# Modelling used cars on Craiglist

This document describes the modelling performed to analyze the prediction capabilities of four features selected from the Kaggle used cars Craiglist dataset.

In order to be able to differentiate between the model, all models were analyzed using the same metric - the root mean squared error (RMSE).

For the purpose of this modelling the first model considered was Linear Regression using year, odometer, type and state as predictors. In this case, the exploratory analysis showed that there is a correlation between year and price, and between odometer and price suggesting that these features might have prediction capabilities. Odometer was normalized before modelling, in order to be able to perform lasso and ridge regressions. The year feature was considered both as a numerical feature and categorical. While using it as a categorical feature enabled using penalized regression, using the feature in numeric form had a better fit for the data. (in the notebook is kept as numerical) From the categorical variables type and state yielded best results in comparison with other features tried, even if the features slightly improved the value of the score. It is worth mentioning that there was no ANOVA type 2 tests performed to analyze a restricted model compared to a full model as an indicator for statistically significant features, but rely on the RMSE score as a way to see differences between different combinations of features.

In order to improve on the Linear Regression model, two penalized regression models were fitted, one with Lasso penalty and one with Ridge. In these two cases, the training dataset was split further, and using cross validation, the best regularization parameter was searched using grid search. The results of the penalized regressions didn't improve on the result, RMSE was close to the linear regression model without penalty.

The best model was a Random Forrest model. Initially one model was fitted with 100 estimators (trees) and subsequently using cross validation and grid search the number of estimators were selected across 4 different values. It's worth mentioning that extra hyperparameter tuning was performed such as tree depth, max nodes but due to the slow fitting this was not included in the notebook.