Case Study: Conversion Prediction

Task at hand: Given a set of attributes pertaining to user sessions, the predictive analysis problem involved in utilizing the available information to predict whether user session led to the booking on the advertisers page or not.

Important insights in data: Following are some of the important insights gathered during the observation of the data:

- Data was highly imbalanced, True targets were extremely smaller than the False and that often leads to situation where the classifier has the only hypothesis to classify everything in the dominant class.
- Few of the information (attributes) in data were not much helpful in formulating a target function.
- Inconsistency of user behaviour leading to booking or not booking was evident.

Steps taken: I always start with checking whether any of the column contains null values (it is helpful to be known beforehand)**.** As the data has a potential class im-balancing problem, therefore, one way is to oversample the rare class with synthetic samples to overcome such a restriction. I applied only one sampling technique due to the time restrictions. Then, I visualized the feature correlations by means of a correlation matrix (evident in the source folder). Then, computed the basic statistics of how much of these feature contributes in explaining the variance. Based on these information I then created my feature set and dropped some of the features would otherwise might effect the performance of the classifier. For preparing the data for training, I split the training data came with this challenge into train and validation with 70% train and 30% test in order to choose the best model. I then experimented over a number of algorithms (ExtraTreesClassifier, Random Forrest etc) and so far found out that the best performing algorithms among these was the ExtraTreeClassifier with **76.37 (**Mathews Correlation Coeffcient (MCC)**)** as average over three runs with different random seeds and with standard deviation of **0.39** score on on the validation set . After finding a model, I used the original train data as a whole for training and original test set for prediction. I also experimented a Neural Network architecture for this classification task and unfortunately, the model did not performed optimally and reported a lower score MCC score (34.42 %) mainly due to large data.

Questions :

Why do you think we are interested in predicting if a user made a booking?

This is because it has a substantial impact in the demand management decisions in the industry say it a hospitality industry or any other one. Its an important tool in terms of revenue management and its performance.

What was the most challenging part of the case study for you?

In my opinion at the first glance of the problem, the data and coming up with a right methodology in a hope to approximate the target function was a bit non-trivial.

What were some of your interesting observations or findings?

So far, the important observations have already been mentioned above in the section „Important insights“.

How would you have approached this problem if you had more time, say 3 months?

Good solutions always takes time and effort to come up with. With more time, I will approach this problem with more extensive experimentations that includes various sampling techniques with a mix of underand over-sampling. Furthermore, I will also would prefer feature engineering (with more information, if available). I would also take a bigger dataset to experiment more with the neural networks as these require alot of data but may lead to better results. Feature ablation study unfolds more information on formulating which features are more important and therefore, I will perform such a study as well. These steps might lead to better performing algorithms and could improve towards adding more value to the business.

If you could ask for any more data, what would it be and why?

Precisely, with more data available and with information about the user session behavior (i.,e more variables) the more it will be helpful to come up with a better algorithm.