# Customer Service Improvement for YourCab.com

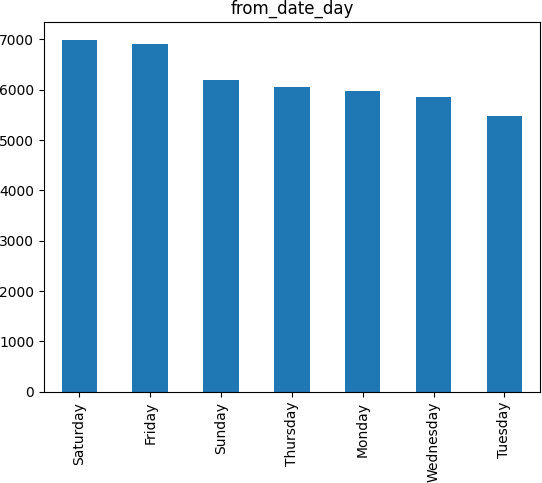
[**Advait**](mailto:chetanwanave@gmail.com) **Ghatge**

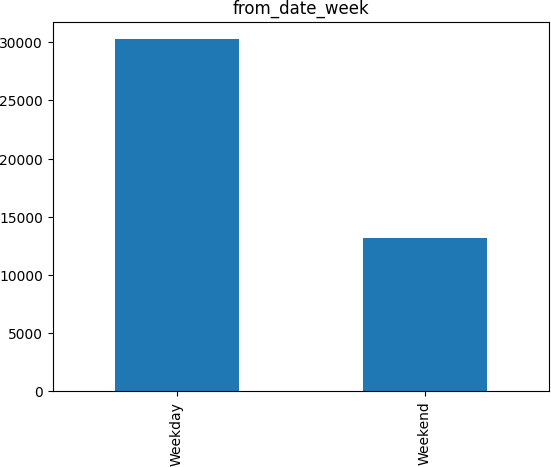
### Dataset used:

### <https://www.kaggle.com/datasets/akashkumar01/yourcabs>

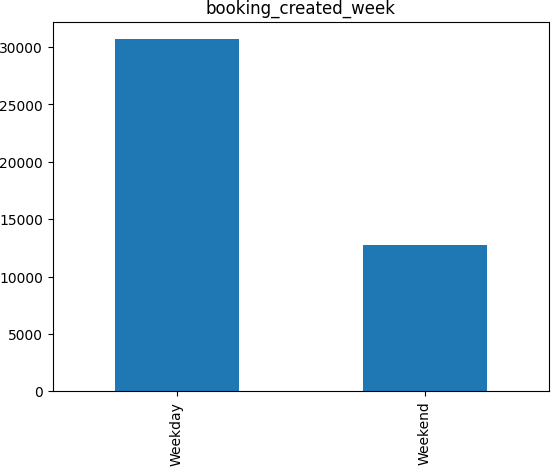
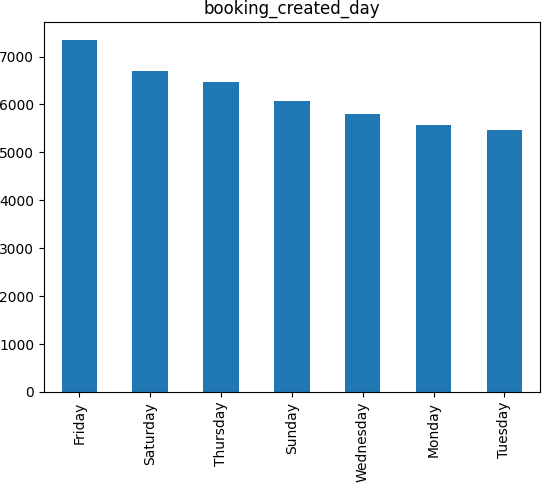


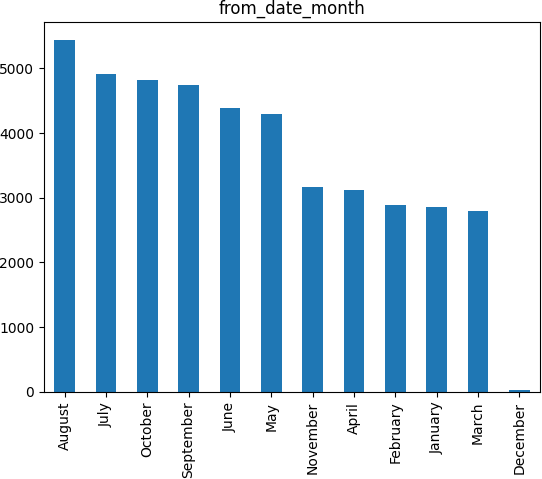
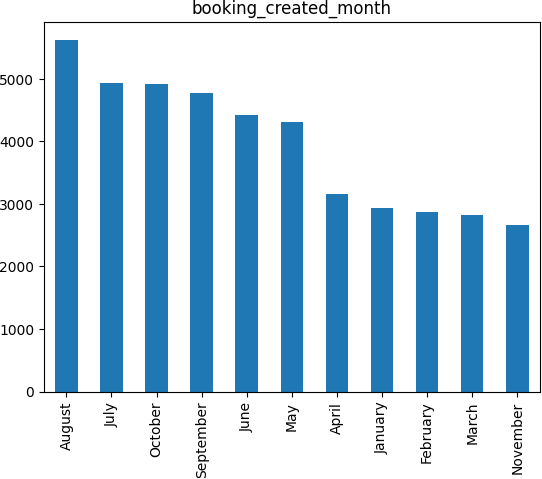
### Visualization

Data visualization is used to make complex data easier to understand, identify relationships and correlations, and communicate insights and findings to others. It also makes data more engaging, which can encourage people to explore it further. Finally, data visualization supports decision-making by providing a clear, visual representation of the data that can help identify trends and patterns that might be missed in other forms of analysis.

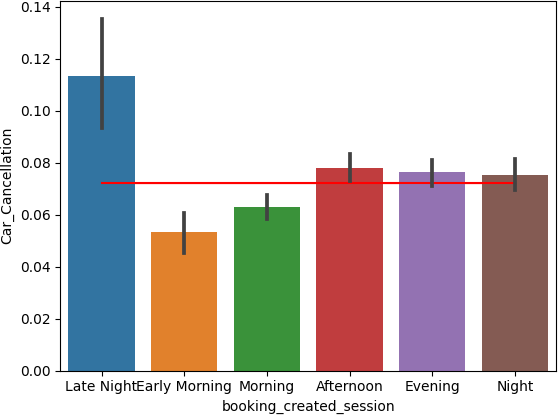


As we can see from the above graph, there are more cancellations on the weekend than there are on weekdays.



We can see that fewer people book cabs on Sundays and yet they cancel their rides more often.

Here, we observe that most people cancel and book their cabs in the monsoon season.



Here, we observe that most people cancel their rides late into the night.

### Splitting of Data

The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications. This method is a fast and easy procedure to perform such that we can compare our own machine learning model results to machine results. By default, the Test set is split into 30 % of actual data and the training set is split into 70% of the actual data. We need to split a dataset into train and test sets to evaluate how well our machine learning model performs. The train set is used to fit the model, and the statistics of the train set are known. The second set is called the test data set, this set is solely used for predictions.



### Model Building

The Machine Learning models can be understood as a program that has been trained to find patterns within new data and make predictions. These models are represented as a mathematical function that takes requests in the form of input data, makes predictions on input data, and then provides an output in response. First, these models are trained over a set of data, and then they are provided an algorithm to reason over data, extract the pattern from feed data and learn from those data. Once these models get trained, they can be used to predict the unseen dataset.

### Logistic Regression

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. It’s referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

**Terminologies involved in Logistic Regression:**

Here are some common terms involved in logistic regression:

**Independent variables**: The input characteristics or predictor factors applied to the dependent variable’s predictions.

**Dependent variable**: The target variable in a logistic regression model, which we are trying to predict.

**Logistic function**: The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.

**Odds**: It is the ratio of something occurring to something not occurring. It is different from probability as probability is the ratio of something occurring to everything that could possibly occur.

**Log-odds**: The log-odds, also known as the logit function, is the natural logarithm of the odds. In logistic regression, the log odds of the dependent variable are modeled as a linear combination of the independent variables and the intercept.

**Coefficient**: The logistic regression model’s estimated parameters, show how the independent and dependent variables relate to one another.

**Intercept**: A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.

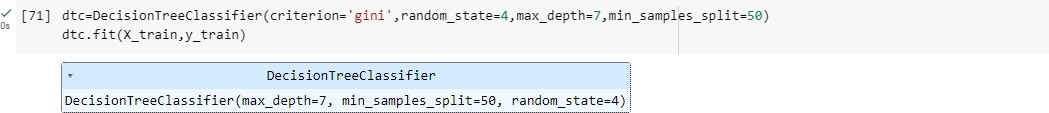
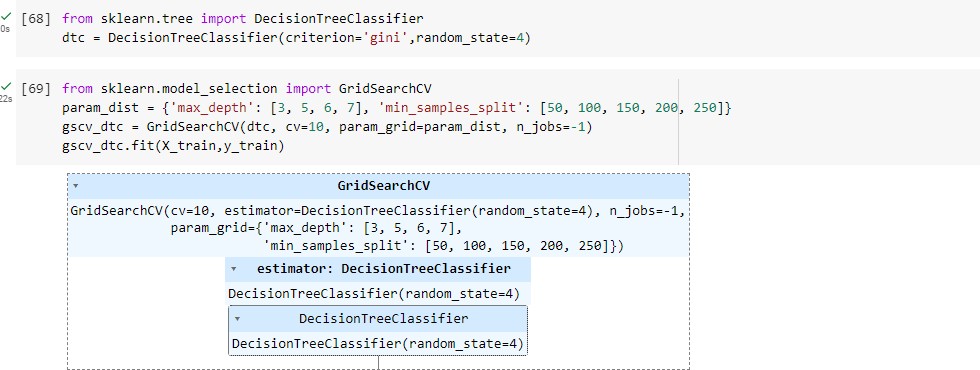
**Maximum likelihood estimation**: The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

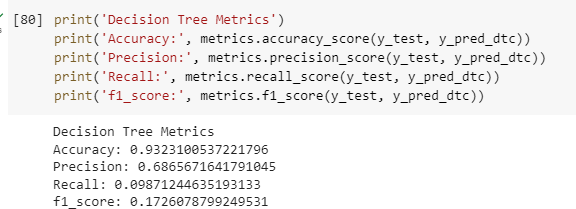
### 

### Evaluation:

### Decision Tree

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees.

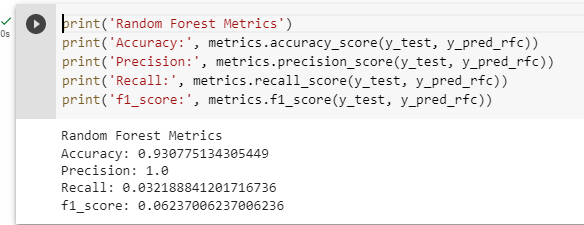


**Evaluation:**

### Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

**Evaluation:**



# Conclusions

## 1.

[**Advait**](mailto:chetanwanave@gmail.com) **Ghatge**

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 company due to unavailability of a car. The challenge is that cancellations can occur very close to the trip start time, thereby causing passengers inconvenience.

**Content**

The goal of is to create a predictive model for classifying new bookings as to whether they will eventually gets cancelled due to car unavailability by the business.

Note: cancellations mentioned are cancellations made by business. Columns details:

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 misclassificaiton is 1. For a cancelled booking, the cost is a function of the cancellation time relative to the trip start time.

\*\*\*

## Github Links

[**https://github.com/Chetan-Wanave/Feynn-MSA**](https://github.com/Chetan-Wanave/Feynn-MSA)[**https://github.com/khushipathak3502/Feynn\_labs**](https://github.com/khushipathak3502/Feynn_labs)[**https://github.com/prateektanwar2/Feynn\_Labs\_Internship/tree/main/Project%202.2**](https://github.com/prateektanwar2/Feynn_Labs_Internship/tree/main/Project%202.2)[**GitHub - ladijeevansai/Online-vehicle-booking-market-segmentation**](https://github.com/ladijeevansai/Online-vehicle-booking-market-segmentation)[**https://github.com/Soumyadwip/Online\_Car\_Market\_Segmentation**](https://github.com/Soumyadwip/Online_Car_Market_Segmentation)

\*\*\*