## **[A21] DA Machine Learning with Python Labs**

PROJECT

Saturday – 25 june 2022

* Stephane Ganhi

Logo

Description automatically generated

**Introduction**

The Business Intelligence application uses the transformed and formatted data to make power and insightful discovery from the data to aid top managers to make quality decision. However, the IT company ServiceSpot receives a lot of calls from their customers which are organized in several file formats CSV, Excel etc and would like to make good use of them in order to study and analyze them. Our goal is to use tool to perform and automate task. Regarding this achievement Python language facilitate the adaptability to deeply understand our good reading data. To do this, 3 steps are necessary:

* + Data analysis
  + Features selection
  + Model training

# 1 Objective

The goal of this project is to use python language to perform machine leaning. Starting from loading the model until we train the data to see if it can give the same response with a large amount of data beside the one we use.

# 2. Data Analysis

Our data are coming from good reading source. The data are in csv file. The first step in our analysis was to download the data. After downloading the dataset which has 11123 rows and 12 columns. For the exploration, we clean the data by removing the semi colon publisher and rewrite number of pages. Moving from this point, we checked for na (not a number) were the proportion is less than 30%. So we decide to remove them. Our data is now 11094 rows and 12 columns. As the aim is to see the impact of the others variable on rating counts. We draw a graph to look for outliers.

Chart, scatter chart

Description automatically generated

As you can see on the picture there are outliers then we decide to remove them.

For our model we plot the correlation matrice for continuous variables.

**Chart

Description automatically generated**

We plot graphs on continuous variables and perform a normalization on the variables. See the notebook (anaconda).

**3. Features selection**

For the sake of our model we choose all the continuous variables which have on impact on the average rating and 3 categorical variables (authors, language and title).

On this features, we apply a deep analysis on categorical variable. We get dummy variables on the variable and assume high ratings as ratings => 100000 ansd OtherRatedBooks.

This strategy applies for the categorical variables to make them numerical variables. And apply the dummy variables on language code.

**4.Model Training**

We split our model in two:

the train model

the test model

Applying the linear regression on our model , we predicted the average\_rating on the test and traning by getting the intercept, coefficients and "average\_rating" predictions for Higly rated books

Text, letter

Description automatically generated

**the actualm predicted table:**

**Table

Description automatically generated**

The model looks reasonable with Root Mean Square Error (RMSE) at around 0.22 and Mean Absolute Error (MAE) around 0.20

Text, letter

Description automatically generated

The second model performs way better with Root Mean Square Error (RMSE) at around 0.35 and Mean Absolute Error (MAE) around 0.23 for other books not rated high

**Text

Description automatically generated**