Predicting cab booking cancellations

Capstone project-1: Milestone Report: Debisree Ray

Introduction:

The business problem addressed here is to improve the customer service for Bangalore (India) based cab company called YourCabs. The problem is that a certain percentage of booking gets canceled by the company due to the unavailability of a car, and the cancellations occur at a time when the trip is about to start. Therefore it causes passengers inconvenience and a bad reputation for the company. So, the challenge is to build a predictive model, which would classify the upcoming bookings as, if they would eventually get cancelled due to car unavailability, or not. So this is a classification problem.

The Data:

The Kaggle hosts the original problem and the dataset in their website as one of their competitions. Here, I downloaded the data from the Kaggle website. Following are the data fields in the dataset, which we are going to read in the Pandas data frame.

- 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. The cost of misclassifying an uncancelled booking as a cancelled booking (cost=1 unit). The cost associated with misclassifying a cancelled booking as uncancelled, This cost is a function of how close the cancellation occurs relative to the trip start time. The closer the trip, the higher the cost. Cancellations occurring less than 15 minutes prior to the trip start incur a fixed penalty of 100 units.

The questions of interest:

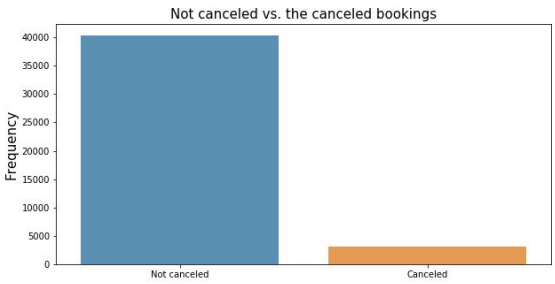
This data analysis and story-telling report is organized around the following questions of interest:

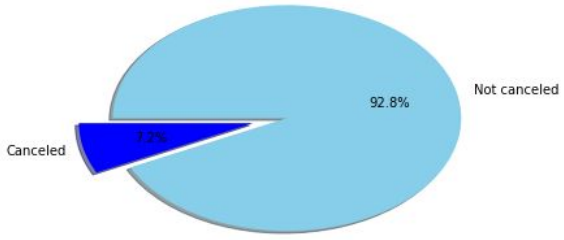
- How many unique users are out there? Are there any returning customers? Did they (returning customers) got their rides canceled?
- What are the different package IDs out there? Is there any relationship with the cancellations?
- What are the different travel types, vehicle IDs and mode of bookings (mobile/website/phone)? How are they related with the cancellations?
- Is there any connection between the drop-off location/city/area ID/latitude-longitude info and cancellations? What about the same with the pick-up locations/city/area IDs
- In which areas/neighborhoods, the cab service is the most popular?
- what is the busiest hour in a day? Does that have any connection with the cancellation?
- Which day of the week is the most popular in the cab users? Is there any connection between the day of the week with the cancellations?

Exploratory Data Analysis:

To start the EDA, here every different features have studied and visually displayed against the cancellations, so as to infer any relationship.

a) **Class Imbalance:** We see, there is a major class-imbalance in the data. Only ~ 7% (only 3132, in total 43,431) of the total bookings have been canceled.

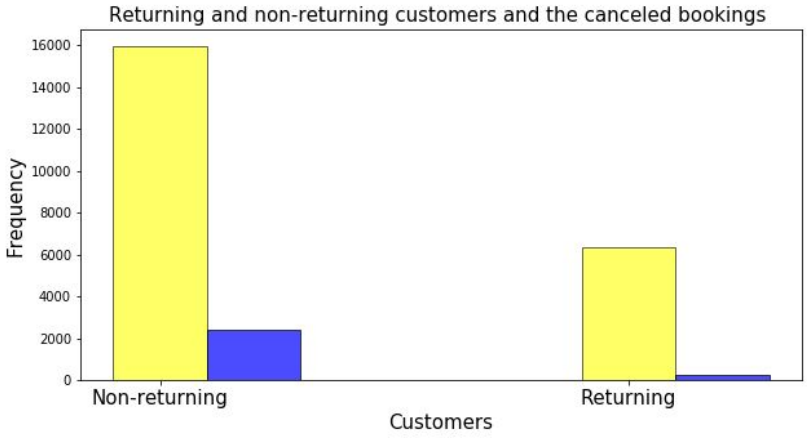




b) **User ID:** Each user has been assigned a unique ‘User ID’ in their booking information. Total 22267 user IDs have been recorded. We can see that,the user with the user_id '29648' is the most frequent user, with frequency 471. So, there are some ‘returning customers’ and, some are the one time users. The no. of one time users (non returning) are: 15935 and that of the returning are: 6332.

There are some gaps/missing data in the user ID column. The most frequent user (user_ID no '29648') got the maximum cancellations too, 55 times! The next most unfortunate user got his/her ride canceled 25 times and so on. To plot these, we have used the log scale so as to make the entire spectrum of the data clearly visible.

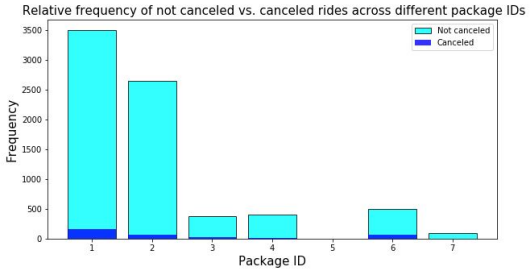
What we see is, that 1049 unfortunate returning customers got their rides canceled. Roughly 16.6% of the total returning customers got their rides canceled. So, 5283 returning customers did not undergo any booking cancellation experiences.

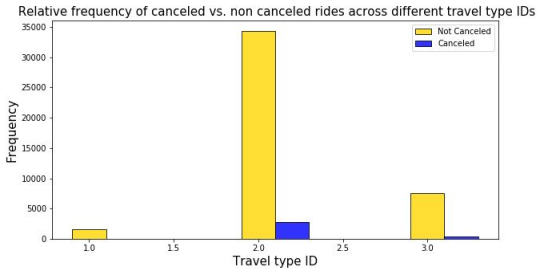


c) **Package ID:** Different package IDs are the different travel (booking) plans, from which customers can choose theirs. We are trying to evaluate if the 'package_id' has much effect on the cancellation or not. So, we have plotted the frequency of the canceled vs.not canceled rides across different package IDs.

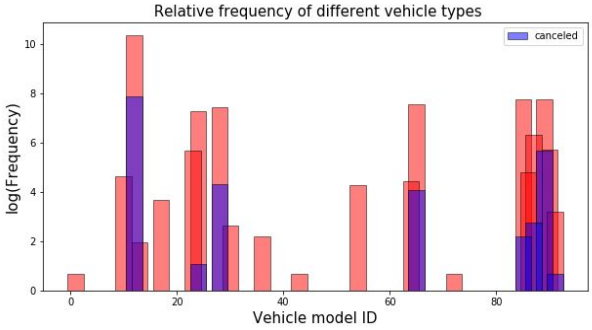
what we see from the plot below is that, people mostly opt for a journey of 4hrs and around 40kms, followed by 8hrs and 80kms. (The description of different package IDs have been given above, in the description of the fields.) And most of the times package_ID no 1 gets cancelled.

Travel type IDs are another feature of similar kind. Here there types of travel types are available to choose. And from the figure it’s evident that the travel type '2' (i.e. for point to point travel) is the most popular.



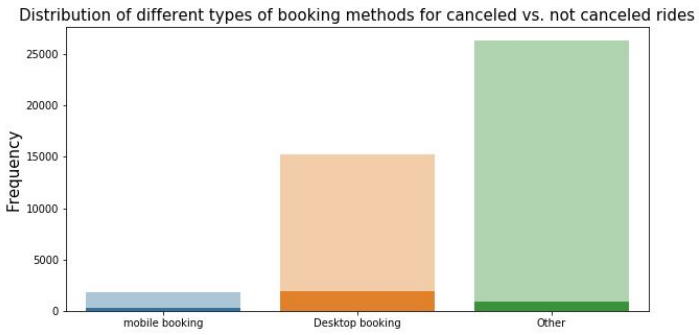


d) **Vehicle ID:** Another feature listed in the dataset is the 'vehicle ID'. 27 different types of vehicles have been listed. The most popular one is the vehicle with the vehicle ID no '12'. It has been used 31859 times. At the same time we see that the vehicle ID no '12' got the maximum number of cancellations (2668 times).
Here the Y-axis have been resized by using logarithmic operation, to get a clear picture of the entire data.

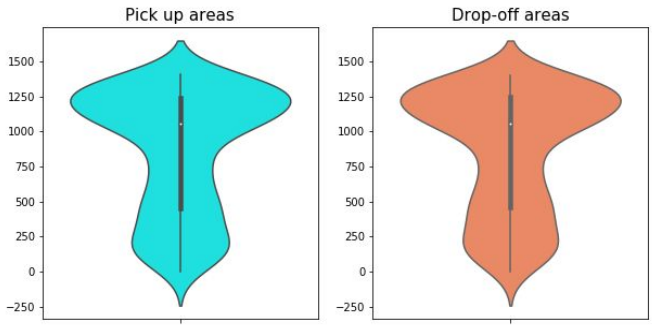


e)**Booking Methods:**There are three different types of 'Booking methods'. Only two types were listed such as, mobile booking and desktop/website booking. So, I concluded the remaining portion of the booking information as 'other method' of booking. We see that, 1878 bookings have been done from the mobile websites, 15270 bookings from the desktop websites, so, 26283 bookings have been done differently! (Total no. of bookings=43431) So, the other methods of booking are mostly popular. though, nothing has been stated about that.

In the same figure we have shown the same plot for the canceled bookings (with deeper shades). Interestingly, this time the maximum frequency of cancellations correspond to the bookings done from the desktop websites.



f) **Pick-up/Drop-off area ID:** There is another feature describing the drop-off and pick-up area IDs. 598 unique origin area and 568 destination area information have been listed. The most popular origin area is the area with area_id no. '393', which is eventually the most popular destination area as well. 559 area IDs are common to both as the pick-up and drop-off locations. The violin-plots show both the pick-up and drop-off area distributions.



This feature has some interesting connection with the cancellations. There are certain pick-up and drop-off areas for which more than 50% of the bookings were canceled. And some routes (from certain pick-up to certain drop-off areas), for which more than 90% of bookings were canceled.

from_area_id	
130.0	80.000000
1148.0	66.666667
1174.0	66.666667
630.0	66.666667
176.0	52.830189
1381.0	50.000000
1160.0	50.000000
1100.0	50.000000
1385.0	50.000000
1276.0	45.454545

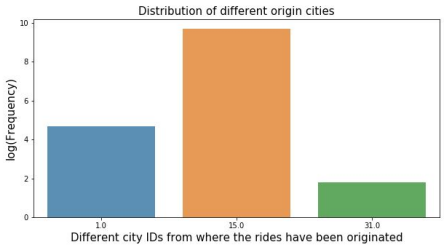
to_area_id	
1247.0	75.000000
677.0	66.666667
355.0	66.666667
1218.0	60.000000
845.0	60.000000
1310.0	50.000000
1311.0	50.000000
1387.0	50.000000
1197.0	50.000000
1225.0	50.000000

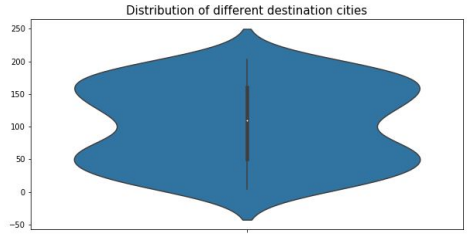
from_area_id	to_area_id	
626.0	122.0	90.909091
1349.0	1052.0	83.333333
1330.0	176.0	80.000000
1052.0	1349.0	78.571429
625.0	452.0	75.000000
1296.0	793.0	75.000000
1365.0	293.0	75.000000
1285.0	61.0	70.588235
122.0	626.0	70.000000
176.0	136.0	66.666667

g) **Pick-up/Drop-off city ID:** Another similar information have been listed in the feature called ‘city ID’. we can see that only 3 origin cities have been recorded. The most popular origin city is the city with the ID no: '15'. Where as, the destination cities are much distributed in numbers. 116 unique

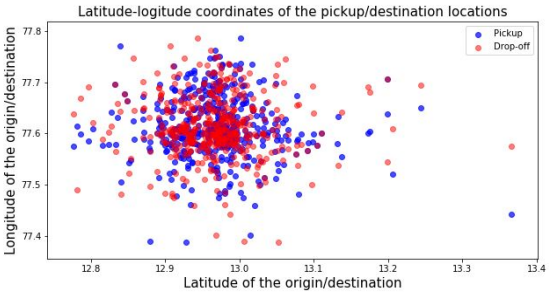
destination cities are there. The most popular destination city is the city with the ID no: '32' (475 rides have their destinations to this city.)

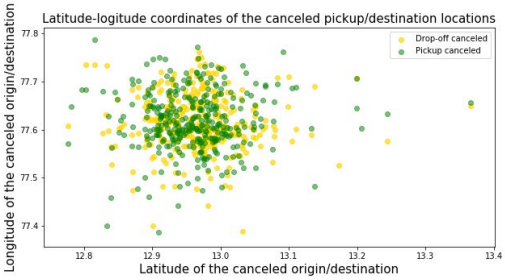
However, we need to remember that, only 16345 non-null values are available in 'from_city_id' information and 1588 non-null values are available in 'to_city_id' information. So, most of the information is missing.





h) **Latitude-longitude information:** Another GPS information about the pick-up and drop-off area locations are given in the form of latitude-longitude coordinates. again we see that, there are certain areas (latitude-longitude combination), for which the pick-up/drop-off cancellations are high.



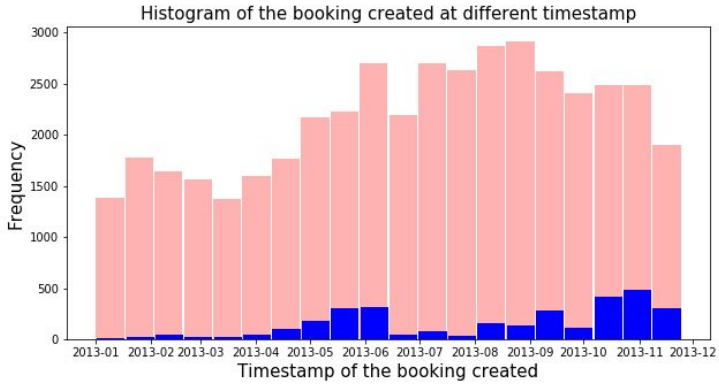


from_lat	from_long	
12.907300	77.695120	52.830189
12.796650	77.386930	50.000000
12.895314	77.461074	50.000000
12.891277	77.458760	50.000000
12.928360	77.683958	45.454545
13.049670	77.604790	44.444444
13.366072	77.683112	40.000000
12.833854	77.400913	40.000000
12.967500	77.608130	33.333333
12.958620	77.696020	33.333333

to_lat	to_long	
12.975390	77.548048	66.666667
12.980360	77.579940	60.000000
12.932229	77.690567	50.000000
12.956163	77.734160	50.000000
12.916941	77.589051	50.000000
12.980470	77.483730	50.000000
12.974967	77.614915	50.000000
12.988325	77.594263	50.000000
12.996820	77.604360	41.463415
12.934312	77.601508	40.000000

i) **Booking Timestamp:** There is one feature column, which records the timestamp of when the ride has been booked. The histogram shows the distribution of the frequencies of the time when the rides have been booked(lighter shade). At the same graph, we have plotted the same for the canceled

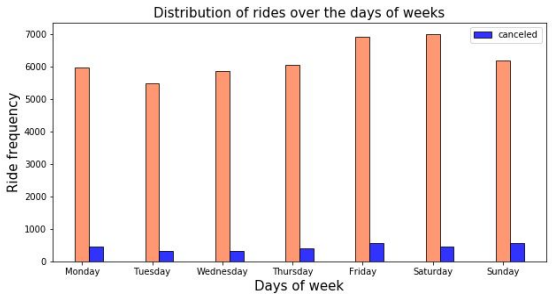
ones.(darker shade). Also, we see that, Maximum no. of bookings are made at a given timestamp:18. And the corresponding date-time is: 2013-10-31 10:30:00.

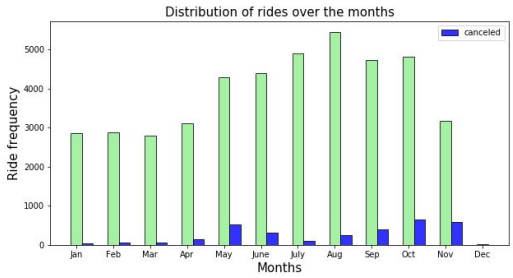


j) **Timestamp of the trip:** This is one of the most important features in the dataset, which can show some connection with the cancellation. Here we have extracted the ride frequency over the days of the week. It seems they are almost equally distributed. Here '0' is equivalent to 'Monday' and so on. We see that the maximum frequency (6990) of rides correspond to day no '5', ie Saturday', followed by the day no '4', which is the 'friday'. So, people book cabs more in the weekends. On the same figure we have plotted the canceled ride frequencies. And they seem to appear equally distributed over the days of the week. However, the maximum cancellations(578) correspond to the day no '4', which is 'friday', followed by the day no '6', which is 'sunday'.

Again, we have extracted the ride frequency over the months of the year. We see that the maximum frequency (5445) corresponds to the month no '8', which is the month of 'August', followed by 'July'.

On the same figure we have plotted the canceled ride frequencies. Maximum cancellation (650) correspond to the month no '10', which is 'October', followed by 'November'.

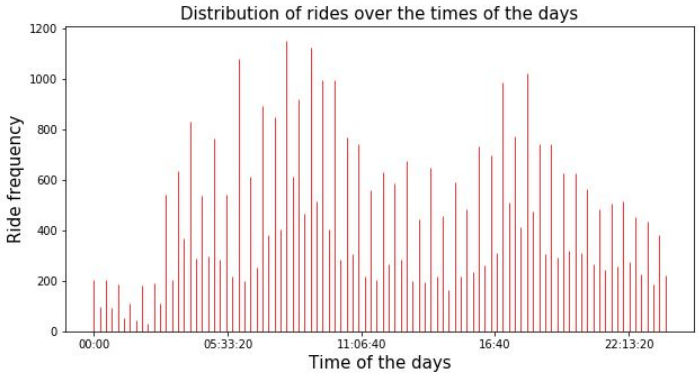


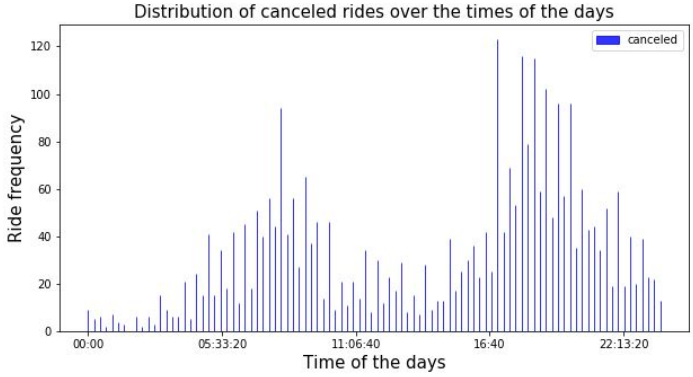


These are the frequencies of the rides across different times of the day. Clearly we can see the two humps/clusters in the distributions of the ride frequencies. So, what we see is that, the maximum rides are booked for two typical time stamps in a given day. One is around the morning and another

for the evening time. Clearly these two are the busiest hours, or mostly what we call as the 'office time' rush' in a day.

The ride cancellation distribution also follows the same trend. Maximum numbers of rides got canceled in these two peak hours. As obvious, these are the times, when rides can get canceled due to unavailability of cars.





k) **<u>'Time difference' (in hours) :</u>** This is a feature created, by taking the difference of the timestamps between the 'booking created' and the 'trip start time', to explore if that has any connection with the cancellations or not. However, looking at the following patterns, it seems that the feature has not much correlation with the cancellations.