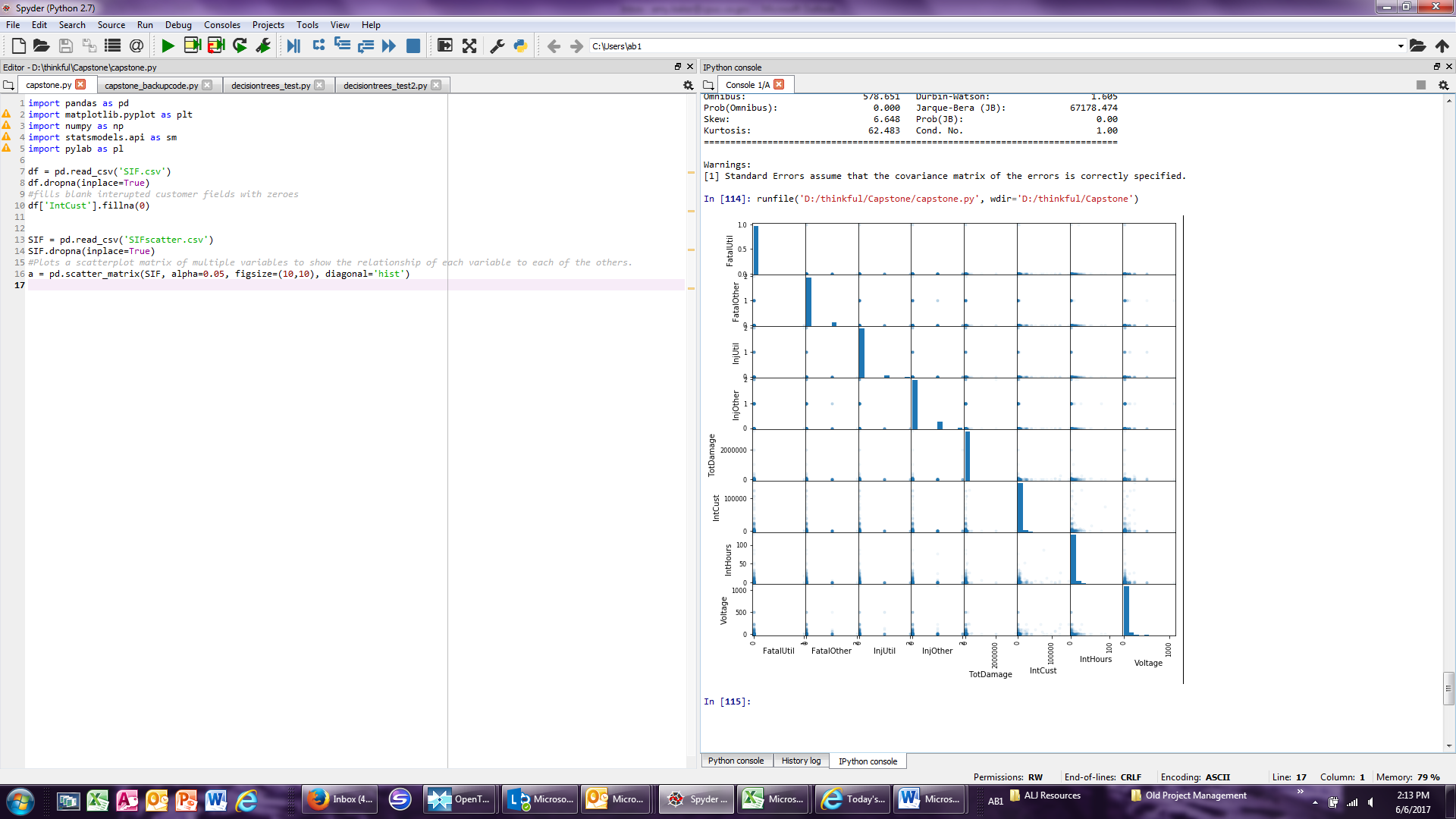
Thinkful Data Science Course Capstone

For my capstone, I’ll be examining mandatory incident reporting data for utility companies from 1997 to 2015. The purpose of the analysis is to determine whether a more robust examination will return any new discoveries from the data, and answer the question, can we predict whether a utility has violated any rule or order for a reported incident?

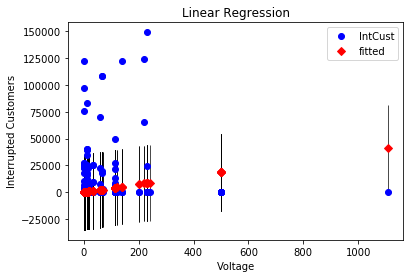
I started my examination by first determining whether there were any obvious linear relationships in the database. We can achieve this by plotting scatterplot matrices of multiple variables in the database to show the relationship of each variable to the others.



Based on the scatterplots, we can see no obvious linear relationships between any of the data points. This is consistent with findings of staff evaluations of the data for individual utilities in previous examinations.

If we look more closely at two of the variables, voltage and interrupted customers, we can confirm there isn’t a strong linear relationship when running a linear regression using stats model ordinary least squares.

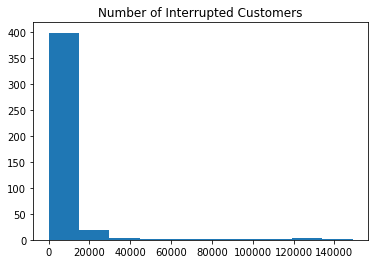
From the chart below, we can see that the model does a poor job of fitting the data.



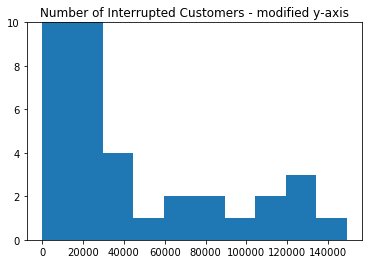
OLS returns a coefficient of 37.1962, a meager R squared value of 0.038 (showing a lot of variance), and a significant P value of 0 (which doesn’t help much given the poor fit here). Voltage has little to no influence on the number of customers interrupted. This is likely the result of a highly redundant and resilient grid.

**Examining Histograms**

Even though there aren’t any linear relationships, we can still look to histograms to inform us in some way. The histogram below shows a plot of the number of interrupted customers. We can see that, for the most part, outage instances skew heavily toward smaller numbers of customers rather than larger, which is positive.

****

It’s difficult to see the other bins, so we can modify the y-axis to examine them a little more closely.



When there are outages of customers in very high numbers, we can use this histogram to demonstrate that these are really outlier events.

**Predicting Violations**

To predict whether a utility has violated a rule or order, we can use a random forest classifier. A random forest classifier is a machine learning technique utilizing multiple decision trees to predict a target value, which in this case is whether the utility violated a rule or order. The data has been split into training, test, and validation sets. In this case the model was tuned to slightly increase performance on the test and validation sets by changing minimum sample leaf size to 70.

Running the classifier results in a mean accuracy test set
score of 0.86 and a
mean accuracy validation set
score of 0.87.

We can also look at the top features for predicting the target value. They are as follows:

1. Ambulance (0.230049)
2. Service Interruption (0.152075)
3. Fatality (0.092873)
4. Injury (0.065420)
5. Voltage (0.056642)
6. Cause - Working Overhead (0.056259)
7. Number of Interrupted Customers (0.050908)
8. Cause - Working Underground (0.048218)
9. Estimated other damage in dollars) (0.041581)
10. Total damage in dollars(0.038556)

You can see from this list that whether an ambulance was called is the most important feature for prediction, followed by services interruption and then fatality and injury.

**Conclusion**

Typically, a prediction accuracy of almost 90% would indicate that our model would be very useful, except that we want to know when there would be violations. A confusion matrix will tell us how many false positives, correct positives, false negatives, and correct negatives the model predicted. In this case, the model only predicted no violations, which is predicted at an accuracy of almost 90%. This means that the model, likely because the data has so few violations, cannot really do want we want it to do, which is to predict violations. This is still very valuable information. Previous staff analysis has indicated that serious injury and fatality events are just not common enough to provide us with any value in terms of predicting future events. This classifier provides us with information that violations in general seem hard to predict, even without fatalities or injuries, even given the collection time of 12 years.