# Capstone Project #1 – Machine Learning

Link: Jupyter Notebook

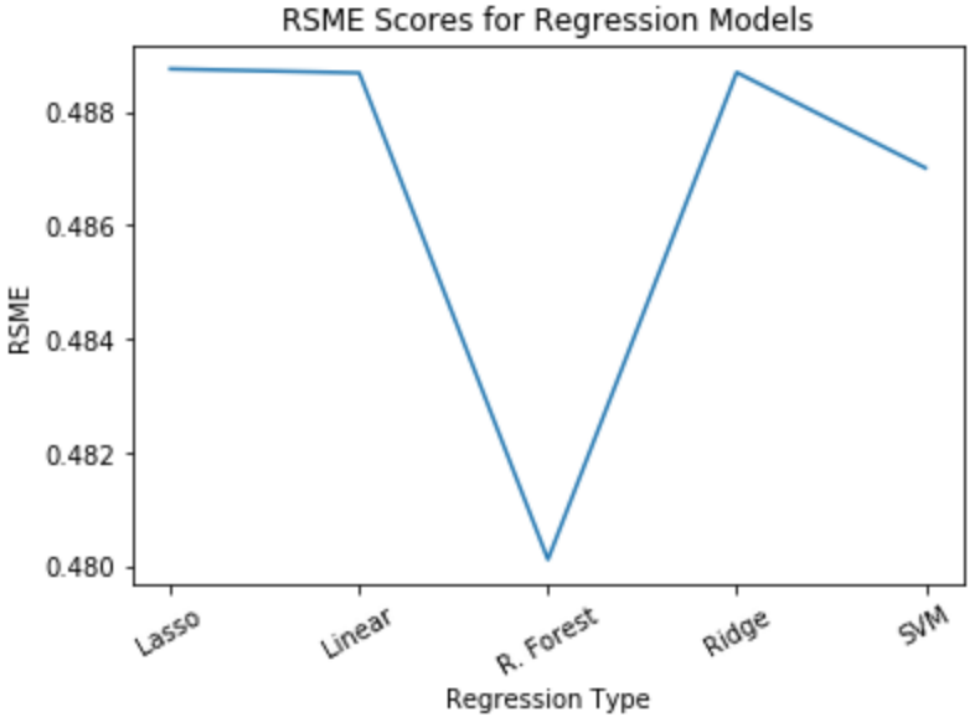
With exploratory analysis complete, some final cleanup steps were necessary to prepare the training set to fit machine learning models.

Since only the numeric variables would be used in the models, all text variables were dropped from the dataframe. In addition, the id column was dropped, as it would not be needed in the analysis, along with the previously created relevance_reduced variable that had been created for visualization purposes.

With this complete, the dataframe was partitioned by separating the dependent variable (relevance) from the other numeric variables. The data set was then split into training and test sets (70/30) in order to fit the models. The overall test set provided from Kaggle is still set aside to produce a final score.

Since the Kaggle competition uses the root mean squared error (RSME) to score the given test set, various regression models will be tried and evaluated using that metric

The regression models that were attempted were Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regression and SVM Regression. A similar process for each model: a pipeline was created that scaled the data and performed a cross-validation grid search to tune the parameters for each model. The models were then fit and an RSME score calculated on each set of predictions. The results are summarized in the following graph:



The model that performed the best was the Random Forest Regression model with an RSME score of 0.48011.

In an attempt to improve the model a Bagging Regressor was fit to the model. The RSME score of this model was 0.48015, which was not an improvement.

Finally, the model was used to make predictions on the provided test set. This was then submitted to Kaggle and received a final RSME score of 0.48614. This is consistent with the results seen previously and a sign that the model was not overfitting and was translating well to new data.