FACULTY OF SOCIAL SCIENCES
Department of Economics
University of Copenhagen

Social Data Science Exam

Exam numbers: 47, 64, tzp831

Pornhub penalized by LASSO

Using machine learning tools to predict number of views for a pornographic video

ECTS points: 7.5

Date of submission: 01/09/2018

Keystrokes: 42,000

Abstract

In this paper we aim to determine the machine learning model most fit for predicting the number of views for a pornographic video. Next, we test if the prediction of the model can be enhanced, when features from thumbnail images are added as features.
We employ public data from www.kaggle.com through the API. A bag-of-words model is used both on the titles of the videos and the categories. Only videos with English titles are used, and we control for the presence of either a male or female name in the title. OLS, LASSO, the elastic net and a random forest regression is used to prediction. The choice of hyperparameters is made based on a 10-folds cross validation.
The best performing model based on both computational time and the mean squared error is the LASSO. However, we are not able to conclude that the performance of the LASSO is enhanced by the thumbnail features.

Contributions
Joint: 1, 6
47: 2.3, 3.2, 3.3, 4.4, 5.2
64: 2.2, 3.1, 4.3, 4.5, 5.1
tzp831 : 2.4, 3.4, 4.1, 4.3, 5.3

Contents

1 Introduction

The Maslow Hierarchy of Needs early on established that sexuality plays a significant role in the well-being of humans. With the increasing access to, and use of pornography, Bridges and Morokoff (2011) argue for the importance of further analysis of the effect of *sexual media*. In general, the field of pornography lacks attention - academically speaking.

Furthermore, it should be noted that the estimated number of visitors on web pages containing pornography is higher than the number of visitors on NetFlix, Amazon and Twitter combined. The pornography industry was in 2016 estimated to have a net worth of more than the Major League, NFL and the NBA combined. That is a net worth of just below 100 billion dollars. The estimated spending per second on pornography is 3,000 dollars. Lastly, 12 pct. of all content on the internet is pornography.[1]

While these may all be fun facts, it also highlights just how major an economy pornography is. And it further stresses the importance of research in this field. Most studies on the subject of pornography is not within the field of economics and is often studies performed on small survey data sets, such as that of Bridges and Morokoff (2011) or Lair and Brand (2016). Lair and Brand (2016) use a sample of 80 male participants and rely solely on questionnaires, which potentially rises generalization issues. Furthermore, due to the delicate nature of use of pornography, the respondents may not answer completely truthfully.

In this assignment we set out to test if different popular methods from machine learning are able to predict the number of views for a pornographic video, and whether we can enhance the performance of the models using features from the thumbnails.

We do not claim that this analysis can explain the human sexuality. But we do show the methods are applicable and obtain reasonable results. And our contribution should mainly serve as a motivation for further analysis in the subject. Public available and relatively easily processed data is employed in this analysis, but more time would greatly enhance the possibilities of matching large data on pornographic videos with geographical data.

While there are some studies available on the effect of pornography on the individual and its relations with others, we are not able to recover any studies describing the economic behavior.

In this assignment we use public available information on videos from www.pornhub.com. The data contain information on the length of the video, the quality, in which categories it belongs, the tags associated with the video and the voting of the video, i.e. the share of people who ‘liked’ the video compared to disliked the video. Furthermore, a URL for the thumbnail is given. This allows us to extract an approximate date for the upload of the video. We aim to predict the number of views for a video. The model employed in the assignment are the OLS, LASSO, elastic net and a random forest regression. LASSO, elastic net and random forest regression greatly outperforms OLS, which performs very poorly probably due to overfitting. Considering computational time, the LASSO outperforms the other models, as the computational time is approximately one fifth of that of the elastic net and less than on tenth of that of random forest regression. Lastly, we investigate if some simple features from the thumbnails are able to enhance the performance of our LASSO model. However, we find that the performance of the model does not change much compared to the performance

[1] https://medium.com/@Strange_bt_True/how-big-is-the-porn-industry-fbc1ac78091b

without features of the thumbnail. We do not present the estimated weights, as a causal interpretation is intended, and the main goal is prediction.

The remainder of the assignment is ordered as follows; section 2 presents the data used in the analysis, how it is obtained, cleaned and features are computed. Section 3 briefly reviews the models employed and section 4 presents the implementation and the main results of the analysis. Section 5 discusses the results and section 6 concludes.

2 Data

2.1 Raw Data

This paper is based on a data set that has been downloaded from www.kaggle.com. It contains 191,521 observations and 10 variables, see table 2.1.

Table 2.1: Variables in data set

Variable	**Description**
Unnamed: 0:	Video ID number
imp_source:	Image source to thumbnail (URL)
length:	Length of the video (in seconds)
nb_views:	Count of video views
quality:	The quality of the video (HD or LOW quality)
title:	The title of the video
video_link:	Video source (URL)
voting:	Percentage of good ratings
categories:	All categories that the video is tagged with.
tags:	Tags to identify what the video is about.

Every observation in the raw data set is a porn movie scraped from www.pornhub.com. The data set was uploaded by the user zelhassn 10 months ago and can be found by this link.[2]

The data set has been obtained by using Kaggle's Public API.[3] Kaggle has made this very easy by wrapping their API in a command-line tool (CLI) implemented in Python. In our paper we illustrate how to work with APIs as well as writing bash commands in our Jupyter notebook. [Pretty kewl.]

From now on, this data set is referred to as our raw data set, and will be employed in the further analysis.

[2] https://www.kaggle.com/ljlr34449/porn-data

[3] https://www.kaggle.com/docs/api

2.2 Data preparation

This section presents the methods applied in the process of cleaning the data.

2.2.1 Language Detection

25 pct. of the observations in the raw data set contains titles in many different languages including German, Spanish, Korean, Chinese etc.. To work around future text cleaning and data wrangling problems this paper is solely working with observations that has a English title. This is implemented by using the Python module *detect_langs*. It leaves us with a subset of 144,661 observations that will be employed in the further analysis.

2.2.2 Bag of words

The data used in this assignment contains text, e.g. title of the video and categories for the video. These variables potentially contain valuable information, which we want to give to the models to better predict number of views. The bag-of-words model allows us to present text as numerical feature vectors fit for machine learning models. The bag-of-words model is quite simple in its form. First, a vocabulary of tokens is created. Here one can choose to remove stop-words or other irrelevant words. Second, by observation a count for each token is generated and included as a feature vector. The *CountVectorizer()* from the sklearn.feature_extraction.text package is used to count the total number of occurrences and generating the vocabulary and the final array.

2.2.3 Age

The variable is the number of days since the video was uploaded measured in days. To compute the age of the video we need to know when the video was uploaded to the web page. However, this information is not available in the data. But the data contain the uploading date of the thumbnail. Therefore, the date is only an approximation of the date the video was uploaded. The date for the thumbnail is a part of the URL for the image source. The dates are extracted by converting the URLs into strings and then using regex. Finally, the age of the video is computed as the number of days since the upload.

2.2.4 Dummies for each category

Each video can be in more than one category. In raw data the categories connected to each video is presented as at string that contains all relevant categories. Therefore, it is necessary to extract the categories from the string by using the regex with the pattern $([a-z]+\backslash s[a-z]+)|([a-z]+)$, and split by "__". Then we created a bag-of-words of categories to generate a dummy variable for each category possible.

2.2.5 Dummies for each word in title

The dummies for each word is generated through the bag-of-words model. Before using bag-of-words some words was deleted. First, the module stopwords from nltk.corpus is imported, and 218 stop words are deleted from the titles. Then all names from the name lexicons (see section below) are deleted from the titles too. The effect of the name will be captured by the dummies for male and female names.

2.2.6 Male and female name dummies

Dummies are generated for the title containing a male or female name respectively. To generate these features, we found a female names lexicon and a male names lexicon.[4] These lexicons are converted into lists. The lexicons are cleaned for names identical to a video category. For example, 'Latina' is both a female name and the name of a category on the web page. We found that also other words in the lexicons might be an issue. For example, the word 'dick' is also in the lexicon of male names. However, we expect it to be used mainly as a noun and not a name in a title pornographic video. To remove some of these cases we printed the matches between the lexicons and the titles for a random sample of $5,000$ titles. From the output 52 words were chosen manually on what names might not be used as a name in the titles. The list of these words can be found in appendix. To generate the dummies a loop is written. The loop loops over the words in each title and appends '1' to the dummy if any of the words in each title matches any of the names in the lexicon. This is done with the male and female lexicon separately.

2.2.7 Quality

The raw data contains a variable with information on the quality of the video.
The get.dummies() function from the pandas package is used to generate dummies for HD and low respectively, and we chose to preserve the HD-dummy indicating a video of higher quality.

2.2.8 Tags

The information from the tags in the raw data is not included in the analysis. Tags describe the videos with subcategories such as wild, train and links to third party contributors. This decision is based on the expected high correlation with the categories, but mainly due to the poor quality of the data after examining the data.

2.3 Descriptive analysis

Figure 2.1 shows the histogram of the log of numbers of view. The distribution is slightly skewed to the right with a mean of 12.15. The mean of the number of views is 420,952. We opt for the log transformation due the highly right skewed distribution of the number of views, where the median observation is less than half the mean. The voting is distributed from 60 pct. to 92 pct., with the majority of videos being rated from 75 pct. and up. The average voting is 77.38 pct.

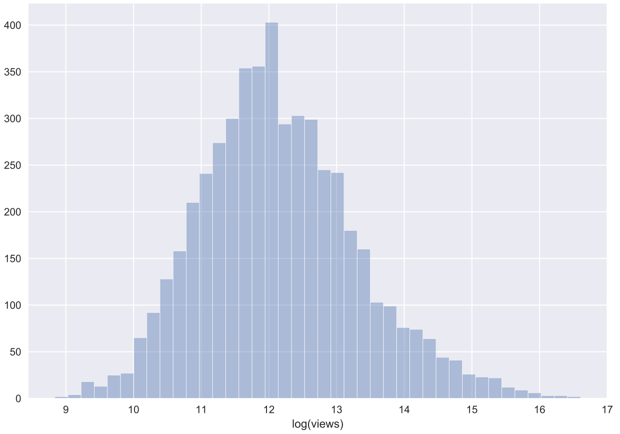
[4] https://www2.census.gov/topics/genealogy/1990surnames/

Table 2.2: Selected descriptive statistic on pornographic videos

	Length	**Views**	**Voting**	**Age**
Count	5,000	5,000	5,000	5,000
Mean	792	420,952	77,38	1,279
Std	750	887,106	6.61	751
Min	7	6,880	60	24
25%	375	85,562	73	646
50%	573	172,040	78	1,239
75%	934	379,912	82	1,841
Max	10,873	16,300,000	92	3,516

Figure 2.2 show the scatter plot between the log of views and days since the video was uploaded and the voting respectively. There does not appear to be any clear patterns in the plot. The relatively newly uploaded videos (< 500 days old) may exhibit slightly greater variation in the log of number of views. The smallest number of views among the videos uploaded more than 2,500 days ago appears larger than the newly uploaded videos.

Figure 2.1: Histogram for log to number of views



Also the variation of log to the number of views across voting does not exhibit clear tendencies. The videos rated from 75 pct. and up seem to have a very similar distribution. It can be noted, that the videos rated to less than 70 pct. seem to have smaller variation and in general have fewer views, which may not come as a surprise.

Figure 2.2: Scatter plots of log to number of views against age and voting of the video

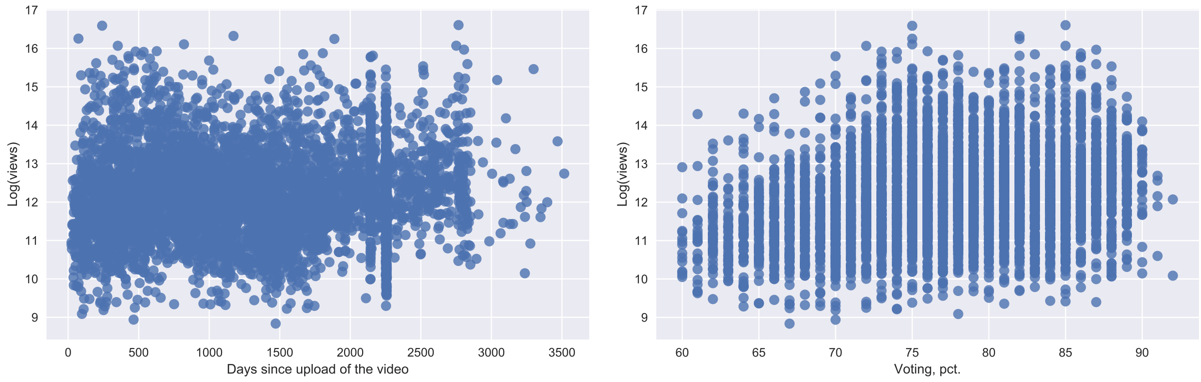


Table 2.3: Male and female names in titles

	Name	**No name**
Female	26.5%	73.5%
Male	3.3 %	96.7%

Some titles include either a male or a female name. In table 2.3 the percentage share of titles with and without names are presented. It seems that it is much more common to include a female name than a male name in the title of a video. The figure states that 26.5 percent of the videos includes a female name in the title while only 3.3 percent of the videos includes a male name in the title.

Table 2.4: Top 10 categories and words in titles

	Word	**Count**	**Category**	**Count**	**Percentage share**
1	Cock	403	German	1,864	37%
2	Hot	371	Parody	1,744	35%
3	Fucked	352	Amateur	1,069	21%
4	Pussy	337	Smalltits	979	20%
5	Teen	326	Bigdick	699	14%
6	Ass	285	Anal	510	10%
7	Big	285	Babe	505	10%
8	Sex	249	Gangbang	504	10%
9	Blonde	248	Massage	441	9%
10	Anal	247	Brazilian	422	8%
Total	Words: 5,158		Categories: 86		

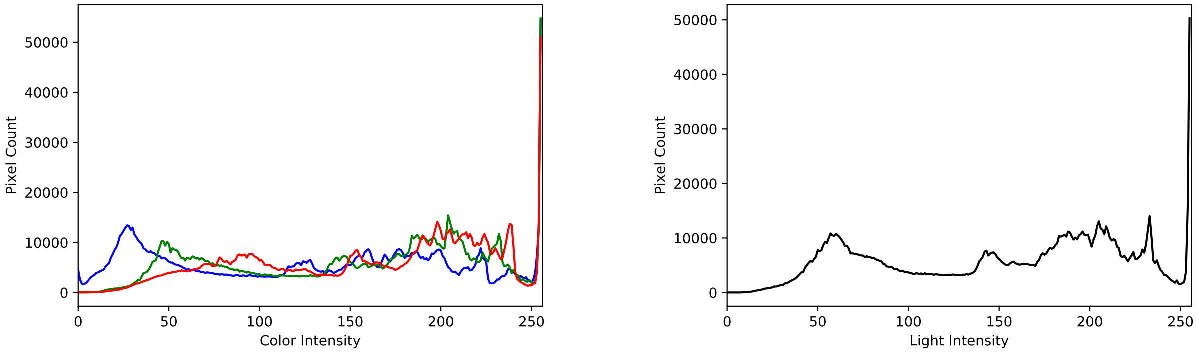
In the sample of 5,000 observations certain categories and words in the titles are used repeatedly. In table 2.4 the top 10 most used categories and words are presented. In the sample 86 different categories are observed, and 5,158 different words are used in the video titles. With a count of 1,864 videos equivalent to 37 percent, ‘german’, is the most common category in the sample, followed by ‘parody’ with a percentage share of 35. With a count of 403 repetitions, ‘cock’ is the most used word in the video titles. Followed by ‘hot’ repeated 371 times.

2.4 Thumbnail Images

The raw data set contains a URL to the thumbnail image for every video. To enrich our data set, new data has been generated from these thumbnails on a sample of 250 random observations.

First a thumbnail has been downloaded. Secondly we have gone through every pixel in the image to extract the RGB color space. 2.3 illustrates how the colors are distributed for each pixel.

Figure 2.3: Thumbnail Analysis



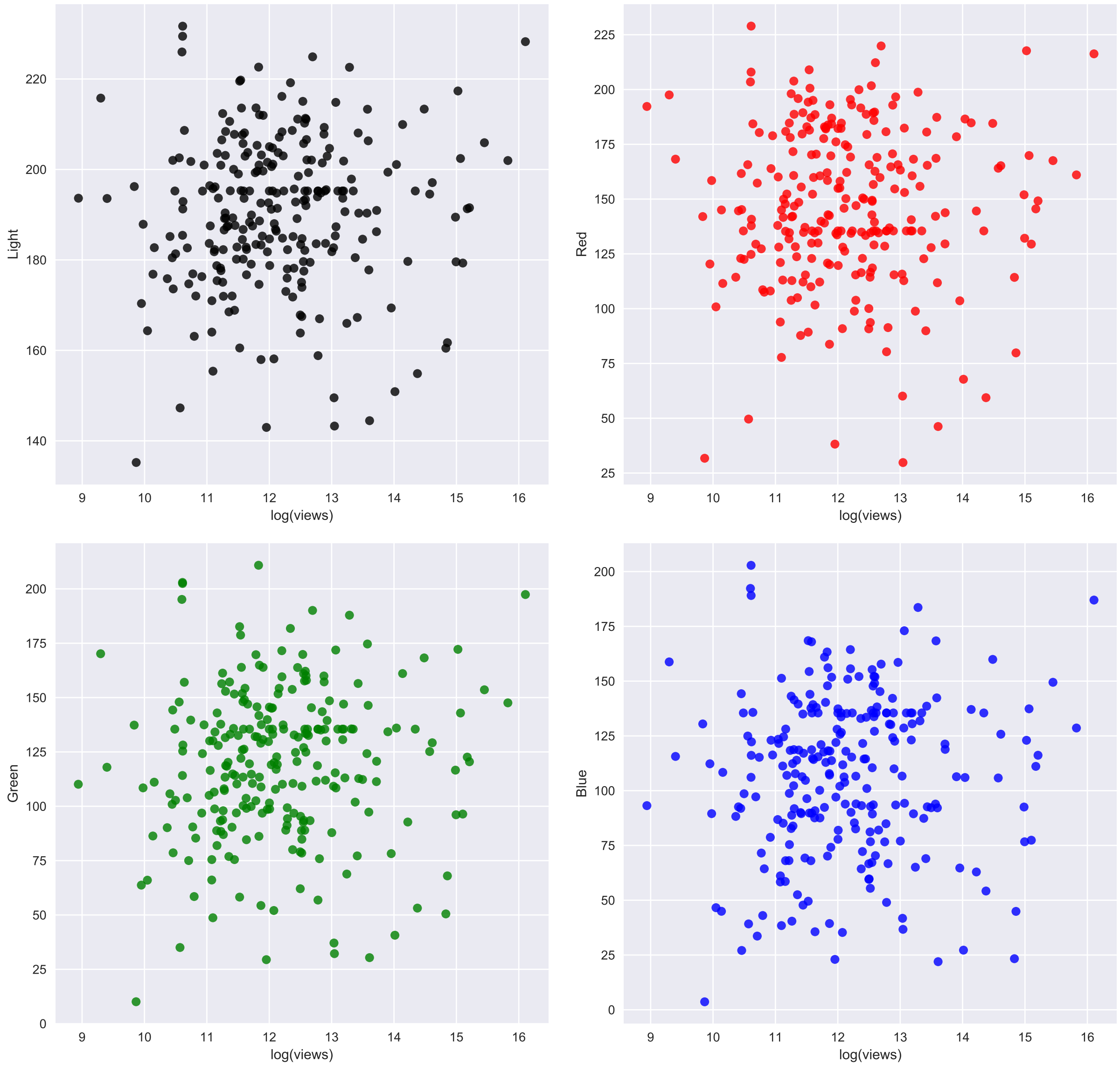
Note: Each color in the RGB color space is an integer in the range from 0 to 256.

Calculating the RGB score for every pixel also makes it possible to calculate the average score for each of the three colors and investigate the correlation between the amount of a color in the thumbnail and the number of views the video got. The three color plots are shown below on 2.4.

The analysed image has also been greyscaled and plotted as shown in 2.3. It is illustrating the correlation between how much light there is in the thumbnail and the number of views. Again it is a simple histogram where pixel to the left are darker and pixels to the right are lighter. The average has been calculated on the histogram and is referred to as the light score or simply how much light there is in the picture on average. This has also been plotted in 2.4 to investigate correlation.

Please notice that the number of views has been transformed with log to accommodate that is skewed to the right.

Figure 2.4: Scatter plot between the average level of light, red, green and blue against log to the number of views



The scatter plot plots the amount of red, green and blue in the thumbnail and the level of light against log to the number of views. The plots do not show a specific pattern. That means that the amount of each color in the thumbnail does not affect the number views, nor does the amount of light.

3 Models

In this section we briefly review the three estimators of linear model, OLS, LASSO and the elastic net. Lastly, the nonlinear estimator the random forest regressor is reviewed. The review is based on *Python Machine Learning* by Raschka and Mirjalili, from here on PML. The linear model to be estimated is of the form

$$Y = X\beta + \epsilon \tag{3.1}$$

Where $Y = (y_1, ..., y_n)'$ is a $n \times 1$ vector and X is a $n \times m$ matrix containing the features. ϵ is the $n \times 1$ vector of errors and β is the $1 \times m$ vector of weights to be estimated. The number of observations is n and the number of features are m including the bias term.

3.1 Ordinary Least Squares

The workhorse estimator Ordinary Least Squares, OLS, will serve as a benchmark estimator throughout the paper, and several of the estimators later employed are extensions of the well known OLS estimator. The objective function to be minimized in OLS is

$$J(\beta)_{OLS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.2}$$

where

$$\hat{y}_i = x_i\hat{\beta} \tag{3.3}$$

Thus, the objective in the OLS is to minimize the sum of squared errors. This problem has an analytical solution where there is no case of perfect multicollinearity between any of the regressors.

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3.4}$$

A mean to evaluate the performance of an estimator when the outcome in continuous, is to compute the mean squared error, MSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.5}$$

Since this is the, slightly transformed, objective of the OLS, OLS will naturally perform well in sample but tends to overfit the data. This can potentially lead to a low MSE in sample, but poor out-of-sample performance. The problem of overfitting, induced by highly complex models, is likely to lead to high variance. The counterpart of the overfitting is underfitting. A problem, which is likely to be caused by using a too simple model, which in turn yields high bias. The trade-off between variance and bias can be balanced via the machine learning method regularization. The problem of of overfitting is further induced by the data employed in the assignment. We opted for data containing text and a bag-of-words approach which leads to a high number of features in the data.

3.2 LASSO

The Least Absolut Shrinkage and Selection Operator is an extension to OLS. LASSO regularization is employed to prevent overfitting, especially motivated by the large number of features in the dataset. LASSO has proven useful to filter out noise and handle a high level of collinearity (PML p. 74) among the features of the model. This is done by introducing a penalty on all non-zero coefficients.

$$J(\beta)_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{m} |\beta_j| \tag{3.6}$$

The penalty is the absolute sum of coefficients weighted by the regularization parameter λ. The penalty in the LASSO regularization yields corner solutions and thus, sparse solutions. This essentially means, that we employ LASSO as a method for selecting the most important features and discard irrelevant features. The size of the regularization parameter is used to balance the bias-variance trade-off. A too small parameter will yield too many included features and thus, a model prone to overfitting. Whereas a too large hyperparameter will punish the weights too harshly and too many weights will be set to zero. This will lead to underfitting, and a model which is likely to suffer from a high bias.

However, the LASSO has some shortcomings; Zou and Hastie (2005) present three main issues to consider. Firstly, the case where the number of features in the model is larger than the number of observations. Here the LASSO will 'saturate' and at most select n features with non-zero coefficients.
Secondly, if the dataset containing features with a high pairwise correlation, the LASSO tends to randomly select one of the features.
Thirdly, the LASSO is outperformed by ridge (not reviewed in this paper) in cases with a high degree of correlation between features, even when the number of observations exceeds the number of features in the model. Zou and Hastie (2005) argue that especially the first two shortcomings make LASSO unfit as a method to select variables in some cases. We have a large number of variables compared to the number of observations used (especially when cross-validation procedure is considered). Furthermore, we expect a high degree of correlation between several of the features, such as words in the title and the category. We opt for the extension of the LASSO, called the elastic net, to overcome some of the challenges.

3.3 Elastic Net

As mentioned, the elastic net is an extension to the LASSO. The objective functions are similar but the elastic net is a combination of the objective function known from LASSO regularization and the objective function employed in ridge regularization.

$$J(\beta)_{ElasticNet} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{m} |\beta_j| + \lambda_2 \sum_{j=1}^{m} \beta_j^2 \tag{3.7}$$

This method was proposed by Zou and Hastie (2005) to overcome some of the pitfalls associated with the LASSO. The objective function now includes two penalties: The first is identical to the one employed in LASSO, this preserves the ‘selection ability’ of the LASSO while encouraging grouping effects (Zou and Hastie (2005)), by introducing a penalty from the sum of squared coefficients. Zou and Hastie (2005) show, that this improves the behavior of the LASSO in cases where data is highly correlated.

3.4 Random Forest Regression

As opposed to the linear models reviewed in the previous sections, lastly, we opt for a model, which is able to deal with non-linearities. The random forest consists of an ensemble of decision trees (PML, p. 339). The decision trees are used to split the samples into smaller samples, where the linear function is fitted. This is conceptually very different from the global linear models. The random forest is generally less prone to overfit the data than a single decision tree. Another of the benefits of the random forest regression is, that is more robust towards outliers in data. Due the (approximately) continuous nature of the outcome variable, the MSE can be employed as the impurity measure and therefore, we can easily compare the performance of the estimators.

4 Results

4.1 Polynomial features

Due to the large number of features in the dataset, we do not include polynomial features for all the features in the analysis, since this would greatly complicate the computations and add on to the computational time. However, to allow for some non-linearities in the linear models employed, we include second order polynomial features of three features describing respectively the length and voting of the video and the number of days since upload (as from 1/1 2018). This allows for convex/concave relationships between (log of) the number of views and the three variables, e.g. the weight on length is positive, but diminishing and turns negative at some point.

4.2 Implementation of the models

Due to time and computational limitations we employ a random sample of 5,000 observations throughout the analysis.

When using features on the thumbnails we employ a random sample of 250 observations. This is due to a limitation on requests to pornhub’s server. All variable are normalized before the models are run. We opt for at split of $70 - 30$ pct., that is, we employ 70 pct. of the data to training data and evaluate performance of the models on the remaining 30 pct. A k fold validation method is used to validate the choice of hyperparameters. This is implemented via GridSearchCV() from the sklearn.model_selection module. We set $k = 10$ as recommended in the course, such that the training data is split in 10 folds, where nine are used to train the model and one is used to evaluate the performance. This is repeated 10 times and the average of the performance is computed and the optimal hyperparameters are returned. Then we train the model on the whole training dataset using the optimal hyperparameters and finally evaluating the performance of each model on the test dataset, which was separated from the training before the validation procedure. The GridSearchCV() requires an input of the values (the grid) for each hyperparameter to be optimized, thus we face a trad-off between computation time and fine-tuning the parameters. The final values for the hyperparameters are in appendix table A3.

4.3 Performance

OLS performs relatively well in-sample with a MSE of 0.68, see table 4.1. This is expected, since the objective is to minimize the sum of squared errors, which is equivalent to minimizing the MSE. However, the issue of overfitting seems to be a very relevant concern when evaluating the out-of-sample performance. The MSE for OLS in the test data is enormous compared to what is obtained by the other models. OLS does a poor job predicting the (log to) number of views, and this may very well be due to overfitting.

Table 4.1: MSE for the models on training and test data

	Data excl. image		Small sample incl. image	
	Train MSE	Test MSE	Train MSE	Test MSE
OLS	0.688	$1.449 \cdot 10^{28}$	-	-
LASSO	0.574	0.942	0.769	1.746
LASSO (small sample)	0.772	1.749	-	-
Elastic Net	0.513	0.937	0.649	1.707
Random Forrest	0.133	0.933	0.155	1.516

Next we implement the LASSO to obtain a sparser model and prevent overfitting. The LASSO outperforms OLS already in the training data, with an MSE of 0.574 setting the penalty to $\lambda = 0.0179$. The true return to the LASSO becomes apparent when examining the performance out-of-sample. The LASSO obtains a MSE of 0.942, which is a huge improvement compared to the OLS performance. Next, we implement the elastic net as an extension to the LASSO, motivated by some of the challenges described previously. Using the 10-folds cross validation method to obtain hyperparameters, where the penalty is 0.126 and the LASSO-ratio obtained is 0.111. The elastic net slightly outperforms the LASSO in both the training and test data with a MSE of respectively 0.512 and 0.937.

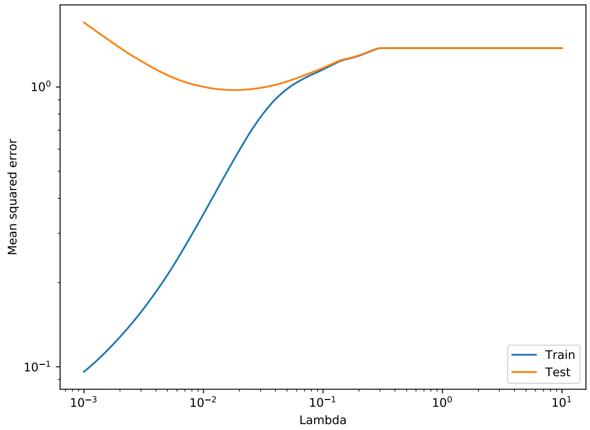
The random forest regressor slightly outperforms both the LASSO and the elastic net employing a set of 2,000 estimators and a max dept of 51 set by cross validation. However, the result of 2,000 estimators is a corner solution, and optimality is therefore not ensured. The estimator performs nicely, however, the computational time greatly exceeds that of both the LASSO and elastic net (more than 10 times slower).

The performance of the models are not greatly enhanced by adding features of the pictures. Evaluating based on the LASSO, which is the preferred model, used in the same small sample, the MSE is slightly smaller. Based on the descriptive analysis of the computed features, this does not come as a surprise, as no clear patterns were found. Again, random forest outperforms LASSO and elastic net, but our conclusion remains unchanged due to computational time. The larger MSE on the data containing images is likely simply due to smaller dataset employed.

4.4 Validation curve

The following figures show the validation curves for the LASSO, figure 4.1, and the elastic net, figure 4.2 using the 5,000 observations data set. The validation curve is used to show the influence of a single parameter on the performance of the model. From figure 4.1, it is very clear, that the model performs well on the training data, but poorly on validation data for low levels of the penalty parameter. This is an indication of overfitting, which is in line with our expectations. Low values of the penalty parameters means a more complex model. Since the LASSO is equivalent to OLS for $\lambda = 0$, see eq. (3.2) and eq. (3.6), it is not surprising, that the model tends to overfit, a classic problem for OLS, when the penalty is low. When the penalty exceeds 0.1, both the training and the validation data yields larger MSE, which indicates a model underfitting the data. This is likely due to the penalty being to large, thus, too many features are deemed irrelevant and the model is too sparse to capture relevant patterns in the data.

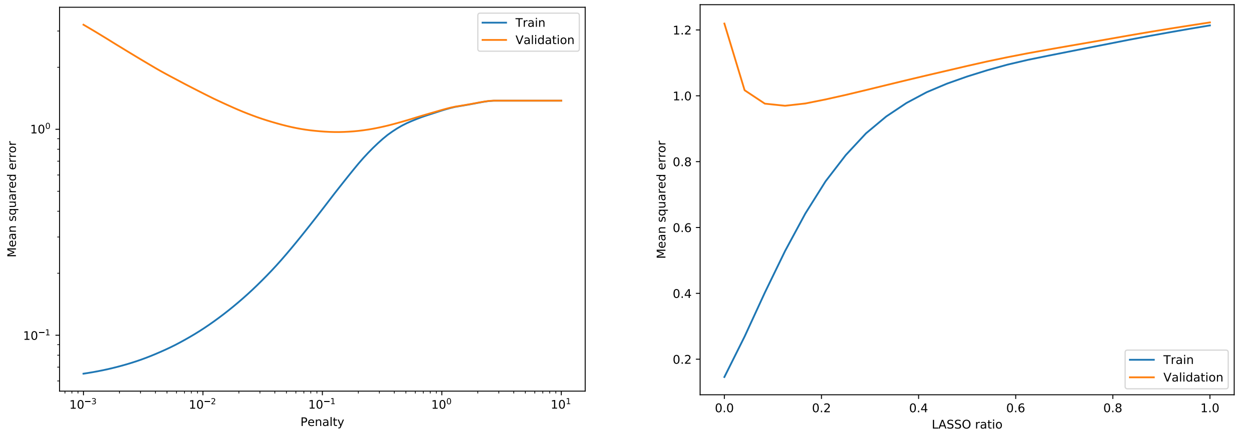
Figure 4.1: Validation curve for LASSO



Note: The figure shows MSE for the test and validation data for different sizes of the penalty parameter using 10-folds cross validation.

Before accessing the look of the validation curves for the elastic net, it should be noted, that the curves are not representative of the minimization problem, which is done over a hyperplane, not merely two separate curves. However, we find that the curves are still useful to access the influence of the single parameters.

Figure 4.2: Validation curves for the elastic net



Note: The figure shows MSE for the test and validation data for different sizes of the penalty or ratio parameter using 10-folds cross validation. For each curve the other parameter is set to the optimal value obtained via cross validation.

The influence of the penalty parameter in the elastic net appears somewhat similar to that of the LASSO. However, the issue of underfitting seems to be a lesser concern, whereas the model is clearly overfitting for low values of the penalty parameter. The LASSO ratio parameter indicates the combination between the LASSO and the ridge regularization. The result of just above 0.1 indicates a solution closer to ridge regularization.[5] That is, more emphasis on shrinking the weights of the features, and less on the feature selection from LASSO. From the validation curve for the LASSO ratio parameter, it appears that larger ratios are associated with underfitting of the model, since both the training and validation data yields larger MSEs. This may be a feature of the LASSO selecting one one feature, if two features are highly pairwise correlated, as previously described.

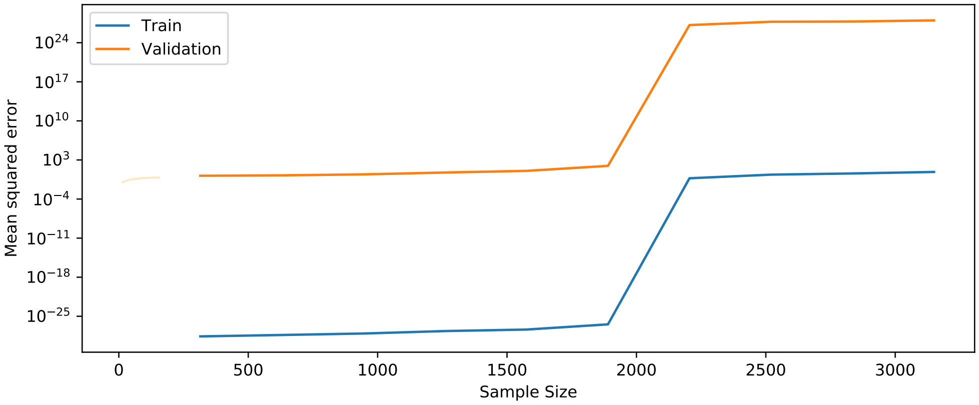
4.5 Learning curve

As previously mentioned, we face a bias-variance trade-off when deciding how complex models to implement. A more complex model is likely to lead to overfitting and large variance, whereas a too simple model will underfit the data and is likely to suffer from bias. A method for addressing this trade-off is the learning curve, where one can examine how the model benefits from more training data. Due to the continuous nature of the predicted variable, the models are evaluated using MSE. We do not have a clear 'goal-value' for this measure (though the smaller the better); neither can we rely on existing literature. Thus, determining bias using this method is challenging in this case. But the learning curve is still useful for accessing convergence between the performance of the test and training data. Thus, we and address the variance. The relevance in our setting is further induced by the fact that we work with a random sample of the total sample. Therefore, more data is available and could easily be given the models, had we had more time. Here, the learning curve comes in handy when addressing which models might benefit from increasing the amount of training

[5] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNetCV.html

data. All learning curves are drawn using 10-folds cross validation as in the estimations. The learning curve for OLS, see figure 4.3, does not look like a typical learning curve.

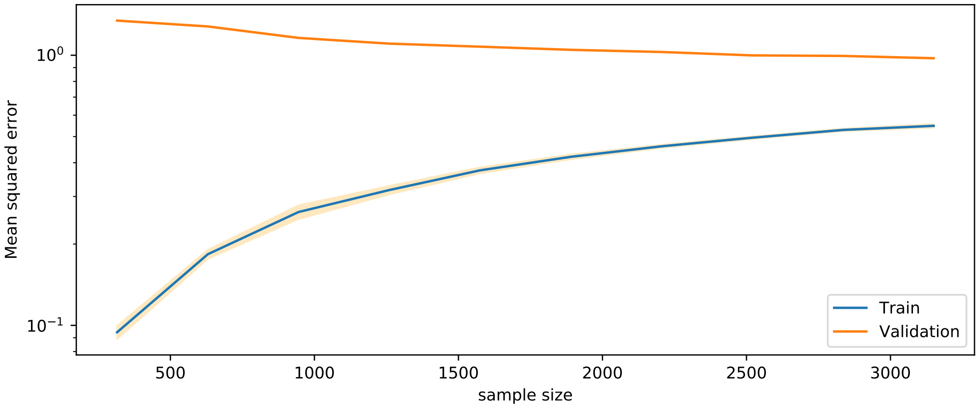
Figure 4.3: Learning curve for OLS



Note: The figure shows MSE for the test and validation data for varying size of the training data using 10-folds cross validation.

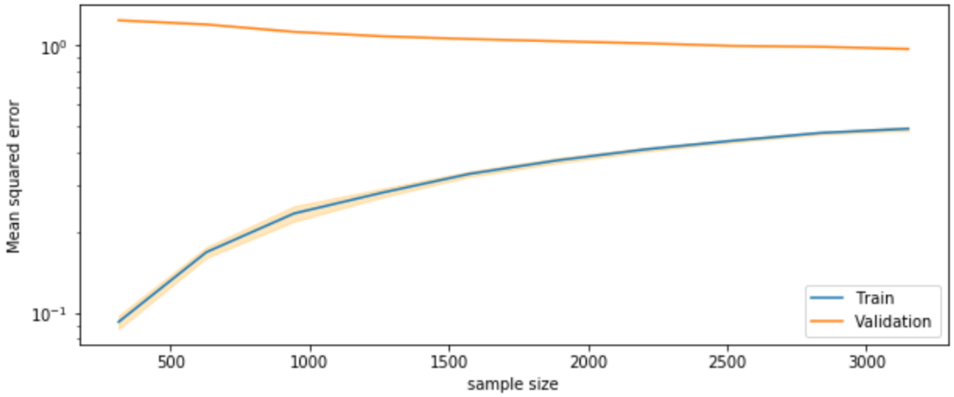
In fact, we observe, that the model performs worse when trained on more data. This may be due to the naturally overfitting nature of the OLS, which is likely to make the predictions sensitive to the training data employed. Further, a comment should be made on the enormous scale of the second axis, which naturally makes it hard to access the behavior of the model. Figure 4.4 is the learning curve for the LASSO using the tuned hyperparameters obtained via cross validation.

Figure 4.4: Learning curve for LASSO



Note: The figure shows MSE for the test and validation data for varying size of the training data using 10-folds cross validation.

Figure 4.5: Learning curve for elastic net



Note: The figure shows MSE for the test and validation data for varying size of the training data using 10-folds cross validation.

While we do observe some convergence, the performance of the model does not seem to be enhanced greatly by a larger training set. A similar conclusion can be drawn from the learning curve of the elastic net, see figure 4.5. The LASSO does seem to converge slightly faster, which can indicate smaller variance and less overfitting, and we do know, that the LASSO will saturate when the number of features is greater than the number of observations, thus selecting maximum a number of features corresponding to the number of observations. However, the difference is so small, especially taking the performance on the test data into account, that no firm conclusions can be drawn from the figures.

5 Discussion

5.1 Data

In this section we will address some of the challenges of working with real world data. A comment must be made on the limitation of our design. Due to computational time, we opt for a random sample of 5,000 videos out of the complete sample. This is not optimal, since more data, and thus more information, is easily available. However, we do find that the optimization, cross validation and cleaning of data is simply too slow to yield any results, if the complete amount of data is used. But for further analysis, this is an obvious and easily implemented extension to the analysis.

The matter of working with text in this context of pornographic videos contains an unforeseen challenge. We applied the bag-of-words model, but as mentioned before, a great deal of work has gone into separating the English videos. Furthermore, we chose to sort both male and female names from the titles. This could have been done differently, especially when considering the ‘hand-made’ changes to the lexicons containing names. The changes were made to prevent to much noise in the dummies for female and male names, driven by confusion of names with nouns. However, since these changes were ‘hand-made’, the data and model may benefit from a more thorough method. Additionally, the analysis may also benefit from further cleaning of the titles. This has proven quite the challenge due to the

non-traditional lingo in pornographic material. Examples of titles posing a challenge are *Tr1n1ty 1* or *Ideepthroat - Heather - BJ, Anal, and Cumshot in Shower!*. We have not been able to recover a lexicon of pornographic words, but further analysis would likely benefit from inclusion of this. We expect benefits from being able to match different forms of the same word. This is a concern both for verbs. But this is also a concern in pornographic material with nouns and their abbreviation, an obvious example from the title above is blowjob and BJ.
But the somewhat lacking spelling in the titles of pornographic videos are hard to handle, and therefore this is beyond the scope of the assignment.
For further analysis a sentiment analysis of the titles could be implemented, though the short nature may pose a challenge.

Another obvious limitation in the data is, that we use the thumbnail's URL to determine the upload of the video. While this serves as a best guess, it may introduce noise, since it is likely, that a new thumbnail has been added for some videos. A method to circumvent this is to scrape comments made to the videos and use these as a sanity check of the thumbnail-dates.

5.1.1 Data ethics

The data this paper is based on is public available data and the data is not individual data. However, there might be some issues to consider given the subject of this paper. Pornography is the topic of an ongoing debate that can be both offensive and violating for some. Pornhub takes corporate social responsibility, by protecting the copyright on the videos uploaded to the website. We have however, no proof that all videos on Pornhub are uploaded with consent from the actors. This would have been a bigger issue if we had been able to do image recognition on the thumbnails, but this paper is not very specific on the content of the videos, as we merely employ the information to predict popularity. Otherwise the data does not contain any data on a personal level, therefore, there is no problem with identification of any individuals. On the website, we are able to locate all the different features that are in the data. Therefore, ethics in general is not a great issue with respect to individuals nor to firms or organizations.

5.2 Method

Again, the computational difficulties working with the data, should be considered. Due to the choice of the bag-of-words model, the feature matrix is of very high dimensionality. This takes a toll on the computers, and the return is likely not proportional to the time wasted. Thus, we would given more time, have opted for a dimensionality reduction of the features matrix. Preferably we would reduce the dimensionality on the matrix from the bag-of-words model applied to the titles. The Latent Dirichlet Allocation technique can be used to model a number of topics from the titles and returns for each observation a vector of probabilities assigned to each topic. The technique is relatively easy to implement using sk.learn.decomposition, however it is not easily implemented on a subset of the feature matrix, which is the motivation for not employing it in this analysis.

A clear limitation of our implementation of the random forest regressor is the setting of hyperparameters. The grid search yielded a corner solution, thus, we are by no means sure that the number of estimators in the regression is optimal. Additionally, the random forest regressor contains several hyperparameters which we have not cross validated. Thus, our model may be prone to overfitting. However, we do discard the random forest regressor in favor of both the LASSO and the elastic net based on computational time. A more thorough search over more hyperparameters would likely be even slower. Thus, we would not recommend further work with this regressor on data similar to this, unless extra computational power is available.

A brief comment should be made on the limitations of the optimization method applied. We opted for a grid search, and this naturally implies, that only values we selected are tested. Therefore, we may not have recovered the global minimum. However, we do not expect that the benefits from a wider and more fine gridded search would change the main conclusions.

5.3 Performance and takeaway

The performance of OLS as a predictor is quite inadequate, and the natural conclusion must be, that OLS is *not* fit for prediction in this case. However, the remainder of the models all perform quite nicely. LASSO generally performs a little poorer than the others, but due to the computational benefits, this is the preferred model. We are not able to greatly enhance the performance of the models adding features on the thumbnails, but we do find a slight decrease in the MSE. Given the validation and learning curves, we believe that the analysis would benefit from more time to reduce the dimensionality of the features, as the models may show some signs of overfitting the data.
Had more time been available, image recognition would have been a very interesting way to go. And especially he effect of faces on the thumbnails would be interesting to investigate. However, this is beyond the scope of this assignment.

6 Conclusion

In this assignment we set out to investigate which machine learning tools applied to data on pornographic videos performs best in predicting popularity.
The data is obtained through the API of www.kaggle.com.
We use only data on videos with English titles and apply a bag-of-words method. A similar approach is taken to the categories of the videos. This yields a feature matrix with a large number of columns, and this greatly increases the computational time spend running the models. If more time were available, a reduction of the dimensions of the feature matrix would be a main priority, as we expect this would enhance the performance of the models.
The OLS, LASSO, elastic net and a random forest regression are tested against each other using 10-folds cross validation to tune the hyperparameters.
We find that OLS performs extremely poor in out-of-sample predictions. However, the performances of the three remaining models are quite similar. The random forest regression

does the best job predicting out-of-sample, but the computational time greatly exceeds that of the other models. This may be due to the large number of features in the data set. Elastic net also predicts slightly better than the LASSO, but again, the computational time becomes an important factor in determining the most fit model, since we can optimize and estimate the LASSO on roughly on fifth of the time spend tuning the elastic net.
Next, we test performance of the models on data containing features of the thumbnail image, to evaluate if the models perform better. The conclusion regarding the preferred model remains unchanged and LASSO stays the preferred model based on computational time. Lastly, we compare the performance of the LASSO with and without image features in a small sample of 250 random observations. The model predicts slightly better when features of the thumbnails are added. However, the numbers are quite similar, and no finite conclusion can be drawn on basis of this. But it does encourage further exploration in the matter of thumbnails influence on popularity.

Literature

1. Brand, M. and Lair, C. (2017), Mood changes after watching pornography on the Internet are linked to tendencies towards Internet-pornography-viewing disorder, *Addictive Behaviors Reports*, Volume 5, p. 9-13
2. Breiman, L. (2001), Random Forests, *Machine Learning*, Volume 45, Issue 1, p. 5–32
3. Bridges, A. J. and Morokoff, P. J. (2011),Sexual media use and the relational satisfaction in heterosexual couples, *Personal Relationships*, Volume 18, p.562-585.
4. Hastie, T. and Zou, H. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B*, Volume 67, Issue 2, p.301-320
5. Mirjalili, V. and Raschka, S. (2017), *Python Machine Learning*, Packt Publishing Ltd., 2. edition

Appendix

Table A1: Deleted words from name lexicons

America
Angel
April
Art
August
Autumn
Bonny
Buddy
Candy
Chanel
Cherry
Dawn
Destiny
Diamond
Dick
Faith
Florida
Forest
France
Ginger
Glory
Golden
Guy
Honey
Hung
Irish
Kitty
Lady
Long
Love
Man
Max
Mercy
Merry
Miss
Moon
Pearl
Princess
Queen
See
Son
Spring
Star
Summer
Sun
Sunny
Tiny
Valentine
Young
Yung

Table A2: Variables in modified data

Variables	**Describtion**
length:*	Length of the video (measured in seconds)
nb_views:*	Count of video views
voting:*	Percentage of good ratings
age:	The age of the video (measured in days)
Dummies for each category	Dummy variable for each category
Female_name_dum:	Dummy for female names in title
Male_names_dum:	Dummy for male names in title
Dummies for each word in title	Dummy for each word in titles
HD:	Dummy for video quality
Dummies for urls in tags	Dummy for each url in tangs
Count_of_tags	count of tags on each video

*Note: * indicates not modification*

Table A3: Hyperparameters

	Estimators	**Max depth**	**Penalty**	**Ratio**
Data excl. image				
LASSO	-	-	0.0179	-
Elastic Net	-	-	0.126	0.111
Random Forest	2000	51	-	-
Data incl. image				
LASSO	-	-	0.126	-
Elastic Net	-	-	0.295	0.333
Random Forest	2000	21	-	-