INDIVIDUAL ASSINGMENT No. 4 **Due Date: April 15, 10:00pm | Weight: 10%**

Business Challenge: Use different types of trees to predict flight departure delays by carrier and airport using records of 1m flights provided in class.

Data Understanding:
File: Flight Delays Data.csv (on Canvas)
Rows:

- Each row = 1 flight occurrence

Columns:

- Year, Month, DayMonth, Day Week
- Carrier(s)
- OriginAirportID, DestinationAirportID
- CRSDepartureTime (hours/24), DepartureDelay (min), DepartureDelay15 (#15min)
- CRSArrivalTime (hours/24), ArrivalDelay (min), ArrivalDelay15 (#15min)
- Cancelled / Not cancelled

Process:

1. Import data
2. Data preparation:
    a. Dependent variable: plane departure delay in number of minutes.
    b. Select columns (eliminate columns that can leak information about prediction targets or not relevant to the analysis).
    c. If any missing values assign dummy variable.
    d. Apply math operations (for ex. minutes to hours, round minutes to next hour)
    e. Edit metadata (force columns to be categorical rather than numerical / vice versa)
3. Randomly partition and sample into training and hold-out samples
4. Run Regression Tree and Random Forest + Classification Tree and Random Forest
    a. Used library algorithms.
    b. Tune models hyper-parameters: tree depth 8, minimum units in each non-terminal node 20, and R^2 minimal improvement per split 1%.
    c. Train algorithms on training data and generate predictor algorithms.
    d. Cross Validate.
5. Performance & Analysis
    a. Report fit for Regression/Random Forest + Classification/Random Forest.
    b. Rank performance between 4 trees using MSE

Deliverables:

1. **Executive Summary:** 10 Lines, non-technical, describing:
    a. Business Understanding: Target business value-add
    b. Data Understanding: Credentials, relevance, and top categories
    c. Modeling: Techniques, parameters, and performance metrics
    d. Performance & Evaluation: Was value-add achieved, areas for further analytics
2. **Technical Methodology:**

STEPS	SPECS
1. Data Management & Workflow	For each row, use technical language to provide step-by-step description of data & modeling decisions, why such decisions, whether performance was as desired, and areas for further analytics. Use business challenge above as model (Max > 20 words per line item).
2. Training & Out-of-Sample Datasets	
3. Model Parameters	
4. Fit & Performance	

3. **Data & Code:** Submit files for following items (in Text, Excel, R, or Python):
    a. Training & hold-out sample datasets
    b. Predictive Algorithms
    c. Interface for prediction & delivery

Evaluation Criteria:

1. **Executive Summary:** Ability to explain and justify in non-technical language your outcome business value-add based on the process and methodology you chose for business understanding, data understanding, modeling, and performance & evaluation.
2. **Technical Methodology:** Ability to use technical language to describe your process and methodology for above table items 1-4 and generate a step-by-step by step guide to duplicate your analytics using the datasets you provided as attachments.
3. **Data & Code:** Usability of datasets and code as provided to duplicate analytics using the Technical Methodology you provided.

Scoring Method

Executive Summary | Clarity of:	**3**
Data Understanding Description	1
Modeling Description	1
Performance & Evaluation Description	1
Technical Methodology | Clarity of:	**4**
Data Classification Description	1
Training & Out-of-Sample Datasets Description	1
Model Parameters Description	1
Fit & Performance Description	1
Data & Code | Duplication:	**3**
Training & Out-of-Sample Datasets Duplicability	1
Algorithms Duplicability	1
Interface for Prediction & Delivery Duplicability	1
Total (Max):	**10**