Group Assignment #2: United Airlines

Background

The airline industry is characterized by very high fixed costs and relatively low marginal costs. For this reason, maximizing capacity utilization on flights is extremely important for the bottom line. One way to increase capacity utilization is to overbook flights, i.e. to sell more tickets than there are seats. However, if more passengers arrive than there are seats, the airline has to offer costly rewards to attract volunteers to give up their seats—or worse, forcibly bump passengers from the flight.

Thus, it would be extremely helpful to an airline to be able to figure out how many passengers on a given flight are likely to "no show." The airline can then sell this number of extra tickets; or it can sell slightly more/fewer, depending on how costly it is to obtain volunteers for a particular flight.

United Airlines has asked you to build a model for them to use to predict the number of no-shows on a flight. They have provided you with data on an O'Hare-Dulles flight for the month of May 2007 (Chicago to Washington, DC). The data are contained in *united_noshow_final.dta*, and a detailed description of the variables is at the bottom of this assignment description. You are free to look up additional information that pertains to these flights or the data, but this is *not* required for full credit, unless indicated explicitly.

Important Pointers

The difficulty of this assignment is a step up from earlier assignments this quarter, reflecting your higher skill level and greater comfort with coding. Use your group members as a resource.

Some important information to help you with the assignment:

- Choice of regressors: Your baseline model must in some way account for, at a minimum: the size of the party, the number of days between the booking and the scheduled departure, the country from which the booking was make, the booking class (economy, first class, etc.), and the connection status of the booking. Besides that, you have discretion in what is included in your model and how variables are defined, and you will be asked in question 3) to assess the impact of some of those discretionary choices.
- Note on categorical variables: Several of the categorical variables have only a handful of observations with some of the less common values. It is important to think about how this might affect your ability to estimate informative coefficients, and think about other ways you can summarize or include the relevant information in your model. Stata's *tab* command and R's *table()* function are useful for diagnosing this. These decisions are also related to whether a distinction between several values of a categorical variable is useful, or whether it would be more meaningful to aggregate them to a smaller number of possible values. You should explain how you are handling this issue in the annotated Stata log file or R notebook that you submit.
- Duplicate records: Sometimes bookings get inputted into the system more than once due to technical glitches, or due to the creation of duplicate records when a change is made to an existing booking. It is your responsibility as the analyst to make sure that error is not introduced

into your model by double-counting such records. Stata's *duplicates tag bookid* and *duplicates drop bookid* commands are useful for this (type *help duplicates* for details). In R, the *unique()* function is useful (type *?unique* for details). You should explain how you are handling this issue in the annotated Stata log file or R notebook that you submit.

- Prediction techniques: If you are familiar with the use of holdout data or training samples, you may use those techniques here. However, since those techniques were not covered in class, that is not required for full credit.
- Calculating total expected no-shows: The total number of expected no-shows on a given flight is equal to the sum of *each individual passenger's probability* of no-show. For example, if there are 10 passengers, each with a predicted probability of no-show of 0.2, then the total expected number of no-shows is 2. In Stata, a useful command for this is something like:

  bysort datevar: egen varname = total(predicted_indiv_prob(daystodep>=5))*

  (and then look at the value of *varname*). In R, the equivalent is something like:

  sum(dataframe$predicted_indiv_prob[dataframe$date=="xyz" & dataframe$daystodep≥5], na.rm=TRUE)

Assignment Questions

1. Build a model to predict no-shows 5 days in advance of the departure date (e.g. if the departure date is May 10, use bookings made up to and including May 5).
2. What is your model's implied effect of adding one additional person to a booking on the party's probability of no-show? (For this question specifically, you can ignore the p-values and just focus on the coefficients.) Report the effect at two sensible baseline probabilities (or values of the covariates) of your choice. Explain why you picked those baselines and interpret the result.
3. Is your answer to question 2) robust to your modeling choices? Pick two of your modeling choices that you think could reasonably have been decided differently and test those. Your model is **robust** if it produces a similar managerial conclusion when you make *sensible alternative modeling choices*, such as excluding some unessential variables, including vs. excluding outliers, changing functional form to another sensible choice, or collapsing categorical variables to fewer categories vs. using the most detailed available categories.
4. Go back to your baseline model from question 2). Predict the total number of expected no-shows on the following dates. Then briefly discuss which date's predictions you trust more.
   a. May 21, 2007
   b. May 28, 2007 (Memorial Day)
5. Now compare your predictions in question 4) to the simplest possible predictive model: assuming the mean no-show rate from the data applies to every observation. How much closer was your model to the actual outcome?

Data Description

Additional information about the original dataset (which has been cleaned by the professor before being given to you):

Flight #	Flight number
Departure Date	Departure date for the flight
Origin	Origin airport
Destination	Destination airport
Mileage Plus Status[1]	Passenger's Mileage Plus status
Assigned Seat Indicator	Assigned seat reservation indicator - whether a seat has been already assigned
Booking Class	Which fare class the passenger was booked in (it is recommended that you look at a summary of these classes, such as the one at https://blog.wandr.me/2011/10/decoding-fare-buckets-on-the-new-united/)
Number in Party	How many people were on this reservation
Segment Modification count	How many times was the itinerary modified
Booking Country	Which country was the itinerary booked from
DaysToDep	How many days before departure was the booking made
E-ticketed Ind (1 = eTicket, 0 = no)	Whether it was an e-ticket or not
Booking CRS Code	IATA codes for the airline from which the booking originated
Booking Agent Duty code	Booking agent's duty code
Upline Cnx	If there was an upline connection, then Y (i.e. if they flew into ORD from somewhere).
Downline Cnx	If there was a downline connection, then Y (i.e. if they had a flight out of IAD).
BookID	United's internal booking identifier code (1 value per booking).
No Show(1 noshow,0 show)	If the party showed up or not.

Mileage Plus categories

missing => No Mileage Plus
J ==> VP Premier Executive Member
K ==> Premier Executive (> 100,000 Miles)
A ==> Premier Executive (> 50,000 Miles)
B ==> Premier Member (> 25,000 Miles)
O ==> Premier Emeritus Member
M ==> Regular Mileage Plus Member
F ==> Any Premier Mileage Plus Member (Specific Category Unknown Due To Schedule Change)
G ==> Any Regular Mileage Plus Member (or Specific Category Unknown Due To Schedule Change)
S and Q are for star alliance members