# Applied Artificial Intelligence

# Project 05

# “Predict Future Sales”

**Background:**

We are provided with daily historical data from Kaggle. The task is to forecast the total amount of products sold in every shop for the test set.  The list of shops and products slightly changes every month.

**Data sets:**

We are given the following datasets:

- sales_train.csv - the training set. Daily historical data from January 2013 to October 2015.
- test.csv - the test set. You need to forecast the sales for these shops and products for November 2015.
- sample_submission.csv - a sample submission file in the correct format.
- items.csv - supplemental information about the items/products.
- item_categories.csv - supplemental information about the item’s categories.
- shops.csv- supplemental information about the shops.

**Data Fields:**

The data fields are as follows:

- ID - an Id that represents a (Shop, Item) tuple within the test set
- shop_id - unique identifier of a shop
- item_id - unique identifier of a product
- item_category_id - unique identifier of item category
- item_cnt_day - number of products sold. You are predicting a monthly amount of this measure
- item_price - current price of an item
- date - date in format dd/mm/yyyy
- date_block_num - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1, October 2015 is 33
- item_name - name of item
- shop_name - name of shop
- item_category_name - name of item category

**Methodology Used:**

(1) The data is fetched and preprocessed.
(2) All the items are grouped with respect to the month they are purchased in. This also contains information about the shop from which the item was purchased.

(3) A train table is created that consists of the item's details and the month in which it was purchased.

(4) The data is trained using any of the following models:

   a. Ridge
   b. DecisionTreeRegressor
   c. AdaBoostRegressor
   d. BoostingRegressor
   e. RandomForestRegressor
   f. GradientBoosterRegressor

(5) The data is then tested based on the models it was trained.

**Steps to run:**

Prerequisites:

Python 3
Packages like numpy, pandas, sklearn

Instructions:

(1) Out of the above mentioned 6 models, at one time only one model could be executed.

```
rgr_ridge = Ridge()
dtr = DecisionTreeRegressor()
ada = AdaBoostRegressor()
bag = BaggingRegressor(verbose=2)
random_forest = RandomForestRegressor(verbose=2)
gboost = GradientBoostingRegressor(verbose=2)
```

Line 1:

```
gboost.fit(X0, Y)
```

Line 2:

```
Y_pred = gboost.predict(X0_test)
```

Line 1 and Line 2 should be replaced with the objects of other models.

Ex:

If we are implementing random forest,

Line 1 would be:

Random_forest.fit(X0, Y)

Line 2 would be:

Y_pred = random_forest.predict(X0_test)

(2) Create a pred.csv file in the same folder as the other input files.

(3) Run the code in python complier.

(4) Check the pred.csv file which would have the item counts.