Predictive Modeling

Hope Albers

March 20, 2017

Project Description: Create a script that employs two or more of the predictive Machine Learning models (preferably more than two) that we have discussed in class to analyze a business situation of your choosing. Your model(s) must be predictive. The script you submit must compute an assessment of the test error rate, preferably across various model types and/or levels of flexibility.

1 Business Problem

Environmental companies are concerned with pollution data because it's essential to their mission and values. With being able to predict pollutant data, environmental companies can more accurately target clients who produce high levels of air pollutants, specifically SO2.

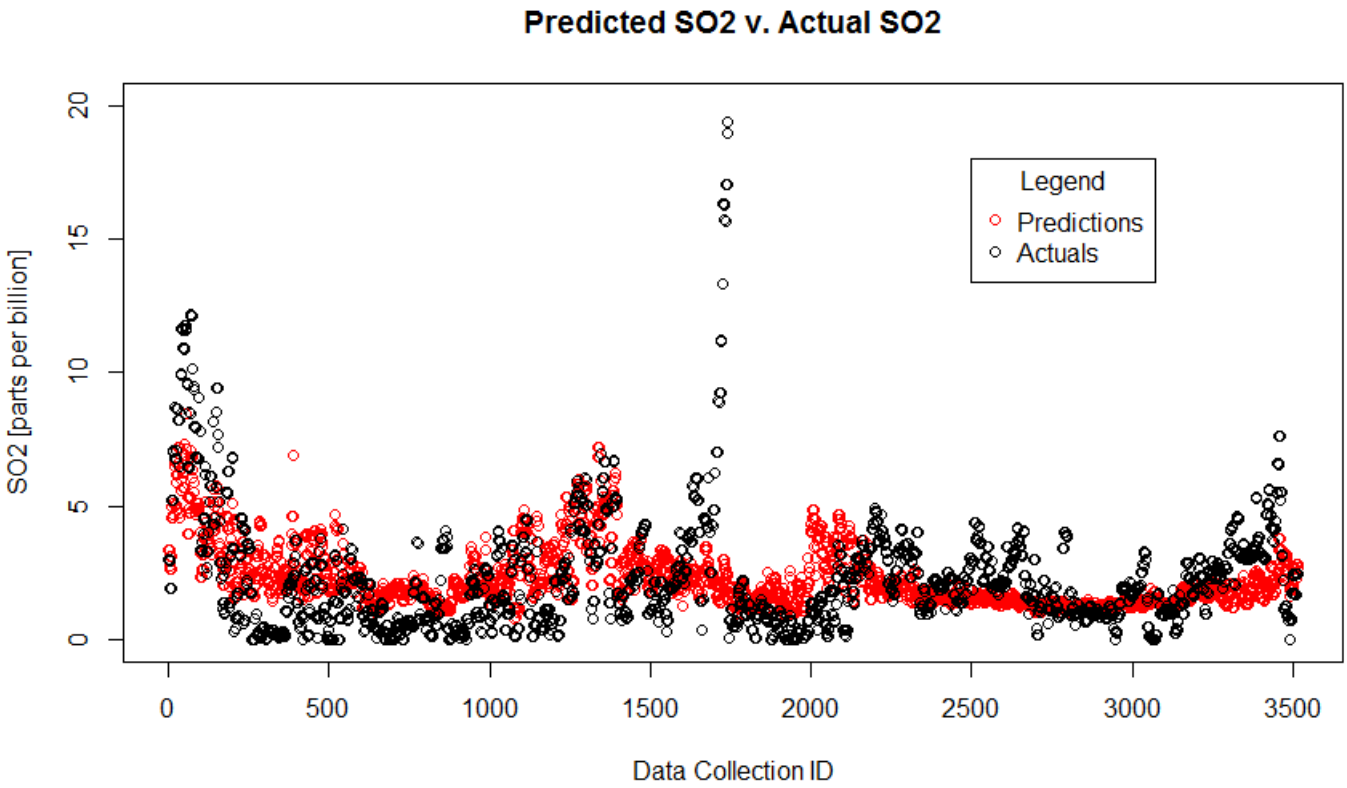
2 Data Collection

The data was obtained through Kaggle.com. The set contained 15 columns of both location and pollutant measures for all 50 states. Since this was a large amount of data the scope of the project was narrowed to just the state of Arizona, it was selected because it was one of the only states with no missing data. The goal of the project is to see if other pollutant (NO2, CO, and O3) are significant predictors in SO2 levels. The measures collected for each pollutant are the mean level and the air quality index (AQI), both in the unit parts per billion. The final data frame contained the columns NO2 mean, O3 mean, SO2 mean and CO mean.

3 Modeling Steps

3.1 Linear Model

Linear modeling is a basic technique used in machine learning. The steps are simple and the results can tell a lot about a set of data. Obviously, the data must follow a linear format for the model to be successful. To begin, the pollutant information is split into training and test sets. 20 percent of the data was withheld to be able to test the accuracy of the trained model. 80 percent of the data is used to train the model then the model is compared to the portion of the data that was left out. Below is a picture of the results of this analysis.



The linear model did fairly well. One thing to note about the chart is the irregular spike just after the 1500 mark. The trained model does not hit any of those points which shows that the model is not over trained.

3.2 LDA and QDA

Linear discriminant analysis and quadratic discriminant analysis have a similar process. As seen in the names, one follows a linear fit and the other follows a quadratic. In the chart above from the linear model the data does not quite follow a linear or a quadratic shape, this is why both models were constructed. The benefit of using these models is that the dimensionality is reduced but the predictive power is preserved.

3.3 Model Comparison and Conclusion

	Success Rate	Error Rate	Type I Error	Type II Error	Power	Precision
Linear	0.6325	0.3675	0.2443	0.4920	0.5080	0.6727
LDA	0.6278	0.3722	0.2514	0.5030	0.4970	0.6462
QDA	0.6136	0.3863	0.1448	0.6479	0.3521	0.6919

The six categories in the chart above are used to analyze the models and which one is best suited for the data. Success and error rate are directly correlated as well as type I and type II errors. The power of the model is used to determine the likelihood of rejecting a null hypothesis. The precision of the model shows how close the points are to each other.

With all the categories taken into consideration, the best model for the data is the linear model. Although each category can provide valuable information on the model the most important category in this case is the success rate, how often is the prediction correct? In the case of answering the question of if NO2, O3 and CO can predict the level of SO2, the best model is linear. In some cases the type I and type II error are more important that it is here. For this model if SO2 is predicted but not actually there then there is no harm done, inversely if the SO2 isn't detected but it is present then the environmental companies lose out on targeting potential clients. These errors are harmful to the profits of the company but are not a matter of life and death as it can be for medical data.