Predicting MotoGP race finish times using linear regression
Ankur Vishwakarma
Metis SF Winter 2018

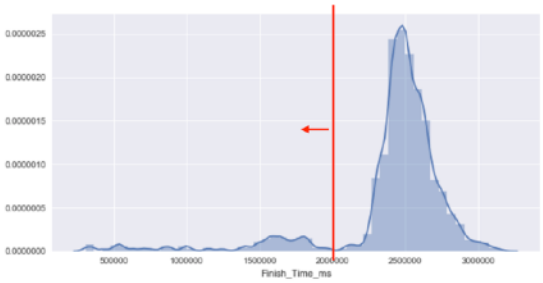
Project Goal - Predict **time to finish a race** based on (1) physical track characteristics, (2) weather data, and (3) class of racing. There were 28 individual features that fell into the three categories above.

1. Scraping

I scraped www.motogp.com for the data and ended up with 19,000+ rows and 30 columns spanning 2005 - 2017.

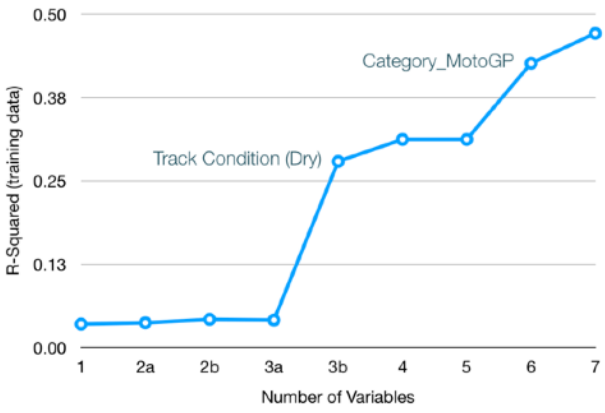
2. Cleaning

After converting the finish time from past races into milliseconds (to get a continuous target value), I plotted the target distribution and discovered it was significantly skewed.



Fortunately, the values on the left of the line represented abnormal races - those that were cut short due to accidents, technical malfunctions, weather conditions, etc. They could be discarded as they were not relevant data points.

3. Modeling



For the first step, I checked the R^2 value on the training data for a simple OLS linear regression with a variety of different inputs, starting with only 1 feature. The preceding chart shows the progression of R^2 as I added more features.

My regression was finally optimized with 16 features. Lasso regression also did not drop any of those 16 so I chose to keep them for all future models. Each model was trained on training data, validated on validation data using 5-fold cross-validation, and finally tested on held-out test data.

4. Results

Model Description (16 features for all models)	R^2 on Test Data
OLS	0.53
Lasso regularization	0.53
Polynomial (degree 2)	0.82
Polynomial (degree 3)	0.95

Best model - Polynomial Regression (deg 3)
R2 = 0.95
RMSE = 37 seconds (1.4% of target mean)

5. Conclusions

It is possible to predict the race finish time using regression. However, a simple linear regression is too simple of a model to do so. Polynomial features are needed.

Even then, the error is too high for this application as 37 seconds spans 50%+ of the riders' finishing times. More info on rider characteristics and skill will probably help the model further.