University of Westminster
School of Computer Science and Engineering

7BUIS008W	Data Mining & Machine Learning
Module leader	Panagiotis Chountas
Unit	Coursework 1/ Prepared by Natalia Yerashenia
Weighting:	50%
Qualifying mark	40%
Description	Students are expected to critically justify the use of effective and novel data mining and machine learning techniques for a specific problem domain and definitely reflect on the knowledge of how different data mining and machine learning algorithms operate in terms of their underlying design assumptions and biases for a given problem domain. Students are expected to methodically analyse the output of data mining and machine learning algorithms by drawing technically appropriate and sound conclusions resulting from the application of data mining and machine learning algorithms to the given problem
Learning Outcomes Covered in this Assignment:	This assignment contributes towards the following Learning Outcomes (LOs): • LO1 critically justify the use of effective and novel data mining and machine learning techniques for Data Science applications; • LO3 critically reflect on the knowledge on how different data mining and machine learning algorithms operate and their underlying design assumptions and biases in order to select and apply an appropriate such algorithms to solve a given problem; • LO5 critically analyse the output of data mining and machine learning algorithms by drawing technically appropriate and justifiable conclusions resulting from the application of data mining and machine learning algorithms to real-world data sets
Handed Out:	24th March 2022
Due Date	21st April 2022 Submission by 13:00 hours
Expected deliverables	Submit on Blackboard *a zip file* containing the required documentation (either in docx or pdf format). All implemented codes should be included in your documentation together with the results/analysis.
Method of Submission:	Electronic submission on BB via a provided link close to the submission time.
Type of Feedback and Due Date:	Feedback will be provided on BB, on 12th May 2021
MSC CRITERIA MEETING IN THIS ASSIGNMENT	• a systematic and methodological way about Data Analytics/ Data Mining issues • develop problem-solving skills and knowledge of various techniques/tools/methods • ability to model and deploy appropriate software tools that satisfy specified requirements, and test their use in a target domain. • independent in-depth analysis of a chosen topic making use of information resources outside a teaching environment • studying the context within which the design of systems for Data Science and Analytics takes place • identifying the security and legal implications of Business Intelligence, Data Science and Analytics applications

Assessment regulations

Refer to section 4 of the "How you study" guide for undergraduate students for a clarification of how you are assessed, penalties and late submissions, what constitutes plagiarism etc.

Penalty for Late Submission

If you submit your coursework late but within 24 hours or one working day of the specified deadline, 10 marks will be deducted from the final mark, as a penalty for late submission, except for work that obtains a mark in the range of 50 – 59%, in which case the mark will be capped at the pass mark (50%). If you submit your coursework more than 24 hours or more than one working day after the specified deadline you will be given a mark of zero for the work in question unless a claim of Mitigating Circumstances has been submitted and accepted as valid.

It is recognised that on occasion, illness or a personal crisis can mean that you fail to submit a piece of work on time. In such cases you must inform the Campus Office in writing on a mitigating circumstances form, giving the reason for your late or non-submission. You must provide relevant documentary evidence with the form. This information will be reported to the relevant Assessment Board that will decide whether the mark of zero shall stand. For more detailed information regarding University Assessment Regulations, please refer to the following website:**http://www.westminster.ac.uk/study/current-students/resources/academic-regulations**

Coursework Description

A. Building an Ensemble

Ensemble learning uses multiple machine learning models to try to make better predictions on a dataset. An ensemble model works by training different models on a dataset and having each model make predictions individually. The predictions of these models are then combined in the ensemble model to make a final prediction.
In this task, you will be using a Voting Classifier in which the ensemble model makes the prediction by majority vote. For example, if we use three models and they predict [1, 0, 1] for the target variable, the final prediction that the ensemble model would make would be 1, since two out of the three models predicted 1.
You will use four different models to put into the Voting Classifier: k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression (LR). Use the Scikit-learn library in Python to implement these methods and use the **Facebook Metrics** dataset.

The data is related to posts published during the year 2014 on the Facebook's page of a renowned cosmetics brand.
This dataset contains 500 rows and part of the features analysed by Moro et al. (2016). It includes 12 features known prior to post-publication and 7 features for evaluating post-impact (see Tables 2 and 3 from Moro et al., 2016). The student has to decide which of the features (1-3 out of 12) to leave as the output feature.

The dataset is uploaded to BB.

The full dataset is available here.

1st Task: Voting Ensemble

You need to construct, train and test a Voting type classifier in Python, based on (kNN, LR, RF and SVM)

a) With the aid of the Python package visualise and justify the properties of the given *Facebook Metrics* dataset. Your analysis must include reference to the following items
   - Identification and treatment of any Missing Values;
   - Identification and treatment of any Outliers;
   - Data Normalisation;
   - You need to try 2 different methods of forming training and test sets.
   - Exploratory Data Analysis: choosing the Input and the Output columns, Filtering the needed columns from the initial dataset, transforming categorical data into integers

b) Build the kNN classifier using the training and test sets generated based on the method tried as part of 1a**;**

c) Build the Logistic Regression (LR) classifier using the training and test sets generated based on the method tried as part of 1a**;**

d) Build the Random forest classifier (RF) using the training and test sets generated based on the method tried as part of 1a;

e) Build the Support Vector Machines (SVM) classifier using the training and test sets generated based on the method tried as part of 1a**;**

f) Input the (kNN, LR, RF and SVM) models into the Voting Ensemble using the training and test sets generated based on the method tried as part of 1a**.**

[20 Marks]

2nd Task: Tuning and Performance Measurement

a) Tune the parameters, using Grid search and check for improvement. Tuning parameters value, for machine learning algorithms effectively improves the Voting Ensemble model performance.

b) Focus on the voting classifier type by introducing and performing seven common performance metrics used in a classification project. Briefly explain the importance of the obtained metric values/graphs. The seven metrics are as below:

1. Accuracy score
2. Confusion matrix
3. Precision
4. Recall
5. F1 Score
6. ROC Curve
7. AUROC

[12 Marks]

REFERENCE:

Moro, S., Rita, P. and Vala, B., 2016. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, 69(9), pp.3341-3351.

B. Predicting the Price of NETFLIX Stock with LSTM Neural Networks

Build a Python program that can predict the price of a specific stock. This project is a great example of applying machine learning in finance.

As mentioned in the subtitle, we will be using the stock price history of **Netflix, Inc. (NFXLX)**, a streaming company.

We can get the data for free at *Yahoo.Finance*. We need the daily stock data from May 23, 2002, till now (April *__*, 2022). See Table 1.

The full dataset can be downloaded from here.

Time Period: May 23, 2002 - Feb 11, 2022 Show: Historical Prices Frequency: Daily Apply

Currency in USD Download

Date	Open	High	Low	Close*	Adj Close**	Volume
Feb 10, 2022	402.10	408.00	396.36	406.27	406.27	8,452,915
Feb 09, 2022	408.65	412.98	398.79	412.89	412.89	7,738,200
Feb 08, 2022	398.18	406.61	395.83	403.53	403.53	6,818,500
Feb 07, 2022	410.17	412.35	393.55	402.10	402.10	8,232,900
Feb 04, 2022	407.31	412.77	396.64	410.17	410.17	7,782,400
Feb 03, 2022	421.44	429.26	404.28	405.60	405.60	9,905,200
Feb 02, 2022	448.25	451.98	426.48	429.48	429.48	14,346,000
Feb 01, 2022	432.96	458.48	425.54	457.13	457.13	22,568,100
Jan 31, 2022	401.97	427.70	398.20	427.14	427.14	20,047,500

Table 1: Preview NETFLIX Stock Data

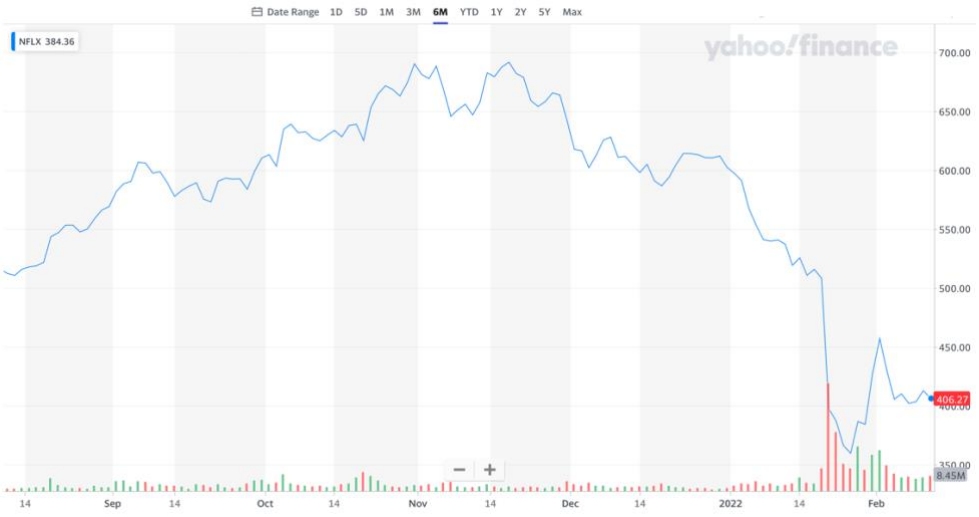

Image 2: NETFLIX latest 6 months stock price chart

Note: Read a CNN article about the recent Netflix stock crush (See Image 1) here.

Tasks

1. **Data Visualisation:** Define a simple line chart to give an idea of the stock price change in the last year**.**
2. **Build the Model:** Define the Long Short-Term Memory model (LSTM)
3. **Define the Train and Test Data:** This step covers the preparation of the train data and the test data
4. **Prediction Function and Result:** In this step, we are running the model using the test data we defined in step four. Visualise the predicted versus the actual stock values for the specific time period

[18 Marks]

Guidelines:

You are required to deliver a report (max 30 pages including all figures) describing the methods adopted and the discussion of achieved results with reference to the tasks listed below. Assume that the report is targeted to a *marketing strategist*, who is interested to learn the business insights inferred in your analysis and to receive suggestions on how to take appropriate actions.

Marking Scheme

Due to the nature of the assessment candidates may come up with more than one equally, good solutions. Thus, marks will be allocated as follows

A. Building an Ensemble

1st Task: Voting Ensemble

You need to construct, train and test a Voting type classifier in Python, based on ((kNN, LR, RF and SVM)

a) With the aid of the Python package visualise and justify the properties of the given Facebook Metrics dataset. Your analysis must include reference to the following items
   - Identification and treatment of any Missing Values;
   - Identification and treatment of any Outliers;
   - Data Normalisation;
   - You need to try 2 different methods of forming training and test sets.
   - Exploratory Data Analysis

[3 Marks]

b) Build the kNN classifier using the training and test sets generated based on the method tried as part of 1a**;**

[3 Marks]

c) Build the Logistic Regression (LR) classifier using the training and test sets generated based on the method tried as part of 1a**;**

[3 Marks]

d) Build the Random Forest classifier (RF) using the training and test sets generated based on the method tried as part of 1a;

[3 Marks]

e) Build the Support Vector Machines (SVM) classifier using the training and test sets generated based on the method tried as part of 1a**;**

[3 Marks]

f) Input the (kNN, LR, RF and SVM) models into the Voting Ensemble using the training and test sets generated based on the method tried as part of 1a**.**

[5 Marks]

[20 Marks]

2nd Task: Tuning and Performance Measurement

a) Tune the parameters, using Grid search and check for improvement. Tuning parameters value for machine learning algorithms effectively improves the Voting Ensemble model performance.

[5 Marks]

b) Focus on the voting classifier type by introducing and performing seven common performance metrics used in a classification project. Briefly explain the importance of the obtained metric values/graphs. The seven metrics are as below:

1. **Accuracy score** [1 Mark]
2. **Confusion matrix** [1 Mark]
3. **Precision** [1 Mark]

4. **Recall** [1 Mark]
5. **F1 Score** [1 Mark]
6. **ROC Curve** [1 Mark]
7. **AUROC** [1 Mark]

[12 Marks]

B. Predicting the Price of NETFLIX Stock with LSTM Neural Networks

Tasks

1. **Data Understanding & Manipulation**: useful as a preliminary step to capture basic data properties. Justify suitable transformation of variables and elimination of redundant variables.

   [3 Marks]
2. **Data Visualisation:** Define a simple line chart to give an idea of the stock price change in the last year**.**

   [3 Marks]
3. **Build the Model:** Define the Long Short-Term Memory model (LSTM) and clearly explain the input features as a function of time lag.

   [4 Marks]
4. **Define the Train and Test Data:** This step covers the preparation of the train data and the test data

   [4 Marks]
5. **Prediction Function and Result:** In this step, we are running the model using the test data we defined in step four. Visualise the predicted versus the actual stock values for the specific time period

   [4 Marks]

[18 Marks]