Project Write Up

Aidan Smith and Chris Tang

Project Introduction and Goals

In this project we are attempting to predict energy usage based on weather conditions in the same place and time. In order to do this we utilized one dataset that tracked energy usage of a house and another dataset

Motivation

The goal of this project was the predict the amount of electricity based on various weather data and the time of the year. This result is important to two potential clients. Homeowners being able to predict their electricity costs for the month is important to their ability to budget. Additionally, weather forecasts for the season can affect the homeowners ability to adjust their upcoming budget. This includes forecasts such as a particularly cold or rainy winter and would allow homeowners to adjust their electricity budget to fit these forecasts.

Being able to predict the amount of electricity used is also important to power companies. For example, if a power company were to expand into a new area, they could predict amount of electricity the area would need in a year based on past weather reports. This could change their plan in building infrastructure such as power lines or power generators.

Ridge Regression

For this project we chose to use ridge regression. Ridge regression is a similar technique to linear regression. The main difference is that ridge regression helps deal with multicollinearity which is when independent variables are highly correlated. In our case, our database used for

the weather does include independent variables that are highly correlated such as ‘temp’ and ‘feels-like’, or what the temperature outside felt like. Variables such as these two are highly correlated and would lead to multicollinearity issues.

Lasso Regression

For this project, we also decided to use lasso regression as a backup technique. Lasso regression is similar to ridge regression in that it should help deal with multicollinearity. However, lasso regression does tend to perform better when only a few predictor variables are significant. We wanted to implement both lasso and ridge regression on our data set so that we could compare the final results.

Datasets

The first dataset that we used was the electricity usage, which had an observation from nearly every minute, and included a large number of values. The value we were primarily interested in were the ‘GlobalActivePower’, since that indicates the power usage across the entire household. We were also curious to see if we would be able to use similar algorithms to predict the ‘SubSystem3’ value, since that was associated primarily with the air conditioning, so it would make sense that it would be weather dependent. However, we were not able to get nearly as good of accuracy when trying to predict that as we did with ‘GlobalActivePower’. This dataset did have a fair number of missing values.

The second dataset we are using is the weather dataset that contained observations taken every hour across a wide range of categories. Many of these included significant amounts of missing values such as the various solar data points, along with there being a lot of missing values in some of the wind related measurements. We also got rid of other variables for reasons that are explained more in the data cleaning section. Some of the particularly important data points that we identified were temperature, windspeed, humidity, and precipitation.

Both of these datasets are measured from December 2006 until November 2010 and take place in or near Paris, France. After merging these datasets, we are left with approximately 34,000 datapoints across these 47 months.

Data cleaning and merging

The first issue we had to solve was the fact that we wanted these two datasets to be merged together, so that we could ensure that every temperature measurement was matched to the correct electricity measurement. In order to do this, we first had to resample the electricity data because of the fact that the electricity data was sampled every minute. We took either the sum or the average depending on the specific measurement of all measurements that happened within the hour before creating the new dataset. After that was performed, we had to make sure that both data sets were using similar formats for their time and date variables. This ended up being tricky because of the fact that the weather data had a strange format, and because the electricity data was originally sampled on such a small interval. However, once we were able to get the datetime variables to be the same, merging the two datasets together was very straightforward.

After merging the two datasets together, we took a harder look at the variables we were still using before feeding the dataset to our ridge regression model. We decided to remove several variables like severerisk (whether there were any weather warnings), station (what station the observation was taken at), winddir (wind direction), uvindex (amount of UV light), preciptype (type of precipitation), and conditions (current conditions, e.g. sunny). These were all removed either because they were entirely uniform across the dataset, were irrelevant to what we were studying, or were categorical data that was better represented by other variables.

This data manipulation was performed in R where it was then exported to a csv where we could read it into our python program. The final part of our data cleaning was making sure there were no missing values so that ridge regression could be performed. In order to do this, we decided to just drop any rows that had missing values.

Results

After testing and comparing both ridge regression and lasso regression, we discovered that they were very similar in their results. However, generally lasso regression seemed to very marginally outperform ridge regression. Additionally, we found that lower alphas and a dataset of 5 variables, including temperature, windspeed, feelslike, humidity, precipitation, and windspeed gave us the best results. To quantify this, while we got an RMSE of 51.77566 for ridge regression, we got an RMSE of 51.77554. However, when using smaller predictor sets, ridge regression generally outperformed lasso regression with an RMSE of 51.78426 for ridge regression while lasso regression had an RMSE of 51.78475. We think these differences are likely due to the fact that the temperature predictor was by far the most depended on, so particularly in the larger dataset, lasso regression came out ahead due to temperature being very significant.