**Executive Summary:**

The business goal of this paper is to build machine learning models that can predict the flight departure delays by using flights records from different carriers and airports. The dataset we used has 1m flight records from year 2013, for each records it has 14 different independent variables including: *Year, Month, DayMonth, Day Week, Carrier(s), OriginAirportID, DestinationAirportID, CRSDepartureTime, DepartureDelay, DepartureDelay15, CRSArrivalTime, ArrivalDelay, ArrivalDelay15, Cancelled.* Only 12 of them are remained after data cleaning including a new variable *FlightTime*. Four machine learning tree models are used in building the model. They are Regression Tree, Random Forest Regression Tree, Classification Tree and Random Forest Classification Tree. For results, Random Forest Regression Tree is the winner out of the four.

**Technical Methodology:**

*Data Management & Workflow* : The dependent variable is *DepDelay.* Negative values of *DepDelay* are changed to zero since it represents early departure. Data points for cancelled flights are deleted from the dataset. The dataset originally has 14 dependent variables. A new variable *FlightTime* is added to the dataset which calculates the difference between *CRSArrTime* and *CRSDepTime*. Then *Cancelled,CRSArrTime, CRSDepTime* are dropped due to redundancy.

*Training & Out-of-Sample Datasets* : For Classification tree, categorical variable 1/0 is assigned for *DepDelay*, 1 for *DepDelay* greater than 0, and 0 for *DepDelay* less than 0. Data points for cancelled flights are deleted from the dataset. The dataset is split in 80%-20% for test/holdout set.

*Model Parameters* : For the 4 types of tree models I used, tree depth is set to 8, minimum units for node is 20, and minimum improvement for split is 1% for R^2.

*Fit & Performance* ：The average cross-validation score for Regression Tree, RandomForest Regressor, Classification Tree and RandomForest Classifier are 0.9269/0.9274/0.7813/0.7561, the MSE for these four models are 89.4430/88.6433/0.2184/0.2161. Clearly, Regression Tree and RandomForest Regressor has better performance than Classification Trees.