BUSS-5802

Capstone Project

MLFlow in Databricks

Group 9

Alejandra

Irving C

Radhika Jayakrishnan

Rohit Mahendran

Content

Introduction

Databricks is a unified analytics platform for building, deploying, sharing, and maintaining enterprise-grade data, analytics, and AI solutions at scale. It has components of Data Engineering, Machine Learning and Data Management. MLFlow is an open-source platform for managing the end-to-end machine learning lifecycle. Databricks has seamlessly integrated MLFlow into its features to manage ML projects at scale. We have leveraged MLFlow in Databricks to forecast the estimated fire area of potential Australian bushfires using weather data from NASA.

Problem Statement

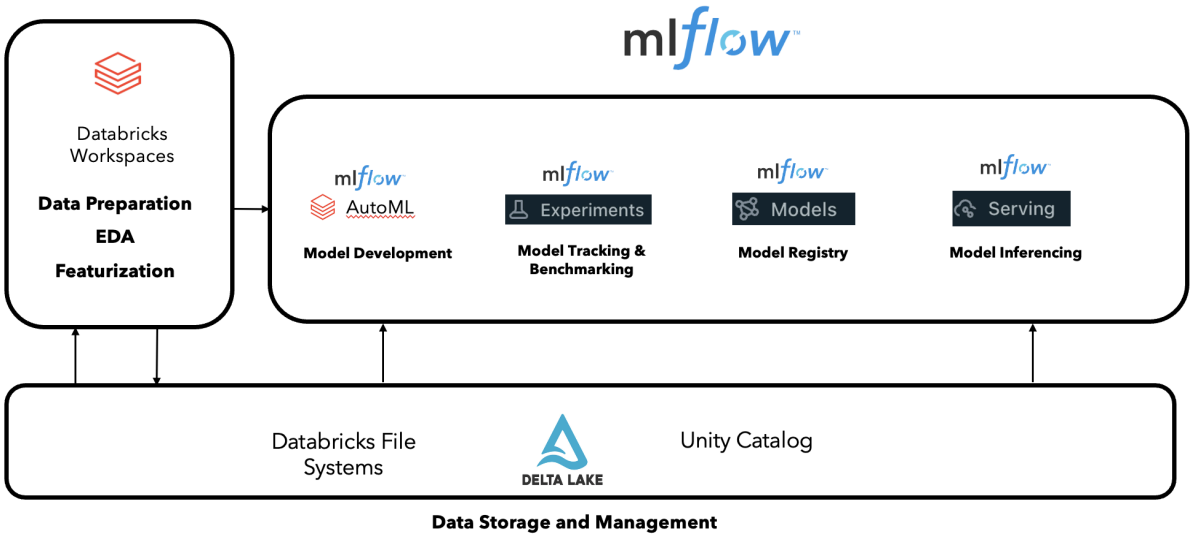
Bushfires in Australia are a widespread and regular occurrence that have contributed significantly to shaping the nature of the continent over millions of years.[1] The 2020 wildfire season, as reported by RGS, stands as a stark testament to the devastating impact these fires can inflict, with tragic statistics including the loss of 28 lives, destruction of over 2,000 homes, an estimated 1 billion animals perished, and a significant depletion in the koala population, with one-third lost.[3] These catastrophic events are often preceded by a combination of extreme high temperatures, low relative humidity, and strong winds, creating optimal conditions for the rapid spread of fires, exacerbating their destructive potential.[3]

Dataset

'Predicting Wildfires with Weather Forecast Data' is a Spot Challenge curated by IBM to predict the area of wildfires. IBM collected weather and wildfire data generated by NASA and curated it for 15 years, from 2005-2020. Our project uses 4 datasets to predict wildfire for Australia in February 2021. The datasets provided are wildfires, historical weather, vegetation index, and land classes. Our target column will be 'Estimated_fire_area'.

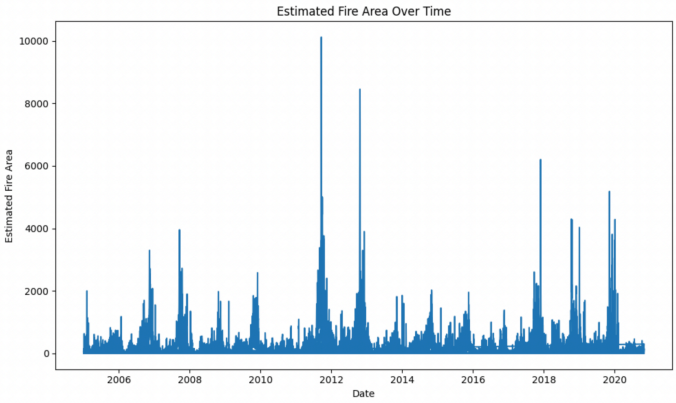
System Architecture

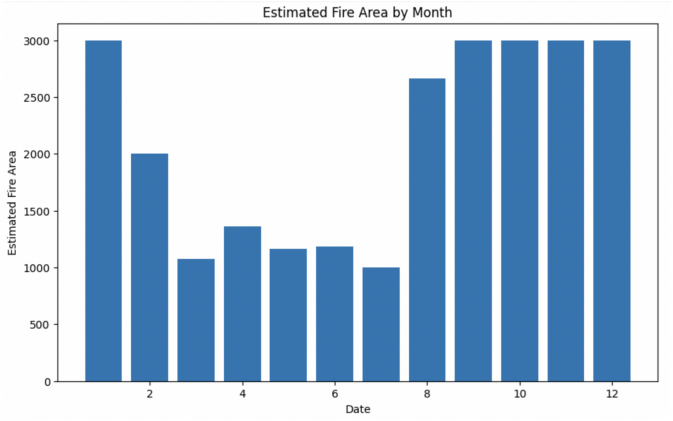
Databricks is a unified platform to manage our end-to-end machine learning flow. The machine learning pipeline was managed by MLFlow. The data was stored in Spark tables in the Delta Lake. It was imported, cleaned and featurized using the Databricks notebooks. The cleaned notebooks were again pushed and stored as a spark table. These tables were used to run AutoML experiments to develop machine learning models. Best performing models were registered and inferred with new data within the platform.

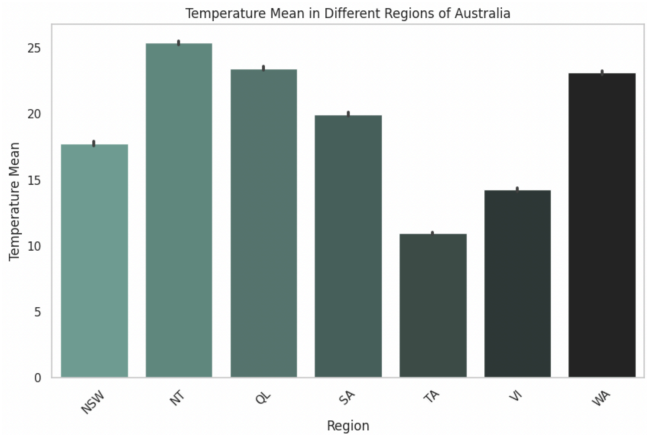


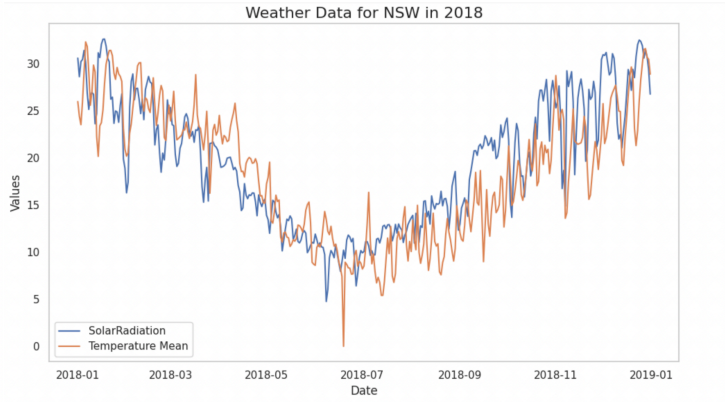
Exploratory Data Analysis

The 4 datasets were imported and initial health checks were performed on them. The datatypes, missing values and duplicate values were checked and rectified. The columns that were not in accordance with the problem statement were dropped and the outliers were fixed. Data exploratory charts were created to better understand the dataset. Here are a few.





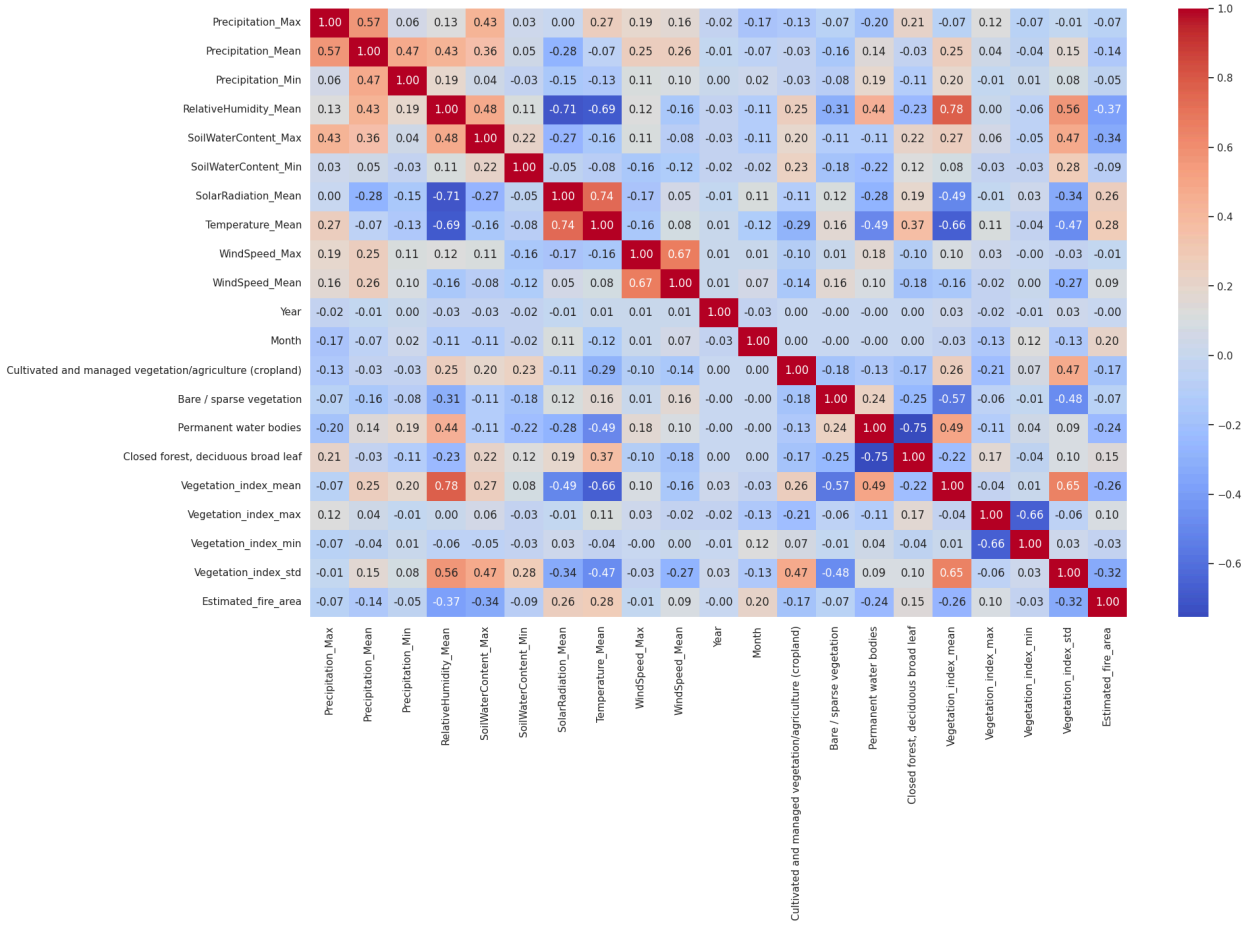




Feature Engineering

1. Correlations

Correlations within individual datasets were checked and columns with multicollinearity were removed. The reduced datasets were then joined and correlations were checked for the merged dataset as well.



2. Granger Causality Test

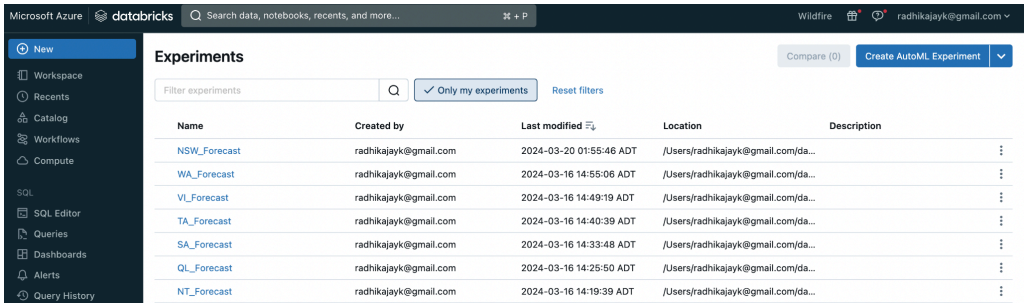
Granger causality is an econometric test used to verify the usefulness of one variable to forecast another. A variable is said to: Granger-cause another variable if it is helpful for forecasting the other variable.[4] Causality tests were performed on all variables to understand which variables may contribute to forest fires and 23 variables were selected that may influence the forest fire area.

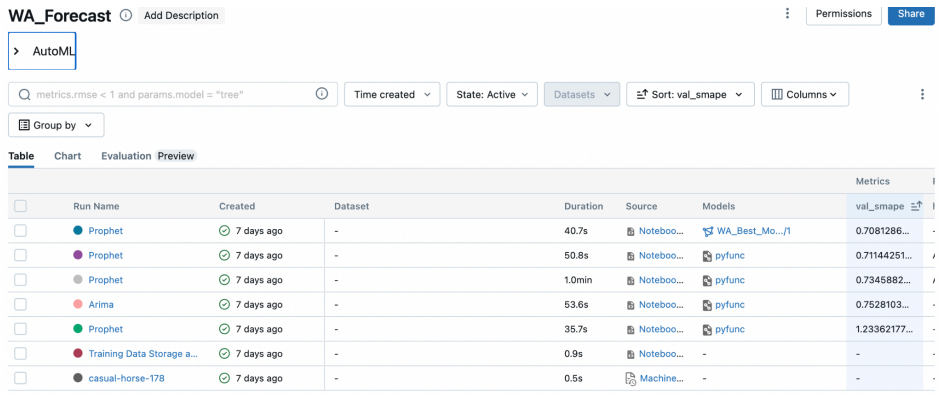
After feature engineering, the dataset was split into 7, as the data had 7 regions. The feature engineered dataframe was then pushed into Delta Lake as Spark tables.

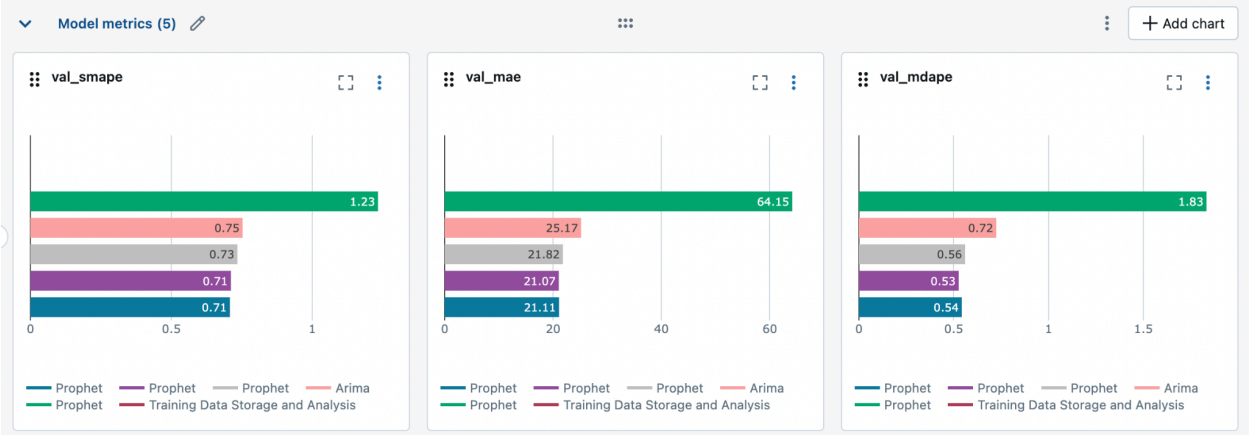
Machine Learning using AutoML

Databricks AutoML streamlines the application of machine learning to datasets by automating the process. Once the dataset is provided and the prediction target is specified, AutoML conducts a series of trials to generate, fine-tune, and assess multiple models. After model evaluation, AutoML displays the results and provides a Python notebook with the source code for each trial run so you can review, reproduce, and modify the code.[5]

Individual experiments were created for the different regions. The experiment was run to forecast the estimated fire area on the timeseries data. AutoML uses ARIMA and Prophet models to forecast. The models within each experiment were compared using the SMAPE metric.

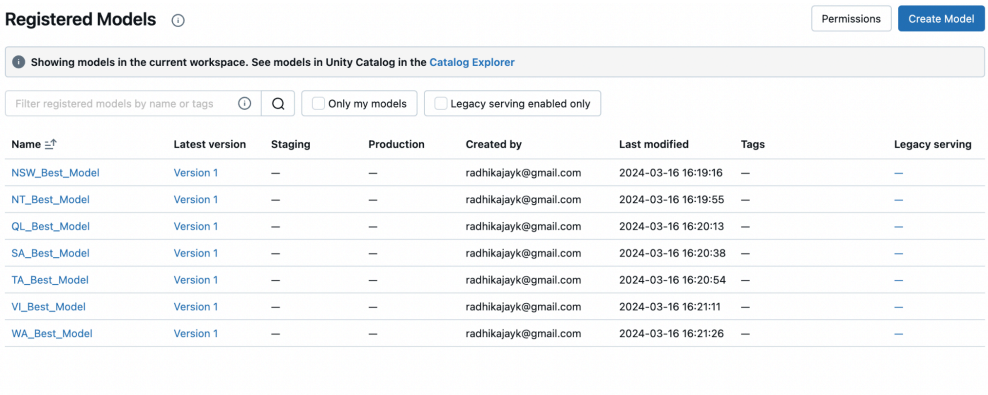






Model Registry

The MLflow Model Registry component is a centralized model store, to collaboratively manage the full lifecycle of MLflow Models. It allows us to track, log models with MLflow and to register models with the Model Registry. The best models were registered from each of the 7 experiments.



Registered Models

Permissions | Create Model

Showing models in the current workspace. See models in Unity Catalog in the Catalog Explorer

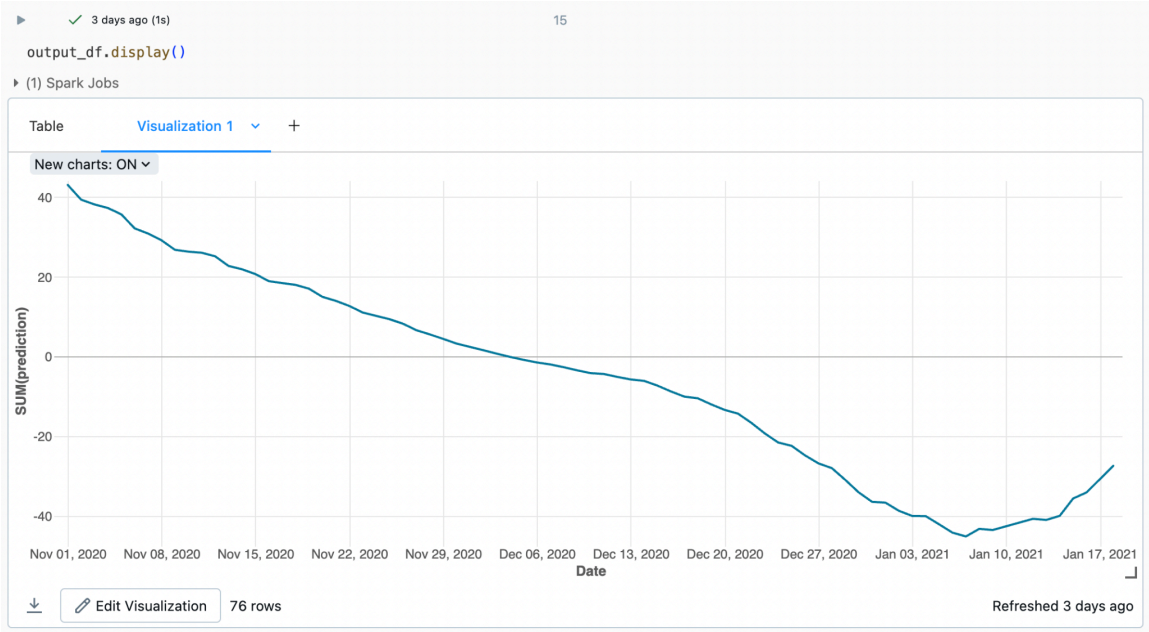
Filter registered models by name or tags | Only my models | Legacy serving enabled only

Name	Latest version	Staging	Production	Created by	Last modified	Tags	Legacy serving
NSW_Best_Model	Version 1	–	–	radhikajayk@gmail.com	2024-03-16 16:19:16	–	–
NT_Best_Model	Version 1	–	–	radhikajayk@gmail.com	2024-03-16 16:19:55	–	–
QL_Best_Model	Version 1	–	–	radhikajayk@gmail.com	2024-03-16 16:20:13	–	–
SA_Best_Model	Version 1	–	–	radhikajayk@gmail.com	2024-03-16 16:20:38	–	–
TA_Best_Model	Version 1	–	–	radhikajayk@gmail.com	2024-03-16 16:20:54	–	–
VI_Best_Model	Version 1	–	–	radhikajayk@gmail.com	2024-03-16 16:21:11	–	–
WA_Best_Model	Version 1	–	–	radhikajayk@gmail.com	2024-03-16 16:21:26	–	–

This allowed us to track and have version control over different versions of model training.

Forecasting using Batch Inference

As the final stage of our project, the best models were used to forecast estimated_fire_area with unseen data. Batch Inference was used for this. Data from Nov2020 - Jan 2021 was used for this.This is the prediction for the Western Australia model.



Results and Conclusion

Databricks is a large platform used for end-to-end machine learning and data analytics. It utilizes MLFlow to run machine learning models on preprocessed data. MLFlow works by running multiple models on the given data, with different hyper parameters and selects the best model based on them. Once registered the best model can be used to forecast/predict on unseen data. The project utilized data curated by NASA and collected by IBM to predict the estimated fire area for Australian bushfires. The data was preprocessed and divided into 7, corresponding to 7 different regions and MLFlow was utilized to run Machine Learning experiments on them. The best model was used to predict data for the year 2021.
The project used features of MLFlow like Unity Catalog, MLFlow Experiments, Model Registry etc to forecast the estimated fire area,

Resources

1. https://en.wikipedia.org/wiki/Bushfires_in_Australia
2. Williams, Liz T. (3 November 2011). "The worst bushfires in Australia's history". Australian Geographic. Retrieved 12 February 2020.
3. https://www.rgs.org/schools/resources-for-schools/australian-wildfires
4. https://www.aptech.com/blog/introduction-to-granger-causality/#:~:text=Granger%20causality%20is%20an%20econometric,for%20forecasting%20the%20other%20variable.
5. https://docs.databricks.com/en/machine-learning/automl/index.html