Syllabus

Data Science for Economists - Econ 148

UC, Berkeley
Department of Economics
Tamer Çetin, Ph.D.
tamercetin@berkeley.edu

Summer 2023
Mon-Thu & 3:00 PM - 5:30 PM
Office: #531 Evans Hall
Office hours: 12:30pm-2:30pm Thursday

Course Description:

Data Science for Economists is a unique field in data analysis. While it refers to core Data Science analysis on one hand, it introduces a still-emerging interdisciplinary approach to applied data analysis on the other hand. This approach has been built where Machine Learning algorithms and economic applications meet. For that reason, in this course we will basically cover four main sections in data analysis: data exploration, regression analysis, prediction, and causal analysis. The purpose is to gain a newly developing and strong analytical tool in applied data analysis. To this aim, we will examine how to analyze all types of datasets using both Data Science and Machine Learning-based Econometrics since over 80% of data today is only applicable to Data Science-based analysis, but not the traditional econometric models. On the other hand, causal analysis still matters since Data Science-based applied analysis focuses on the prediction of expected value of conditional y on conditioning x while causal inference econometrics is interested in the effect of x on y. In order to provide more reliable results in data analysis, Data Science for Economists will help students uncover such relationships between y and x. After completing this course, students will:

- better understand how to generate and prepare real-world data for applied analysis using extremely useful Data Science techniques,
- be using Data Science tools from exploratory data analysis to Machine Learning algorithms and their usage in Econometric models in both predictive and causal analyses.

Prerequisites:

Data C8\CompSci C8\Info C8\Stat C8
or
Stat 20 + familiarity with Python recommended

Lectures:

There will be 10 hours of lecture per week - 4x2.5-hour lectures for six weeks. When appropriate these lectures will include data explorations and live coding.

Course Format

Three homework assignments/problem sets, two midterm exams, and one final exam are designed to motivate and evaluate students' learning. Students attend two-and-half-hour lectures from Monday to Thursday. Also, please expect quizzes and recitations on Thursdays. The quizzes will be written in-class and will not be graded. The course material, which will be delivered on Mondays, will be part of homeworks and exams.

Assignments:

Data Science is about analyzing real-world data sets, and so a series of projects involving real data are a required part of the course. Students may work with a single partner on each individual project/homework, and you are allowed to work with a partner in your data/code challenge section. Each student must submit each homework independently, but you are

allowed to discuss problems with other folks.

Biweekly Individual Project/Homework: Four biweekly homework/project will be done using Python, when necessary, individually.

Final: The final exam might be a final project that would be carried out as a group project based on a reproduction analysis of a published journal article. We will talk about that!

Grading:

Biweekly Individual Homework (4): 40% - These will be due on the last class day of the week at 5pm a week after they are released. Basically, you will have one week to submit your homework after you receive it.

Midterms (2): 40% - Midterms will be on the last class days of the second and fourth weeks.

Final: 20% - Final will be on the last class day of the last week.

Late Policy:

Students are allowed to submit homework late for a 50% penalty until the next Thursday at 5 pm after the predetermined submission deadline for homework.

Disabled Students Policy:

UC Berkeley is committed to creating a learning environment that meets the needs of its diverse student body including students with disabilities. If you anticipate or experience any barriers to learning in this course, please feel welcome to discuss your concerns with me.

If you have a disability, or think you may have a disability, you can work with the Disabled Students' Program (DSP) to request an official accommodation. The Disabled Students' Program (DSP) is the campus office responsible for authorizing disability-related academic accommodations in cooperation with the students themselves and their instructors. You can find more information about DSP, including contact information and the application process here: dsp.berkeley.edu. If you have already been approved for accommodations through DSP, please meet with me so we can develop an implementation plan together.

Students who need academic accommodations or have questions about their accommodations should contact DSP, located at 260 César Chávez Student Center. Students may call 642-0518 (voice), 642-6376 (TTY), or e-mail dsp@berkelely.edu (link sends e-mail).

Learning outcomes:

Learning outcomes will be:

a hands-on knowledge of how datasets are created and explored using the most common data analysis tools,

and an understanding of how data can be used to answer economic research or business questions using the appropriate data analysis techniques.

This course is based on the elements of the data science life cycle; from data exploration and feature engineering to formulating questions, visualization, and modeling. By looking at a variety of datasets, students will be exposed to a broad range of Data Science applications to economic questions. Students will get a combination of skills and tools needed to be a successful data scientist, researcher, analyst, and/or (applied) economist.

Materials & Resources: *References will be drawn from the following according to the topics:*

Books:

Bekes, G. and G. Kezdi, 2021. *Data Analysis for Business, Economics, and Policy,* Cambridge University Press.

Cleff, T., 2014. *Exploratory Data Analysis in Business and Economics: An Introduction Using SPSS, Stata, and Excel,* Springer.

Hastie, T., R. Tibshirani, and J. Friedman, 2016. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Second Ed., Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2021. *An Introduction to Statistical Learning: with Applications in R*, Second Ed., Springer.

Kuhn, M. and K. Johnson, 2013. *Applied Predictive Modeling*, Springer.

Vanderplas, J., 2017. *Python for Data Analysis*, O'Reilly.

Suggested Readings:

Athey, S. 2021. 'The Impact of Machine Learning on Economics', Eds. A. Agrawal, J. Gans, and A. Goldfarb, *The Economics of Artificial Intelligence,* Ch. 21, University of Chicago Press, 507-552.

Breiman, L., 2001. 'Statistical Modeling: The Two Cultures', *Statistical Science*, 16(3), 199-231.

Ludwig, J, and S. Mullainathan, 2021. 'Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System', *Journal of Economic Perspectives*, 35(4), 71–96.

Mullainathan, S. and J. Spiess, 2017. 'Machine Learning: An Applied Econometric Approach', *Journal of Economic Perspectives*, 31(2), 87–106.

Wager, S. and S. Athey, 2019. 'Machine Learning Methods That Economists Should Know About', *Annual Review of Economics,* 11, 685-725.

Further Readings:

Chan, F. and L. Matyas, 2022. *Econometrics with Machine Learning*, Springer.

Semenova, V. and V. Chernozhukov, 2021. 'Debiased Machine Learning for Conditional Average Treatment Effects and Other Causal Functions', *The Econometrics Journal*, 24(2), 264-289.

Semenova, V., 2023. 'Debiased Machine Learning of Set-Identified Linear Models'. *Journal of Econometrics*,

Wager, S. and S Athey, 2018. 'Estimation and Inference of Heterogeneous Treatment Effects using Random Forests', *Journal of the American Statistical Association,* 113, 1228-1242.

SECTION I: INTRODUCTION TO DATA SCIENCE: DATA EXPLORATION
Technology for Data Science
Origins of Data
Big Data, Statistical Inference, and External Validity
Data Wrangling: Preparing Raw Data for Analysis
Exploratory Data Analysis and Feature Engineering
Model Selection, Evaluation, and Validation
Regularization, Parameters, and Hyperparameters

SECTION II: REGRESSION

Testing Hypotheses
Comparison and Correlation
Introduction to Regression
Linear Regression in ML vs Linear Regression in Econometrics (with application in Python)

Time Series Forecasting (with application in Python)
Predicting Probabilities
Logistic Regression in ML vs Logistic Regression in Econometrics (with application in Python)

SECTION III: PREDICTION: ML ALGORITHMS

Supervised and Unsupervised Models
Prediction
Regression
Classification
Clustering
Experimental Design, Reinforcement Learning, and Multi-Armed Bandits

SECTION IV: CAUSALITY IN DATA ANALYSIS

Data Science and Machine Learning
Data Science and Economics
Econometrics and ML
Causality, Intervention, and Variation
The Setup: Intervention, Treatment, Subjects, and Outcomes
Potential Outcomes Framework
Causal Maps (DAGs) to Uncover Causal Structure
Controlled Experiments
Randomized Experiments

Calendar (Tentative)

Week	Topics	Assignments
Week 1: INTRODUCTION TO DATA SCIENCE: DATA EXPLORATION **- Technology for Data Science** **- Origins of Data** **- Data Structures: Structured vs Unstructured** **- Collecting Data** **- Selection Bias vs Random Sampling** **- Reading Data** **- EDA I**	Lecture 1- Introduction to Course, Overview of Course Goals, Introduction to Technology for Data Science Python Common Libraries/Modules SQL GitHub Lecture 2 - Origins of Data What is Data? Data Structures: Structured vs Unstructured Data Types Data Quality Sources of Data and Collecting Data	Data/code challenge 1 - Introduction to Notebooks EDA Frequencies and Probabilities Visualizing Distributions Data/code challenge 2 - Extreme Values: Missing Values, Outliers, and Sparsity EDA for Numerical and Categorical Variables Summary Statistics

	Data Sampling: Selection Bias vs Random Sampling Data Science Flow for Reading Data Big Data, Statistical Inference, and External Validity Lecture 3 - Data Wrangling: Preparing Raw Data for Analysis Types of Variables Types of Observations Linking Relational Data with Tables Organizing Data Tables for a Project Exploratory Data Analysis Frequencies and Probabilities (with application in Python) Visualizing Distributions (with application in Python) Lecture 4 - Extreme Values: Missing Values, Outliers, and Sparsity (with application in Python) Some Graphs for Data Visualization (with application in Python) EDA for Numerical and Categorical Variables (with application in Python) Summary Statistics (with application in Python)	
Week 2: INTRODUCTION TO DATA SCIENCE: DATA EXPLORATION – EDA II **- Fundamentals of Feature Engineering** **REGRESSION - Testing Hypotheses** **- Comparison and Correlation**	Lecture 1 - Bias-Variance Tradeoff Overfitting and Underfitting Splitting Data into Training and Test Sets with Application in Python The Curse of Dimensionality Dimensionality Reduction by Feature Extraction (with Application in Python) Lecture 2 - Dimensionality Reduction by Feature	Data/code challenge 3 - Dimensionality Reduction by Feature Extraction Dimensionality Reduction by Feature Selection Data/code challenge 4 - Principal Component Analysis Feature Generation and Encoding

	Selection (with Application in Python) Principal Component Analysis (with Application in Python) Feature Generation and Encoding (with Application in Python) Model Selection, Evaluation, and Validation; Regularization, Parameters, and Hyperparameters; Lecture 3 - Testing Hypotheses The Logic of Hypothesis Testing Null vs Alternative The t-test and the p-Value Making a Decision: False Negatives vs False Positives Testing Hypotheses with Big Data Comparison and Correlation The γ and the x Describing Patterns of Association Conditioning Conditional Probabilities Conditional Distribution and Expectation Dependence, Covariance, and Correlation Sources of Variation in x Lecture 4 – Midterm I	Project 1 – Missing Values, Outliers, and Sparsity EDA for Numerical and Categorical Variables Visualizing Distributions, Frequencies, and Probabilities
Week 3: REGRESSION - **Introduction to Regression** **- Multiple Regression** **- Time Series Forecasting** **- Predicting Probabilities**	Lecture 1- Introduction to Regression Simple Regression Regression in Prediction Regression in Causal Inference Non-Parametric and Parametric Regression Assumption vs Approximation in Regression Analysis Regression Coefficient OLS	Data/code challenge 5 - Linear Regression in ML vs Linear Regression in Econometrics Logistic Regression in ML vs Logistic Regression in Econometrics Project 2 - Selected data project - import, clean,

	Predicted Values and Residuals Lecture 2 - Multiple Regression Case of Two Regressors Visual Representation Many Explanatory Variables Non-Linearity Using Qualitative Variables Generalizing Regression Results Linear Regression in ML vs Linear Regression in Econometrics (with application in Python) Lecture 3 - Time Series Forecasting Time Series Specialties Data Preparation Aggregation What is Special in Time Series? Stationary Serial Correlation: Order and Sign Serial Correlation: Magnitude Non-Stationary: Random Walk Unit Root Tests Seasonality in Time Series An application in Python Lecture 4 - Predicting Probabilities Binary Events Linear Probability Model (LPM) and its Interpretation Predicted Values in LPM Maximum Likelihood Estimation Predictions for LMP, Logit, and Probit Logistic Regression in ML vs Logistic Regression in Econometrics (with application in Python)	describe, and explore data Data Preparation Data Exploration Feature Engineering

Week 4: PREDICTION: ML ALGORITHMS - **Framework for Prediction in ML** **- Machine Learning Algorithms** **- Regression Models in ML**	Lecture 1 - Supervised and Unsupervised Models Framework for Prediction in ML Prediction Setup Predictive Analysis: What is New? Regression and Prediction Prediction Error Decomposing Prediction Error Interval Prediction for Quantitative Target Variable Loss Functions Squared Loss Adding Up – Mean Squared Error (MSE) MSE Decomposition: Bias and Variance Lecture 2 - Model Selection External Validity, Avoiding Overfitting, and Model Selection Comparing Models: Overfit vs Underfit Finding the Best Model using Best Fit and Penalty: The BIC Model Fit Evaluation Finding the Best Model using Training and Test Samples Finding the best model by cross-validation 5-fold cross-validation External validity and stable patterns Lecture 3 - Machine Learning and the Role of Algorithms Machine Learning Algorithms Regression Business question and defining γ Steps of prediction in ML Sample Design: Filtering Label Engineering: Defining Target	Data/code challenge 6 - Predicting Airbnb Prices Data/code challenge 7 – Post-prediction diagnostics Selected Performance Matrices ROC Curve Feature Importance Project 2 - Selected data prediction project – EDA and Linear Regression Prediction

	Log vs Level Feature Engineering What Features In What Functional Form What to Do with Different Types of Variables Predicting Airbnb Prices (with application in Python) Lecture 4 – Midterm II	
Week 5: PREDICTION: ML ALGORITHMS - **Classification** **- Clustering** **- Reinforcement** **SECTION IV: CAUSALITY IN DATA ANALYSIS - $\widehat{\boldsymbol{\beta}}$ vs $\widehat{\boldsymbol{\gamma}}$** **- Potential Outcomes Framework**	Lecture 1 - Model Building Evaluating the Prediction using a Holdout Set Selecting Variables in Regressions by RIDGE and LASSO Penalized/Generalized Regression Models Regression Trees and Forest Post-prediction diagnostics (with application in Python) Feature Importance (with application in Python) ROC Curve Lecture 2 - Classification (with application in Python) Logistic Regression Decision Trees Random Forests Boosting Super Vector Machines K-Nearest Neighbors Deep Learning and Neural Networks Lecture 3 – Clustering (with application in Python) K-Means K-Modes K-Prototypes Hierarchical Clustering DBSCAN Experimental Design, Reinforcement Learning, and Multi-Armed Bandits	Data/code challenge 8 - Classification and Clustering Models Project 3 - Code to Cleaning to Visualization to Outputs in Classification and Clustering Models

	Lecture 4 - Data Science, and Machine Learning (ML) Data Science and Economics Econometrics and ML $\hat{\beta}$ vs $\hat{\gamma}$ Association vs Causation Model Identification vs Algorithmic Data Modeling Causal Questions Measuring Causality Requires Intervention and Variation The Setup: Intervention, Treatment, Subjects, and Outcomes The Causal Question Again	
Week 6: CAUSALITY IN DATA ANALYSIS - Controlled Experiments **- Natural Experiments** **- Observational Data** **- Presentations and Telling a story with Data** **- Data Management** **- Reproduction and Reproducibility**	Lecture 1- Potential Outcomes Framework Individual Treatment Effect Heterogeneous Treatment Effects Average Treatment Effect ATE as Average / Expected ITE Average Effects in Subgroups and ATET ATE When Quantitative Causal Variables Quantitative Causal Variables ATE and Quantitative Causal Variables Causal Maps (DAGs) to Uncover Causal Structure Causal Maps: Simplest Case DAG: Mechanisms Random Assignment Random Assignment and ATE Random Assignment, ATE, and ATET Sources of Variation in the Causal Variable An Exogenous and an Endogenous Source of Variation in x Good and Bad Sources	Data/code challenge 9 - Making a Presentation using Data and ML Models Project 4 – Selected Topics

	Lecture 2 – Experimenting versus Conditioning: 1 Controlled Experiments Controlled Experimental Variation in x Experimenting versus Conditioning: 2 Natural Experiments Experimenting versus Conditioning: 3 Conditioning Confounders in Observational Data Three Types of Confounders Common Cause Confounder Mechanism of Reverse Causality Unwanted Mechanism Confounders in Practice: Selection from Latent Variables to Measured Variables Omitted Variable Bias The Three Types of Bad Conditioning Variables Lecture 3 - An Overview: Data Science, Machine Learning, and Causal Inference Presentation and Telling a Story with Data and ML Models Lecture 4 - Final	