Sample_Lab2_Python

February 8, 2020

1 ML for Finance

1.1 Fall 2019

1.2 Lab

Here you have to resolve 5 Data Science problems (each of 2 points in total). Please, find out the material, placed in the book Introduction to Machine Learning with Python : A Guide for Data Scientists and the notebooks for each chapter. Navigate also here: https://mlcourse.ai/.It should be helpful.

```
In [ ]: # YOUR NAME
```

1.3 1. Linear Regression

Use `train.csv` data from Sberbank Russian Housing Market competition (you need to register on Kaggle in order to have an access to these datasets).

TASKS 1.1 Use `train.csv` dataset. Perform two linear regression models for `price_doc` prediction.

1.2 Assess the quality of your models on `train.csv` dataset (split it 80/20). Use RMSE metric. Briefly describe wheter or not this quality is appropriate.

1.4 2. Logistic Regression Classifier

Use data sample from Sberbank Data Science Jorney Contest 2016 (Task A). * `transactions.csv` * `customers_gender_train.csv`

```
customer_id  id of client;
tr_datetime  day and time of transaction;
mcc_code  special code of transaction;
tr_type  transaction type;
amount  sum of transaction with sign: +  inflow transaction, -  outflow transaction;
term_id  ATM code
```

Preamble. You need to predict the probability to be the male-gender person (`gender = 1`) based on train sample data (`customers_gender_train.csv`) for those bank clients, whose gender is not defined in the dataset (`transactions.csv`). The quality of prediction is assessed as the area under ROC curve (AUC-ROC) between real and predicted data.

Hint. You can use baseline solution, implemented with Gradient Boosting, in order to start your calculations.

TASKS 2.1 Use `transactions.csv` and `customers_gender_train.csv` dataset. Use only data with non-empty `gender`. Split this data to test and train samples. Implement Logistic Regression Classifier to predict the gender.

2.2 Assess the area under ROC curve (Quality metric) for your test sample data. Briefly describe wheter or not this quality is appropriate.