Vodafone Data Science Hackathon 2021

The Sons of Analyst

Luigi Noto – Bernardo Principi – Manuel Serra

Data pre-processing

We joined the "full_dataset.parquet"and "train.csv"files by *session_id*, in order to use the information contained in the parquet file for training the model.
We filled the missing values in the columns *event_category_idx_lv2*, *event_category_idx_lv3* and *event_category_ idx_lv4* with the values *Not_applicable_lv2*, *Not_applicable_lv3*, *Not_applicable_lv4*, respectively, in order to make sure to consider the missings when training the model.
In order to take into account the temporal component and the event-based structure in the data before aggregating multiple rows into a single row for each session id (primary key in the final dataset), we created the following variables. For each session we computed the following:

- *max_time_delta*: time difference (in days) between the oldest event and the interaction with TOBi. We noticed that there is an upper bound at about 60 days.
- *min_time_delta*: time difference (in days) between the most recent event and the interaction with TOBi. As written in the instructions, this variable was encoded in a fictitious event, having *event_category_idx_lv1* = 1_5. However, we noticed that some sessions missed this event, therefore for such sessions we computed this variable manually using the available timestamps.
- *mean_time_delta* and *mean_median_diff*: mean of the time difference (in days) between each event and the corresponding interaction with TOBi, and difference between this variable and the median of this distribution, respectively. The mean is used as measure of centrality of the event-interaction time-difference distribution inside a session, but we consider also the mean-median difference to take into account the fact that such distribution is skewed for some sessions and to capture some correlation, since we noticed an average difference in the mean-median difference between sessions with different labels.
- *weekend*: dummy variable equal to 1 if the interaction with TOBi occurs in the weekend, 0 otherwise. In the exploratory data analysis phase, we checked whether there were some links between the week day of the interaction and the session label. We noticed that for each label, the number of sessions with that label occurring in the weekend was generally lower than the number of the ones occurring in the business days. However, such a difference was lower for the sessions with label 4 than for the sessions with any other label.
- *week_number_0*, ..., *week_number_17*: for each week, dummy equal to 1 if the interaction with TOBi occurs in that specific week, 0 otherwise, where the weeks are encoded progressively from 0 to 17, starting from the first week of June (oldest sessions) to the last week of September (most recent sessions).
- *# of events*: number of events in the session. The dataset was aggregated by session id in order to count the number of previous events for each session.
- Frequency of each concatenation of values of the variables *event_category_idx_lv1*, *event_category _idx_lv2*, *event_category_idx_lv3*, *event_category_idx_lv4*. In order to take into account the hierarchical structure of the event category index levels 1, 2, 3 and 4, we sequentially concatenated

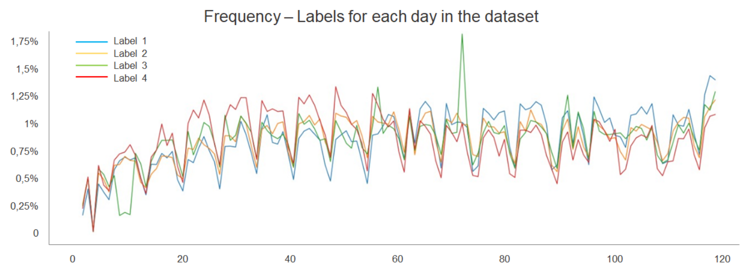
the values of these 4 variables for each event and then computed the relative frequency of each such concatenation among the events of the session. Before doing this, in order to potentially filter out some noise and improve the performance of the prediction, for each event category level, we grouped the categories with less than 5000 occurrences in a single variable. Moreover, for the same reason, we grouped the concatenation with less than 10000 occurrences in a single variable.

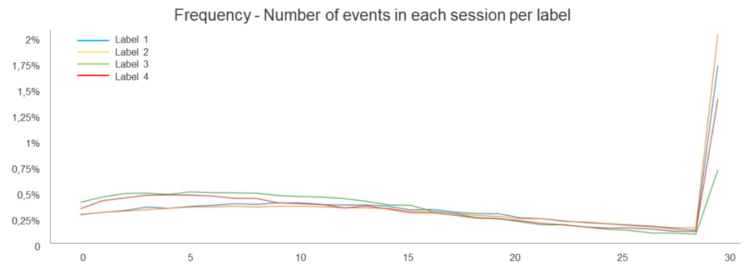
- *other_sessions_label_1*, ..., *other_sessions_label_4*: for each label, dummy variable equal to 1 if at least one of the other sessions of the user involved in the session in question has that specific label, 0 otherwise.
- *gini_index*: Gini index of the concatenation of values of the variables *event_category_idx_lv1*, *event_category_idx_lv2*, *event_category_idx_lv3*, *event_category_idx_lv4* across the events of the session. We used this variable to measure the heterogeneity of the relevant events in the session leading to the interaction with TOBi.
- *customer_label_0*, ..., *customer_label_8*. In order to incorporate more information about the customers in the dataset, we used unsupervised learning to cluster the customers according to the averages across all of their sessions of the following variables: number of events, *max_time_delta*, *min_time_delta*, *gini_index*, *mean_time_delta* and *mean_median_diff*. We used the K-Means algorithm with number of clusters k=9 as input, optimally chosen through a cluster analysis based on the silhouette index. Then, for each cluster, labeled from 0 to 8, we defined a dummy equal to 1 if the customer of the session belongs to that label, 0 otherwise.

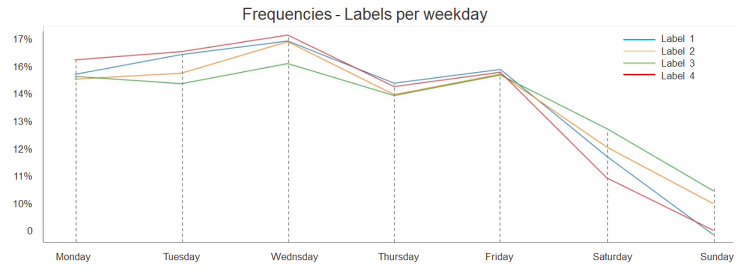
Model

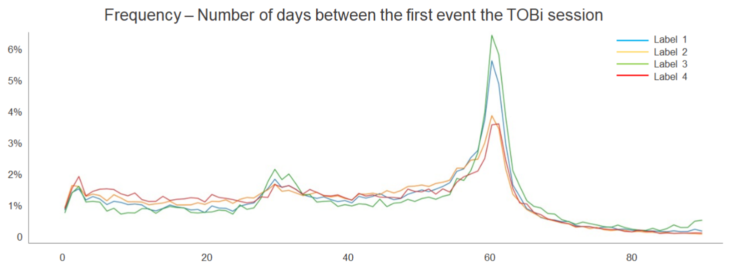
In order to preserve interpretability while still aiming to achieve the highest accuracy, we preferred to consider trees bagging or boosting over a neural network: initially, the chosen model was a gradient boosting implemented through Scikit-learn library, but we later proceeded to implement an XGBoost (eXteme Gradient Boosting) using the xgboost package, a model based on the same principle but with a better regularization function that is capable of better handling overfitting, while also enabling the users to perform computations using a GPU, thus significantly decreasing training time and allowing for a wider cross-validation.
Moreover, once again with the goal of preserving as much interpretability as possible, we would like to specify that no transformation has been applied to the training set, such as standardization or PCA, which would instead have been necessary in order to fit different types of models.
In particular, after a careful cross-validation process, the parameters of the final model are the following: 150 trees where 70% of the available features are sampled and considered for each split, with a learning rate of 0.5 and maximum tree depth of 8. The model achieved a quite stable accuracy of about 53%-54% during the validation process, and we have also been able to identify the most important features that allowed us to achieve this result, as it can be seen in the figure in Appendix.

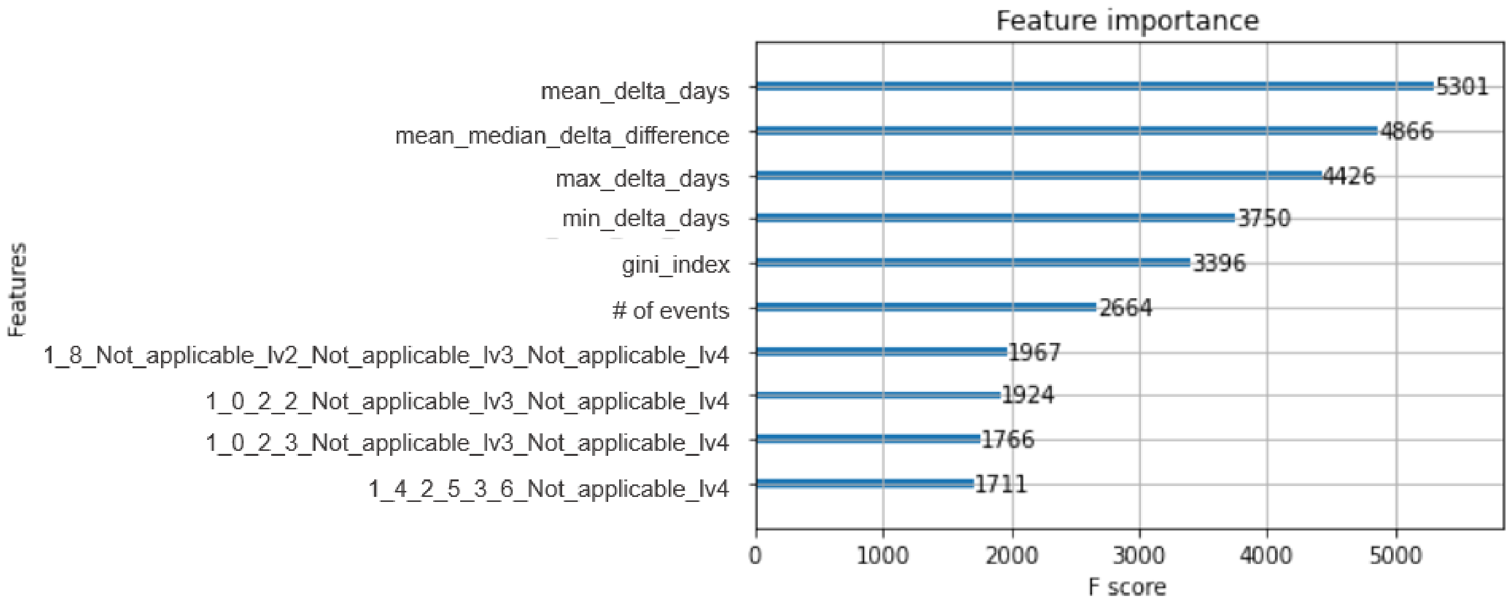
Appendix - Data Exploration and Feature Importance


Frequency – Labels for each day in the dataset
Label 1
Label 2
Label 3
Label 4


Frequency - Number of events in each session per label
Label 1
Label 2
Label 3
Label 4


Frequencies - Labels per weekday
Label 1
Label 2
Label 3
Label 4
Monday
Tuesday
Wednsday
Thursday
Friday
Saturday
Sunday


Frequency – Number of days between the first event the TOBi session
Label 1
Label 2
Label 3
Label 4


Feature importance
Features
mean_delta_days
mean_median_delta_difference
max_delta_days
min_delta_days
gini_index
# of events
1_8_Not_applicable_lv2_Not_applicable_lv3_Not_applicable_lv4
1_0_2_2_Not_applicable_lv3_Not_applicable_lv4
1_0_2_3_Not_applicable_lv3_Not_applicable_lv4
1_4_2_5_3_6_Not_applicable_lv4
F score