ML Regression Assignment

Introduction

Video content is the most important data type on the information traffic across the Internet. This deluge of data together with resource limitations (as network bandwidth or device computation capabilities) demands the use of compressed video formats. Video transcoding is the process of modifying the compressed video representation from one format to another and represents a key element when quality standards must be ensured.

Transcoding is a daunting task in terms of computational resources. Therefore, the estimation of transcoding processing time is a piece of valuable information on developing suitable strategies for video traffic, and in consequence on the improvement of user experience.

In this assignment, you will use regression techniques to predict the transcoding time of a video based on several parameters (duration, bitrate, size, codec, etc) of both, input and output compression format.

Dataset Description

You will receive two datasets containing a list of videos with their respective information.
There is a total of 17836 records and 22 explanatory variables divided into two datasets.

- `model.csv:` the dataset contains the information of 12000 videos with the respective target variable. You must use this data to create and evaluate your model.
- `predictions.csv:` the dataset contains the information of 5836 videos without the target variable. You must provide the predictions for this set of records.

The task is formulated as a regression problem. Your grade will be based on the Mean Absolute Error (MAE) of your predictions together with the modeling process presented in your report.

Target :

The target attribute is the total time for transcoding and called `utime`.

Attribute Information:

n	Attribute	Type	Values
1	id	numerical	Video ID
2	duration	numerical	duration of video
3	codec	categorical	coding standard used for the video
4	height	numerical	height of video in pixels
5	width	numerical	width of video in pixels
6	bitrate	numerical	video bitrate
7	category	categorical	YouTube video category
8	frame rate	numerical	actual video frame rate
9	i	numerical	number of i-frames in the video , i-frames are complete images.
10	p	numerical	number of p-frames in the video, p-frames contain only the differences from the previous frame.
11	b	numerical	number of b-frames in the video, b-frames store just the differences from previous/following frame.
12	frames	numerical	number of frames in video
13	i_size	numerical	total size in byte of i videos
14	p_size	numerical	total size in byte of p videos
15	b_size	numerical	total size in byte of b videos
16	size	numerical	total size of video
17	o_codec	categorical	output codec used for transcoding
18	o_bitrate	numerical	output bitrate used for transcoding
19	o_framerate	numerical	output framerate used for transcoding
20	o_width	numerical	output width in pixel used for transcoding
21	o_height	numerical	output height used in pixel for transcoding
22	umem	numerical	total codec allocated memory for transcoding
23	**utime**	**numerical**	**total transcoding time for transcoding**

Submission Instructions

1. Model Training Data Release: 22 March 2021, 20:00.

2. Description of analysis on the training set and model identification: 31 March 2021 20:00.
You are asked to kindly send an email with the following supporting information:

a) A **brief report** of the step-by-step methodology (i.e. pre-processing, visualization, training, testing, etc.) that you have followed to develop your model, this document must illustrate the motivation behind your selected approach.

• File Format: .pdf • Filename: surname1_surname2_surname3.pdf (e.g. orsenigo_soto.pdf)

b) **The commented python code** that you used in your model. Comments in the code must ensure that the code is easy to follow.
• File Format: .ipynb, .py • Filename: surname1_surname2_surname3 (e.g. orsenigo_soto.ipynb or orsenigo_soto.py)

3. Prediction Data Release: 01 April 2021 18:00.

4. Prediction Submission: 05 April 2021 20:00.
You are kindly requested to strictly follow the described submission guidelines:
• File Format: .csv
• Filename: surname1_surname2_surname3 (e.g. orsenigo_soto.csv)
• Column Format: **A single** column named "target"
• Row Format: Your predictions with **the same number of rows and in the same order** as the prediction test set.

Further Instructions

• **Any submission that does not respect the guidelines (submission after deadline, empty file, wrong student code) will not be graded**.