Week 7 Assignment:

Assignment: This is an optional assignment. If you decide to complete this assignment you can earn up to an additional 100 points for your homework assignments.

Due Date: Saturday before 9 am, July 11th.

Data File:	**diamonds_train.xlsx**
Test File:	**diamonds_validate.xlsx**
Data Dictionary:	**DiamondsDictionary.pdf**
Python Template:	**GA_OilProduction.py**

The python template, GA_OilProduction, is designed to explore the use of Genetic Algorithms for feature selection when the target is interval. In this case, the target identified in the template is "log_cum_production", the log of the first year of production.

You assignment is to modify this to explore feature selection with the diamonds data. The template is designed to compare linear regression models designed from:

1. GA Feature Selection
2. Stepwise Feature Selection
3. All Feature Model (no feature selection)

This template is designed for GA Feature Selection for Interval forecasting and Binary classification. You can identify whether you are using linear or logistic regression. You can also select whether to use the sklearn linear regression or statsmodels. Currently sklearn is not using regularization. You should see similar results, but the runtimes can be very different.

You can also optimize your model for BIC, AIC or the adjusted R-squared. The first two are minimized and the last is maximized.

Finally you can select from two initialization methods. The 'star' method initializes the generation zero by turning on only one feature in

each candidate individual. This is basically starting with simple linear regression and then adding features iteratively.

The second method, 'random', randomly selects features for the first generation. With this approach, approximately half of the available features are selected for the first generation.

These are two extremes. The 'starts' method stars with few features and builds on these learning as it goes forward. The 'random' method starts with many features and then can either build on these or reduce this number.

There is no need to upgrade AdvancedAnalytics. However you will need to install DEAP, available using:

```
conda install -c conda-forge deap
```

If you attended our last Q&A session, you might have completed this install. If not, give this a try and let me know if you encounter any problems.

Your assignment is to modify this template to apply the same technique to the diamonds_train.xlsx data. Use that data to identify the best features from GA, stepwise and full features models.

As a final test, validate your models against the hold out data in diamonds_validate.xlsx. This is a small data set selected to run reasonably fast on your computers.

Since we are validating our model, please use BIC as the fitness metric.

We will talk about this in the Q&A Session, Monday July 6.

Python Solution Upload: Please upload your code, the .py file. Also upload a pdf file with the output from your program. Include a summary of your impression and what you learning from this exercise.

The Python Expert Challenge: OK, I couldn't resist. Karen asked what about classification targets. Good question, I morphed the GA selection

code to handle classification. It is working well for binary classification, but there are a few remaining issues with nominal classification.

I uploaded an example template for the binary classification data, sonar.xlsx. This is the data that uses 60 sonar frequencies to classify an object as either a mine or rock on the bottom of a harbor.

I uploaded the template and the data. I also uploaded a challenge, the voice classification data. The target is gender, and the challenge is to identify whether the speaker is male or female from these data.

See if you can morph the sonar template to work with the voice classification data. If you can get this going and demonstrate your work, you'll get a few extra points on this assignment.

In this case, there are only 4 speakers. Please develop a validation sample using 70/30 partitioning for each of the 4 speakers. That is, use random segmented sampling, segmenting on speaker.

Feel free to experiment with the fitness metrics, initialization and model (sklearn or statsmodels)