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

Deep learning @ INAF – Pula 16 - 19 September 2019

Stefano Fadda - CRS4 - stefano@crs4.it

Agenda

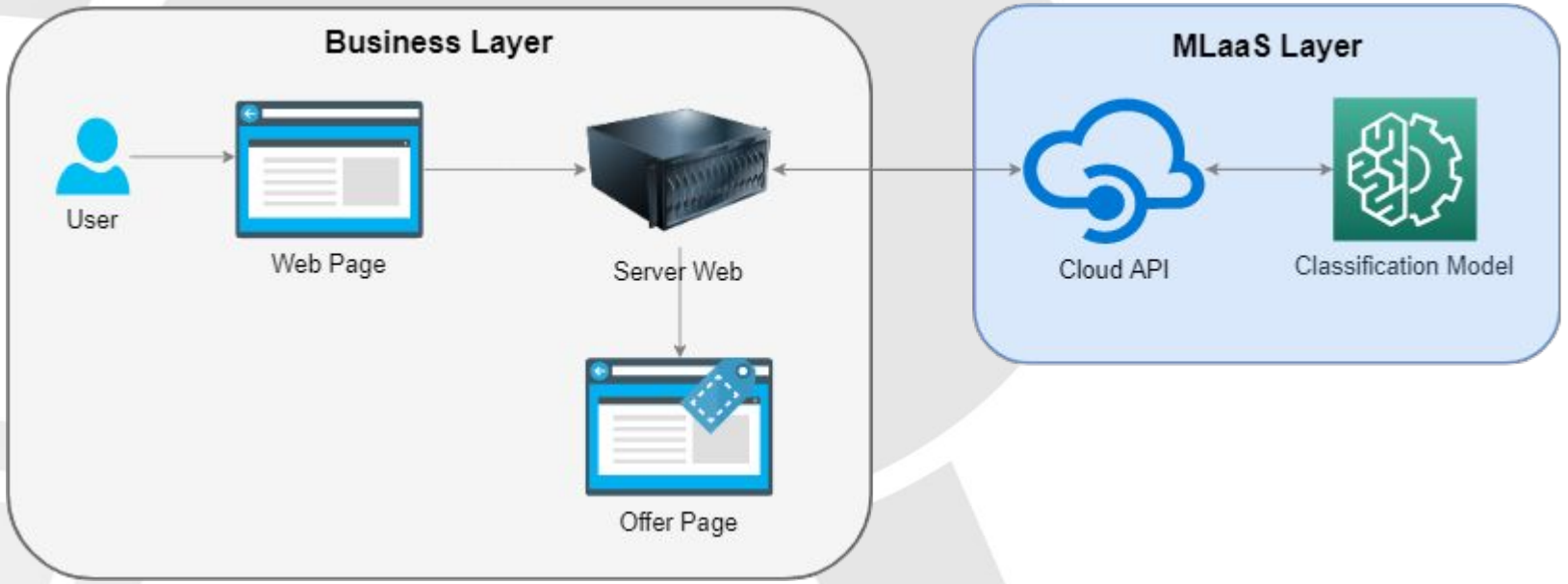
- context
- purchase prediction use case
- solution and methodology
- dataset
- models
- web application (MLaaS)
- conclusions

Context

- statistical data from the partner agency about hotel/holydays structure
- user behaviour and satisfaction comprehension
- historical user navigation sessions to understand user behaviours

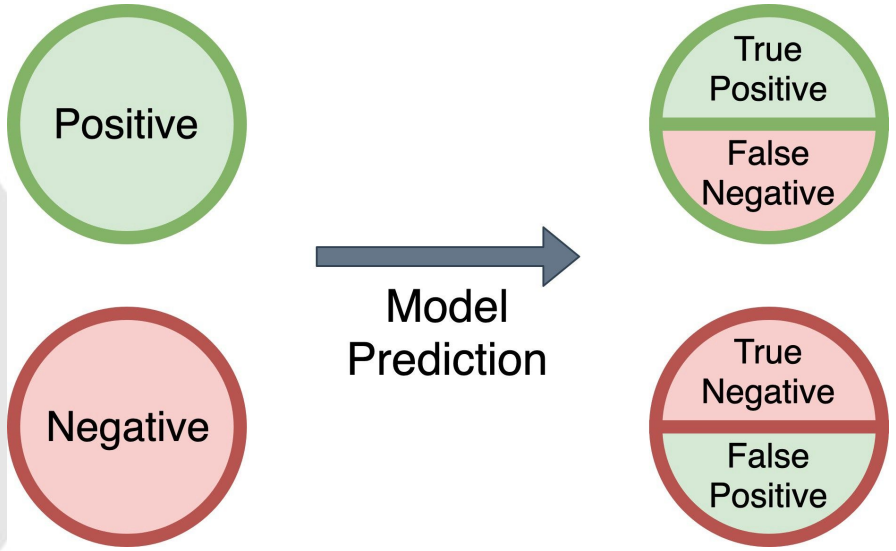
Use case

- automatic prediction of the user's purchase will
- possibility to perform the prediction in real time



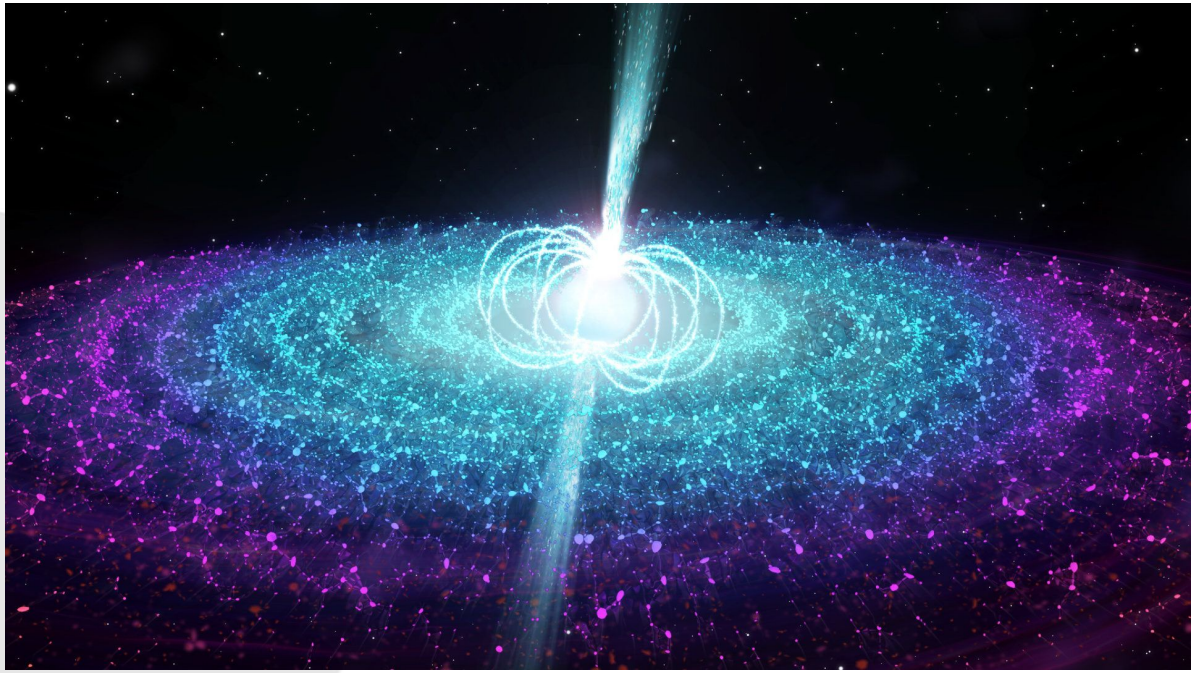
Solution and methodology

- the problem can be addressed as a binary classification, so the data available can be cataloged as belonging to the category of who buys (positive) or not (negative)
- selection of all data about user navigations that are related to the two categories



Pulsar detection

- similar methodology has been applied to classification and identification of Pulsar
- a lot of false positives
- oversampling techniques to balance the dataset



(Punia, Akhil, Ashish Sardana, and Monica Subashini. "Evaluating advanced machine learning techniques for pulsar detection from HTRU survey." 2017 International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2017.)

Dataset

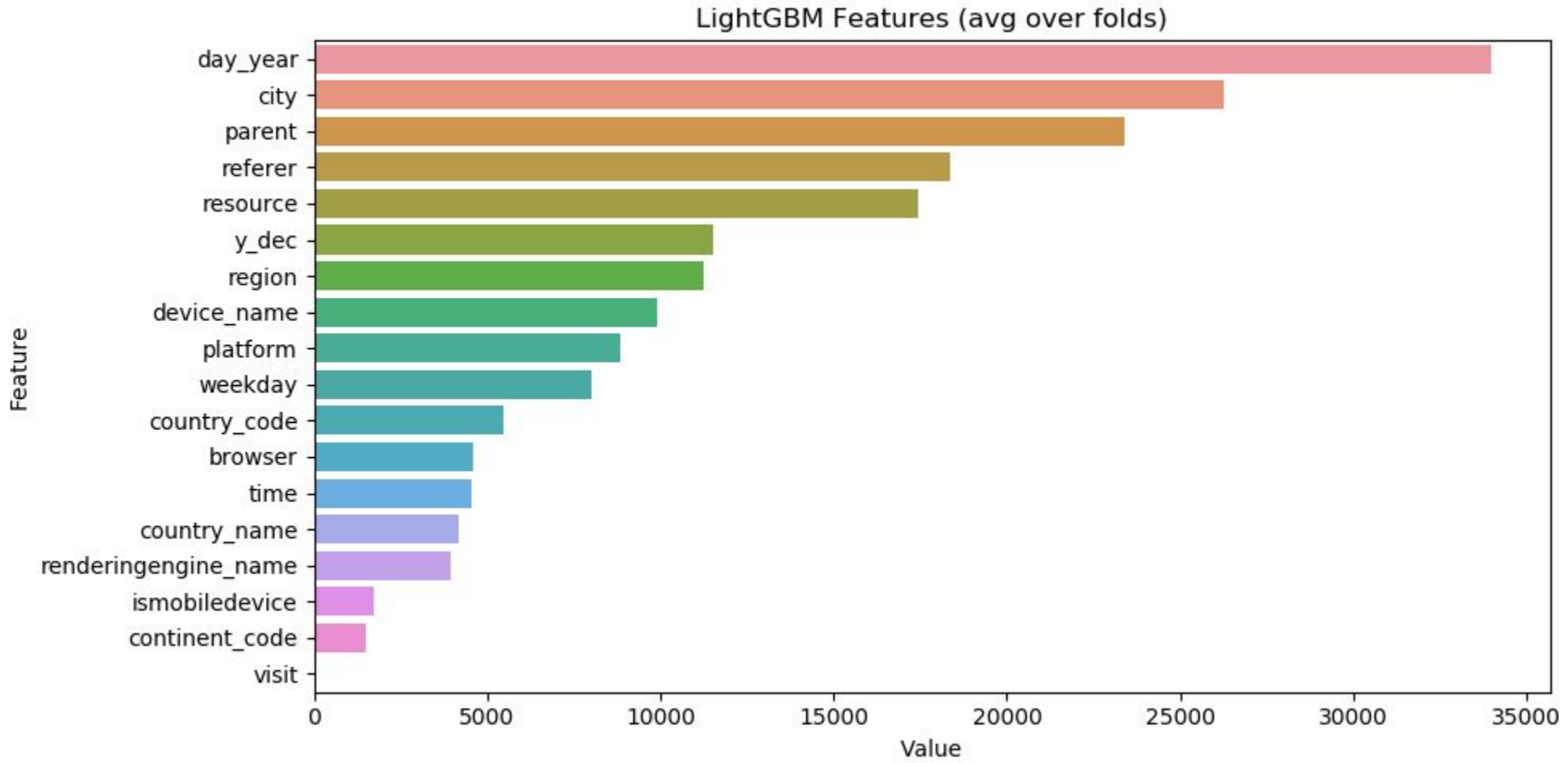
- positive user sessions represent 4% of the total
- dataset is extremely unbalanced on the negatives
- undersampling algorithm like RandomUnderSampler
- its obtained a balanced dataset of the two categories with the same weight
- dataset preprocessing phase to manage null data, duplicates and utility transformation

Models

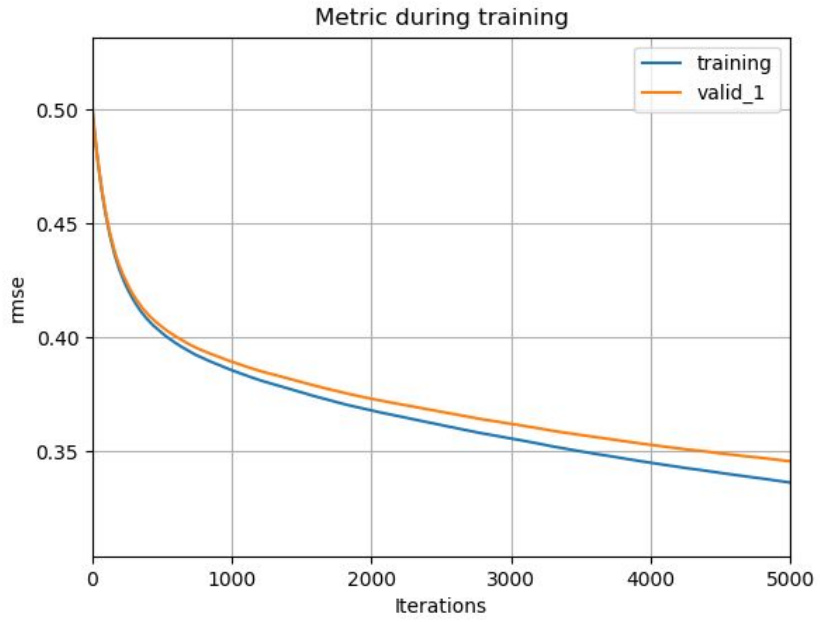
- analysis of the state-of-the-art of the main python libraries about binary classification: LightGBM, XGBoost, Scikit-learn
- RMSE is a measure of accuracy, to compare forecasting errors of different models for a particular dataset
- LightGBM has been identified as the faster in the training phase

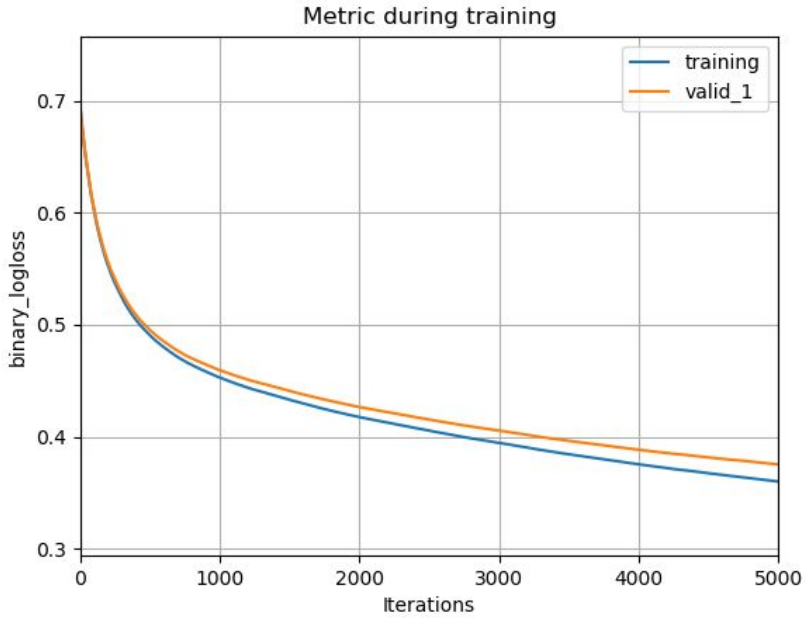
	RMSE-train	RMSE-test	Training-time (sec.)
LightGBM	0.20426	0.22897	**3.846010**
XGBoost	0.23423	0.25058	20.176368
Scikit-learn	0.25721	0.26421	11.143738

Models - features importance

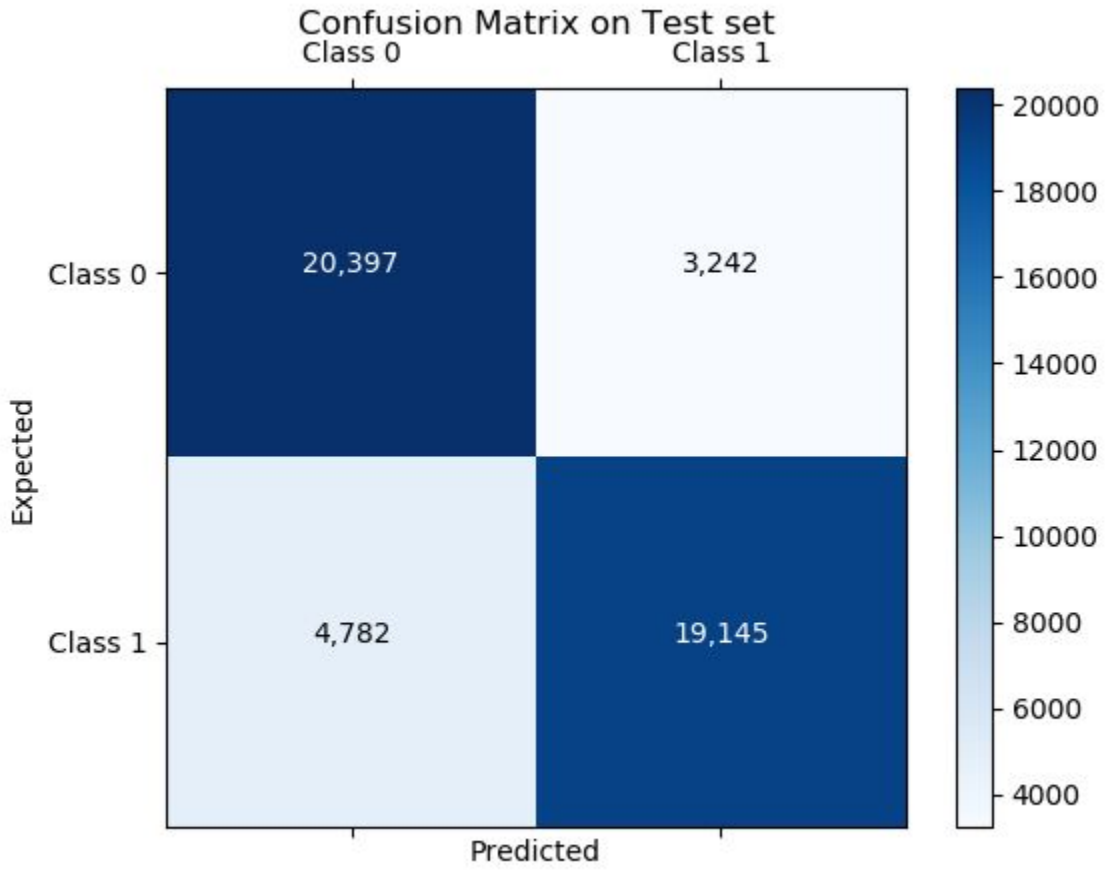


Models - overfitting





Models - confusion matrix



Web application (MLaaS)

- Web service as Restful API interface to execute the prediction on real-time navigation user data
- facilitate the integration with third party systems
- Web frontend to test the ML service, monitor the service and see the results

Web application (MLaaS)

POST http://hma-service.crs4.it/api/v1/predict Send

none form-data x-www-form-urlencoded raw binary GraphQL BETA JSON (application/json)

```
[
    {
        "commit": 1,
        "resource_id": 658566,
        "session_id": 172660,
        "visit": 0,
        "resource": null,
        "time": 32,
        "created": "2018-08-09 10:14:42",
        "referer": "https://www.google.it/",
        "label": "ita",
        "continent_code": "EU",
        "country_code": "IT",
        "country_name": "Italy",
        "region": "10",
        "city": "Morrovalle",
        "postal_code": "63854",
        "parent": "Firefox 61.0",
        "platform_version": "6.1",
        "renderingengine_name": "Gecko",
        "ismobiledevice": null,
        "device_name": "Windows Desktop"
    }
]
```

```
{
    "status": 200,
    "message": "true",
    "body": {
        "prediction": {
            "commit": 0.0,
            "predicted": 0.0,
            "Accuracy": 0.8421052631578947,
            "MSCLS": 0.1578947368421053,
            "Precision": 0.0,
            "F1_score": 0.0
        }
    }
}
```

Conclusions

The project its actually a work in progress, we evaluate the system with real live data in the near future

The approach and the technology is immediately applicable in other context