**Proposal for Data Science Project**
**Deliverable 1: September 4, 2014**

*Akhil Chugh*
*Jasmine Touton*

**Project Title:** Predicting whether a cab booking will get cancelled

**Data Retrieved From:** Kaggle: Predicting cab booking cancellations for a cab company in Bangalore: https://inclass.kaggle.com/c/predicting-cab-booking-cancellations

**Potential Data Fields:**
**id** - booking ID
**user_id** - the ID of the customer (based on mobile number)
**vehicle_model_id** - vehicle model type.
**package_id** - type of package (1=4hrs & 40kms, 2=8hrs & 80kms, 3=6hrs & 60kms, 4= 10hrs & 100kms, 5=5hrs & 50kms, 6=3hrs & 30kms, 7=12hrs & 120kms)
**travel_type_id** - type of travel (1=long distance, 2= point to point, 3= hourly rental).
**from_area_id** - unique identifier of area. Applicable only for point-to-point travel and packages
**to_area_id** - unique identifier of area. Applicable only for point-to-point travel
**from_city_id** - unique identifier of city
**to_city_id** - unique identifier of city (only for intercity)
**from_date** - time stamp of requested trip start
**to_date** - time stamp of trip end
**online_booking** - if booking was done on desktop website
**mobile_site_booking** - if booking was done on mobile website
**booking_created** - time stamp of booking
**from_lat** - latitude of from area
**from_long** - longitude of from area
**to_lat** - latitude of to area
**to_long** - longitude of to area
**Car_Cancellation (available only in training data)** - whether the booking was cancelled (1) or not (0) due to unavailability of a car.
**Cost_of_error (available only in training data)** - the cost incurred if the booking is misclassified. For an un-cancelled booking, the cost of misclassification is 1. For a cancelled booking, the cost is a function of the cancellation time relative to the trip start time (see Evaluation Page).

**Business Problem Addressed:** Cab companies all have to endure cancellations from customers at the last minute. YourCabs.com, a cab company in Bangalore, India is trying to address booking cancellations by the cab company due to unavailability of cars. The challenge is that cancellations can occur very close to the trip start time, thereby causing passengers inconvenience.

**Use Scenario of Result:** We will analyze how distance, to and from locations, method of bookings all impact the number of car cancellations there are with the cab company. This will help us deduce the ability for the cab company to predict how many bookings they can afford to do knowing that a certain percentage will cancel. In turn, this will help alleviate the issue of booking cancellations due to unavailability of cars (overbookings).

**Data Instance and Useful Features:** The main data variables we will include in our instance are package_id, travel_type_id, to_city_id, from_city_id, Cost_of_error, and vehicle_model_type.

**Target Variable:** Our target variable we will use in our classification problem is Car Cancellation – whether the booking was cancelled or not due to unavailability of a car.

**Added Business Value:** The added business value is the ability for cab companies to improve on their customer service.

Sanghee Lim 9/4/2014 5:50 PM
**Comment [1]:** How is it defined? The number of cancellation in a day? In a week? In a month? Think carefully about what the level of analysis is.

Sanghee Lim 9/4/2014 5:53 PM
**Comment [2]:** A similar issue. Do you have information about the total number of cars available? How will you determine the shortage? Recall the case of credit card default. The model predicts customer default based on the known features (e.g. income, balance) of customers. What you may want to get here may not be the cancellation due to unavailability of a car.