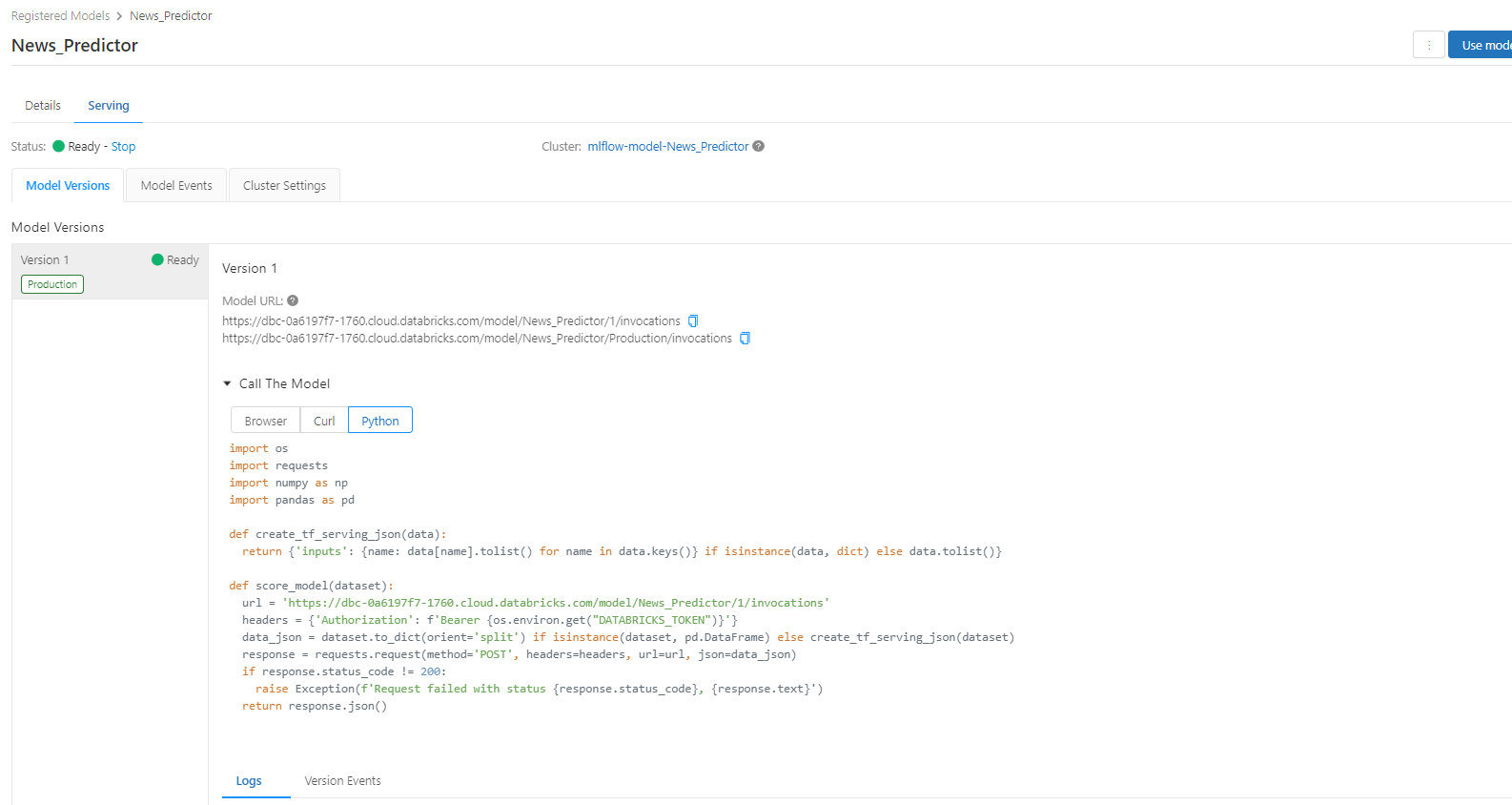
1. **ML Model view:**

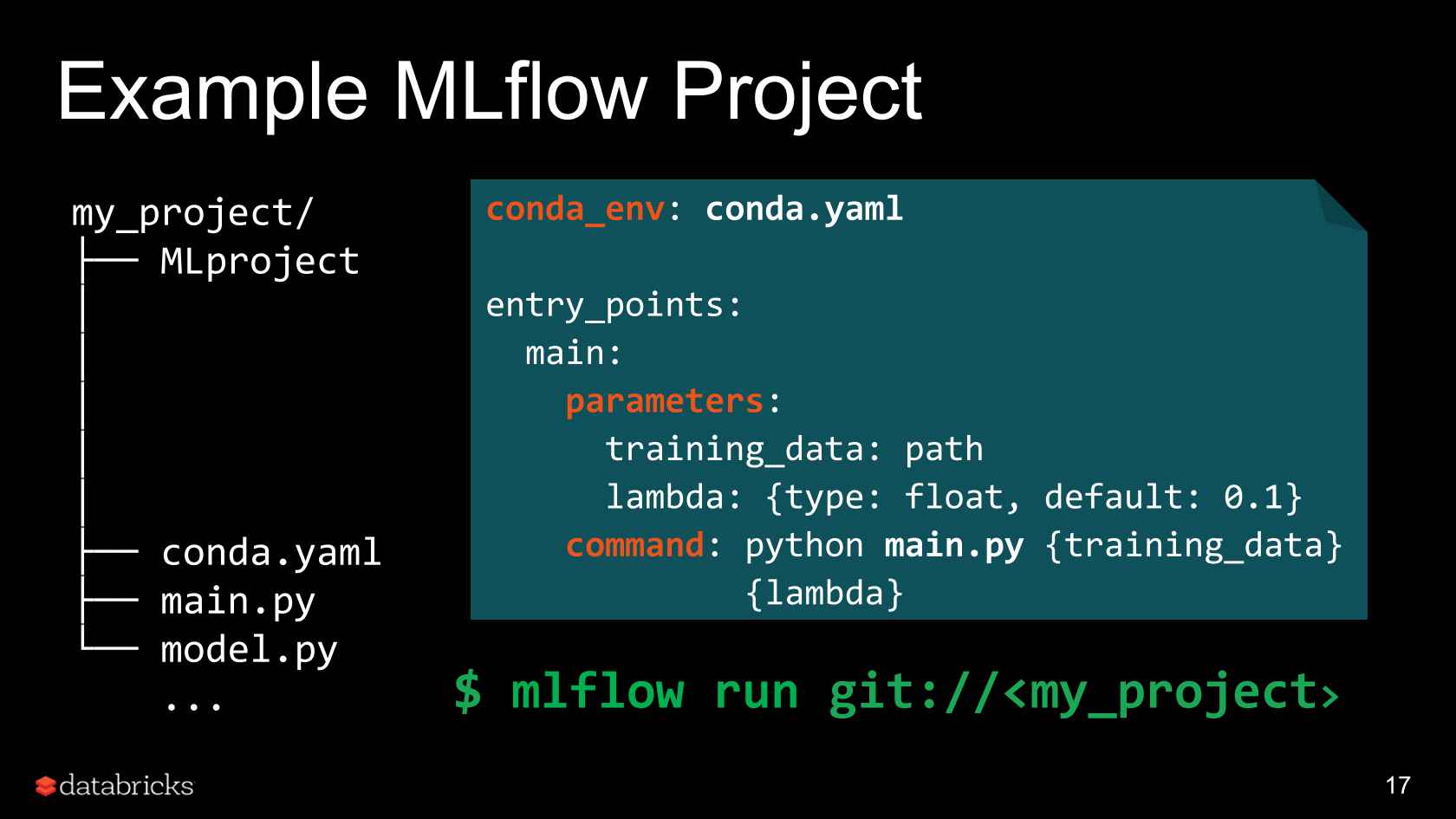


1. **Conda.yaml view:**



1. **Deploy a ML Flow Model for real-time serving**





1. **GitHub Link / URL:**

<https://github.com/bksindhu/Capstone---News-articles-classifier/>

<https://github.com/DayakarKodirekka/Capstone---News-articles-classifier>

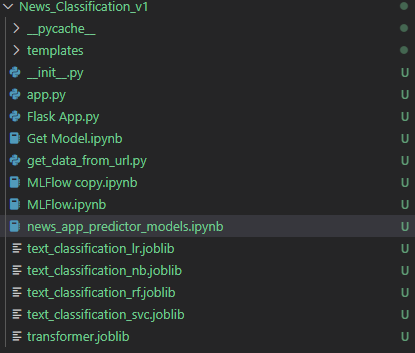
1. **Challenges:**

ML results are difficult to reproduce

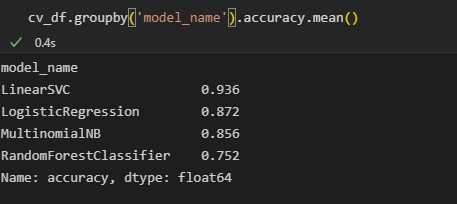
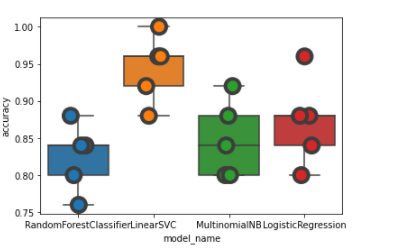
Wordcloud



After generating model for different models, select the best model and use that in the web application development



Model accuracies for the models are as below



Once we get the model files saved

Run the flask api

Notes:

As the data we are getting from fastapi the classification of the topic is not good or accurate . in most of the cases the data is miss classified.

As this capstone project is not focused on accuracy of the model we are diving deep into the problem to improve the accuracy of the model.

Rather we are building an whole architecture to develop an user interface with flask api where user can enter the text and model will predict the type of news,

The type of news is from the trained data only, if we need the prediction accuracy exactly match, we need to refine the model.