Prediction of Cab Cancellation

1st Komal Kumari
MT19124
IIITD
Delhi, India
komal19124@iiitd.ac.in

***Abstract*—Communication is one of the most vital aspect in today world.With modernization there has been significant advancement in making communication as easy as possible.To ease travelling in one very important aspect of communication improvisation.To aid in this direction, birth of several cab service offering startup or ventures have really made our world lot easier for commuting from one location to the other.Their service has allowed people to spread out their wing and fly to any place they want.To aid more to this , their service booking online or over mobile phone has made things a lot more easier for the people.However, even though their service continues to make things easy for us, but the same also has an associated bane. The bane is in regards cab cancellation after accepting the booking by the cab providers. This lead to quite a huge inconvenience on the part of the customer.The report is based on how can predicting cab cancellation at an early time can help customer as well as the cab providers to plan their steps accordingly.**

I. Introduction

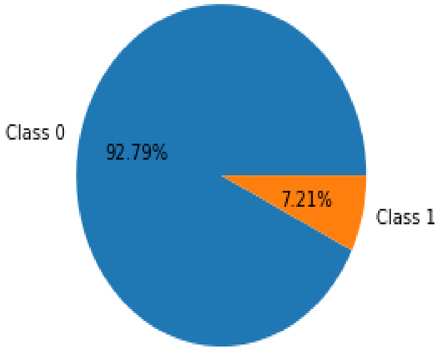
Riders have to go through a lot of inconvenience when their booking is cancelled by the cab providers.Things go worse when the cancellation happens close to their trip start time. Some of those booking might get cancelled by the company due to unavailability of the cab or to say overbooking. Hence the aim of the project is to render a predictive model that can be helpful for both customers and the cab service providers.Customers through this predictive model can plan their trip accordingly and the cab service providers can take action by bringing up more cabs on road.

The project is a supervised task aiming at finding relationship between dependent variables and cab cancellations.As part of building this predictive model, the data-set has been taken from Kaggle website.The data has been collected by the YourCabs, a Bangalore based cab startup company. The processes involved in the analysis starts with data visualization, proceeding with data cleaning , followed by data preprocessing and finally building up the predictive model for the test data.To ease my analysis, pre defined data mining techniques of classification like Support Vector Machine,Random Forest Classifier have been made use.This is binary and supervised classification problem

II. Implementation

A. Data Visualization

Knowing your data is the first and the most crucial step in any data mining project.The data used in the project was taken from the Kaggle website. The shape of the data is 43431 rows and 20 columns.The data is combination of both categorical and numerical. The data is a mixture of float , object and integer types. To proceed with data visualization, firstly the date column have been converted to their correct data type that is data-time.Index has been build to allow easy referencing of data points. An important observation that could be made seeing the data was the presence of class imbalance.To deal with the same up-sampling and down-sampling methods have been used.



Several columns relation with the target variable has been visualized to get a rough look of the data.Relation of package-id column with the target reveal that shorter distance and shorter duration routes are more preferable by people. Similarly , point to point travelling is more prevalent among folks. Weekdays and Fridays are the days when booking made are quite large. Also festival months like October, August receives quite a number of bookings.Another important observation is online or mobile booking are quite famous among people in contrast to offline booking.Moreover, area wise cancellation relation has shown that some areas are not preferred by cab drivers may be due to unhealthy environments or bad roads.Also,no significant impact on cab cancellation due to difference between booking and trip start has been observed. Although, the dependent variables which has close relation with the target, those are also a major player in cabs cancellation.

B. Data Cleaning

Cleaning of data is another important step to get the right and good accuracy predictive model.This steps involves detection and removal of missing values , noise , outliers if present and duplicate values. The data consisted of missing values and noise.No outliers found and since data is taken from one source no duplicates are there. As most of the missing values were present in the categorical attributes , the same has been dealt with by creating separate class to consider them under noise class.Another attributes have been removed under pre-processing task , thus getting rid of both noise and missing values.

C. Data preprocessing

Pre-processing of data aims at making the data ready for the model creation. Techniques like aggregation, feature subsetting, feature creation and dimension reduction and sampling has been used here. As part of aggregation booking and trip start date has been removed to merge into new column that provides their difference.For feature subset method like SelectKBst,Recursive Subset Elimination, feature importance and Boruta package has been used.Filter method using chi as is metric has been used in SelectKBst.Also to check for multi-collinearity among the dependent variables heatmap has be plotted and attributes like latitude,longitude, from-city id , to-city is has found a good correlation.Also user-id,from-city id , to-city is has been rated the least attributes important by most of the above mentioned methods based on their relation with the target attribute.Feature creation involved creating new attributes from-weekday, to-weekday , from-booking day, to-booking day, to give a better correlation Thus removing the unnecessary attributes and with the completion of this phase out of around 24 attributes half of them has been removed and thus leaving with 12 attributes to go for the classification phase.Boruta packages employs the use of appending the shadow features to the original dataset to compare the connection between the increase in dimensionality with the shadow object and thus helping in getting both the best as well as the worst fit features.

D. Boruta Package

The Boruta method is one of the wrapper methods built around random forest.It aims at capturing all important features of the given data-set.Shadow features are created which are appended with the data-set and the z-score both of the features are compared at each iteration when the modified data-set is given to the random forest classifier The method used by it are as below:

- Duplicate features are created by shuffling the values in each of the columns.The randomness so created creates a new set of features called the shadow features.
- the data-set together with the original and the shadow features is given to the random tree classifier.The accuracy mean so calculated is kept track of
- the program then sees if the original features have a higher importance or not. Meaning do the real features have a better Z-score than the maximum Z-score of the shadow features than the best of the shadow features
- At each iteration the program compares the Z-scores between the shuffled copies and the original features to check if the latter performed better than the former. If so happens ,the algorithm will mark the feature as important

The advantage of using Boruta package over other traditional feature selection algorithm is that the former captures all features which in some circumstances is relevant to the target variable.But the later relies on a smaller subset of the features that could yield a minimal error for choosen classifier.Thus Boruta finds all features which are either strongly or weakly related to the target and not just focuses on getting the optimal or minimal subset that could work the best for the given classifier.

E. Data Classification

In this section the problem of class imbalance has been dealt first.The methods used for this purpose are up-sampling and down-sampling.Up-sampling is a method where the minority class is randomly sampled with replacement to bring the count of the minority class same as the majority class.Whereas in down-sampling, the majority class is randomly sampled without replacement to bring down the count of the majority class down to the minority class.Now since in either of the methods the count of both the class is brought to same level as each other, the dataset becomes class balanced.

The data is split into training and test data using train-testSplit method.Initially the dataset was tested without solving class imbalance solution the accuracy as expected was high.Later after the imbalanced was removed by the up-sampling and the down-sampling methods the SVM showed a better change in accuracy and also the precision, recall and f-score.This is due to the fact the class imbalance leads to presence of dominance of one class over the other. This affects the accuracy as well as the error score of the model.As the number of positive points are more hence the accuracy for that is more.

III. Analysis

As the dataset had class imbalance problem , the same has been handled graciously. This can be seen from the increase in cv-score from 73 percent to around 90 percent in case of SVM.Random forest has inbuilt capability of detecting class imbalances so perform good in both the cases after and before class imbalance.

IV. Future Scope

The data had a mis-classification error cost associated with which implied the cost incurred if the data which has to be classified as cancelled is is wrongly written as not cancelled.Moreover, SVM and random forest has been used currently , would like to test the combinations of various other classifiers to test their behaviour and accuracy and errors score on the given dataset . Apart from this, as the dataset was not to large and not that good with providing observation of various kinds , would like to work on a better and larger ,if possible real-time datasets.Besides, could have an improvement on better data exploration and feature engineering.

References

- https://medium.com/datadriveninvestor/choosing-the-best-algorithm-for-your-classification-model-7c632c78f38f
- https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e
- https://towardsdatascience.com/feature-selection-in-python-recursive-feature-elimination-19f1c39b8d15
- https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/
- https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/
- https://www.scikit-yb.org/en/latest/api/model-selection/rfecv.html
- https://ieeexplore.ieee.org/document/7916845
- https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
- https://pandas.pydata.org/pandas-docs/version/0.16.2/generated/pandas.core.groupby.DataFrameGroupBy.plot.html
- https://www.youtube.com/
- https://www.jstatsoft.org/article/view/v036i11/v36i11.pdf
- https://www.geeksforgeeks.org/python-programming-language/
- https://www-users.cs.umn.edu/kumar001/dmbook/index.php
- https://medium.com/@indreshbhattacharyya/feature-selection-categorical-feature-selection-boruta-light-gbm-chi-square-bf47e94e2558
- https://www.analyticsvidhya.com/blog/tag/boruta-package/
- http://danielhomola.com/2015/05/08/borutapy-an-all-relevant-feature-selection-method/