# **Rules of the competition**

* Only **One** Kaggle account to be created per team as mentioned in the email (You cannot sign up to Kaggle from multiple accounts and therefore you cannot submit from multiple accounts.).
* Submission limits - Only a **maximum of 5 submissions per day** are allowed for the competition (You can try five test data submissions per day during the time of the competition window is open).
* No private sharing of the response variable of the test dataset allowed. This would lead for disqualification.
* Team mergers are not allowed in this competition.
* Competition timeline –

***Start date – 17th of September 2019 at 5.00 a.m***

***End date – 21st of September 2019 at 11.59 p.m***

# **Description of attached files**

* Trainset.csv - the training data set. (including the response variable Revenue)
* xtest.csv - the predictors of the test set.
* sample\_submission.csv - a sample submission file in the correct format. This should contain both the ID and predicted outcome (0- No revenue generated, 1- Revenue generated) and should have the headings as "ID”, "Revenue".

# **About the dataset**

The dataset consists of observations belonging to 12,330 sessions.

The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The dataset consists of 10 numerical and 8 categorical attributes. The ‘Revenue’ attribute can be used as the class label.

Citation: Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018)

# **Data fields**

Online Shoppers Purchasing Intention Dataset

|  |  |
| --- | --- |
| **Variable** | **Description** |
| ID | An anonymous id, unique to a given customer (only included in the test dataset for the purpose of predictions). |
| Homepage | Number of different types of pages visited by the visitor in the Home page section. |
| Homepage \_Duration | Total time spent in the Home page section. |
| Aboutus | Number of different types of pages visited by the visitor in the About us section. |
| Aboutus\_Duration | Total time spent in the About us section. |
| Contactus | Number of different types of pages visited by the visitor in the Contact us section. |
| Contactus\_Duration | Total time spent in the Contact us section. |
| BounceRates | Percentage of visitors who enter the site from that page and then leave (“bounce”) without triggering any other requests to the analytics server during that session. |
| ExitRates | From all pageviews to the page, the percentage that were the last in the session. |
| PageValues | The average value for a web page that a user visited before completing a transaction. |
| SpecialDay | The closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine’s Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8 |
| Month | The month of the year (Nominal) |
| OperatingSystems | Operating system used (Nominal) |
| Browser | Browser used (Nominal) |
| Province | Province of the visitor (Nominal) |
| TrafficType | The number of times the visitor visited the site |
| VisitorType | Visitor type as returning, new visitor or other (Categorical) |
| Weekend | Visited during a weekend or not, (a Boolean value- Categorical) |
| Revenue | Revenue will be generated or not  Response variable (Desired Target)  Categorical: 0 – No revenue generated, 1- revenue generated) |

# **Task**

Predict the class of the type of customer as to whether the customer would be a revenue generating customer or not, by using **the revenue variable** as the dependent variable. The rest of the variables would be independent variables.

Upload the predicted outcome of the test set according to the format provided, to obtain the accuracy of the prediction.

Good Luck!