Nearest Neighbor Correlation in production of Airplane-Induced Pollution

George Durendal

Abstract

We present and discuss a nearest neighbors correlation in the production of airplane-induced air pollution. A k-nearest neighbors (k-NN) regression consistently produced a higher r^2 values than multiple other types of regressions as a function of time. There are therefore clusters of airplane-induced air pollution high-production and low-production days. Private planes are a potential source of lead pollution, so this can have a major impact on public policy.

1 Research

An AutoML regression analysis of data closely correlated with the amount of airplane-induced air pollution produced by airlines consistently yielded the result that k-nearest neighbors (k-NN) regression consistently produced a higher r^2 values than multiple other types of regressions. The regressions were plotted as a function of day index in the year, so Jan 1st is 1, Dec 31st is either 365 or 366 during a leap year. The data is based on U.S. composite federal data, not localized to any particular region of the US.

This suggests that high airplane-induced air pollution (ApIA pollution) on one particular day is correlated with high levels of ApIA pollution produced on nearby days. We note that this is produced ApIA pollution, i.e. it makes no statement on lingering pollution. Hence the nearest day correlation is not due to lingering pollution from nearby dates.

We posit that this nearest neighbor correlation is due to popular travel dates such as holidays, peak business dates, weekends, and peak seasonal travel dates unrelated to common holidays (e.g. summer). Weekends introduce a common and consistent occurrence of

One competing algorithm is NuSVR, which beat k-NN in some tests, although never reaching as high a r^2 as those achieved by some k-NN regressions. NuSVR is here defined as Support Vector Regression (SVR) with a parameter Nu controlling the number of support vectors.

2 Discussion

One notable point is that the k-NN beats regression algorithms of numerous types. This implies that, in regard to production of ApIA pollution, production levels from nearby dates are a significant indicator of current production levels. Moreover, this nearest neighbor relationship is strong enough to outperform other regressions attempted. For instance, historically the US sees elevated travel levels at the end of the year, during its major holiday seasons. This positive trend was not strong enough to make some regressor which might show some positive trend towards the tail end of the dataset outperform the k-NN regressor. Thus, in terms of air travel and associated ApIA pollution, the nearest neighbor relation seems to be more important than any A natural corollary of this nearest neighbor relation is that ApIA pollution levels are concentrated around specific clusters of dates, particularly holidays and high-travel dates. Lingering ApIA pollution is only compounded by elevated production levels on adjacent dates. The data used here is generalized to that of the entire US, so it makes no prediction about ApIA levels in any specific region or locality. Although we may naturally posit that ApIA pollution levels are highest around public and private airports and the areas where those planes fly over or where the pollution from them disperses to. An additional corollary is that ApIA production is lowest during clusters of dates with lower levels of airplane travel. Given the fact that private aircraft are known to emit toxins such as lead or lead composites, this may be applicable to public policy. It might, for instance, be prudent to sched-

ule outdoor activities for vulnerable demographics – such as children living near airports – on dates with lower levels of air travel and which might be expected to see lower levels of air travel on nearby dates, based on historical data. These sources of ApIA pollution lead may be found to be generally less harmful than lead levels suffered from other sources. Similarly, ground level uptake of ApIA pollutants may be below given thresholds explicitly known to cause medical harm. It may, however, be exceedingly prudent to minimize levels of lead and other ApIA pollution-related toxin exposure in vulnerable groups such as children.

3 Regression Type Comparison Plots

Figure 1: Plot 1

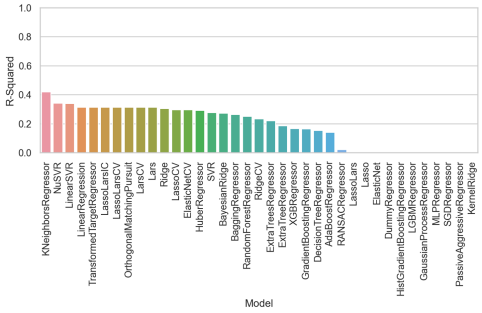

Figure 2: Plot 2

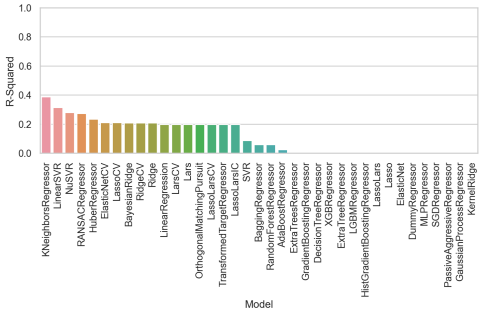

Figure 3: Plot 3

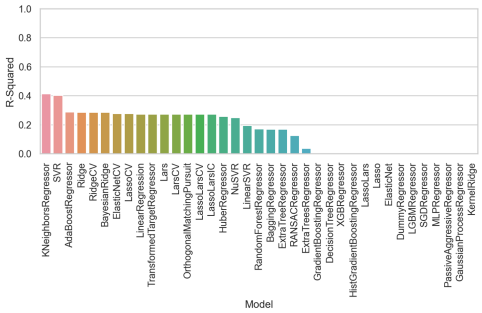

Figure 4: Plot 4

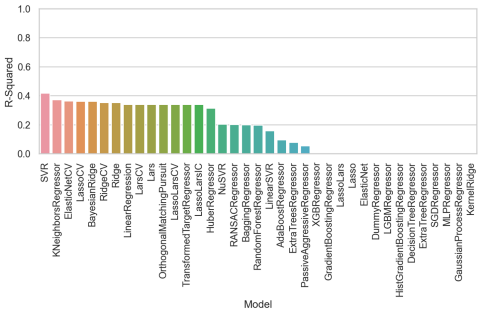

Figure 5: Plot 5

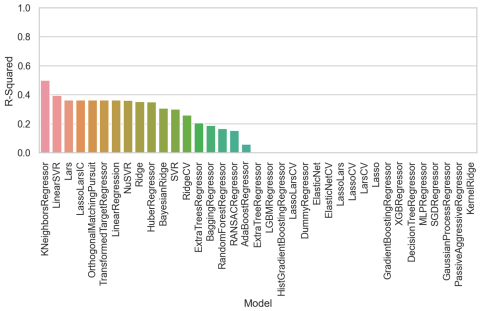

Figure 6: Plot 6

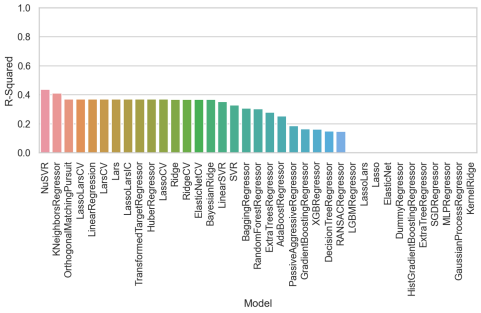

Figure 7: Plot 7

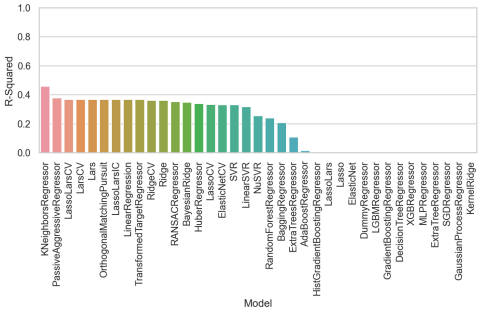

Figure 8: Plot 8

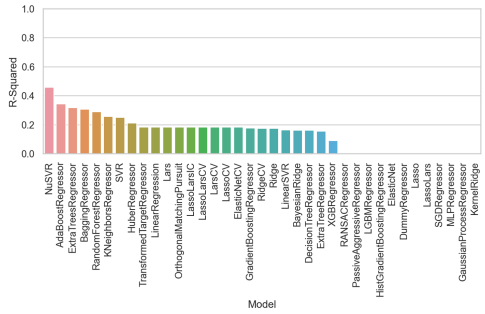

Figure 9: Plot 9

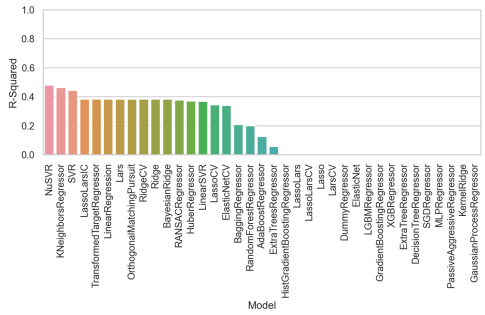

Figure 10: Plot 10

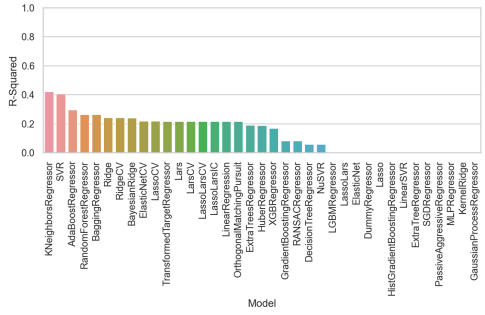
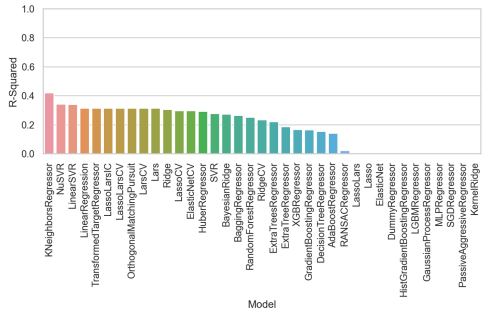
Figure 11: Plot 11