Name: Ghodrat Rezaei
Personal Code: 952128
Date: 16/12/2020

Machine Learning - Regression Assignment Report

Politecnico di Milano
Prof. Carlo Vercellis

Introduction

This report is aimed at explaining the methodology that was chosen to analyze and do the regression for a data set concerning video transcoding. Regression techniques were used to predict the transcoding processing time. The approach that was followed in order to reach the training and test results are briefly presented in the following pages, along with how the data was processed.

Preprocessing and Data Preparation

As the first step in the data processing and preparation, rows with missing data were checked to be removed, but no missing rows were found. Then numerical and categorical data were split. Categorical data are just codec, the category and the o_codec. All their values were counted and plotted against utime, which is our target parameter to understand how much they are correlated. Then all categorical parameters were treated as dummies.
All other parameters are numerical data, and they were presented in histograms. An outlier detection function was applied to identify and remove any outliers, but none were found. From the histograms we can see that the duration, i, frames and p follow a log shape. Hence, the log of these parameters was taken, and their original parameters were dropped along with the id. The correlation between all variables was applied. Any two variables having a correlation more than 0.95, one of them was removed (fig. 1).

It ended up that log_p, height, size and o_width were dropped.

Finally, as the last step of preprocessing all variables were standardized along with the target variable (utime) and after that utime was also dropped.

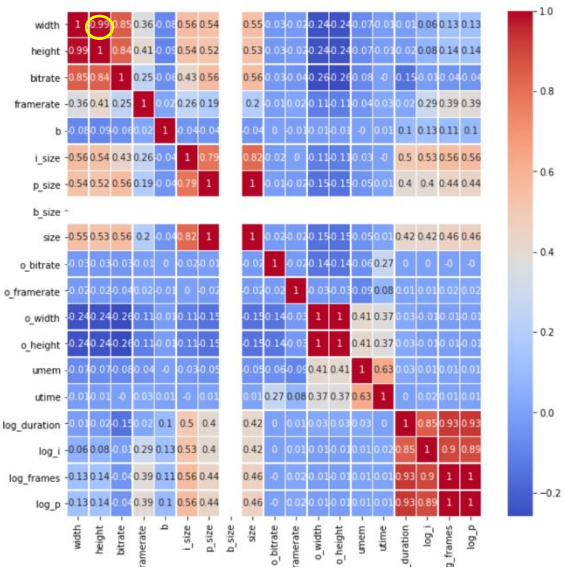

Figure 1

Training-Test Sets

The data was split then into a training and test set to apply the different classification methods and test their accuracy. By default, the test set was set to be 30% of the total data. For each model, we defined the Model Structure itself and Parameterization space by which the greedy search(gs) approach can iteratively examine through the defined space of parameters and pick up the best parameter which can give us the minimum amount of error (Objective_Function(J)). By defining the cross-validation (cv) equal to 3 in greedy search, probability of over-fitting will decrease significantly.

Regression Models

Table 1

Model	Mean Absolute Error (Train Set)	Mean Absolute Error (Test Set)
Random Forest	2.098	2.997
SVM	3.032	3.819
Gradient Boosting	3.927	4.218
Decision Tree	4.896	5.254
KNN	5.329	6.755

Ridge Regression	8.975	9.067
Ada Booster	8.987	9.150
Lasso Regression	9.050	9.072
Linear Regression	9.302	9.370

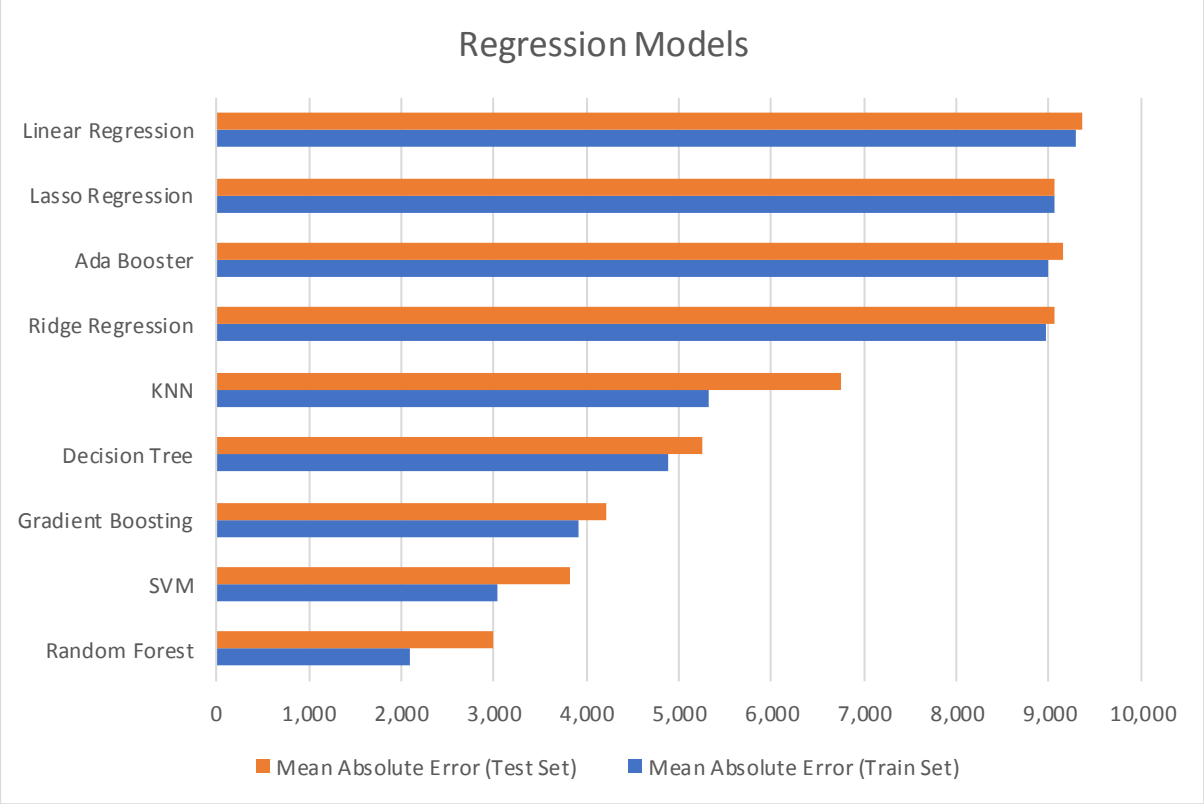

Figure 2

Different regression models were applied to the final standardized data. Looking at the above results we can see that the random forest model is giving us the lowest possible mean absolute error (MAE), while the leaner regression is giving the worst results. The negative MAE was applied to directly see which is the highest results. Due to the slight difference between the train and test results in the random forest there's a slight overfitting, which can be neglected.

Linear Regression Analysis

Plotting the error against the normal distribution for the linear regression, we can see that the error is not following a perfect normal distribution and heteroscedasticity is present (fig. 3, fig. 4 and fig. 5), which is a result of the large range of observed values or due to the presence of outliers. The model that was followed to remove outliers was from the 0.5 to the 0.95 quantile. Maybe increasing the scale for the outlier removal from the 0.25 till the 0.75 quantile would fix the residuals to an extent. Due to the presence of heteroscedasticity, the p-values are relatively small (6.191366558121906e-78 for the Kolmogorov-Smirnov Test and 0 for the D'Agostino Test).

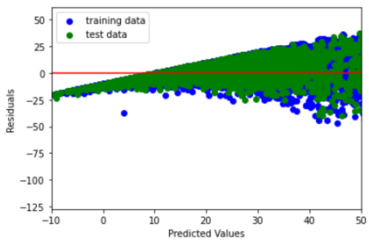

Figure 3

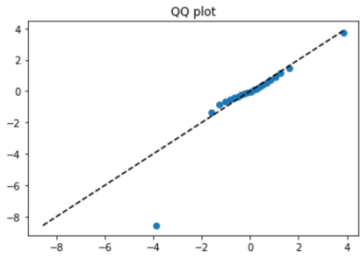

Figure 4

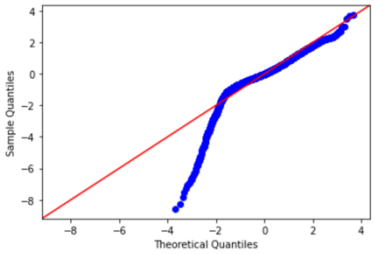

Figure 5

Also looking at the confidence interval for the different variables we can see that some of them can have a value that ranges from negative to positive values, which is not a very determined coefficient with a specific effect toward the target (fig. 6).

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.3284	0.405	23.012	0.000	8.534	10.123
codec_flv	4.5430	0.477	9.517	0.000	3.607	5.479
codec_h264	1.4145	0.436	3.243	0.001	0.559	2.269
codec_mpeg4	0.1772	0.559	0.317	0.751	-0.919	1.273
codec_vp8	3.1937	0.529	6.032	0.000	2.156	4.232
category_Autos & Vehicles	0.3186	0.765	0.417	0.677	-1.180	1.818
category_Comedy	0.4389	0.624	0.703	0.482	-0.784	1.662
category_Education	-0.4480	0.825	-0.543	0.587	-2.065	1.169
category_Entertainment	-0.0377	0.477	-0.079	0.937	-0.972	0.897
category_Film & Animation	0.8597	0.811	1.060	0.289	-0.729	2.449
category_Gaming	-0.2890	0.506	-0.571	0.568	-1.282	0.704
category_Howto & Style	1.6437	0.998	1.648	0.099	-0.312	3.599
category_Music	-0.0030	0.446	-0.007	0.995	-0.877	0.871
category_News & Politics	0.0105	0.770	0.014	0.989	-1.499	1.520
category_Nonprofits & Activis	-0.8840	1.385	-0.638	0.523	-3.598	1.830
category_People & Blogs	0.5594	0.393	1.425	0.154	-0.210	1.329
category_Pets & Animals	0.5622	0.939	0.599	0.550	-1.279	2.404
category_Science & Technology	-1.0982	1.055	-1.041	0.298	-3.166	0.970

Figure 6

Random Forest Analysis

In Random Forest model the residuals follow more the normal distribution shape, except for a few points, where they are far (fig. 7, fig. 8 and fig. 9). This explained by the same concepts explained above in the linear regression model analysis. Also, the p values were not that high (8.536372510233588e-184 for the Kolmogorov-Smirnov Test and 0 for the D'Agostino Test), because some data have not been scaled properly during the standardization.

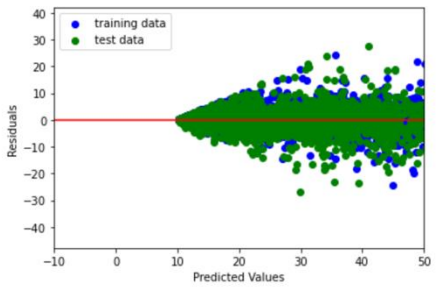

Figure 7

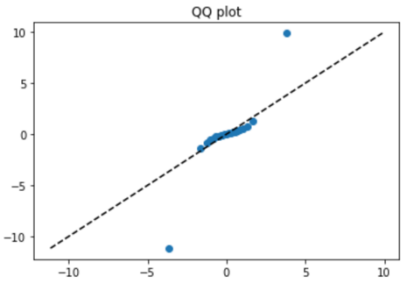

Figure 8

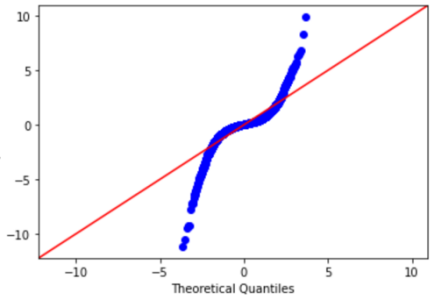

Figure 9