James Griffin

Stat 536

River Flow Report

**Introduction**

Since the beginning of human history, people and animals have depended on the flow of rivers as an essential part for sustaining life. In the Rocky Mountains of the United States, rivers provide the water for farmers for irrigation ditches and others fill reservoirs for other purposes. This analysis seeks to understand the various factors that impact water flow in this region of the country, thereby understanding how to limit negative effects which are under human control and try to maximize the flow of water.

The data for this analysis had 102 observations from different rivers in West from the year 2000 to 2015 in five-year increments, tracking different biological and human factors. The total of features in our data set was 98. However, lots of our data was very correlated, for example we had populations per square mile for each five-year observation, and those will be highly correlated one to another. We also have the mean water fall for each month and the cumulative water amount for each month in each year of the data. This is illustrated using some seaborn correlation heatmaps:

A screenshot of a computer

Description automatically generated with low confidence Graphical user interface, application, calendar

Description automatically generated

Given this cereal correlation, the data is not conducive to linear techniques, which assume independence in the feature set. However, this data set is also somewhat limiting in some more robust machine learning techniques as we also do not have complete independence of our rows. Another challenging aspect of the data set is the difference in how the variables are measured. Some were in milometers, others percent’s, some in meters, kilometers and some kilometers squared. The features themselves were also not linear in nature in relationship with the target, here are some examples. While the relationship between bio15 and the metric is more linear, bio15 is the seasonal precipitation, so it will also be highly multi colinear with the precipitation variables for each of the four seasons.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**Methodology**

Having established the co-linear nature of the data, but still wanting the interpretability of regression, we decided to proceed with the approach of Lasso Regression. Lasso preforms the L1 regularization, where we introduce a penalty to prevent from over fitting, which is equal to the absolute value of the magnitude of the coefficients.

In other words, the minimization objective does not only include the residual sum of squares (RSS), but it also considers the sum of the absolute value of the coefficients. Mathematically the lasso minimization objective becomes:

Which equals:

Where (lambda) provides a trade-off between balancing the RSS and magnitude of the coefficients across the different observations and columns , where is the number of columns in the data and the number of rows. If then the minimization is the same as linear regression, and as we will get the same coefficients as logistic regression, and anywhere in-between is a balance of the two, as increases, bias increases, and as it decreases, the variance increases.

As directed by the prompt, we did not choose our feature set, however, L1 does this implicitly, similar to ensemble tree methods, which we will compare our Lasso results to,

L1 regression zeros out coefficients that are not significant in prediction, this is part of the penalty in the equation above. This penalty helps us avoid over fitting and generalize to outside data better.

As mentioned above, the difference in measurements is something that we need to be aware of for Lasso regression. We thus had to scale and normalize our data, this is because the scale of the variables affect the how much regularization will be applies to specific variable, so we did this for all of the data.

**Model Assumptions and Evaluation**

Having satisfied our linearity assumptions, we still needed to confirm that the other assumptions for linear models are satisfied before proceeding with the Lasso model. Traditional linear models also require that the residuals form a standard normal distribution. If this is satisfied, then the residual distribution is spread around mean 0 with the same variance as our data. Here is a histogram of our residuals, to show that the assumption is satisfied:

**Chart, histogram

Description automatically generated**

As mentioned above, we scaled and normalized our data to filfil the equal variance assumption, and we are able to proceed with the evaluation.

As prompted by the questions for this analysis, we wanted to break out the columns into the different categories of impact, “Network”, “Climate”, and “Human”, to see how each category of features impacted the metric in question. After selecting the columns particular to each group, we ran a Lasso regression using cross validation and using different values. Lambda is a hyper parameter, we are allowed to test different values as we seek to improve our testing scores and reduce our root mean squared error (RMSE).

For our climate features, we got an of 0.49 in both our testing and our average cross validation across five folds. We wanted to make sure our testing and training scores were comparable. If our training scores far outpace our testing scores, then we are over fitting. For this group our lambda value was 0.005 and we had 12 features, shown below:

Table

Description automatically generated

Not surprisingly, many of our non zero coefficient were related to the fall of water both on average and cumulatively. This would prove to be our most predictive category of features.

Next, we looked at the “Human” impact of river flow. These scores did not do as good of a job at predicting the metric with only a 0.14 score in training and a .06 in testing, this time with a lambda level of 0.01. This could be interpreted in one of a few ways, perhaps the human impact on the outcome metric is not very substantial, or these data are simply not good features in predicting. Whatever the case may be, these features were not super predictive for this analysis.

Table

Description automatically generated

Lastly, we looked at the Network group of features from the data, using a lambda of 0.05, our training and testing scores were -0.05 and -0.08 respectively, and we only used 1column, *gord* with a coefficient of 0.011777.

**Results**

The prompt wanted us to see how well the above features did in predicting the outcome metric. Here is what we found. Our results were actually very comparable to the random forest regressor I had initially done as my approach.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

The Lasso had more uniform errors throughout, and even though it had slightly higher rmse and mae, which were 0.53 and 0.36 respectively, compared to a rmse of 0.41838, and a mae of 0.2974 from the random forest, it is more consistent in its errors. This itself does provide a certain amount of utility, it is this fact, that leads directly to the higher .

Overall, the data do a fair job in both models of predicting the flow of rivers in the Western US. It puzzles me that the human impact was not as strong as the others, although another possible explanation is that the human impact is colinear with climate variables as more pollution is put out into the world, which have potentially damaging effects on the climate.

For our models to improve, we need more data, either in terms of more years of data, or data from different regions in the world with similar climates, geographic features, population densities, etc. Having only 102 observations was tricky, enough to work with, but not ideal, especially when we are preforming not quite as well as we would want.

The limitations of our analysis are mean that our external validity is low, we can’t say much about rivers in different climates or in different parts of the world. We also don’t know if the years that this data was given were high, low, or moderate in their rain fall. It is possible that they were all high, low or very average, which could be potentially biasing our data and limiting our validity.