FInal Project 2-Devesh Last Checkpoint: 25 minutes ago (autosaved)

Python 2

No kernel

* [File](http://localhost:8889/notebooks/FInal%20Project%202-Devesh.ipynb)
* [Edit](http://localhost:8889/notebooks/FInal%20Project%202-Devesh.ipynb)
* [View](http://localhost:8889/notebooks/FInal%20Project%202-Devesh.ipynb)
* [Insert](http://localhost:8889/notebooks/FInal%20Project%202-Devesh.ipynb)
* [Cell](http://localhost:8889/notebooks/FInal%20Project%202-Devesh.ipynb)
* [Kernel](http://localhost:8889/notebooks/FInal%20Project%202-Devesh.ipynb)
* [Help](http://localhost:8889/notebooks/FInal%20Project%202-Devesh.ipynb)

CodeMarkdownRaw NBConvertHeading-

CellToolbar

Edit Metadata

**## What's the project about? What problem are you solving?**

​

Edit Metadata

The business problem tackled here is trying to improve customer service for YourCabs.com, a cab company in Bangalore. The problem of interest is booking cancellations by the customer. The challenge is that cancellations can occur very close to the trip start time, thereby wasting the resources of the company

​

Edit Metadata

**- Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?**

In [1]:

Edit Metadata

The goal **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

​

File "<ipython-input-1-f73d29a26180>", line 1

The goal is to create a predictive model for classifying new bookings as to whether they will eventually get cancelled due to car unavailability.

^

SyntaxError: invalid syntax

Edit Metadata

**- What kind of impact do you think it could have?**

In [ ]:

Edit Metadata

The problem describes the scenario when the booking **is** cancelled by the customer .

​

Using the predictive approach the company :

1)could get new insights into customer satisfaction

2)Depending on the accuracy of the model the company can save money by calling customer **and** checking **if** they still need the cab

​

Edit Metadata

**- What do you think will have the most impact in predicting the value you are interested in solving for?**

In [ ]:

Edit Metadata

looking at the dataset the following features could most likely have more weights than the others:

1)from\_Area**/**To\_Area**-**This could reveal new insights into the areas **from** where the bookings are taken **and** are eventually cancelled

2)Vehicle\_model\_Id**-** If there **is** a particular model **for** which the bookings are cancelled

In addtion the following could give great insights:

1)User id**-** If the booking **for** a single user gets cancelled multiple times, then the company could lose that customer forever

​

Edit Metadata

**Datasets**

Edit Metadata

**\*\*Variable\*\***|**\*\*Description\*\***|**\*\*Type\*\***

---|---|---|

id |booking ID|integer

user\_id|the ID of the customer|unique integer(based on mobile number)

vehicle\_model\_id|vehicle model type|Integer

package\_id|type of package(A range that depicts the no of hours the vehicle is hired|range(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|Depcicts the city

to\_city\_id|unique identifier of city|Applicable for intercity travel

from\_date|time stamp of requested trip start|Data field

to\_date|time stamp of trip end|Data field

online\_booking|if booking was done on desktop website|Boolean

mobile\_site\_booking|if booking was done on mobile website|Boolean

booking\_created|time stamp of booking|Date field

from\_lat|latitude of from area|Float

to\_lat|latitude of to area|Float

to\_long|longitude of to area|Float

Car\_Cancellation|Response variable-indicating whether the booking was cancelled|Boolean

Cost\_of\_error|the cost incurred if the booking is misclassified.For an un-cancelled booking, the cost of misclassificaiton is 1. For a cancelled booking, the cost is a function of the cancellation time relative to the trip start time |Integer

​

Edit Metadata

**Domain knowledge**

Edit Metadata

**- What experience do you already have around this area?**

In [ ]:

Edit Metadata

The domain seems to be closely related to the average users who use the ride sharing services of the companies like UBER etc.

Edit Metadata

**Does it relate or help inform the project in any way?**

In [ ]:

Edit Metadata

Domain knowledge definitely helps **in** understanding the data sets **and** the business problem that we are trying to solve

Edit Metadata

**- What other research efforts exist?**

In [ ]:

Edit Metadata

This problem was published on Kaggle **in** 2013(https:**//**inclass.kaggle.com**/**c**/**predicting**-**cab**-**booking**-**cancellations)

**and** a few submissions exist on the internet **and** the findings.

The benchmarks exist on the below URLS where the other teams have submitted the solutions using multiple models **and** regression approaches.

​

http:**//**www.slideshare.net**/**rajzeest**/**predicting**-**cab**-**booking**-**cancellations**-**data**-**mining**-**project

http:**//**www.galitshmueli.com**/**sites**/**galitshmueli.com**/**files**/**BADM\_Group**%**20A9**%**20Final**%**20PPT.pdf

Edit Metadata

**Project Concerns**

Edit Metadata

**- What questions do you have about your project? What are you not sure you quite yet understand?**

Edit Metadata

1)Datasets-Not 100% sure on which columns to retain and which to drop

2)How to divide ths training set into Training and validation

3)Kind of problem- Although Kaggle says that this is a classification problem ,a few users have used regression to predict the cost of the error

4)How many approaches and models have to be used on the problem, it seems that some of the submissions exist with KNN, Naive Bais,Simple regression

5)Can a simple logistic regression be used on this problem

6)this problem also talks about a Cost Function- The cost of missclassifying a booking. Can we assume a cost function?

7)Some of the users on Kaggle have mentioned mentioned about the skewedness of the data i.e a very few bookings have been categorized as a classification error. Is that an issue in training the data

8)What is the rationale behind including the timestamp on the booking data. Is that relevant?

9)There are close to 43K rows in the Training dataset. Is that considered to be high? If yes? How can we decide how many rows should be included in the training data

10)Latitude and Longitude data is up to 4 places of decimal ? will it have an impact on the model perf and speed.

11)Is there a need to create binary variables

Edit Metadata

**- What are the assumptions and caveats to the problem?**

Edit Metadata

1)There is a cost associated with a missclassified booking 2)There are ways to mitigate these costs 3)The formula to calculate these cost is error free 4)There are other means available to offset the misclassified costs 5)It is assumed that the company has the necessary infrastructure to capture the features of the model like online\_booking,mobile\_site\_booking,latitude,longitude

Edit Metadata

**What data do you not have access to but wish you had?**

Edit Metadata

I wish that the cost function was unambiguous and data evenly distributed

Edit Metadata

**What is already implied about the observations in your data set?**

Edit Metadata

The Data set has been published on Kaggle and has been complied by IDRC- a data sciences consultancy focused on Quantitative Modeling, Data Analytics, Scientific Computing, and Data Visualization/Infographics.

Edit Metadata

**- What are the risks to the project? Or the Benefits of the Model being Right**

Edit Metadata

A high accuracy could provide a predictive approach to predict a cancelled booking , which could cost the company a lot of direct and indirect cost in terms of labor,fuel and other missed booking that could have been served .

Edit Metadata

**- Is any of the data incorrect? Could it be incorrect?**

Edit Metadata

The Data is collected by a consultancy firm providing services in the areas os Data science and also the problem is published on Kaggle , so it is assumed that the data is correct.

Edit Metadata

**# Outcomes**

Edit Metadata

**## What do you expect the output to look like?##**

Edit Metadata

The output should contain a prediction i.e. car\_calcellation to classify if a booking is being cancelled or not and associated cost for misclassified booking

Edit Metadata

**## - What does your target audience expect the output to look like?**

Edit Metadata

Output should contain two columns: id and Car\_Cancellation. Car\_Cancellation is a columns with classifications, which includes only values of 0 (not cancelled) or 1 (cancelled).

In Ideal scenarios users would also like to look at the cots associated with the misclassfied booking.

​

​

Edit Metadata

**## - What gain do you expect from your most important feature on its own?**

​

Edit Metadata

SOme of the most important features could be :

1)from\_area- The company could be able to predict the areas that is resulting in the most cancellations

2)to\_area- The destinations to which the bookings are getting cancelled

3)Vehicle Id- The kind of vehicles for which the bookings are cancelled, with this insight they can decide to keep the fleet as per the business model

4)package id- The package or the booking type that gets cancelled and decide to revise the offerings

​

​

Edit Metadata

**## How complicated does your model have to be?**

Edit Metadata

Not sure at this point in time

Edit Metadata

**## - How successful does your project have to be in order to be considered a "success"?**

​

Edit Metadata

approx 90% of the bookings should be correctly classified.

NOTE- 90% is an estimate and need to check with the instructional team on how to measure the success