Contribution ID: **10**  Type: **not specified**

# A real time approach targeted at buyer prediction, using behaviour classification in Tourism Web Analytics

*Wednesday, 18 September 2019 17:10 (25 minutes)*

Given the continuously growing importance of digital marketing in the tourism sector, understanding customer's purchase behaviour is a critical issue for the online competition. Competitors who are able to identify prospective patterns can create a value to lift business.

This work presents a supervised machine-learning (ML) approach, focused on the buyer prediction, in the context of online behavioural targeting to analyze in real time the user session, and predicting precisely if she/he would buy or not.

The appropriate selection of data samples is important for effective analysis and prediction based on behavioural patterns, for this purpose as use case is used a huge dataset representing the user experiences in a hotel/holydays structure to train and test an ML model. The dataset is organized in 669.653 user sessions defined as observation that can be brought back to users that finalize a purchase (positives) and user session that not purchase (negatives). The data analysis highlighted the great unbalancing of the observations as positives (4%) versus negatives (96%) that is solved by applying the Under Sampling (Scikit-Learn) algorithm to obtain a leveled dataset to train the model. A Feature engineering analysis is done by selecting the best features that produce less errors and best accuracy. Our training method is based on a hybrid combination of the Gradient Boost Classifier (XGBoost and LightGBM) and Decision Tree Classifier (Scikit-Learn) to perform the binary classification.

We performed our experiments developing a prediction module as a RESTFull service, that the has been connected in the retail application. When a user session is collected in real-time, the module make a prediction that can be evaluated to put in place some action to engage the user. The evaluation experiment of the RESTFull service in production environment took place over a period of six months and the collected data were further analyzed. During the experiment ~240.000 real-time predictions were generated in total, 5% of these are positive predictions of which 87% are correct (accuracy > 70%), 95% of these are negative predictions of which 80% are correct (accuracy > 70%).

Our suggested approach compared to the last six months, before the start of the test, seems capable of dealing with more complex online advertising models. In particular we evaluate the impact in the marketing business noting a 70% decrease of management time and 90% increase in viewability of the proposed product to sell.

The obtained results are promising and encourage us to continue experimentation with more sophisticated models or other algorithms to further improve the performance of the system.
In the next steps we are planning experiments to improve the model prediction with more information obtained from third party data providers (eg. work-calendar, national-holidays, events…) and than introducing the temporal dimension to our model to apply time series analysis techniques.

**Primary author:** DERIU, Massimo (CRS4)

**Co-authors:** Dr MANCHINU, Andrea (CRS4); Dr FADDA, Stefano (CRS4)

**Presenter:** DERIU, Massimo (CRS4)

**Session Classification:** Session 2