Phase 3 Presentation

Feature Engineering and Hyperparameter tuning

W261 Machine Learning at Scale, Fall '22
Luis Garcia Lizaran, Nic Brathwaite, Himabindu Thota
Group 1 Section 1

Berkeley
UNIVERSITY OF CALIFORNIA

Berkeley
UNIVERSITY OF CALIFORNIA

Question: Predict whether the flight will be delayed by 15 mins or more

Target Audience: Airline Travelers

Feature Engineering

Feature Engineering - Data

- Airlines Data
    - Removed perfectly correlated and highly correlated features in each subsection of the flights dataset.
    - Removed rows associated with canceled flights (Tail_num null)
    - Removed Diverted_4 and Diverted_5 columns which have all null values.
- Weather Data
    - We've noted many null values in the weather dataset.
    - 2 phase imputation
        - Phase 1: Imputed with station-wise average.
        - Phase 2: remaining nulls with global mean values

Feature Engineering - Graph and Time based features

- Airline previous voyage
    - Flight tracked by tail number landing on the same day
    - Departing and flight performance of the previous voyage.
    - Weather information at the source airport of the previous voyage

Numerical & Categorical Features

- Categorical features are string indexed and represented using one-hot encoding vectors.
    - Categorical features include ORIGIN, DEST, CARRIER, STATE, Weather Station Name etc.
    - Note: Numerical features which were represented as strings in the weather data were converted to double type.
- Normalization
    - Strategy: Standard scalar.

Data Join

Status 2 hours before flight departure

- Joined all the daily (or previous day) weather timestamps of the closest station for each origin airport
- Removed weather information not available 2 hours before the flight departure
- Taken the most recent data

History of the airplane

- Used information about previous performance of the same airplane

  (accounting for when each outcome is known)

  - Previous departure delay
  - Previous arrival delay
  - Previous diverted flights

Modeling

Feature Transformation

- Standardization
  - Standard Scaler : 0 mean
    - Normal distribution per attribute
- Withheld Features
  - Timestamp information

Pipeline

- Filtering
  - Set a year for experimentation
- Sampling
  - Take a sample of 20% from the given year
- Stages
  - Identify, Index, and One-Hot Encode our category variables
  - Vectorize
  - Standardize
- Fit & Transformation
  - Fit and Transform our dataset

Logistic Regression

```
The accuracy score of our Logistic Model is: 0.8198117872322518
The precision score of our Logistic Model is: 0.8211729114605484
The recall of our Logistic Model is: 0.8198117872322519
The F1 score of our Logistic Model is: 0.7447914141792943
```

Parameters: regParam = 0.1, maxIterations = 50, elasticNetParam = 1

Experimentation

- 2015 joined dataset
- Pipeline Initialization
- 20% of the total entries

Time/Duration

Data Imputation	~8 mins
Join Data	~24 mins
Logistic Regression (20% sampled data)	~12 mins

Next Steps (Phase IV)

- More Feature Engineering
- New Pipeline(s)
- Additional Models
- Generalization