# Regression with 200 variables

## Exploration & missing value handling

After data loading pre-processing was conducted:

* Checked for duplicates.
* Filled in the missing values using the mode of the categorical variables and the median of the continuous variables.

After all the missing values had been filled in, the relevant features were selected from the 200 available variables.

* The continuous variables were filtered using correlation, and the ones between -0.05 and 0.05 were dropped.
* The categorical variables were filtered using the p-value of an ANOVA between each variable and the predicted variable. The significance level used was 0.05.
* The multi-level categorical variables were checked using a post-hoc Tukey test, and the irrelevant levels were removed.

After the feature selection, the potentially influential outliers were identified using 2 times the mean cook’s distance of 2, and then these were removed.

## Model training & evaluation

The available data was split into training and validation sets, various models from the caret package were evaluated using the RMSE and repeated-cross validation.

* elastic net - RMSE : 0.4678055
* kknn - RMSE : 0.5302197
* svmLinear - RMSE : 0.4853201
* svmRadial - RMSE : 0.4875742
* randomforest - RMSE : 0.4876961
* xgbLinear - RMSE : 0.5127372
* **xgbTree - RMSE : 0.4572067**

At the end, the xgbTree algorithm worked the best, even though the best 3 algorithms were put into an ensemble model; however this had a worse performance.

After the model selection, a random forest algorithm was trained on the entire dataset and the errors from the cross-validation were compared:

* xgbTree using 80% (1600) of the data, the training RMSE was: 0.4663071
* xgbTree using 100% (2000) of the data, the training RMSE was: 0.4624294

Using 25% more data (+ 400) during training, resulted in a slight performance improvement. Therefore it would be interesting to see the performance using a sample of a 100K dataset.