Data Mining: 36-462/36-662
Final Project

Airplanes!

Deliverables and deadlines

Here are some key deliverables and deadlines upfront:

- Your predictions are due Saturday, May 6 at 11:59pm, submitted on blackboard.
- Your writeup is due Friday, May 12 at 5pm, submitted on blackboard.

Each team only has to submit one of everything. Hence you can read "you" throughout as "your team".

Introduction

Your final project concerns flight delays at the Pittsburgh airport. The data set is an extract from the Airline On-Time Performance Data made available through the Bureau of Transportation Statistics of the U.S. Department of Transportation. It reflects commercial flight activity to and from Pittsburgh International Airport (airport code PIT). For each flight departing from Pittsburgh, you are interested in predicting whether or not there will be a flight delay. This will be predicted only from information available before the departure of that flight.

Your project has three parts:

1. Data exploration and unsupervised learning.
2. Building and validation of a predictive algorithm for delays.
3. Actual submission to a prediction contest.

You will be given data on every flight into and out of Pittsburgh in 2015 by commercial airlines. Your focus will be on the departing flights, with the goal being prediction of departure delays. The arriving flights are just there in case you want to use them to build features. Remember to separate the two kinds of flights when cleaning your data.

More details

Download the file "flights2015.csv" from blackboard and load it into your R session. This is the data that you will use for model building. You are trying to predict the `DEP_DEL15` variable, a binary variable indicating whether the flight was delayed at least 15 minutes.

You can read about all the variable definitions at the following site (where the data are from):

`https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236`

(Note: Downloading the missing 2016 flights and using this to tune your model is considered cheating. Do not do this. It's more obvious than you think and will result in a zero if detected.)

You are allowed to gather additional data about 2015/2016 from around the internet and use it for feature creation: weather, plane information, etc. (Don't use things like news reports that would not be available by the day of the flight; this is supposed to be prediction after all. If you are unsure whether an information source is fair, just ask.)

The test data is in flights2016.csv. For each day in 2016, you receive data on flights up to some random point in time (in case you want to use information from earlier that day in your predictions). After that time, you will see no more flight information. Instead, you will be asked to provide predictions about delays for a small number of flights occuring after the cutoff.

The file `flights2016.csv` has all these flights (both the ones you see and the ones you must guess). `flights2016_visible.csv` has the portion of these that you see fully. `flights2016_guess.csv` has the portion you will need to guess on. (The first file is the union of the other two.)

For the test points, multiple fields will be blanked out (replaced with NAs). These are all the fields that are equivalent to knowing whether the departure was delayed (for example, the fields giving the reason or duration of the delay.). You only need to predict whether it was delayed at all: `DepDel15` variable (whether it was delayed 15 minutes or more).

The following fields are missing from the flights you must predict: `DEP_TIME, DEP_DELAY, DEP_DELAY_NEW, DEP_DEL15, DEP_DELAY_GROUP, DEP_TIME_BLK, TAXI_OUT, WHEELS_OFF, WHEELS_ON, TAXI_IN, ARR_TIME, ARR_DELAY, ARR_DELAY_NEW, ARR_DEL15, ARR_DELAY_GROUP, ARR_TIME_BLK, CANCELLED, CANCELLATION_CODE, DIVERTED, ACTUAL_ELAPSED_TIME, AIR_TIME, CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY, LATE_AIRCRAFT_DELAY, FIRST_DEP_TIME, TOTAL_ADD_GTIME, LONGEST_ADD_GTIME`

Making predictions

How can you make your predictions? You can use any of the techniques we have discussed in class. You can use any of the variables in the data set, and you can also consider constructing new variables by combining or transforming the variables that are present in the data set. Consider creating features using all of the fields provided, including information on other flights that day and even about arriving flights. All of this information is available in the test set, since you see all flights before the time cutoff.

You can also seek out other information on the internet (e.g., weather) to create features and improve your predictions. This is not required.

Submitting predictions

You will submit a single RData file with your predictions. This file should contain three variables: variables `delay.guesses`, `performance.guess`, and `teamname`. These should have:

- `delay.guesses`: A single vector of 0s and 1s, where 1 indicates a flight delay and 0 indicates no flight delay. This vector should be as long as the number of test cases in `flights2016_guess.csv` (and in the same order).
- `performance.guess`: A single number indicating your best guess at your performance on the new test set. This will give an idea how well your validation strategy has worked in setting your expectations.
- `team.name`: A string with your team's name.

To make this file, if you have the appropriate `delay.guesses`, `performance.guess` and `team.name` variables in your workspace, you can type

```
save(list=c("delay.guesses","performance.guess","team.name"),file="stat462final.RData")
```

This will create `stat462final.RData` file, which you can upload on blackboard. (Please rename the file to include your team name before sending.)

Write-up

You will submit a write-up, as a hard copy, on **Friday May 12 at 5pm**. This write-up should be a polished report, with figures and snippets of R code as you deem helpful. You don't need to submit your R code in its entirety. Your report should have the following sections (you can of course add subsections if you want), and should be no more than 8 pages.

Introduction: Describe your data set. What is the problem you are trying to solve? This can be quite brief.

Exploration and unsupervised analysis: Exploration of your data, including unsupervised analysis. You don't need to do the typical "exploratory data analysis" that you might do in 36-401, but you should provide proper motivation for your work and explain any insights and exploration that led to your features and models. This can include unsupervised approaches that we learned in the last part of the term if you detect any interesting structure with them. (If you don't find anything interesting, then just describe what you tried. You don't need to artificially manipulate the data to find something that's not there.)

Supervised analysis: How did you make your predictions? Describe this process in detail. Again, you can use any of the classification techniques that we learned in the first half of the

course, or any other techniques as long as they are adequately described. What predictor variables did you include? How did you engineer features from the data? What technique did you use for prediction, and why did you choose it? If there were tuning parameters, how did you pick their values? Can you explain anything about the nature of the relationship between the predictors in your model and the predictions themselves?

Evaluation

Your predictions will be evaluated against the true delays. You will be judged based on your misclassification rate.

The misclassification rates will be revealed on the scheduled "Exam" day (May 8). The top 5 undergraduate teams and top 3 graduate teams (with the lowest misclassification rates) will be given extra credit. These teams will be asked to describe their prediction approaches and what worked. (Comments from everyone else are also welcome!)

For the grading of the write-up, the exploration and unsupervised analysis will be worth roughly 33%, and the supervised analysis worth roughly 67%.

Cheating

Don't cheat. We know that there are ways to cheat on this final project. If we suspect you of cheating (e.g., if you have a remarkably low misclassification rate, but your method is not really statistically motivated), then we reserve the right to give you a 0.