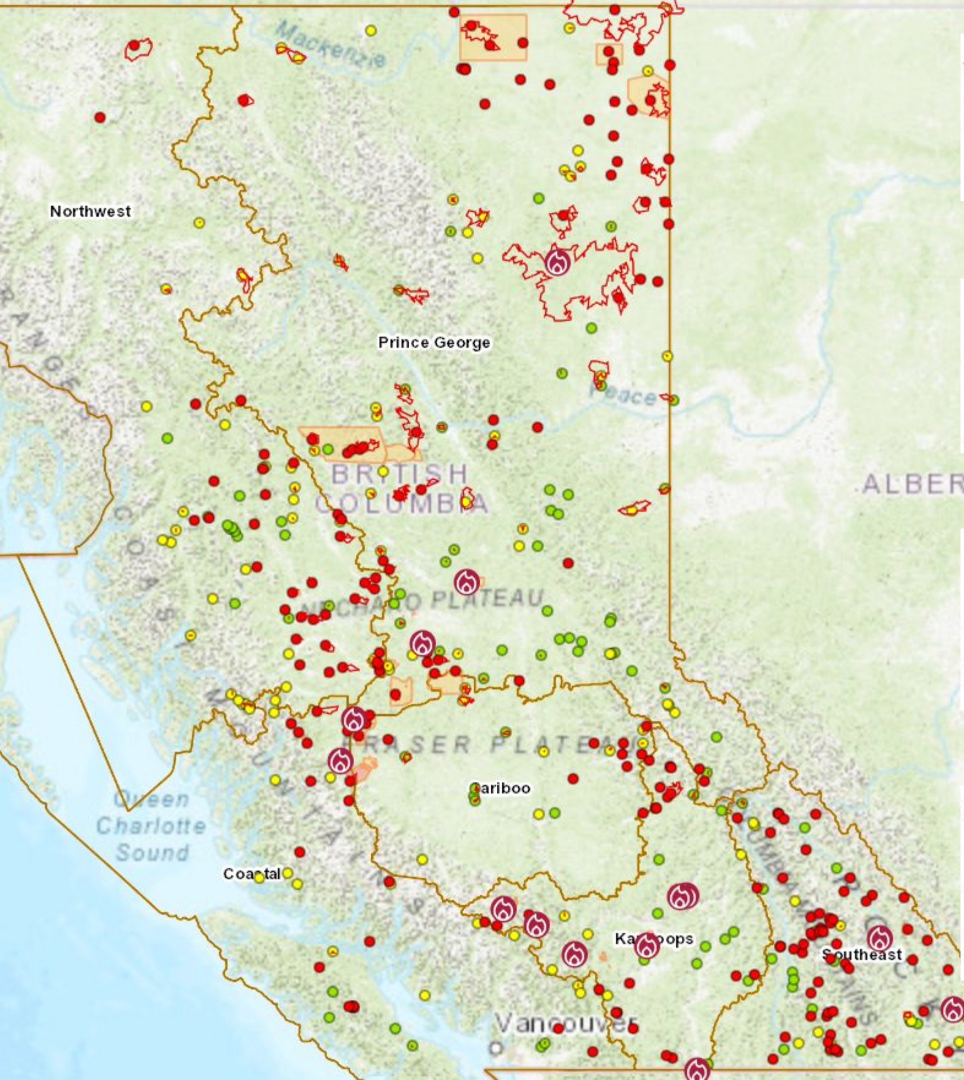
Wildfire Seasonal Feature Importance

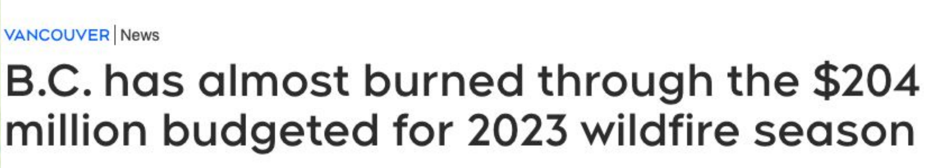
Huizi Wang, Jinda Zhang, Yanting Zheng

Agenda

- Motivation
- Problem Statement
- Algorithms
- Simulation Setup
- Results
- Conclusion and Limitations
- Future Works

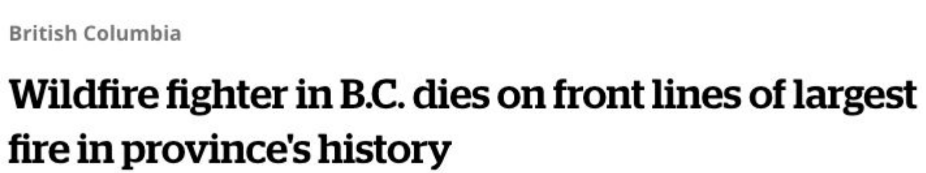
Motivation





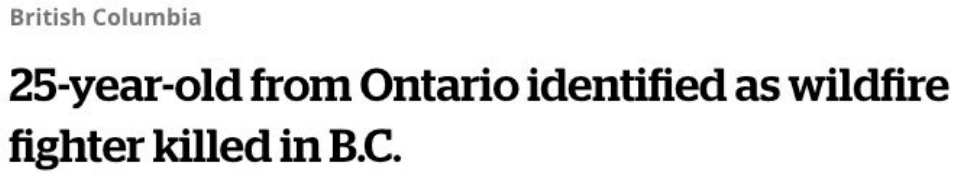
VANCOUVER | News

B.C. has almost burned through the $204 million budgeted for 2023 wildfire season



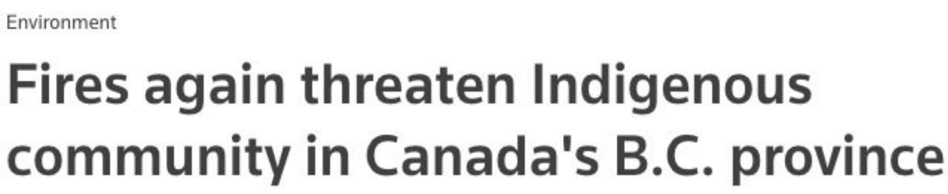
British Columbia

Wildfire fighter in B.C. dies on front lines of largest fire in province's history



British Columbia

25-year-old from Ontario identified as wildfire fighter killed in B.C.



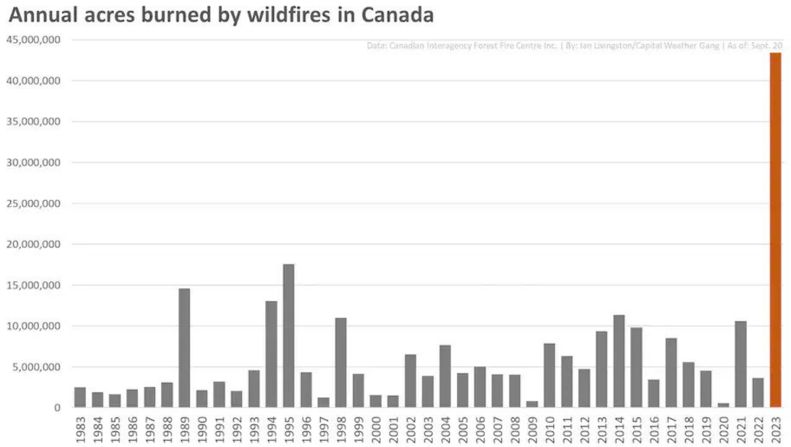
Environment

Fires again threaten Indigenous community in Canada's B.C. province

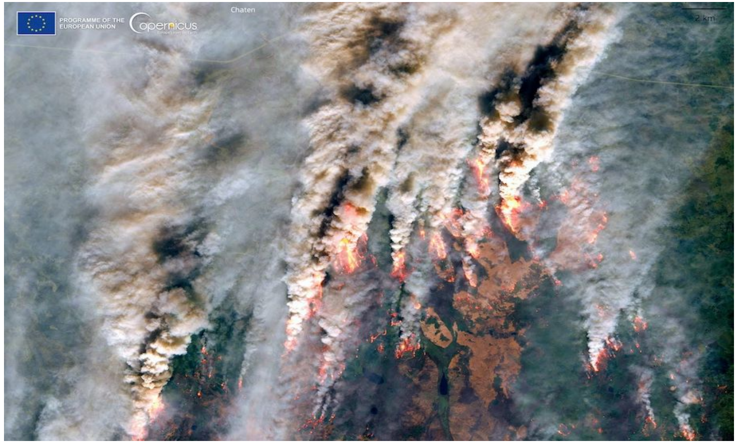
Motivation

- Bring more people’s attention to the severity of wildfires.

- Use machine learning techniques to solve real meaningful problems.


Annual acres burned by wildfires in Canada

Data: Canadian Interagency Forest Fire Centre Inc. | By: Ian Livingston/Capital Weather Gang | As of: Sept. 20



Problem Statement

Feature Importance in Wildfire Prediction

Though lots of research has been conducted to build wildfire prediction model, we didn't find any that focuses on analyzing feature importance across different seasons.

Our goal is to investigate seasonal differences in wildfire occurrences. Our plan can be divided into these steps:

1. Build prediction models for spring, summer and fall.
2. Use metrics to select models that score top in each season.
3. Analyze feature importance with these selected models.
4. Compare differences among the three seasons and make conclusion.

Algorithms

- Feature importance analysis
- Each team member’s algorithms

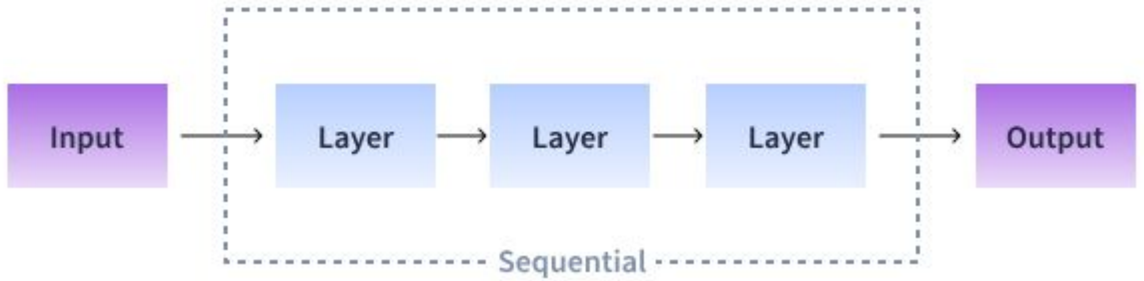
Feature Importance Analysis

Permutation feature importance

1. Calculate the error between prediction and actual labels as baseline error.
2. For each input feature, shuffle all test records of it and maintain the correct order for other features.
3. Predict with shuffled test data and measure the error after the shuffling.
4. Sort the error from large to small, then we get the order of each feature's importance.

The idea behind this method is, the more important a feature is, the more impact it has on the result. Shuffling such features will break the link between features and labels, thus causing error to increase.

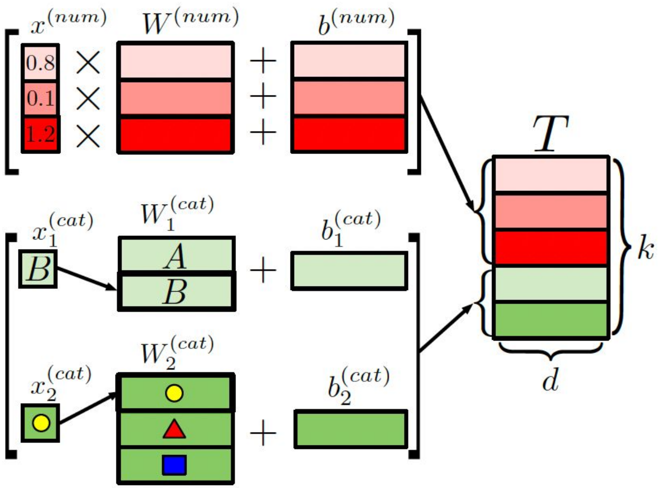
Deep Learning

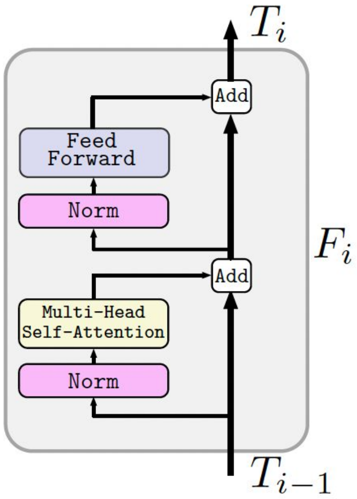


Tabular datasets are the last “unconquered castle” for deep learning, with traditional ML methods like Gradient-Boosted Decision Trees still performing strongly even against recent specialized neural architectures.

– A review paper for deep learning model with tabular datasets

Approach





1. Model 1: Baseline Model
   - Input + 2 hidden layers + output
   - Preliminary result
2. Model 2: Deeper Baseline Model
   - More units and more layers
   - Added Batch Normalization, Dropout
   - Improved result compared to baseline model but not much
3. Model 3: FT-Transformer
   - From “Revisiting Deep Learning Models for Tabular Data”, competitive performance with GBDT models such as XGBoost.
4. Fine-tune Model 2 & 3

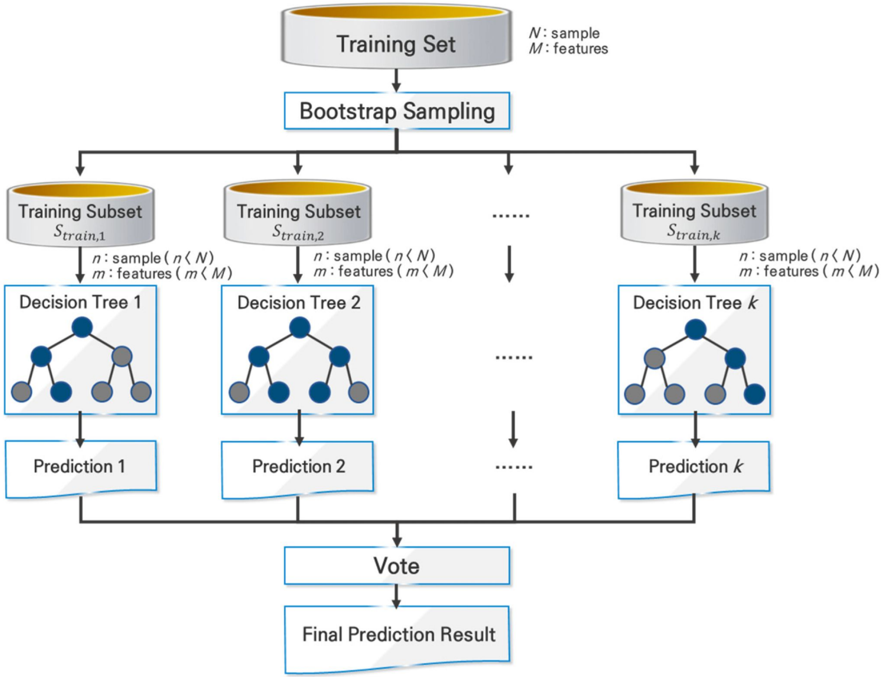
Approach: Random Forest

Ensemble learning method that combines multiple decision trees.

Offers higher accuracy through bagging and feature randomness.

Handles large data sets with higher dimensionality.

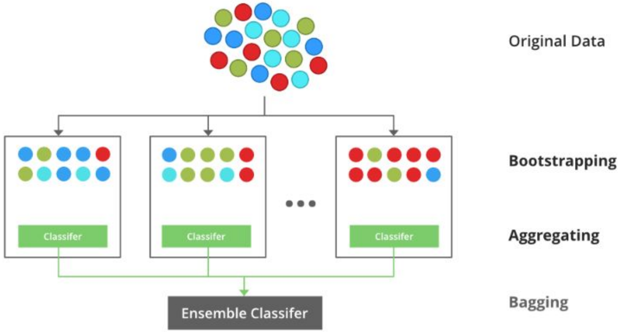
Can model nonlinear relationships.



Reference: https://www.researchgate.net/figure/Architecture-of-a-random-forest_fig1_347075904

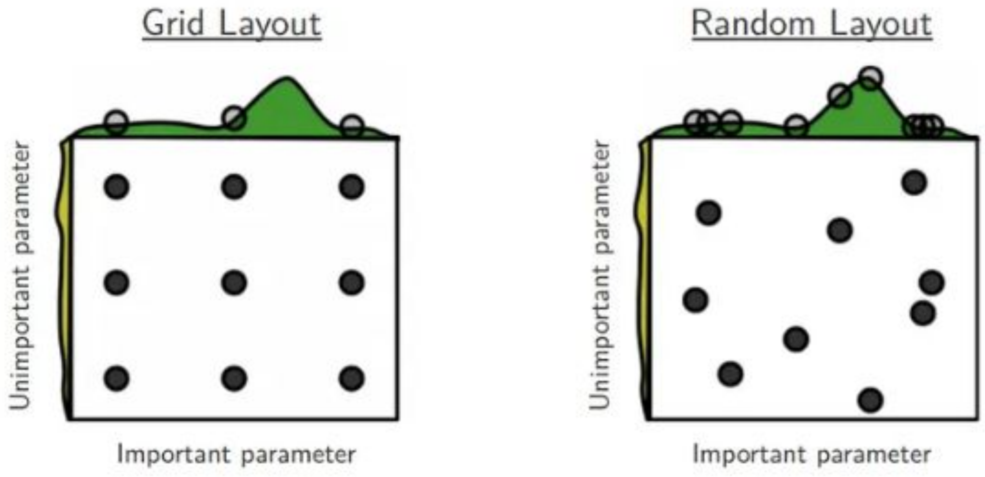
Justin Starts

Approach: eXtreme Gradient Boosting (XGBOOST)



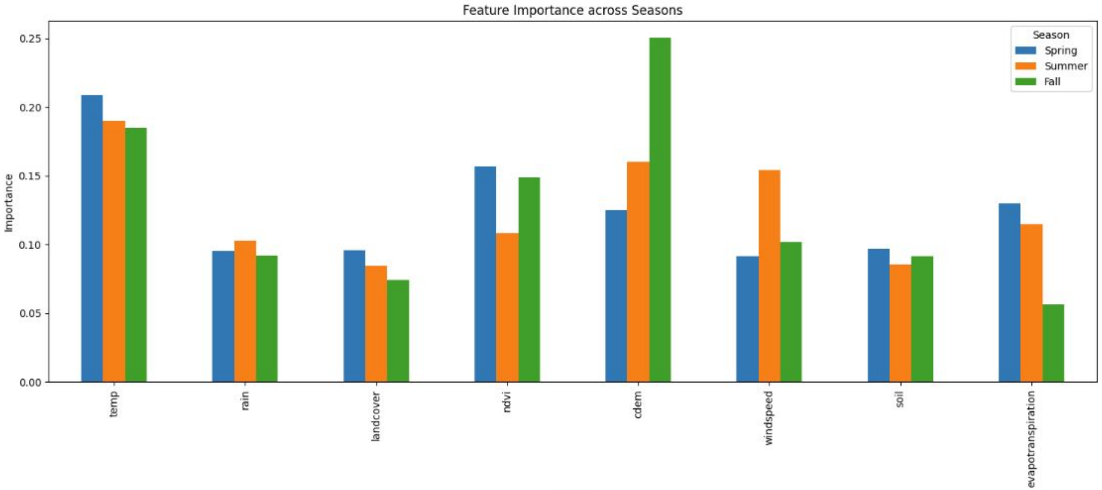
- XGBoost is a scalable, distributed gradient-boosted decision tree method.
- XGBoost has gained prominence for its exceptional performance and wide range of applications in machine learning.

Hyperparameter Tuning: Randomized Search



- More efficient way for **parameter search**
- With grid search, nine trials only test three distinct places. With random search, all nine trails explore distinct values.

XGBOOST Feature Importance



- Model.feature_importances_ API in XGBOOST library
- Seasonal Variations: Investigate the changing importance of predictors across different seasons, bar plot to visualize the results

Approaches

- Kernel SVM: support vector machines with kernel tricks
- LSTM: recurrent neural network, captures temporal dependency
- KNN: Nearest neighbors for classification
- TabNet: attention mechanisms for tabular data
- Logistic Regression: linear classifier
- Naive Bayes: probabilistic models
- 1D-CNN: applies convolutional layers for feature extraction from sequential data

Simulation Setup

Dataset

The fire season defined by BC officials ranges from April to November. So we define the temporal scope at the scale of season as: **Spring (3~5), Summer (6~8), Fall (9~11)**.

Given the temporal overlap of input maps available on Google Earth Engine, we choose training data from 2018-2020 and test data from 2021.

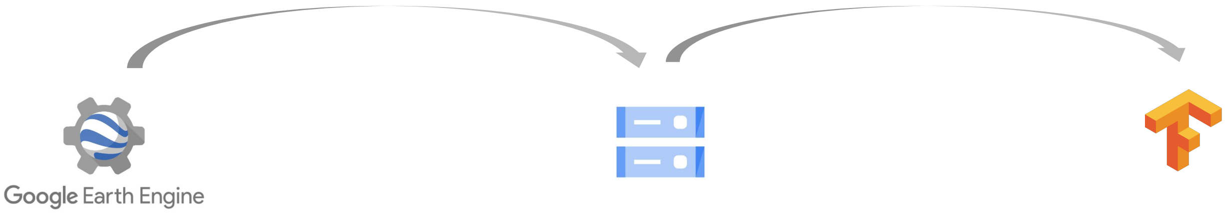
We stratified-sampled **800 points in each month** for training and test.

In summary:

- Training dataset contains 7200 points for each season in three years.
- Test dataset contains 2400 points for each season in 2021.

With stratified sample we make sure the dataset is balanced.

Process: Training



Export **feature collection** in TFRecord format

Parse **TFrecord** patches to get a **tf.data.Dataset** as training input

Sample over **features and label** Dataset in target training regions.

Each file contains features and fire/no fire labels for certain region at certain time.

-Create ML model, fit with training data.
-Use RMSE to evaluate model performance.

Method - environment, tools and config

Pipeline on gcloud platform

Virtual machine

Google Compute Engine

- ➔ GCP: Cloud computing power with T4 GPU for training
- ➔ Research Cluster: k80 GPU
- ➔ Cloud Server(GPU: 3*1080Ti)
- ➔ Colab: T4 GPU

Data buckets and folders

Cloud Storage

- ➔ Created designated data buckets for this project
- ➔ TFrecord format data, Reading from and writing to the bucket
- ➔ Service accounts and permissions

ML package & framework

TensorFlow

- ➔ The data format that we use to record the data - TFrecord
- ➔ Combined with GEE functions to transform EE image to TF data for training
- ➔ Foundation of the unet model with Keras

One-stop shop for ML

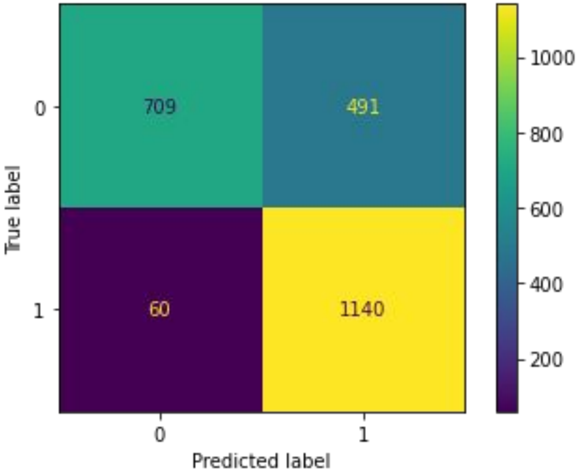
Vertex AI

- ➔ Hosting Jupyter notebook file on VM
- ➔ Future pipeline and deployment

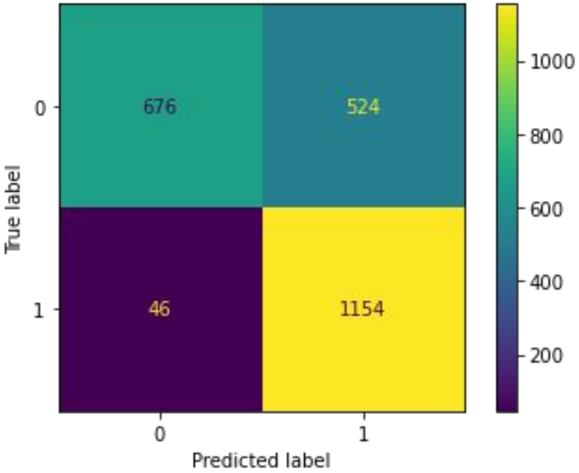
Results

Fine-tuned deeper ANN

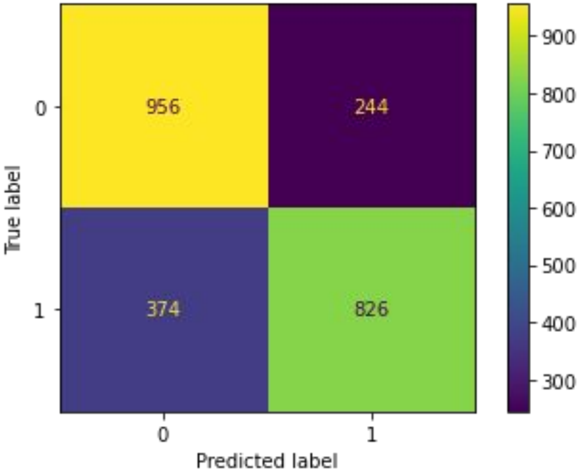
Spring
Recall: 0.95
Accuracy: 0.77



Summer
Recall: 0.96
Accuracy: 0.77

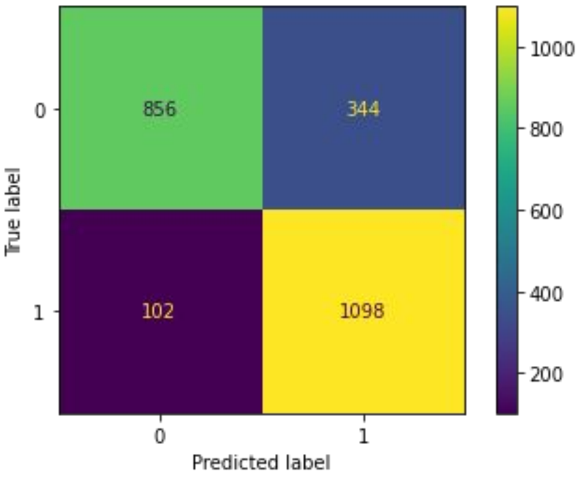


Fall
Recall: 0.87
Accuracy: 0.76

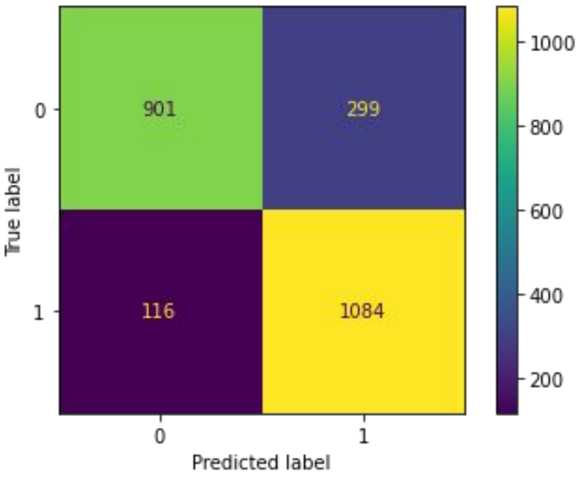


Fine-tuned FT-Transformer

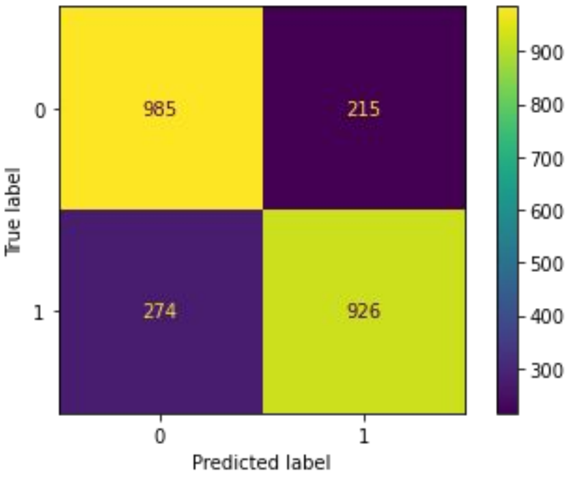
Spring
Recall: 0.92
Accuracy: 0.81



Summer
Recall: 0.91
Accuracy: 0.82

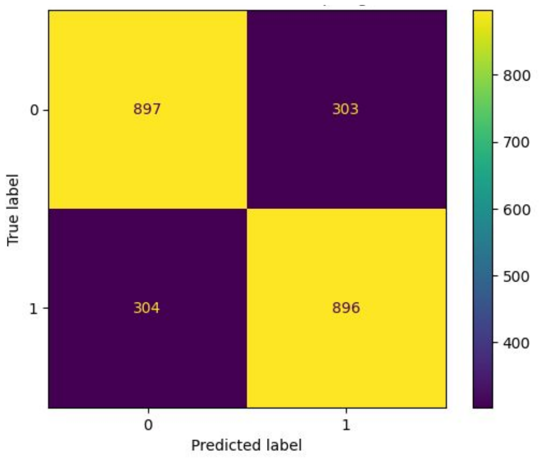


Fall
Recall: 0.86
Accuracy: 0.81

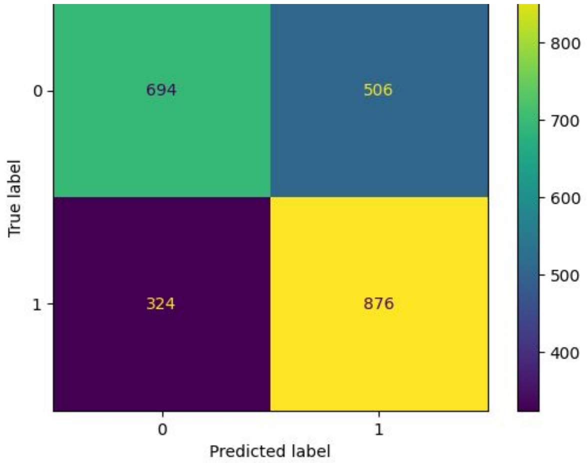


Fine-tuned Random Forest

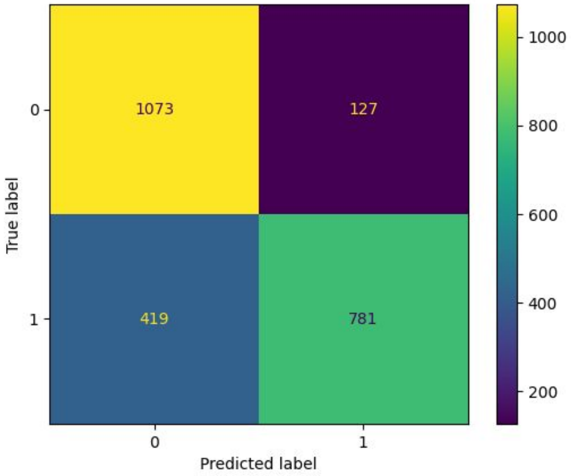
Spring
Recall: 0.88
Accuracy: 0.88



Summer
Recall: 0.87
Accuracy: 0.88

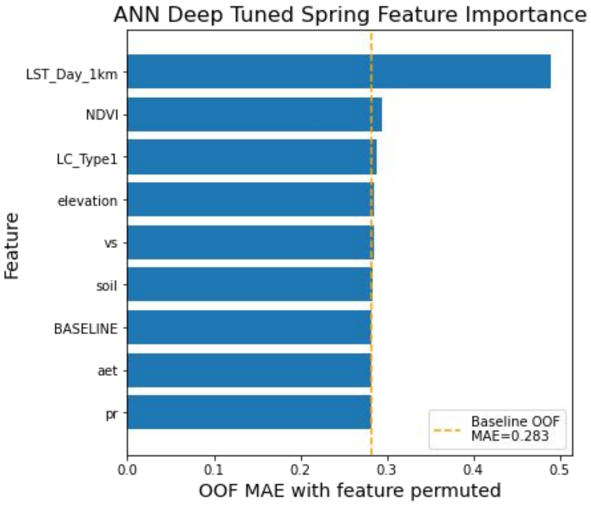


Fall
Recall: 0.90
Accuracy: 0.90

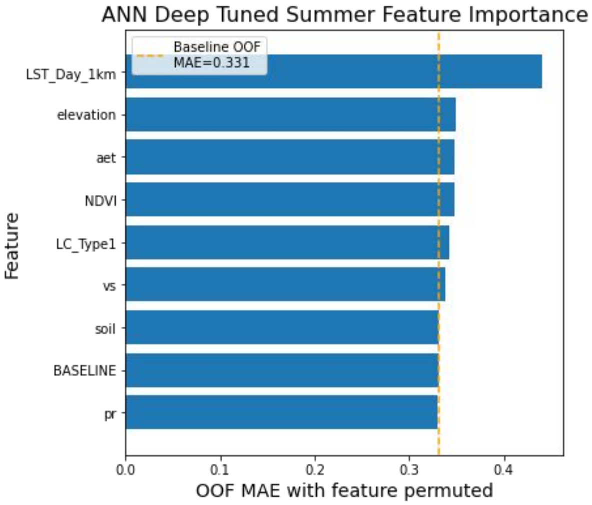


Fine-tuned deeper ANN

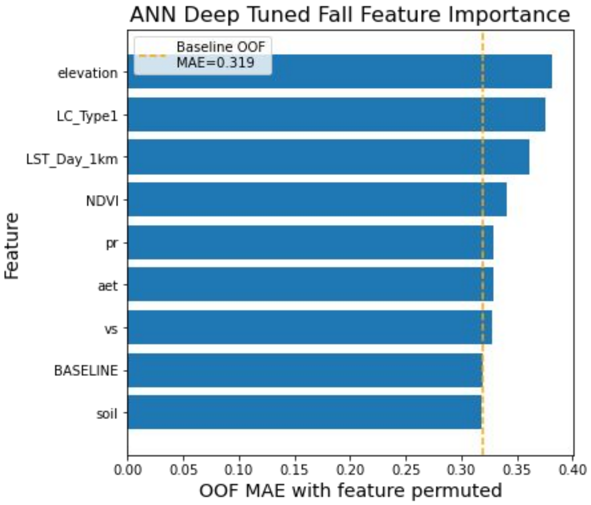
Spring


ANN Deep Tuned Spring Feature Importance

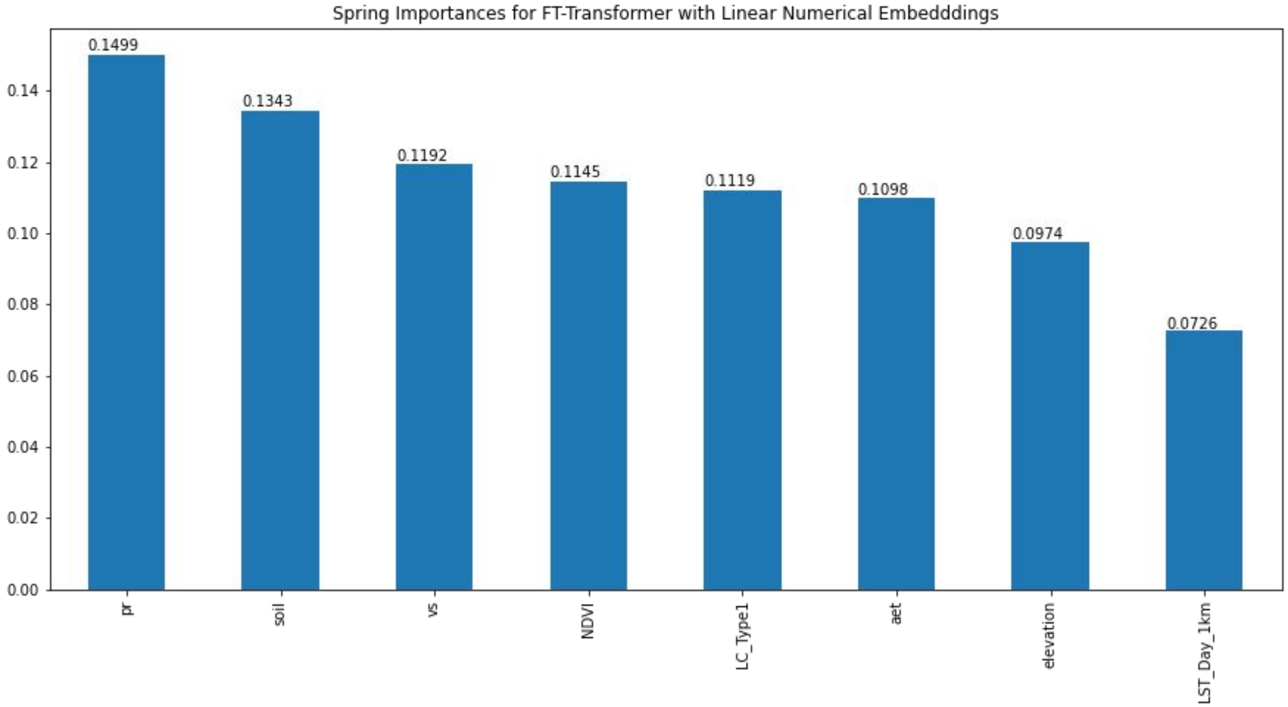
Summer


ANN Deep Tuned Summer Feature Importance

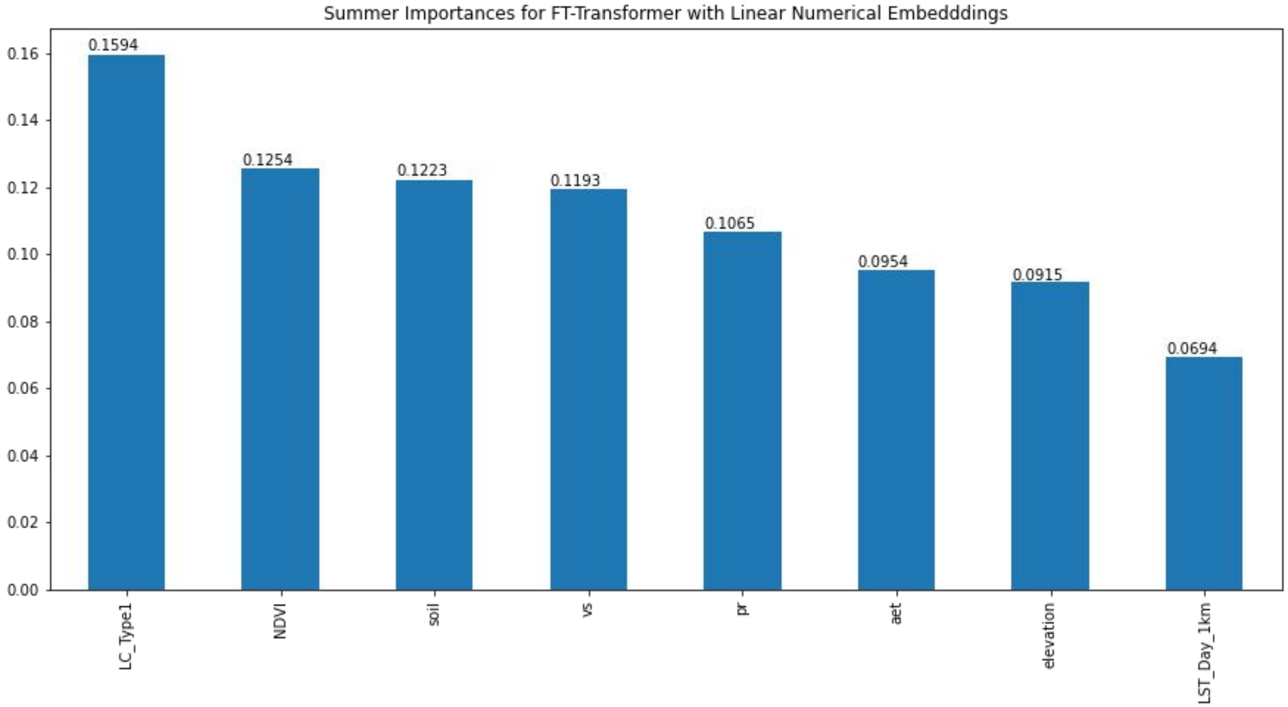
Fall


ANN Deep Tuned Fall Feature Importance

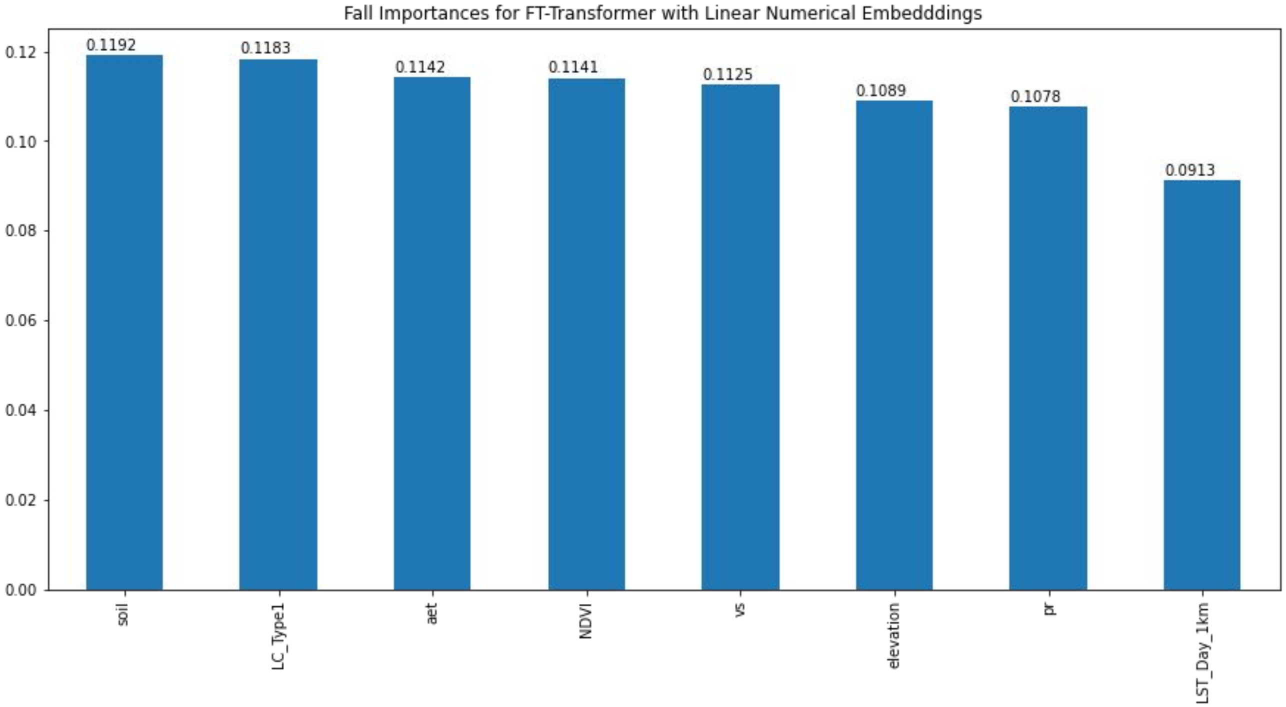
Fine-tuned FT-Transformer



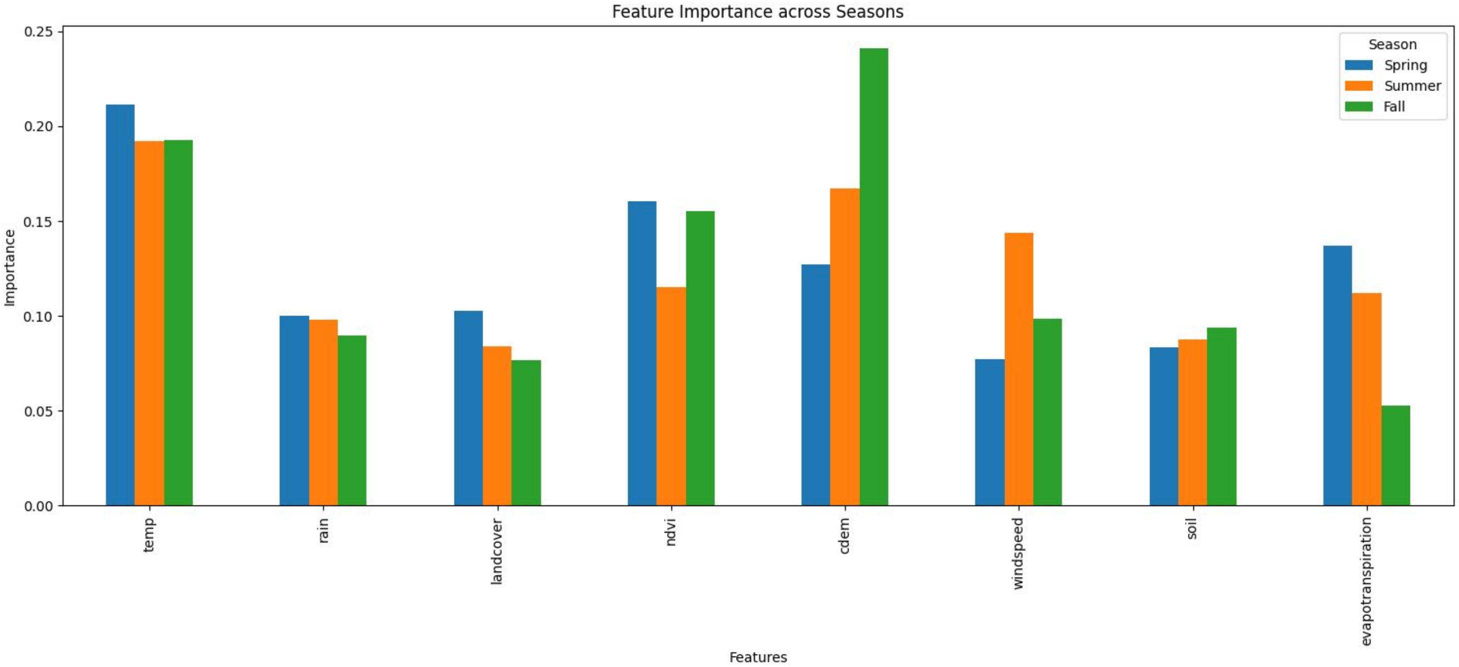
Fine-tuned FT-Transformer



Fine-tuned FT-Transformer

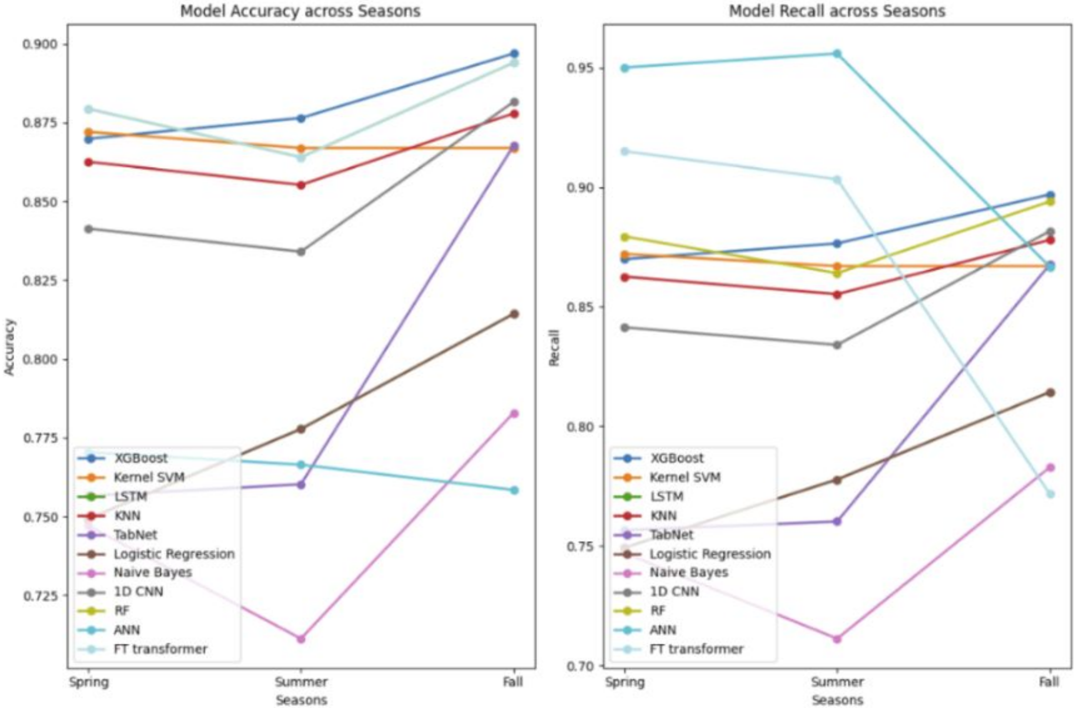


Fine-tuned Random Forest



Model Selection

- Select based on **Recall** and **Accuracy**
- For Spring, choose **fine-tuned Artificial Neural Network** have high accuracy(0.95), with accuracy(0.75).
- For Summer, choose **Random Forest** for its consistent high accuracy and recall(0.86).
- For Fall, choose **XGBoost** for its consistent high accuracy and recall(0.89).

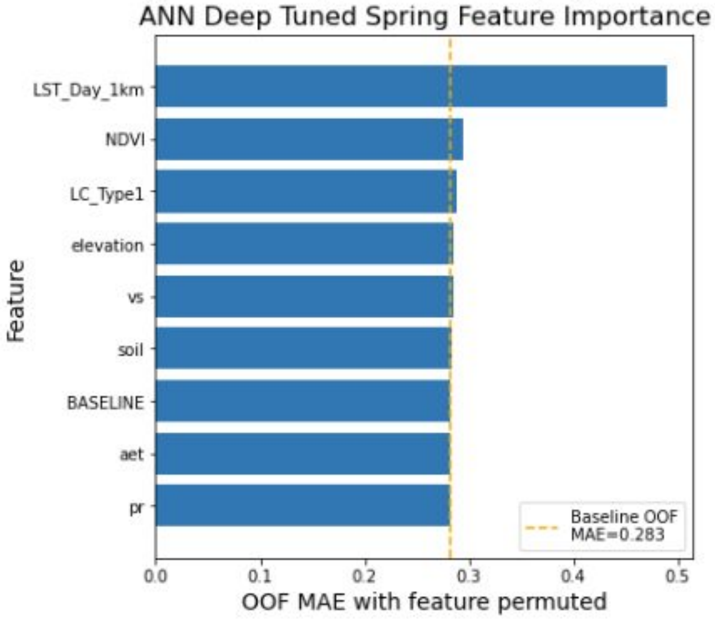


Feature Importance Results for Spring

- the most important feature for Spring model(Fine-tuned ANN) is Land Surface Temperature(LST_Day_1km)

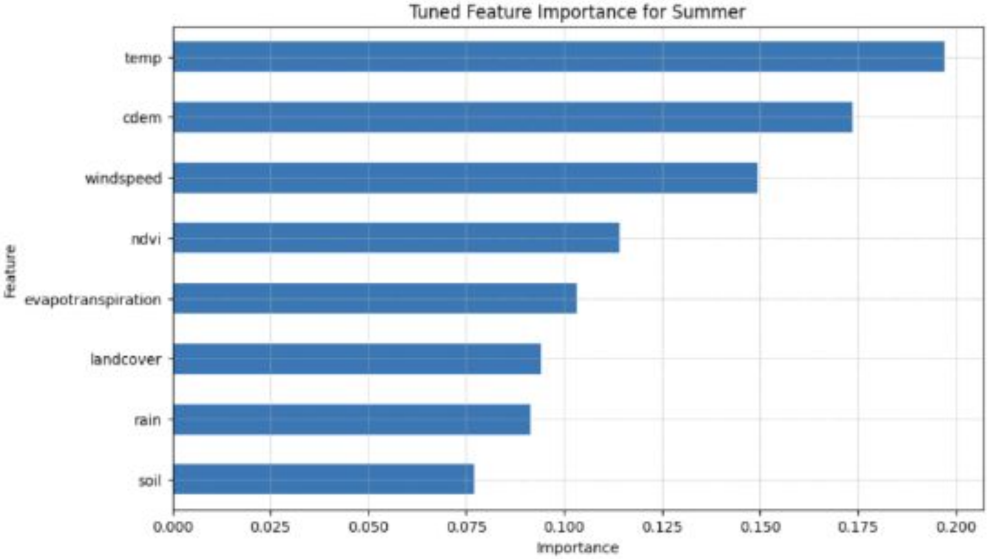
- Temperature plays important role

Fig. 12. Feature Importance Results for Spring



Feature Importance Results for Summer

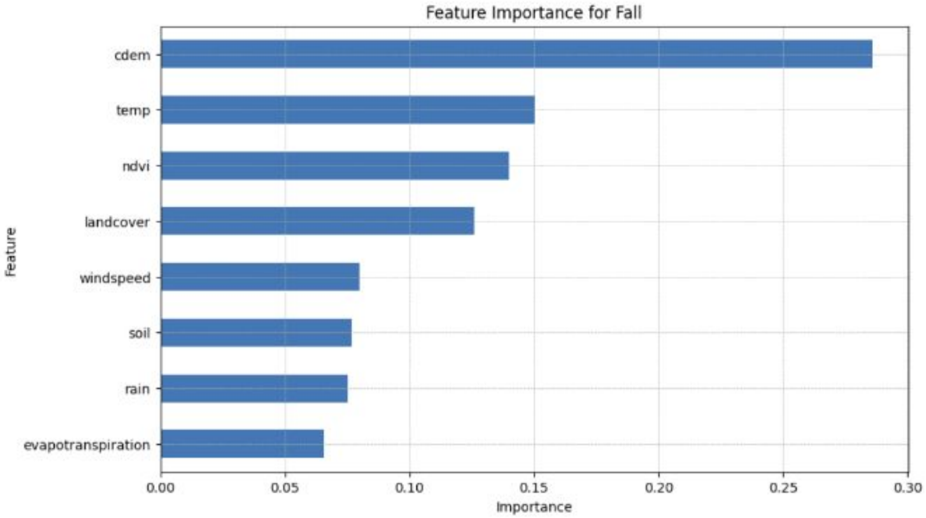
- the top 3 important feature for Fall (Random Forest) is elevation, temperature and Normalized Difference Vegetation Index(NDVI)



Feature Importance Results for Fall

-Top 3 important feature for Fall is elevation, temperature and Normalized Difference Vegetation Index(NDVI).

-Higher elevations tend to have cooler temperatures, which can affect vegetation. Lower NDVI values may indicate decreased vegetation cover and increased vulnerability to wildfires.



Discussion

- In Spring, Land Surface Temperature emerges as the most critical feature, aligning with the understanding of how temperature impacts vegetation flammability and fire spread.

- Moving into Summer, temperature, the combination of high temperatures, strong winds, and varying elevations contributes to conditions for rapid wildfire propagation.

- In Fall, top features are elevation, temperature, and Normalized Difference Vegetation Index (NDVI). Lower NDVI values indicate decreased vegetation cover and heightened vulnerability to wildfires.

Limitations & Conclusion

Limitations

- **Lack certain features**

In some literature for wildfire prediction, researchers found that anthropological factors also play an important role in accurately predicting wildfire occurrences. We didn't include such features because they are inaccessible from GEE.

- **Lack of exploration for seasonal performance differences**

For certain models, the performance can vary significantly across the three seasons. We didn't find enough information that can explain the differences. More training data may be needed to ensure the models can learn enough patterns for the poorly-performed seasons.

Future Works

Future Works

- **Excluding less important features**

A potential future research direction is conducting experiments where features with low importance are excluded.

- **Exploring beyond the default architectures**

In this study, we borrowed models from existing research without applying any changes to the architectures. For example, in FT-Transformer, there are other options to embed the numeric features other than the piecewise linear method we used, such as periodic encoding.

- **Application**

Getting decent results is just the first step. What makes the results meaningful is figuring out what we can do in wildfire management. To achieve this goal, more studies will be needed outside the realm of machine learning.

Conclusion

- Seasonal Variability in Wildfire Predictors
- Key Seasonal Insights
- Data Limitations and Future Research
- Practical Implications
- Paving the Way for Precision Forecasting

Thank you