# ETX2250/ETF5922 Data Visualisation and Analytics

Total of 100 points equating 15% of the assessment in ETX2250/ETF5922

## Business Case

In late 2013, the taxi company Yourcabs.com in Bangalore, India was facing a problem with the drivers using their platform-not all drivers were showing up for their scheduled calls. Drivers would cancel their acceptance of a call, and, if the cancellation did not occur with adequate notice, the customer would be delayed or even left high and dry.

Bangalore is a key tech center in India, and technology was transforming the taxi industry. Yourcabs.com featured an online booking system (though customers could phone in as well) and presented itself as a taxi booking portal. The Uber ride sharing service started its Bangalore operations in mid-2014.



Yourcabs.com had collected data on its bookings from 2011 to 2013, and posted a contest on Kaggle, in coordination with the Indian School of Business, to see what it could learn about the problem of cab cancellations.



The data presented for this case are a randomly selected subset of the original data, with 10,000 rows, one row for each booking. There are 17 input variables, including user (customer) ID, vehicle model, whether the booking was made online or via a mobile app, type of travel, type of booking package, geographic information, and the date and time of the scheduled trip. The target variable of interest is the binary indicator of whether a ride was canceled. The overall cancellation rate is between 7% and 8%.



**Assignment:**

1. How can a predictive model based on these data be used by Yourcabs? Describe how you would approach this problem in practice. 150 – 200 words **(10 points)**
2. Explore, prepare, and transform the data to facilitate predictive modeling. Provide the code of your final results how you created your predictor variables and print first 6 rows as a table. **(30 points)**

   Here are some hints:

   - In exploratory modeling, it is useful to move fairly soon to at least an initial model without solving *all* data preparation issues. One example is the GPS information-other geographic information is available, so you could defer the challenge of how to interpret/use the GPS information.

- How will you deal with missing data, such as cases where NULL is indicated?
- Think about what useful information might be held within the date and time fields (the booking timestamp and the trip timestamp).
- Think also about the categorical variables, and how to deal with them. Should we turn them all into dummies? Use only some?

3. Fit a classification tree: Does it provide information on how the predictor variables relate to cancellations? The data-set is imbalanced, make use of the upSample function to balance the classes. Prune the tree to not have more than 3-4 predictors. Provide the code and plot the pruned tree as a figure **(30 points)**. Describe the predictor variables, do they make sense? 150-200 words **(10 points)**
4. Report the predictive performance of your model in terms of error rates (the confusion matrix). How well does the model perform overall? Which plot do you choose? Can the model be used in practice? 100 words max **(10 points)**
5. Examine the predictive performance of your model in terms of ranking(lift). Which plot do you choose? How well does the model perform? Can the model be used in **practice?** 100 words max **(10 points)**



Submission

A pdf document should be submitted. The code should be nicely formatted and be in a fix size font, e.g "Courier New". There is no need to include code for loading libraries. Maximum length of 6 A4 pages.