*I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.*

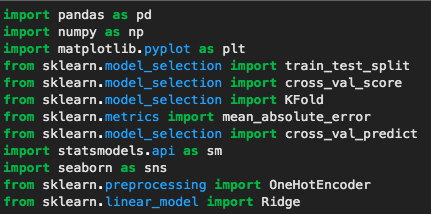
**INTRODUCTION**

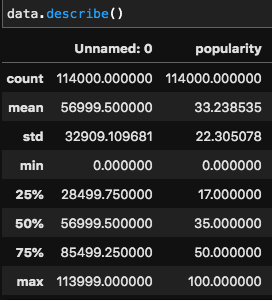
The main aim of this project, was to build ridge regression for the dataset, in order to predict the popularity of the songs. The dataset contained 114 000 observations, that characterized all kinds of music, and 15 variables. With the aim of creating a ridge regression, one-hot encoding was used on variables that weren’t numerical. Ridge regression was performed from the scratch and then using sklearn package cross validation with 5 k folds was implemented.

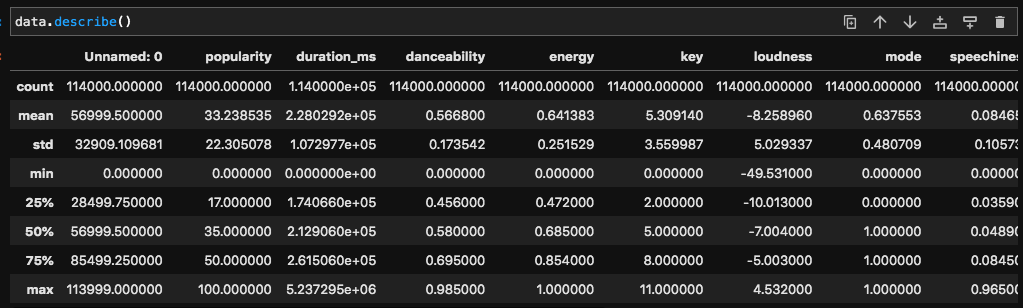
**THEORY**

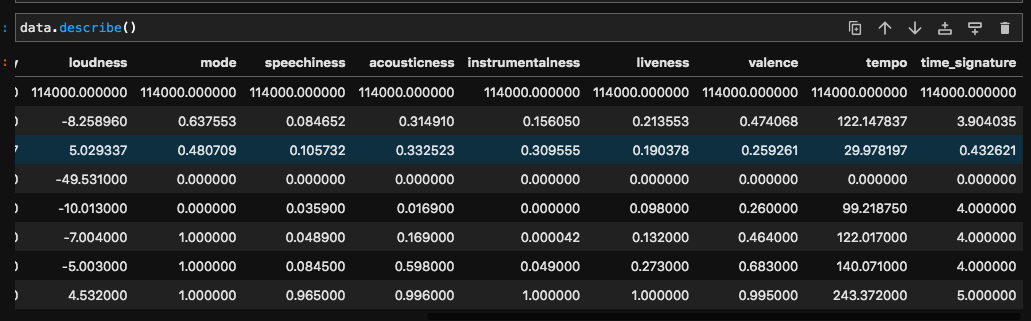
Ridge Regression is used when model deals with multicollinearity(2). Dataset includes too many variables with strong ecoefficiency, that undermines the real power and statistical significancy of the variable and can lead to overfitting. Ridge regression adds to a model a penalty, that puts a similar constraint on the coefficients (1). Penalty, regularizes the regression coefficients by shrinking them towards zero. The estimator is biased and has low variance. In ridge regression there is also used a tunning parameter alpha, that controls the impact of the shrinkage penalty on the created model (5).

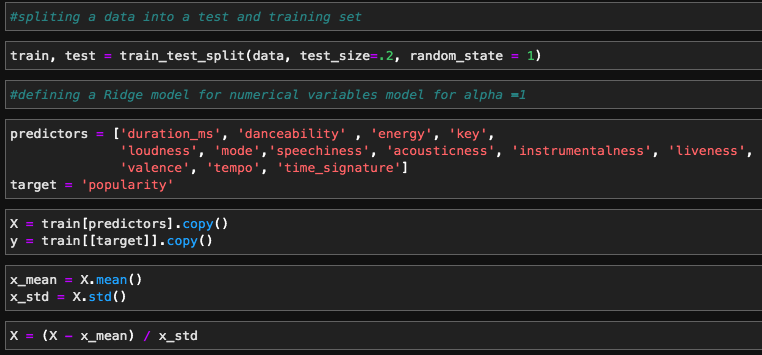
Cross validation using K-folds is used to evaluate the created prediction model(4). Dataset is divided into a k subset and then trained k times, while using a different fold for validation every time. The final value of the model is obtained by averaging performance metrics from every fold. In a projects K folds cross validation was used to obtain test Mean Square error to estimate the liability of the model on an unseen data. MSE measures how close the ridge regression falls from the set of data points. The lower the value, the better the model.





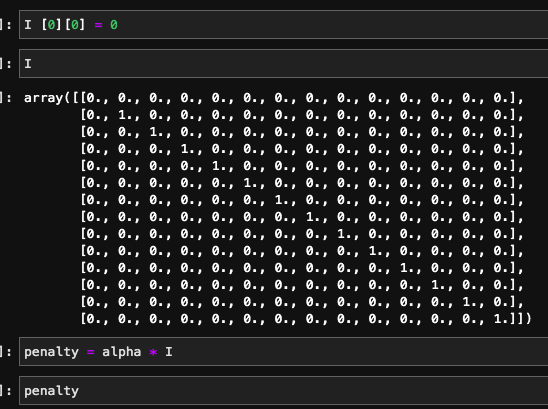


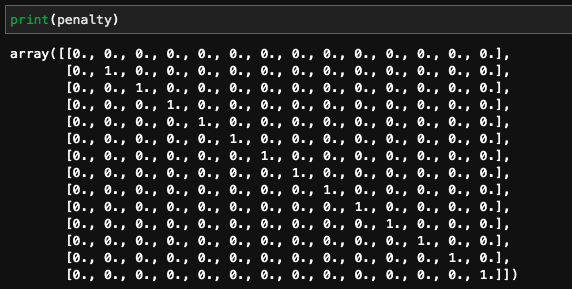


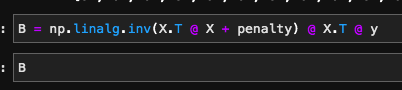


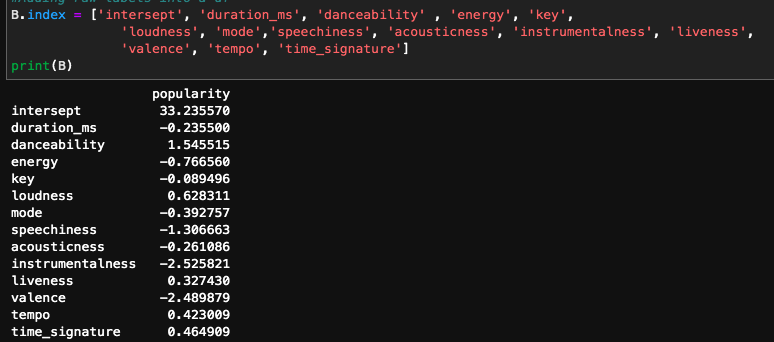


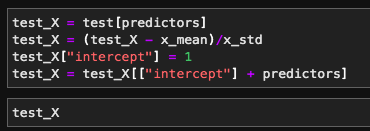


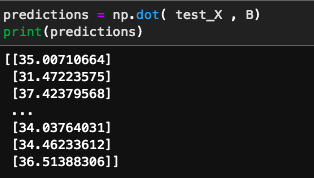


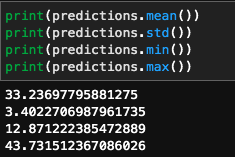


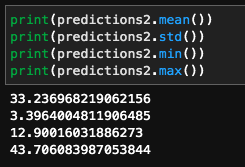


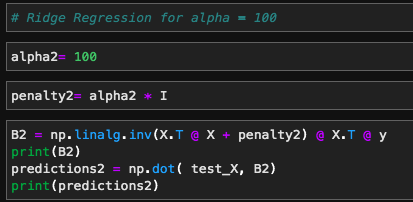


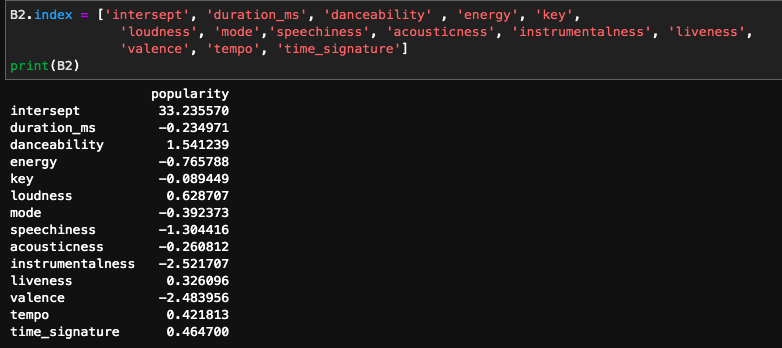


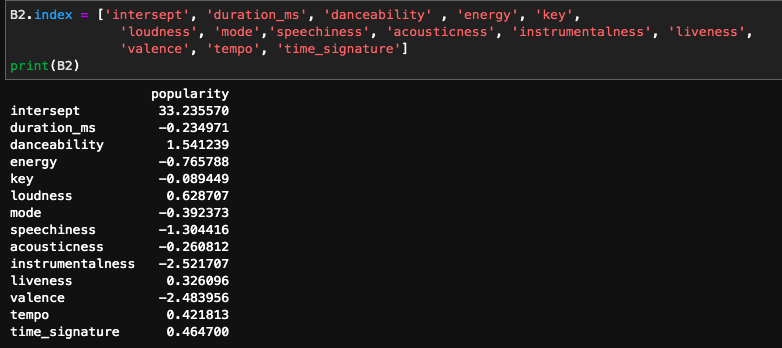


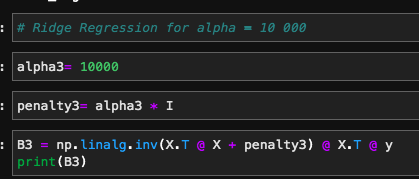


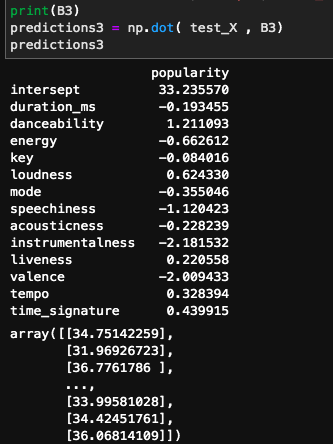


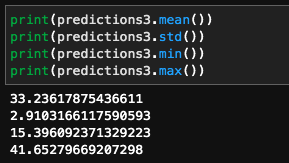
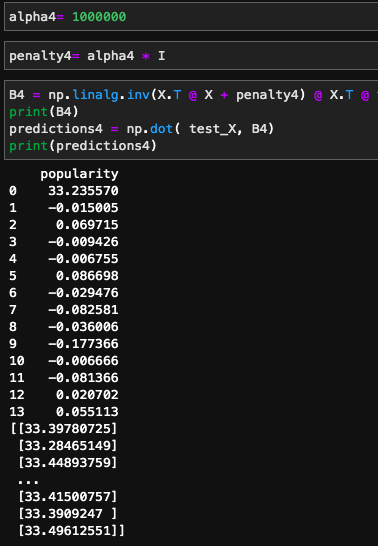


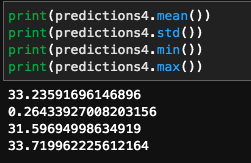




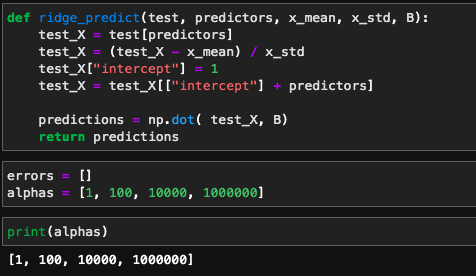


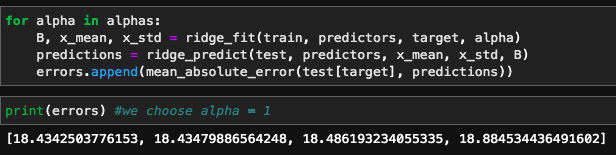


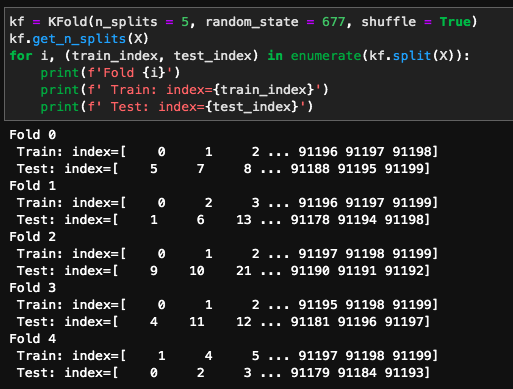


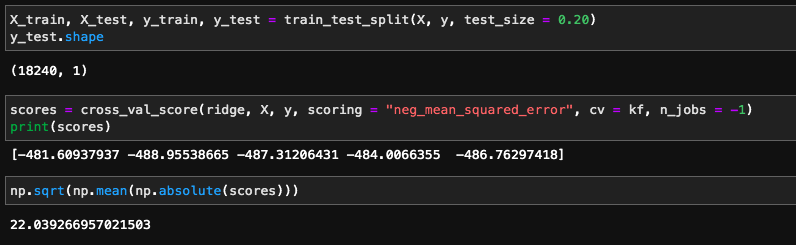




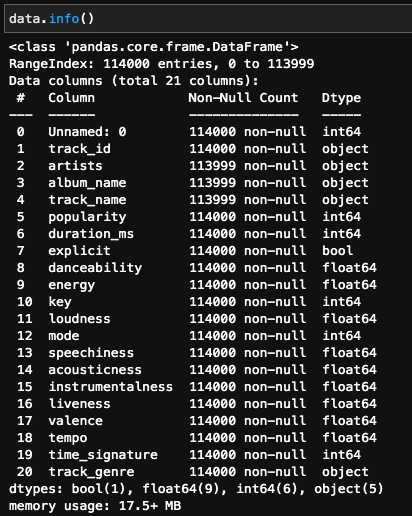


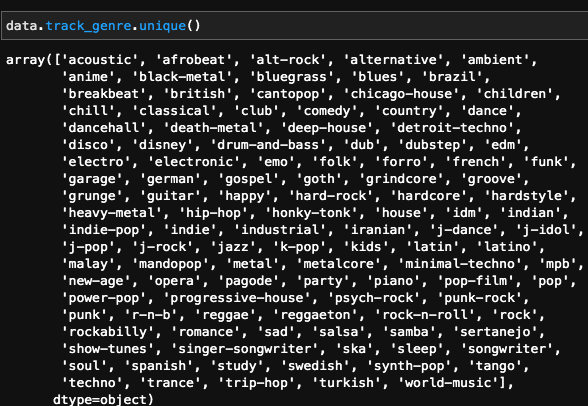


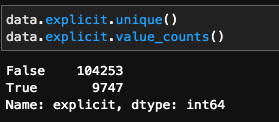


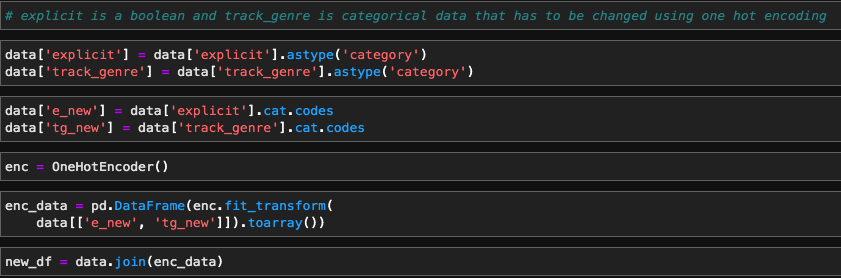


DRUGI MODEL

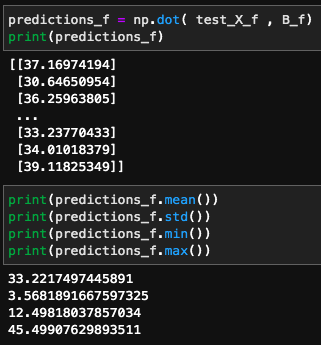


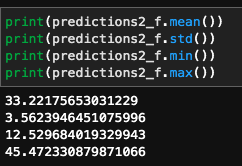
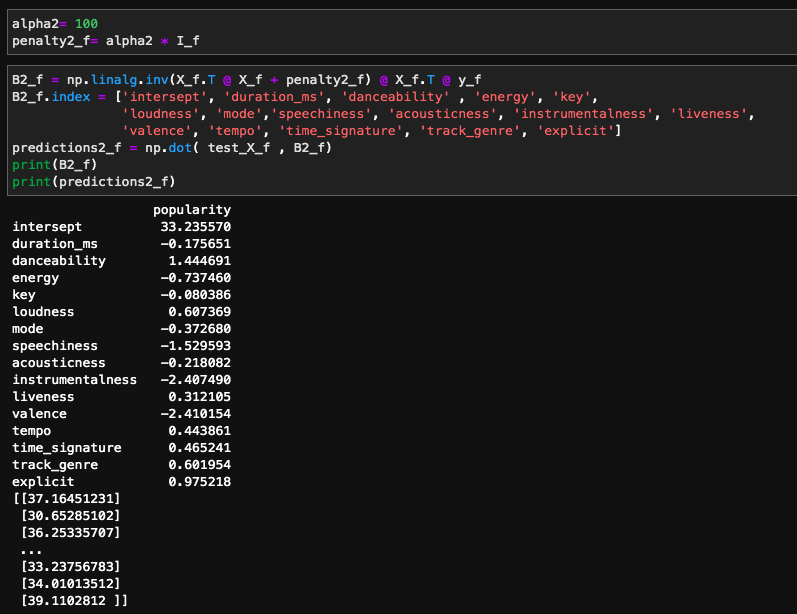


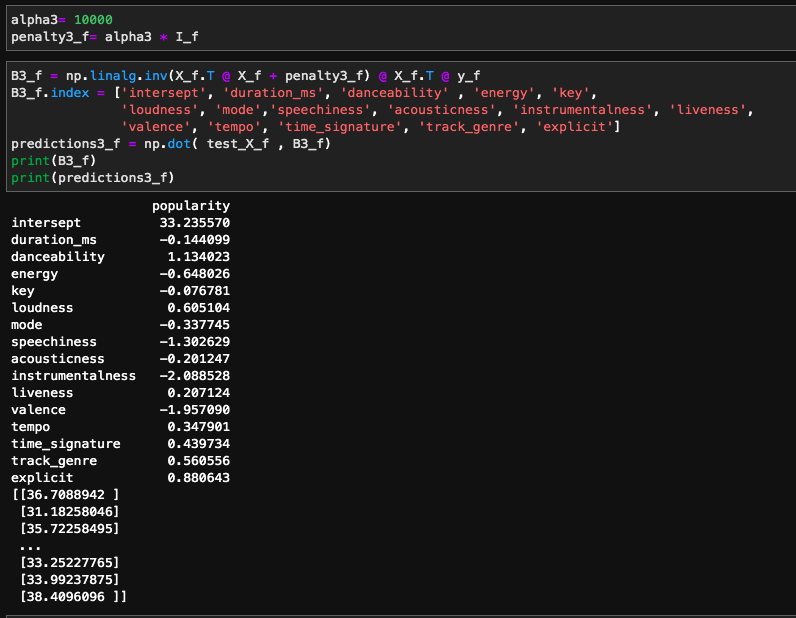


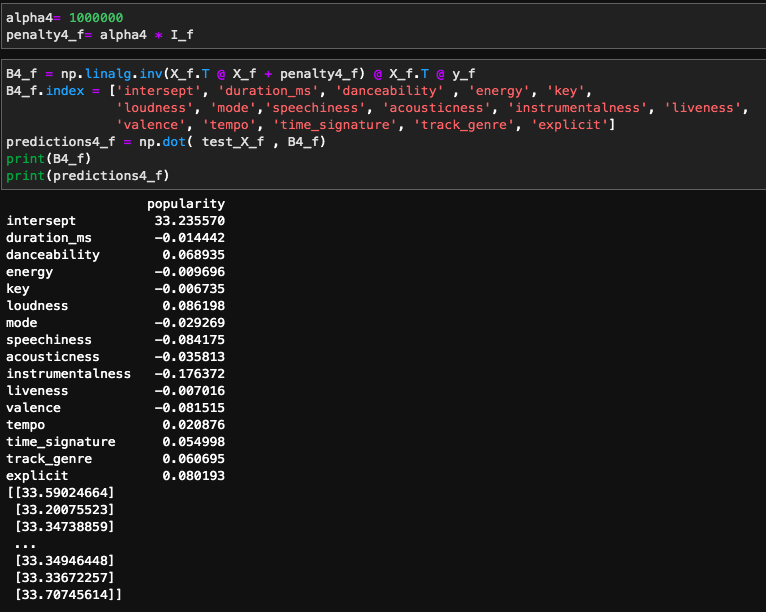
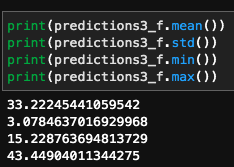


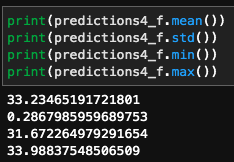


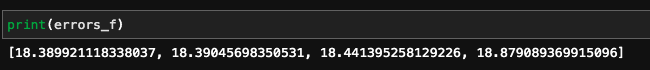


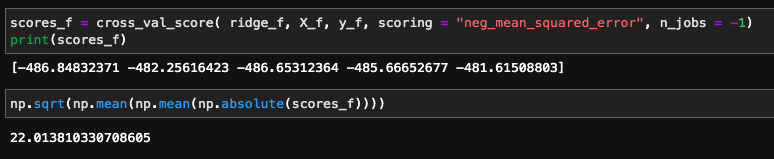












Source:

1 <https://machinelearningmastery.com/k-fold-cross-validation/>

2 <https://www.statlect.com/fundamentals-of-statistics/ridge-regression>

3 <https://www.engati.com/glossary/ridge-regression>

4 <https://towardsdatascience.com/what-is-k-fold-cross-validation-5a7bb241d82f>

5. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. Second Edition. New York, 2013.