In-depth Analysis

In Capstone project 1, our problem is a regression problem. That is to say we would like to predict the total seismic events based on the total injection volume. In addition, I also predicted the maximum seismic magnitude based on the total injection volume in this analysis.

The main difficulty in this project is that there is limited data for regression analysis. Even though there are totally 8000+ seismic data and 10000+ injection volume data between 2008 and 2017 in Oklahoma, we grouped them into two groups based on the locations. As a result, we have totally 2 groups of data, with 10 data points in each group, illustrating the seismic and injection data in each year. In order to tackle this issue, we utilized the cross-validation method. Cross-validation is a resampling procedure used to evaluate machine learning models based on limited number of data. Among various models of cross validation, leave one out cross validation (LOOCV) is utilized in this study. For the progression model, I tried linear regression and decision tree model.

The code is submitted and the procedure of analysis in coding will not be discussed in detail. For instance, the following is the code to analyze seismic and injection data in region 2 with decision tree and LOOCV.

```
# region 2, injection volume - total seismic number, decision tree
# import library
from sklearn.model_selection import LeaveOneOut
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeRegressor

lm = DecisionTreeRegressor()
cv = LeaveOneOut()

# transform ot numpy array
Inj_sum_2_array = np.array(Inj_sum_2)
Seis_count_2_array = np.array(Seis_count_2)

# normalize
norm = np.linalg.norm(Inj_sum_2_array)
Inj_sum_2_array_s = Inj_sum_2_array/norm
norm = np.linalg.norm(Seis_count_2_array)
Seis_count_2_array_s = Seis_count_2_array/norm

# model, cross validation
model = lm.fit(Inj_sum_2_array_s.reshape(-1,1), Seis_count_2_array_s)
scores = cross_val_score(model, Inj_sum_2_array_s.reshape(-1,1), Seis_count_2_array_s,
scoring='neg_mean_absolute_error', cv=cv, n_jobs = -1)

print("Folds" + str(len(scores)) + ", MSE:" + str(np.mean(np.abs(scores))) + ", STD:" +
str(np.std(scores)))
```

I collected cross validation scores and calculated the means and standard variations of the score values of the cross validation to evaluate the accuracy of linear regression and decision tree method. The following table illustrates the summary of means and standard variations of scores. The results indicate that the accuracy is quite low and standard variations is relative high. Unfortunately, linear regression and decision tree method does not provide good accuracy of the data analysis in this case.

			MSE	STD
Region 1	Seis Number vs. Injection volume	Linear Regression	0.162	0.141
		Decision Tree	0.228	0.221
	Seis magnitude vs. Injection volume	Linear Regression	0.142	0.084
		Decision Tree	0.206	0.134
Region 2	Seis Number vs. Injection volume	Linear Regression	0.173	0.126
		Decision Tree	0.153	0.126
	Seis magnitude vs. Injection volume	Linear Regression	0.060	0.036
		Decision Tree	0.072	0.045