JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Prediction of Cab Cancellation

**N. Rishitha[1], K. Dhanush[2], P. Vinay Reddy[3]**

[1](Computer Science and Engineering, Anurag Group of institutions, India)[2](Computer Science and Engineering, Anurag Group of institutions, India)[3](Computer Science and Engineering, Anurag Group of institutions, India)

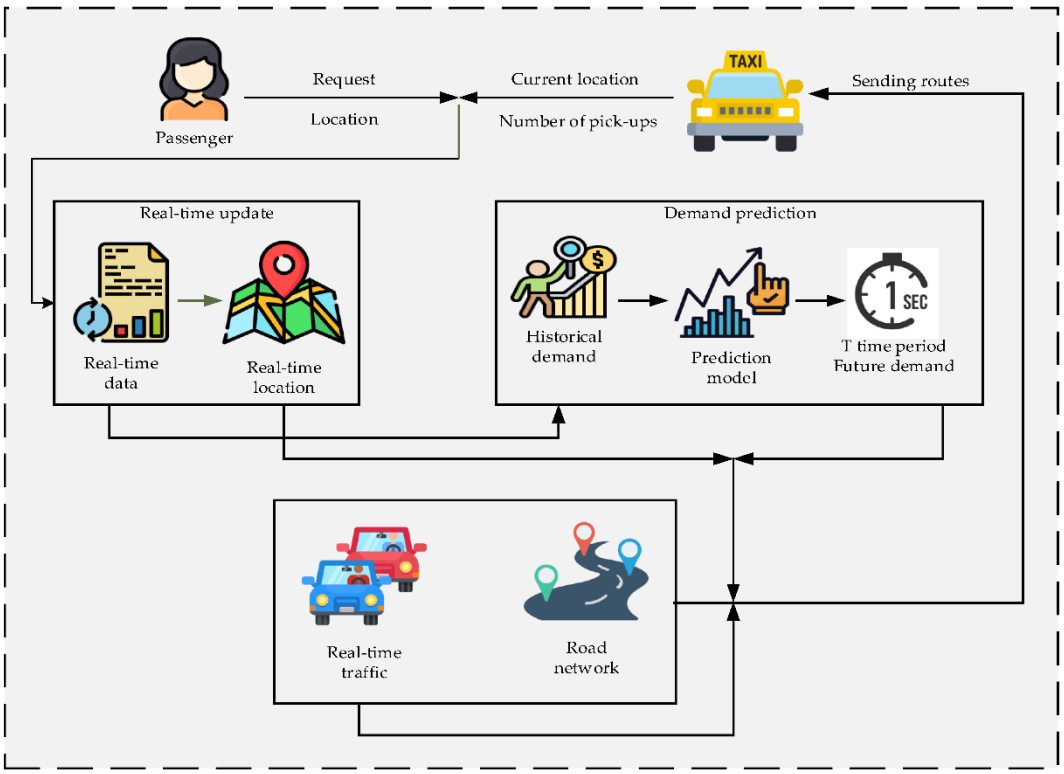
**Abstract:** Cab booking cancellation have been one of the most common issues being faced in transport services. To understand the actual reason for the cancellation, it's important to understand the behaviour of driver and the customer by analysing their activity patterns from existing data which contains the ride history of his previous bookings, pickup location, drop location, distance to pick- up, and ride acceptance rate. Using this existing data and BI tools or data representation languages like python or R we can predict the cab cancellation rate, which helps in understanding driver's behaviour to prevent further cancellations and match drivers to customers where the probability of a successful ride is high.

**Key Word**: Python, R, Cab Cancellation

## I. Introduction

There were numerous occasions of not getting cab or being cancelled and many reasons when booked it from many services and felt that may be some kind of functionality which tells us about the cab cancellation rate so that we may not wait for the cab which may get cancelled or due to no cabs availability. So, we will try to predict possible cancellations made by the drivers. By predicting possible cancellations an hour before the pick-up time, cab companies will be better able to manage their vendors and driver's up-to-date information about customer cancellations and reducing the cost incurred from sending a cab to a booking location that has been cancelled by the driver. Accurate prediction of customer cancellations will lead to a reduction in company cost and also improves the customer experience by reducing cab cancellations which saves much of customer's time. There is a booking option available to book in advance but for some people who want to travel urgently and not getting a cab on time, will go through hell and face a lot of struggle. So, from many incidents that happened, it is evident that this functionality in all leading service providers will be a game changer and helps many people across the world who depend on this kind of transportation.

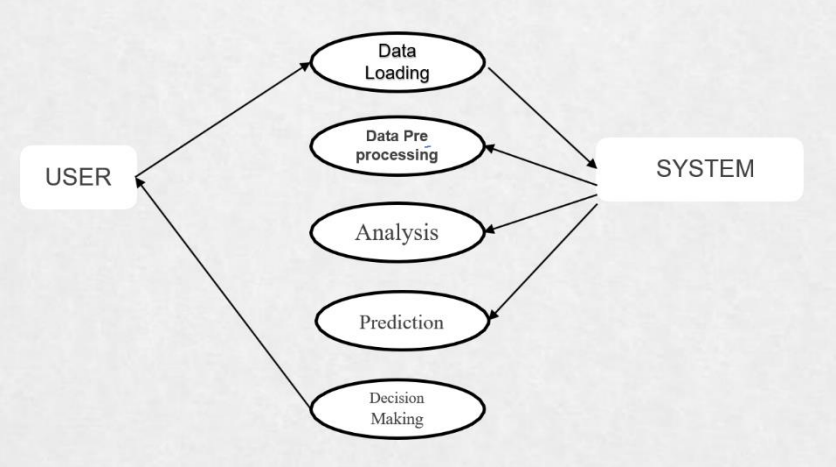
## II. Methodology



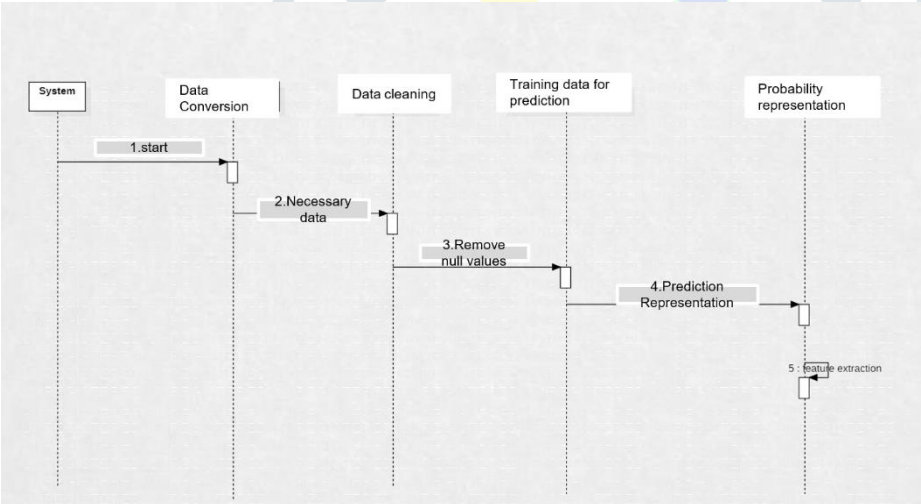

1. Visualize the data to understand the categories of each attribute and their influence on the dependent variable.
2. Data pre-processing (handling the imbalances in the dataset, feature selection etc.).
3. Build the classification models (logistic regression, decision tree algorithms, random forest) and find the best model.
4. Compare and evaluate the AUC for all models.
5. Explaining the influence of each independent variable for the target variable.

# Interpretation

- Cab Cancellation Data Cleaning ---> Proper prediction ---> lower 'cab_cancellation'--> lower 'cost_of_error'.
- The cab company can reduce the loss by correctly cancelling the booking which were supposed to be cancelled.
- The firm could do further analysis by back tracking the 'cab_ cancellation' to see the corresponding 'user_id' -there -by-figuring out how prediction is happening.




| User | Application | Predictor |
|---|---|---|
| • Loading Datasets()<br>• Predicting results<br>• Decision Making | • Data Preprocessing<br>• Evaluate training set<br>• Validate test set<br>• Representation of Predicted result | • Accuracy<br>• Predict<br>• Result |



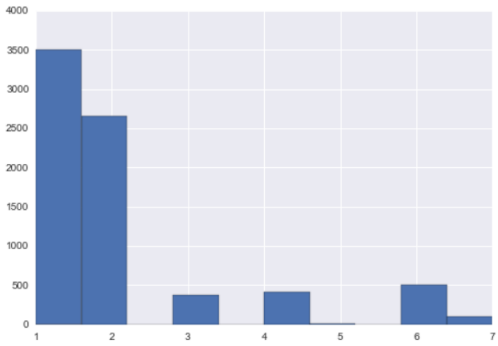
# III. Result

**Major class imbalance, very few cancellations as compared to large amount of non-cancellations.**

In [10]:
```
# Lets see the distribution of package_id
cars_cancel_train.package_id.value_counts()
```

Out[10]:
```
1    3503
2    2651
6     502
4     412
3     375
7     101
5       6
dtype: int64
```

In [11]:
```
cars_cancel_train.package_id.hist();
```




**Most of the packages that people opt for are for a journey of 4hrs and around 40kms, followed by 8hrs and 80kms.**

**Fig3.1:** Exploratory analysis distance and time

Out[15]:

| Car_Cancellation | 0 | 1 |
|---|---|---|
| from_area_id | | |
| 2 | 27 | 4 |
| 6 | 7 | NaN |
| 15 | 6 | NaN |
| 16 | 5 | NaN |
| 17 | 2 | NaN |

In [16]:
```
g['percent_cancelled'] = (g[1] / (g[1] + g[0])) * 100.
```

In [17]:
```
g.percent_cancelled.order(ascending=False).iloc[:20]
```

Out[17]:
```
from_area_id
130     80.000000
1148    66.666667
1174    66.666667
630     66.666667
176     52.830189
1381    50.000000
1160    50.000000
1100    50.000000
1385    50.000000
1276    45.454545
211     44.444444
1372    40.000000
356     40.000000
987     40.000000
626     34.375000
1258    33.333333
34      33.333333
326     33.333333
177     33.333333
833     33.333333
Name: percent_cancelled, dtype: float64
```

**So as you can see there are certain areas ( from area ) for which more than 50% of the bookings were cancelled.**

**Fig3.2:** Exploratory analysis on area

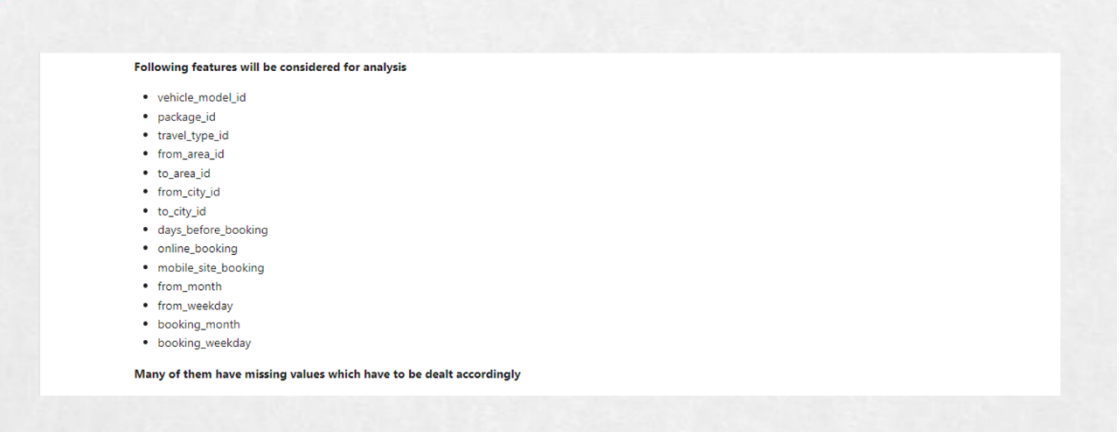



**Fig3.3:** Data preparation

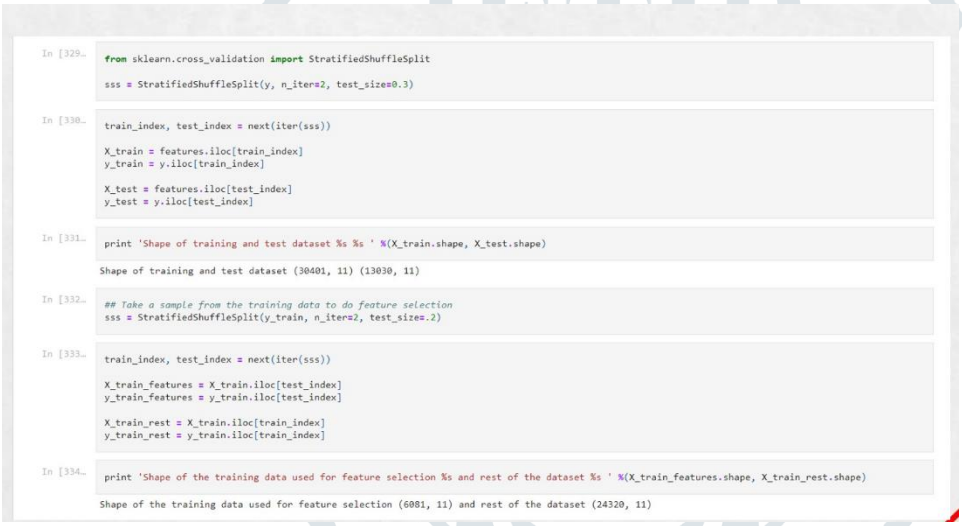



**Fig3.4**: Feature extraction

```
In [290…  from sklearn.learning_curve import validation_curve

In [294…  C = np.linspace(0.1, 50, 5)
          val_train, val_test = validation_curve(LogisticRegression(penalty='l1', class_weight='auto'),
                                                 Xtr, ytr, 'C', C, cv=5, scoring='roc_auc')

In [296…  def plot_with_err(x, data, **kwargs):
              mu, std = data.mean(1), data.std(1)
              lines = plt.plot(x, mu, '-', **kwargs)
              plt.fill_between(x, mu - std, mu + std, edgecolor='none',
                               facecolor=lines[0].get_color(), alpha=0.2)

          plot_with_err(C, val_train, label='training scores')
          plot_with_err(C, val_test, label='validation scores')
          plt.xlabel('C'); plt.ylabel('ROC AUC score')
          plt.legend();
```

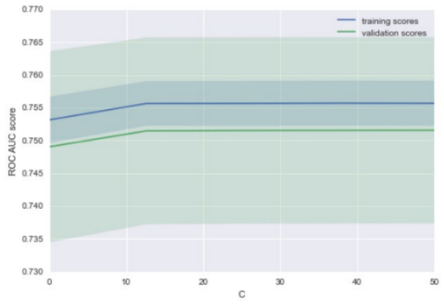



**Fig3.5:** Validation

```
In [411…  # knn.fit(Xtr, ytr)
          # logreg.fit(Xtr, ytr)
          gbc.fit(Xtr, ytr)
          # rf.fit(Xtr, ytr)

Out[411…  GradientBoostingClassifier(init=None, learning_rate=0.1, loss='deviance',
                        max_depth=6, max_features=None, max_leaf_nodes=None,
                        min_samples_leaf=1, min_samples_split=2,
                        min_weight_fraction_leaf=0.0, n_estimators=200,
                        random_state=None, subsample=0.8, verbose=0,
                        warm_start=False)

In [436…  # ytest_pred_knn = knn.predict(Xv)
          # ytest_pred_logreg = logreg.predict_proba(Xv)[:, 1]
          ytest_pred_gbc = gbc.predict_proba(Xv)[:, 1]
          ytest_preds_threshold = [1 if pred > 0.10 else 0 for pred in ytest_pred_gbc]
          # ytest_pred_rf = rf.predict(Xv)

In [437…  from sklearn.metrics import roc_auc_score

In [438…  print 'ROC AUC score on the test dataset ', roc_auc_score(yv, ytest_preds_threshold)

          ROC AUC score on the test dataset  0.746771205593
```

**Fig3.6:** Performance on test- set


**Fig3.7:** Model- training- on- full- dataset

## IV. Conclusion

Our model uses existing infrastructure, under utilized drivers and the existing data of cab cancellations. Implementation of our model into the daily operations would be simple and minimal amounts of training. The results could be easily confirmed via data and the cost of trial would be minimal and have no negative effects on the company or for the customer. It built good relation between the customer and the driver and saves the time of customer. In addition, we believe this model could be further improved and the savings increase in line with natural growth of the company.

### References

[1]. Abhishek sharma , Cab Booking Analysis, 2016- Kaggle activity

[2]. Data hosted by Kaggle

[3]. https://www.python.org/

[4]. https://github.com/

[5].https://scikit-learn.org/stable/