

CIS 600 / CSE 691: Course Project

Instructions: This is a group project. You will work in teams of 3-4 students. There are multiple due dates corresponding to various activities. As a final document you need to upload a project report as a PDF file, your final predictions as a CSV file, as well as a PDF version of the python code you used including the outputs generated. As in the assignments, make sure the PDF documents are readable.

- **Team formation:** By Tuesday March 28 11:59pm send an email with the names of all students in your team to me ngautam@syr.edu, Cc the TAs (Yifu: yqiao08@syr.edu and Amith: amadhava@syr.edu), as well as ALL the team members. Only ONE member of the team needs to send an email but they MUST copy the others. Teams should have 3 to 4 students each.
- **Problem Description:** United Airlines has direct flights into Syracuse (SYR) from four cities: Chicago (ORD), Denver (DEN), Newark (EWR), and Washington (IAD). There could be more than one flight from each of those cities into SYR each day. The goal is to predict 1-4 days in advance if each flight's arrival time into SYR would be early, on-time, delayed, or severely delayed.
- **Data:** We are not imposing any restrictions on the data you can use. One source of data is from the Bureau of Transportation Statistics <https://www.transtats.bts.gov/ontime/> but there could be others.
- **Additional Data:** You are welcome to read about the factors that cause airline delays. One factor is weather and there are many sites that provide weather predictions on an hourly basis for the next few days.
- **Final Predictions:** We will provide a CSV file with an empty column for predicting the arrival time (early, on-time, delayed, or severely delayed). The other columns would be date (April 21-24 to be predicted on April 20), origin airport, and flight number.
- **Ground Truth:** This will be based what is posted on the United Airlines website. If the flight is more than 10 minutes early, we will call it *early*; if it is within 10 minutes (plus or minus) of the scheduled time, we will call it *on-time*; if it is more than 10 minutes late but up to 30 minutes late, we will call it *late*; anything beyond 30 minutes late, we will call *severely late*. There could be a discrepancy between the above BTS website and United, so if you are using BTS, you may want to calibrate.
- **Methodology:** While we do not want to prescribe a method, our only requirement is that it is **not** be based on neural networks or deep learning. Any approach using regression or classification that we did in class, as well as any statistical analysis including time series would be reasonable choices.
- **Deadlines and Expectations:** All are 11:59pm on
 1. Tuesday March 28, 2023: Email ngautam@syr.edu (with Cc to team and TAs) team information (names of team members)

2. Tuesday April 4, 2023: Email ngautam@syr.edu (with Cc to team and TAs) project status information (in one paragraph explain what you have done so far and what you have observed)
3. Tuesday April 11, 2023: Email ngautam@syr.edu (with Cc to team and TAs) initial prediction as a CSV file for April 12-15
4. Thursday April 20, 2023: Email ngautam@syr.edu (with Cc to team and TA) final prediction as a CSV file for April 21-24
5. Thursday May 4, 2023: Upload on blackboard the same final prediction as a CSV file by adding the actual values observed, PDF file of report (describing what was done in the project), and PDF file of the code with output. **Every student must submit all three files on Blackboard so we can assign a grade. It is not enough if one member of a team uploads it.**

Out of the 25 points for this project, each of the deadlines met are worth 1 point (0.5 points for late submission until the next deadline, and 0 points after that). Mark your calendars and instructions to be sure expectations and deadlines are not missed.

- **Report:** Your report must be well written so that you can use it to show what you did for the class, say to a prospective employer. You must explain in some detail so the graders can understand what you did without looking into the code. Also, you must cite sources of data and make references to any information you used (such as journal articles). Even though each team member submits a separate report, the reports for students in a team can be identical.
- **Grading:** Grading will be based on the creativity used in the approach taken and the justification provided in the report. Any obvious errors in terms of the code will be penalized. While it is important to get accurate predictions, that will not be the main focus. However, as an incentive, *we will take the winning team to lunch.*