Misael Morales
MATH 7553 – Statistical Learning
Due: May 5, 2020

Project:
Predicting Production from Multivariate Unconventional Reservoirs

In recent years, unconventional resources have become a major player in the oil and gas industry. In the US, shale reservoirs represent a majority portion of the hydrocarbons produced. However, these formations have quite different properties from conventional ones, and therefore require a lot of study and analysis in order to optimize their production.

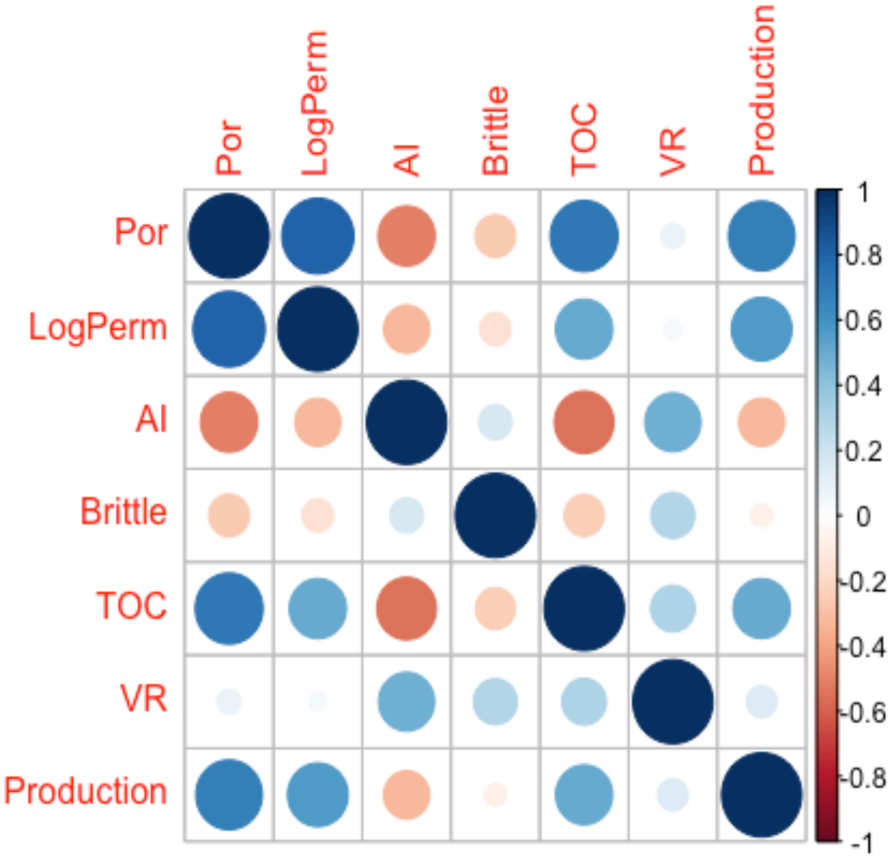
We now explore a multivariate data set consisting of 1,000 different wells in unconventional formations in a certain oil and gas basin of the US. This data set contains 8 different measurements of rock properties, including Porosity (void space in rock matrix), Log Permeability (ability of fluid to move through rock), Acoustic Impedance (density of the rock), Brittleness (easy of fracturing rock), Total Organic Carbon (hydrocarbon content of rock), Vitrinite Reflectance (maturity of the hydrocarbon), and Production (in thousand cubic feet of gas per day). The goal is to predict Production based on the other petrophysical rock properties, making this a regression problem.

The different statistical learning methods used in this predictive modeling are decision trees, linear regression, LASSO regularization, boosting, random forest, and principal component analysis. We ultimately see that most of these methods agree with the prediction of production in this unconventional data set, where Brittleness and Porosity are the main predictors constantly.

We start by loading the data set and removing the first column corresponding to the Well Index. This column is a simple vector going from 1 to 1000, but we are not interested in the exact well number for this scenario. We now observe the summary of each feature:

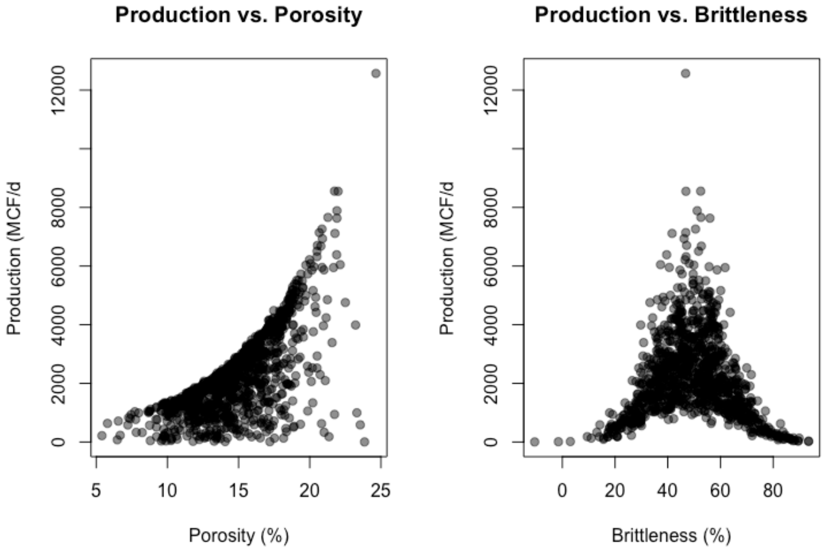
```
##       Por           LogPerm            AI           Brittle      
##  Min.   : 5.40   Min.   :0.120   Min.   :0.960   Min.   :-10.50  
##  1st Qu.:12.86   1st Qu.:1.130   1st Qu.:2.578   1st Qu.: 39.72  
##  Median :14.98   Median :1.390   Median :3.010   Median : 49.68  
##  Mean   :14.95   Mean   :1.399   Mean   :2.983   Mean   : 49.72  
##  3rd Qu.:17.08   3rd Qu.:1.680   3rd Qu.:3.360   3rd Qu.: 59.17  
##  Max.   :24.65   Max.   :2.580   Max.   :4.700   Max.   : 93.47  
##       TOC               VR          Production      
##  Min.   :-0.260   Min.   :0.900   Min.   :    2.714  
##  1st Qu.: 0.640   1st Qu.:1.810   1st Qu.: 1191.370  
##  Median : 0.995   Median :2.000   Median : 1976.488  
##  Mean   : 1.004   Mean   :1.991   Mean   : 2247.296  
##  3rd Qu.: 1.360   3rd Qu.:2.172   3rd Qu.: 3023.594  
##  Max.   : 2.710   Max.   :2.900   Max.   :12568.644  
```

Next we make a correlation plot to see the relationship between the features and Production in the data set. We can see that Production is highly correlated to Porosity, as well as Log Permeability, and last correlated to Brittleness.

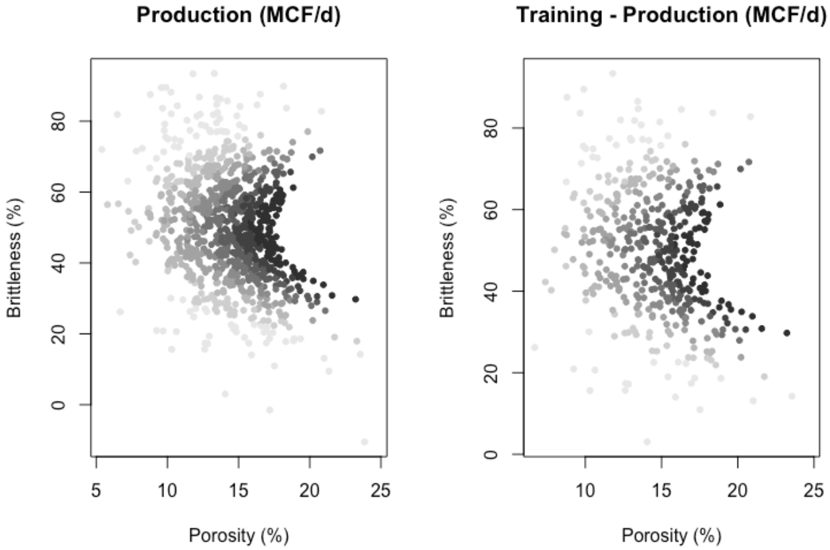


Nonetheless, further study will reveal that Brittleness is actually a major player in the prediction of Production from these unconventional formations, and thus this plot is informative, but not ultimately decisive in regression modeling of the data set.

We therefore decide to observe the plots of Production vs. Porosity and Production vs. Brittleness. We see the clear relationship between the first two variables, and then we see a strange relationship between the latter two, where there is an ideal spot for brittleness percentage to optimize production

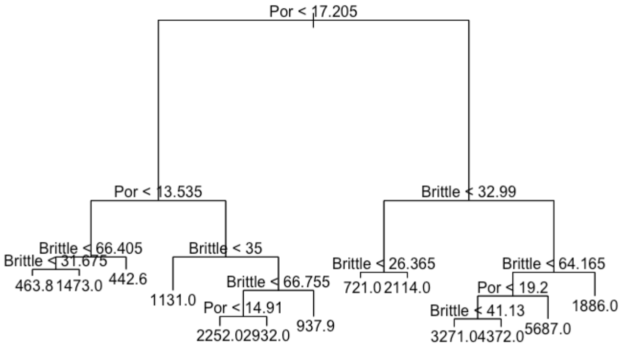


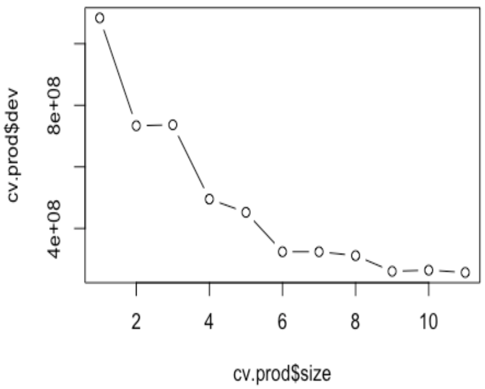
For the statistical learning methods, we create a training and testing data set, in order to cross-validate our models. The split is a 50-50 split, sampling 500 points without replacement from the original data set. For this, we plot Production as a function of Porosity and Brittleness both, and then plot the full data set and secondly, we plot only the training data set. We observe that the train-test split is good since the training set is still representative of the full data set.



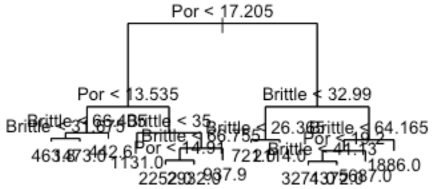
We run a simple bootstrap model for the simple purpose of proving that we can easily replicate the data into a bigger sample. We use 50,000 bootstrap replicates and plot the histogram and cumulative density function, to show that other researchers with smaller data sets might still use these methods to gather more insight into their multivariate unconventional reservoir data.

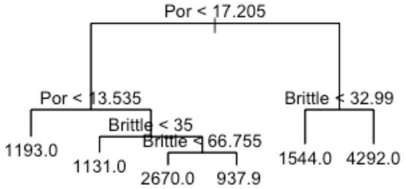
The first method used is a **Decision Tree**. We create a tree model with Production as a function of all other predictor variables, with some simple tree control parameters. Surprisingly, we observe that this tree from the full training set is only a function of Brittleness and Porosity.

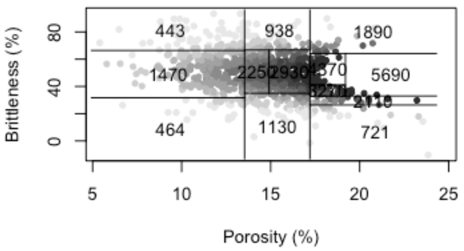


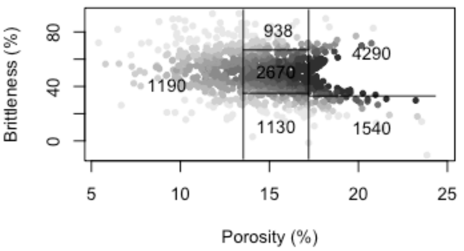


Therefore, we decide to prune this tree to find the best possible split, considering that we only care primarily about those two variables. A cross-validation model for the full tree shows that the best pruning is with 6 terminal nodes. We now show the comparison of both trees and both splitting regions, but still confirm that only Brittleness and Porosity are predictors.







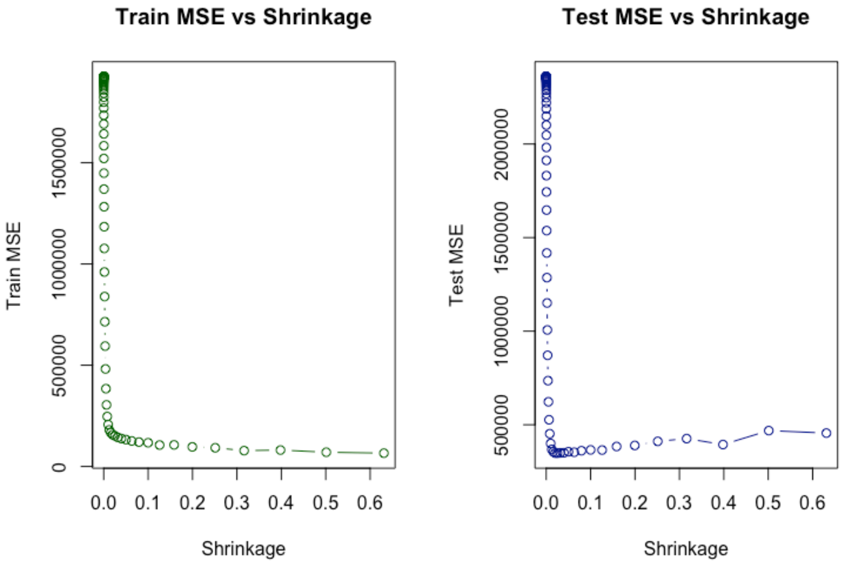


From the pruned tree model, we predict using the testing data set, and calculate the Mean-Squared Error (MSE) in order to later compare with other statistical learning methods.

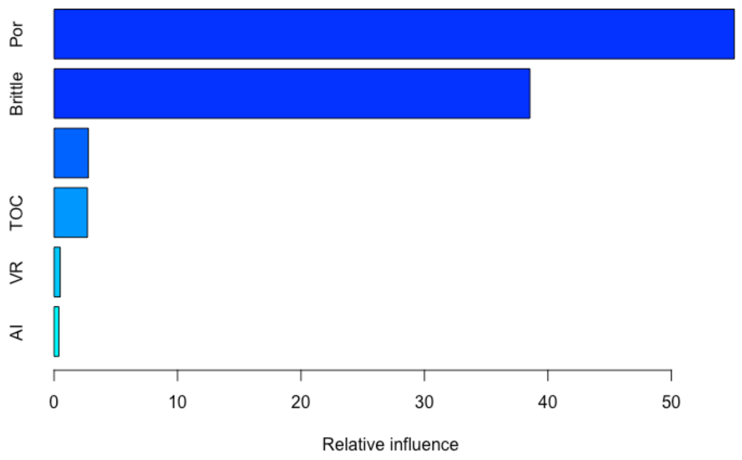
The next methods used are **Linear Regression** and **LASSO Regularization**. There are simpler models, can more easily find the MSE for each model. Using the standard function, we create a liner model for the Production data as a function of all other predictors. Here we can see that Porosity and Vitrinite Reflectance are the variables with largest magnitude coefficients. On the other hand, we use the *glmnet* function, we explore the LASSO model for Production as a function of all other predictors, ensuring that alpha is 1.

For both models we validate against their testing sets and calculate the MSE for each in order to later decide which statistical learning methods yield the smallest MSE.

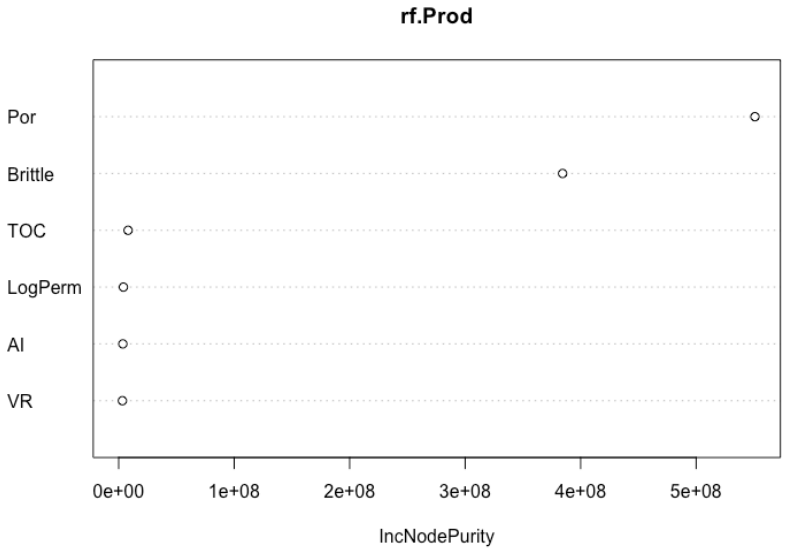
The next method implemented is **Boosting**. Here we have 99 lambda values, and by creating a gradient boosting model for Production for each lambda, we can then calculate the mean error of the model against the testing set. We then plot the testing and training MSE against the shrinkage parameter and observe an ideal lambda value to later use for optimizing the MSE.



A boosting model is then created for the best shrinkage parameter, and we find that this method predicts Porosity and Brittleness to be the most important parameters in predicting Production.



Next, we move to **Random Forest** to predict Production in the training data set. Here we decide to use 500 trees. The see the importance plot for the model showing that Porosity and Brittleness are again the most important predictors by a factor of almost 100 each. We also calculate MSE for the model with regards to the testing set and store it to compare with all other models.



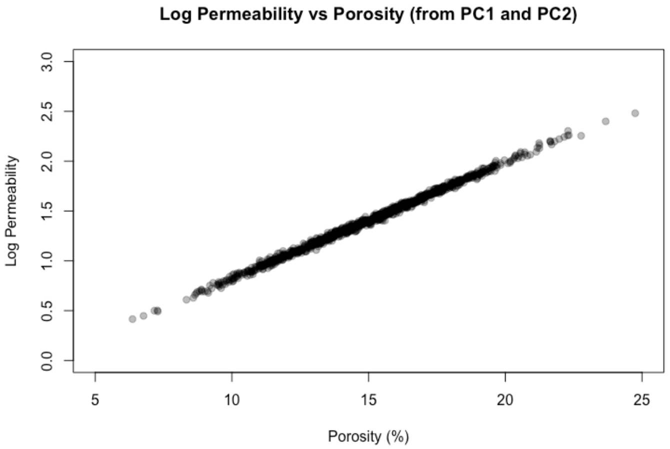
The next thing we do is to compare the MSE for all of the previous models: Decision Trees, Boosting, Linear Regression, LASSO Regularization, and Random Forest. The results show that the method with the smallest MSE is Random Forest, followed by Boosting and Decision Trees. Linear Regression and LASSO Regularization yielded the highest MSE. These values might seem elevated at first, but one must recall that Production is in the units of thousand cubic feet per day, so the mean-squared errors are not that far off in reality.

	METHOD				
	Decision Tree	Boosting	Linear Regression	LASSO Regularization	Random Forest
MSE	$8.7 * 10^5$	$4.0 * 10^5$	$1.2 * 10^6$	$1.2 * 10^6$	$1.3 *10^5$

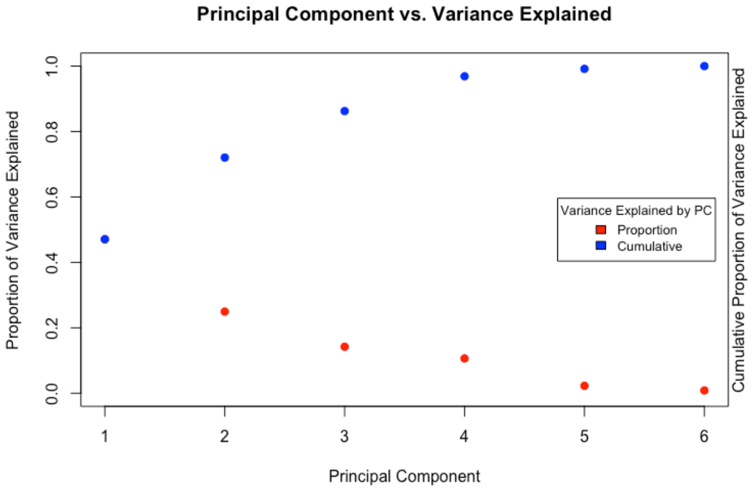
On another note, we perform **Principal Component Analysis** (PCA) on the data set to see the top predictors. We generate the PCA object from the scaled complete data set, and obtain the center, scale, and rotation matrices. The biplot is not quite informative, so we decide to only observe the first two principal components of the data set.

```
PCA$center #the means substracted from variables

##      Por   LogPerm       AI   Brittle      TOC       VR
## 14.95046  1.39888  2.98261 49.71948  1.00381  1.99117

PCA$scale  #the standardization factor

##        Por     LogPerm         AI     Brittle        TOC          VR
##  3.0296343  0.4059657  0.5776287 15.0770064  0.5049778  0.3081940
```

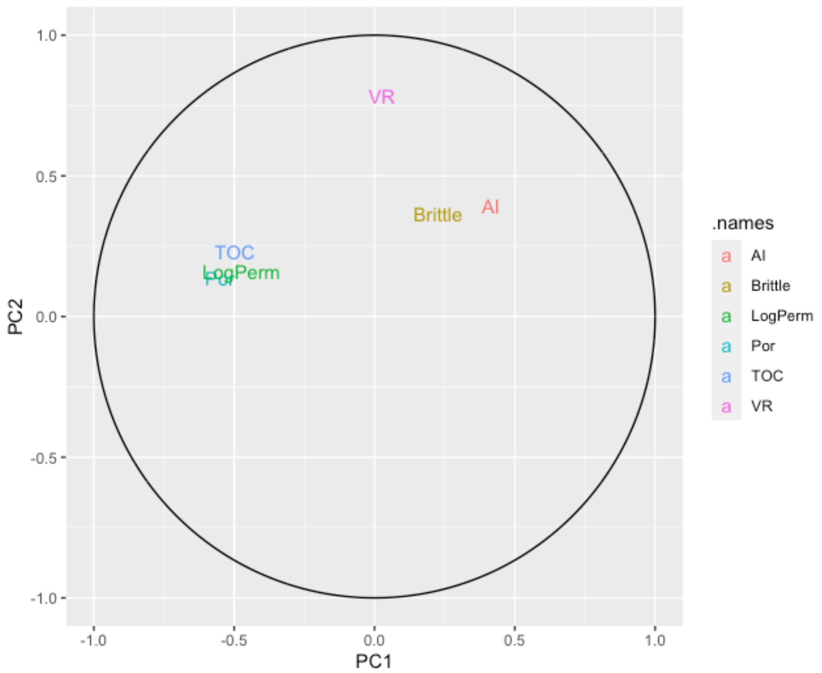
We see from PCA that Log Permeability and Porosity are closely related to each other, almost perfectly linearly. However, from the center and scale vectors we can analyze that Porosity and Brittleness are the most important predictors, again agreeing with our previous analysis.



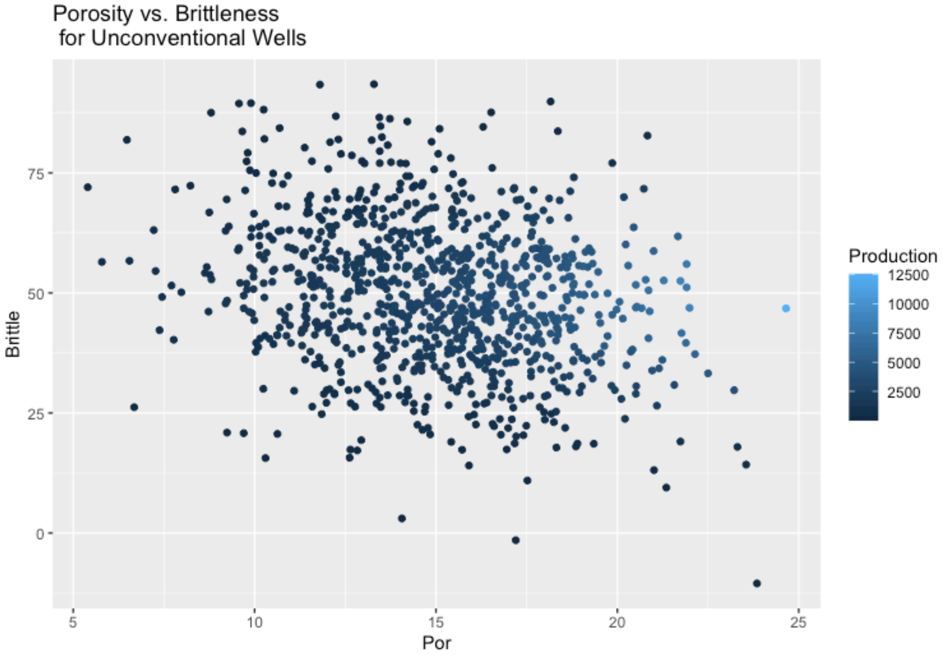
By plotting the Proportion of Variance Explained and the cumulative proportion of variance explained as we increase the number of principal components, we can conclude that 1 or 2 principal components are sufficient to describe the majority of the data set.



Therefore, we can also conclude through PCA that Porosity and Brittleness are the most important predictor features for Production in this data set. With a parametric plot we understand how Porosity is so closely related to Log Permeability in terms of principal component, and also TOC surprisingly; and can also see that Brittleness is in itself a main feature for the second principal component.



Therefore, we now plot Porosity and Brittleness against Production, and see that the behavior is not particularly linear or follows a specific trend, but it is the best way to predict the oil and gas production from and unconventional well.



All methods therefore demonstrate that Production is best described by Porosity and Brittleness of the rock, two unrelated petrophysical properties, but crucial for mineral exploration and production.

In conclusion, we used several methods to analyze a multivariate data set with petrophysical properties of unconventional wells and predict gas production. Initial guesses hinted at the idea of P0orosity and Permeability being the main factors in production, but all methods agreed that the two best predictors for Production are actually Porosity and Brittleness. The best statistical learning method based on the minimization of MSE was Random Forest, followed by Boosting and Decision Trees, and lastly Linear Regression and LASSO Regularization. PCA confirmed the results, and therefore we could use this confidently to predict Production as a function of predictor variables Porosity, Log Permeability, Acoustic Impedance, Brittleness, Total Organic Carbon, and Vitrinite Reflectance.

From a petrophysical perspective, porosity and permeability are highly correlated. A high value in one of these properties almost always means a high value in the other. Other properties such as TOC, AI, and VR, are also important factors, but with modern hydraulic fracturing technologies and refining processes we can still produce a lot of gas from formations that do not exhibit maximum values of these. Nonetheless, the Brittleness of a formation is crucial, especially in unconventional production. Too brittle of a rock and then gas will not be properly transported to the production tubing and therefore produced, and no brittleness would mean that it is nearly impossible to hydraulically fracture the rock and produce its mineral contents.

Future study could include a comparison between conventional and unconventional plays, as well as incorporating more petrophysical properties into the feature space. Another future consideration could be costs efficiency and include into the model the optimization of tangible and intangible expenses for oil and gas production. Moreover, another option to select the best statistical learning method would be to use a different criterion for model selection, such as AIC for example. Finally, *K*-fold cross-validation, random seeding and averaging, and different training-testing splits could also help for future study.

Therefore, we conclude that the combination of high porosity and optimal brittleness yield the best possible formation for unconventional oil and gas production. When reservoir engineers perform analysis on potential formations for production, the main characteristics that should be considered are these two. This has been proven several times with Decision Trees, Linear Regression, LASSO Regularization, Boosting, Random Forests, and Principal Component Analysis.