Applied Data Analytics for Public Policy

TRAINING

Providing professionals the opportunity to develop key computer science and data science skill sets to improve public policy through training programs and research.
Goals of the Course

1. **Train existing workforce** in rigorous and modern computational data analysis methods and tools for decision-making

2. **Develop new data products** for government agencies

3. **Create new integrated data** to address cross-agency challenges

4. **Establish new networks** across agencies and geographies to address shared problems
Training Content

Data
- Ex-Offenders Welfare Recipients
- Housing and Transportation
- Earnings and Employment Outcomes

• Analytical Design
• Databases
• Record Linkage
• Machine learning
  • Predictive Analytics
  • Fraud Detection
• Information visualization
• Inference
• Privacy, confidentiality, and ethics

Results
- Trained staff
- New products
- New networks
- New understanding

Data Results
“If you work in social science and would like to explore the power of big data, this book is clearly for you...This book is complete and comprehensive. It covers all necessary steps to finish a big data project; collecting raw data, cleaning and preprocessing data, applying various modeling tools to analyze the data, evaluating results, protecting privacy, and addressing ethical problems...All the important topics concerning big data are covered, making this book a good reference that you should always keep on your desk.” (2017) Book Reviews, *Journal of the American Statistical Association*, 112:518, 878-882. DOI: 10.1080/01621459.2017.1325629
Content Example:
Machine Learning

Problem Formulation

Exercise 2

We have only scratched the surface of what we can do with our model. We've only tried one classifier (Logistic Regression), and there are plenty more classification algorithms in sklearn. Let's try them!

clfs = {'RF': RandomForestClassifier(n_estimators=50, n_jobs=-1),
        'ET': ExtraTreesClassifier(n_estimators=10, n_jobs=-1, criterion='entropy'),
        'LR': LogisticRegression(penalty='l1', C=1e5),
        'SGD': SGDClassifier(loss='log'),
        'GB': GradientBoostingClassifier(learning_rate=0.05, subsample=0.5, max_depth=6, n_estimators=100),
        'NB': GaussianNB()}

sel_clfs = ['RF', 'ET', 'LR', 'SGD', 'GB', 'NB']

max_p_at_k = 0
for clfNM in sel_clfs:
    clf = clfs[clfNM]
    clf.fit(X_train, y_train)
    print(clf)
    y_score = clf.predict_proba(X_test)[1,1]
Model Evaluation

In this phase, you take the predictors from your test set and apply your model to them, then assess the quality of the model by comparing the predicted values to the actual values for each record in your testing data set.

- **Performance Estimation**: How well will our model do once it is deployed and applied to new data?

Now let's use the model we just fit to make predictions on our test dataset, and see what our accuracy score is:

Python's `scikit-learn` is a commonly used, well documented Python library for machine learning. This library can help you split your data into training and test sets, fit models and use them to predict results on new data, and evaluate your results.

We will start with the simplest `LogisticRegression` model and see how well that does.

You can use any number of metrics to judge your models (see `model evaluation`), but we'll use `accuracy_score()` (ratio of correct predictions to total number of predictions) as our measure.

```python
# Let's fit a model
from sklearn import linear_model
model = linear_model.LogisticRegression(penalty='l1', C=1e5)
model.fit( X_train, y_train )
print(model)
```
Networks: to date ~215 participants from ~75 agencies have completed 49 projects
Collaborative secure environment: NYU Administrative Data Research Facility (ADRF)
“Love the Jupyter notebooks!! ... I love how the code snippets and explanations are set up in the Jupyter notebooks. The format of going through it individually and discussing questions/challenges in your group, with the experts available when needed, worked really well for my learning style.” – Danielle Fulmer, Director of Business Analytics for City of South Bend, IN
clayton.hunter 11:48 AM

hi folks - for anyone using IDHS data in their projects we have a bit more info on programs to help welfare recipients find stable jobs (thanks to Susan H for posing question and Rick Hendra for a great response!) - this doc will also be linked on the class website: https://docs.google.com/document/d/1GTnuPAWxxtw3CUnCx238cWwVbzx6FAdhl5O1pXsuNg/edit?usp=sharing

clayton.hunter 11:48 AM

shared this file:

Job assistance programs for welfare recipients

Document from Google Drive

Job assistance programs for welfare recipients

Question posed:
We are trying to add some context to our project and I wondered if you had a contact person at the Illinois DHS that could help fill in some questions about programs available to TANF/benefit recipients. I looked on the DHS website and while they do have some information, there’s not much on programs available to help recipients move to stable jobs. For instance, there’s a program called EPIC directed towards SNAP recipients, but I haven’t found much else.

Response from Richard Hendra, MDRC:
Yes, we have very specific guidance as we worked on this particular issue there. The ERA evaluation had a site in Chicago that was focused on providing TANF recipients with stable jobs. The short term report had more detail about the program, the implementation, and the interim effects. Note that the 111 data had major coverage issues with the segment of the TANF caseload that we were working with. The final results are in this report. I’d suggest the interim (shorter) report. We used various measures of enrollment stability. A common measure is the extent to which individuals worked in 4 consecutive months.
For more information: https://coleridgeinitiative.org/training

Providing professionals the opportunity to develop key computer science and data science skill sets to improve public policy through training programs and research.

Email: dataanalytics@umd.edu