Predicting Cab Booking Cancellations

CAPSTONE -1 Project Proposal - by Debisree Ray

The Problem: The business problem addressed here is to improve the customer service for a cab company. The problem of the cab company is the booking cancellations (by the company), due to unavailability of a car, which can occur very close to the trip start time, thereby causing passengers inconvenience. The goal is to create a predictive model for classifying new bookings as to whether they will eventually get cancelled due to car unavailability or not.

The Client: The cab company called YourCabs.com (a cab company in Bangalore, India). The problem is originally listed in Kaggle competition.

The Data: the data is taken from the following link: https://inclass.kaggle.com/c/predicting-cab-booking-cancellations/data**.** The variables in the data set are as follows:

id	Booking id
user_id	The ID of the customer (based on mobile no)
vehicle_model_id	Vehicle model type
package_id	Type of package (1=4 hrs & 40kms, 2=8 hrs & 80 kms, 3=6 hrs & 60 kms, 4= 10 hrs & 100kms, 5=5 hrs & 50 kms, 6=3 hrs & 30kms, 7=12 hrs & 120 kms)
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_cancelation (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	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

The Business problem and the Approach: The goal of the project is to create a predictive model for classifying new bookings as to whether they will eventually get cancelled due to car unavailability. This is a classification task that includes misclassification costs. The first step is the data wrangling and the pre processing. Next, is to apply a right ML model for the classification problem. The associated business problems which can be addressed are as follows:

1) What are the important factors which cause the car unavailability?
2) What is the busiest hour in a day (maximum demand)?
3) In which city the cab is the most popular (or which neighbourhood)?
4) What is the night time demand of the car?
5) How the demand changes in the weekends?

Deliverable: The deliverables will be the code(s) on GitHub in the form of Jupyter Notebooks, and a slide desk. This will include a report and I intend to write a blog explaining the code and the results.